Email Spam Detection using Naïve Bayes and Particle Swarm Optimization

Nandan Parmar[1], Ankita Sharma[2], Harshita Jain[3] , Dr. Amol K. Kadam[4]
[1, 2, 3,4] *Department of Computer Engineering, Bharati Vidyapeeth College of Engineering, Pune, India*

***Abstract*- For sharing of important and official information, email is used as a default medium of communication. Most of the institutions and companies prefer to use emails over all other mediums as it is one of the cheapest, easy to use, easily accessible, most official and reliable way of sharing information. It is used widely as it also provides the confidentiality of the data shared. But with the pros also comes the cons, as many people misuse this reliable and easy. way of communication by sending unwanted and useless bulk messages for their own personal benefits. These unwanted emails affect the normal user to face the problems like flooding of the mail box with unwanted emails making it harder to look for the useful once, even sometimes one may skip through important and useful emails because of all these unwanted emails. So, this gives rise to a need of a strong email spam detector which can filter maximum amount of spam emails with a greater accuracy so that a genuine email does not get filtered as spam. In this paper an integrated approach using Naïve Bayes algorithm along with Particle Swarm Optimization is used for email spam detection. Naïve Bayes algorithm is used for learning and classification of email as spam and ham. Particle Swarm Optimization is a stochastic optimization technique and is used for heuristic global optimization of parameters of Naïve Bayes. For experimentation Ling Spam dataset is considered and result is evaluated in terms of precision, f-measure, accuracy and recall.**

***Index terms*- Text Mining(TM), Email Spam Detection, Machine Learning (ML), Naïve Bayes(NB), Particle Swarm Optimization (PSO).**

I.INTRODUCTION

Electronic mails are opportune way used by specific users and business concern for the purpose of communication of useful information. But because of large volume of spam and junk emails received, it makes it difficult for the users to filter out the important ones. It has been found that each user receives daily about 40-50 spam emails, which means that about 60%-70% of the total emails received daily are spam emails. Such large volume of spam emails lead to a lot of other problems such as consumption of lot of resource and time, cost shifting, fraud, identity theft etc. Many researchers are already working on spam filtering techniques, but accurate spam detection is considered a difficult task due to many reasons including subjective nature of spam, processing overhead and message delay, type of language used and irregular cost of filtering errors. Text mining approach is used for the classification of mail as spam and non spam. Different machine learning algorithms have been used by different authors for the detection and classification of spam mails [1], discussed in section II.

In this paper, we propose a framework for spam email detection using an unified approach of machine learning based Naïve Bayes (NB) algorithm [2] and computational intelligence-based Particle Swarm Optimization (PSO) [3], [4]. Bayes theorem has strong independence property and it gives the probability of an event based on the prior knowledge of a related event. PSO is based on the concept inspired by the social behavior of flying birds. A dataset consisting of 1000 emails are considered for experimentation, out of which 500 are used for the purpose of training and the other for testing. The evaluation of experimental results is done on the basis of precision, recall, accuracy and f-measure.

The existing work on email spam classification is presented in Section II. The basic concepts of Naïve Bayes and Particle Swarm Optimization are presented in Section III. Section IV presents the proposed algorithm for email spam classification along with flowchart. Section V presents the experimentation results and Section VI concludes the paper.

II. RELATED WORK

The methods and techniques that have been previously used by the researchers for filtration of spam emails are presented in this section. Table I gives the information about different spam filtration techniques used by various authors over the years which includes research title, author name, year of publication, method used, evaluation parameters and remarks. Some of the techniques are able to detect both textual and image data format while some can only detect textual data format. Different strategies are tailored by totally different completely different authors with experimentation on different datasets. Some authors have worked on the detection of spam email in each the matter and image data formatting. Harisinghaney et al. (2014)[5] and Mohamad & Selamat (2015)[6] have used the image and matter dataset for the e-mail spam detection with the employment of various strategies. Harisinghaney et al. (2014) have used strategies of Naïve Bayes, KNN algorithmic program and Reverse DBSCAN algorithmic program with experimentation on Enron Corpus's dataset. For the text recognition, OCR library is employed however this OCR doesn't perform well. 2 experiments are performed with and while not preprocessing steps with analysis parameters of accuracy, specificity, sensitivity and preciseness. Overall Naïve Bayes perform with efficiency with accuracy of eighty-seven. Further, Mohamad & Selamat (2015) thought-about feature choice hybrid approach of TF-IDF (Term Frequency Inverse Document Frequency) and Rough pure mathematics. during this experimentation, a manual knowledge set of 169 emails is generated containing each the text and pictures primarily based data. Authors have used the RSES (Rough Set Exploration System) tool for the removal of expendable words and for the principles generation. Overall thought is compared in terms of accuracy to TDIDF-Decision Tree and shows economical results.

Most of the researchers have targeted solely on the text {primarily based} email spam classification as image based spam are often filtered at the initial stage of pre-processing. For matter data processing, major use of machine learning primarily based Support Vector Machine (SVM) is recorded. Either a number of the authors have used SVM separately (Renuka and Visalakshi, 2014)[7] or some have used SVM in integration with another ideas like SVM-NB (Feng et al., 2016)[8], SCS-SVM (Kumaresan and Palanisamy, 2017)[9], and SVM-ELM (Olatunji et al., 2017)[10]. In 2014, Renuka and Visalakshi have used Support vector Machine (SVM) for the classification of Email Spam detection beside the employment of Latent linguistics compartmentalization (LSI) for feature choice. TF-IDF is employed for the feature extraction. Here, planned SVM-LSI is compared with SVMTFIDF while not victimization LSI, PSO and Neural Network. From the thought-about strategies, SVM-LSI performs higher in terms of accuracy as compare to alternative existing ideas. In 2016, Feng et al. have planned SVM-NB algorithmic program for the e-mail spam filtration. Authors have combined the SVM algorithmic program with NB approach wherever NB will handle massive dataset and SVM is ready to make hyper-plane primarily based separation between completely different feature classes.

ELM is machine learning approach that was planned to beat the perennial downside of feed forward neural network and is employed as learning approach for single layer primarily based neural network. Results of ELM and SVM as compared on the idea of Accuracy and Time taken for the e-mail spam classification from same dataset. In terms of Accuracy, SVM performs higher with 94.06 considered compare to ELM having accuracy 93.04 %. Except for every case, SVM consumes longer as compare to ELM. So, ELM is healthier than SVM in terms of your time taken.

Table I: Existing work related to Email Spam detection

TITLE OF RESEARCH	AUTHOR & YEAR	METHOD USED	EVALUATION PARAMETERS	REMARKS
Email Spam detection using Support Vector Machine and Latent Semantic Indexing	Renuka and Visalakshi (2014)	Support vector Machine (SVM) with Latent Semantic Indexing (LSI)	Precision, recall and accuracy	Reported better results in terms of precision
Spam mail Detection using Naïve Bayes, KNN	Harisinghane y et al. (2014)	Naïve Bayes, KNN algorithm and	Precision, specificity, sensitivi	Reported high precision but very slow

Algorithm and reverse DBSCAN algorithm.		Reverse DBSCAN algorithm	ty and accuracy	performanc e, was able to detect image scan
Spam Email Classification using Term Frequency Inverse	Mohamad and Selamat (2015)	Term Frequency Inverse and Rough Set Theory	Accuracy, Classification	Performed well in terms of feature extraction.
Improving Knowledge Based Spam Detection Methods: K-means Clustering and Artificial Neutral Network[11]	Tuteja and Bogiri (2016)	K-means Clustering and Artificial Neutral Network	Precision and recall	Observed better results in precision with respect to previously existing techniques.
Improving Spam mail detection by Integrating Support Vector Machine Navie Bayes	Feng et al. (2016)	Integrated approach using Support Vector Machine and Naïve Bayes	Precision, recall and execution time	Integrated approach results in increased precision and accuracy than individual SVM and NB approaches
Email Spam Classification using Stepsize Cuckoo Search and SVM	Kumaresan & Palanisamy (2017)	Stepsize Cuckoo Search and SVM	Accuracy, sensitivity and specificity	Processing speed is high as compared to others.

III. BASIC CONCEPTS

In this section we will be discussing the basic concepts of Naïve Bayes algorithm and Particle Swarm Optimization as these are the algorithms used for Email Spam Classification.

A. Naïve Bayes

Naïve Bayes algorithm is based on the concept of conditional probability given by Bayes theorem. Bayes theorem is based on statistical machine learning based approach which has properties of strong independence, probability distribution and ability to handle datasets. In NB, the evaluation of probability distribution is done from the frequency distribution of dataset. NB is used to assign different objects to classes. The class having highest posterior probability value is chosen by the classifier. The Bayes Rule can be defined using Equation (1)

$$P(y/x) = \frac{P(x/y)P(y)}{P(x)} \qquad \ldots\text{Equation (1)}$$

Where, x is any feature vector set and y are the class variables with m possible outcomes. *P(x/y)* is any particular class, *P(y/x)* stands for posterior probability and is dependent on *P(x/y)*, the evidence depending on known feature sets is given by *P(x)*, the prior probability is denoted by *P(y)*. So Naïve Bayes classification model consist of class conditional probability, set of probabilities of prior probability and posterior probability.

B. Particle Swarm Optimization

Particle Swarm Optimization (PSO) is a swarm intelligence concept given by Eberhart and Kennedy in 1995. It was inspired from the social behavior of the flying birds or fish school. It is a computational method which iteratively tries to improve the solution to optimize a problem. In PSO a global solution is obtained from a set of local solutions, it moves closer to the best solution in every iteration as each particle share their experience.

In PSO, according to the interaction between different particles velocity of each particle gets updated. Vector particle position and particle velocity are the two main dynamics of PSO algorithm. Each particle shares its experience with other particles and also changes its trajectory in accordance to the experience of the other particles to achieve better solution. PSO is basically used for the purpose of global optimization of solution. In this research PSO is used to optimize the parameters of the Naïve Bayes algorithm for Email spam classification.

IV. PROPOSED ALGORITHM

In this section spam classification using integrated approach of Naïve Bayes and Particle Swarm Optimization is discussed. NB works on the concept of conditional probability distribution and classifies the emails into classes i.e. spam and ham based on the content of the email. To further optimize the

parameters of NB approach PSO is used, it improves the accuracy and search space of the classification process. Correlation based feature selection method is used for reducing dimensionality and selecting the relevant features from the data on the basis of which classification is performed. Following are the steps involved in classification of emails.

Step 1: Consider the email in raw format from the dataset.

Step 2: Perform the following preprocessing steps on the email in the raw format.

- Tokenization: Break the stream of text in the email into tokens of individual words.
- Removal of stop words such as 'a', 'an', 'the' etc.
- Lemmatization: perform lemmatization which is the process of grouping together derivationally related words with similar meaning by morphological analysis.
- Stemming: Perform Stemming on the list of tokens obtained from the previous step which is the process of bring a word to its root format by removing the suffixes and prefixes

Step 3: Apply correlation based feature selection approach on the pre-processed data to reduce dimensionality and select only relevant feature words from the data. Let us consider a subset S consisting of k features, the CFS for which is defined below.

$$CFS = max_{sk}\left[\frac{r_{cf1} + r_{cf2} + \cdots + r_{cfk}}{\sqrt{k + 2\left(r_{f1f2} + \cdots + r_{fifj} + \cdots + r_{fkfk-1}\right)}}\right]$$

…Equation (2)

Step 4: Find Probability distribution of the tokens with selected feature using Naïve Bayes Approach. The formula for calculating Probability distribution is given below

$$P\left(y\middle|(f1, f2, f3, \ldots. fn)\right) = \frac{P((f1, f2, f3, \ldots fn)|y)\, P(y)}{P((f1, f2, f3, \ldots. fn))}$$

…Equation (3)

Here, *f* is feature vector set *(f1,f2,f3,... fn)*, y is define as class variable form- possible outcomes. *P(y | x)* means posterior probability and *P(y | x)* is dependent for any particular class of *P(x | y)* ,*P(x)* is evidence depending on feature, *P(y)* is prior probability.

Step 5: Our next step is to apply PSO approach to optimize the above outcome.

- Let us consider all the tokens as particles. Initially these particles will randomly fly and search for food source as the best feature match for tokens and then it will search for Local and Global solution.
- The performance of each particle will be dependent on the similarity from which feature has to be optimized.
- Here each particles flies over n- dimensional search space and will update the following information:
  - *Xi* - current position of particle x,
  - *Pi* - personal best position of particle x,
  - *Vi* - current velocity of the particle x.

- Velocity updates in PSO will be calculated as:

$$V_{i(t+1)} = \omega V_{i(t)} + c_1 r_1 (P_{i(t)} - X_{i(t)} + c_2 r_2 (P_g - X_{i(t)})$$

…Equation (4)

- Now Vi is new velocity, then the position of the particle updates with velocity as given below:

$$X_{i(t+1)} = X_{i(t)} + V_{i(t+1)}$$

…Equation (5)

- Last step in PSO is to update the position for each particle, and stores global best solution.

Step 6: On the basis of evaluated feature similarity using PSO, classification of tokens will be declared as spam or non-spam.

Step 7: At last, our final classification will be performed, in which we will be evaluating probability of spam or non-spam tokens in sentences.

- If the probability of spam token is more then consider it as spam email.
- Else consider email as non-spam.

Step 8: Finally, during this step we are going to store the e-mail as spam or non-spam, repeat the given procedure for all the emails.

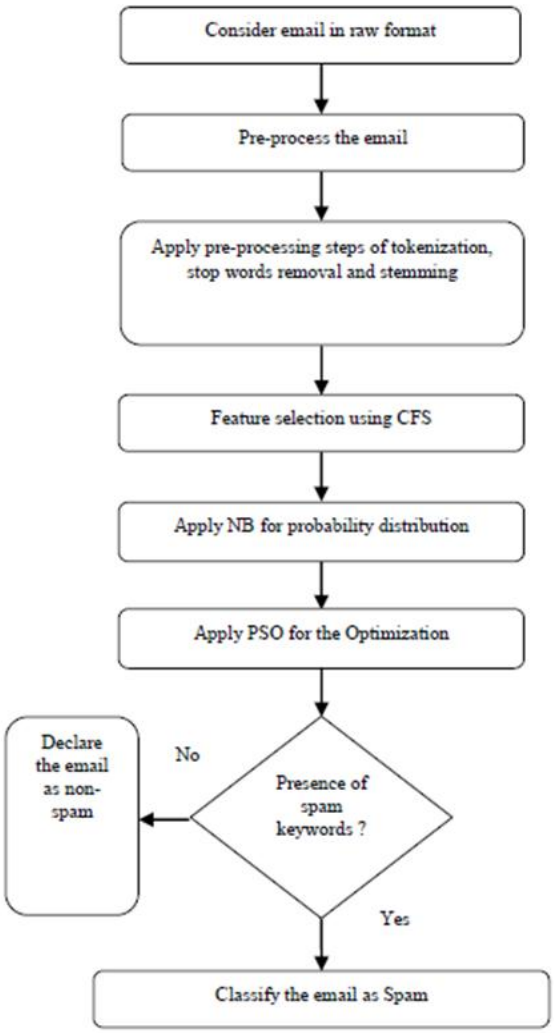

Figure 1: Flowchart of email spam detection

V.EXPERIMENTAL RESULTS AND DISCUSSION

In this section presents the used dataset, evaluate performance measures, evaluated results and with the individual Naive Bayes approach.

A. Dataset Used

The proposed concept was experimented on the dataset of ling Spam dataset. From the ling spam dataset 1500 arbitrarily chosen emails are used. From the given data 1500 emails, 900 emails are used for a training purpose, and 600 emails are used for a testing purpose by maintaining its ratio 60:40. out of 900 emails, 450 emails are spam and 450 are non-spam. In an equivalent manner, out of the 600 testing emails, 300 emails are spam and 300 are non-spam. Initially, training step is performed using 1) Naïve Bayes and 2) proposed integrated approach of Naïve Bayes and Particle Swarm Optimization. Then based on the testing emails results are evaluated on the basis of above technique.

B. Evaluation Parameters

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

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

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

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

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

Precision is that the magnitude relation of properly expected true positive spam email detection with true price to the entire expected true positive observation. It also defines the effectiveness of classifier. It is formulated as:

$$P = \frac{TP}{TP+FP} \quad \text{…Equation (6)}$$

Recall is the ratio of correctly predicted true positive spam email observation to the allobservation in actual email spam. It also defines the sensitivity of classifier. It is formulated as:

$$R = \frac{TP}{TP+FN} \quad \text{…Equation (7)}$$

F-Measure can be defined as the overall performance of the classifier. It is evaluated from the Precision and Recall values as mentioned below:

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

Accuracy is outlined as the of positive expected values to total information values. It can be evaluated as:

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

C. Results and Comparison

On the basis of mentioned formulation presented above, values of precision, recall, f-measure, accuracy are evaluated. Here values of TP, TN, FP and FN are important as precision, recall, f-measure, accuracy are calculated based on values of TP, TN, FP and FN.

The calculated values of TP, TN, FP and FN using individual Naïve Bayes and integrated proposed concept are shown below in table II. Further calculated values of precision, recall, f-measure, accuracy for individual Naïve Bayes ad integrated proposed concept are shown below in table III.

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

EVALUATION MEASURES	NAÏVE BAYES	PROPOSED INTEGRATED CONCEPT
TP	174	185
TN	27	12
FP	179	194
FN	22	07

Table III: 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%

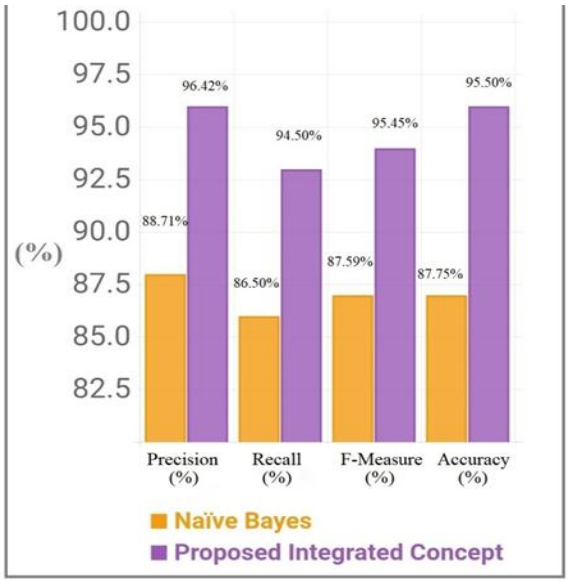

Figure 2: Comparison of proposed integrated concept with individual Naïve Bayes

From the comparison figure two, it will be seen that integrated approach of NB and PSO outperformed as compared with the individual NB approach. The classification accuracy for the individual NB approach lack with 7.75% from proposed integrated approach of NB and PSO concept. The main advantage of this integrated concept is the availability of optimization technique of PSO that have the power to optimize the solution with the global search solution space.

VI. CONCLUSION

Due to the increase in number of spam emails by the users email spam has become one of the most demanding research topics. Various methods are used by different authors for spam email classification, discussed in section 2. We have used the concept of Naïve Bayes and Particle Swarm Optimization for spam email detection. Naïve Bayes works on the concept of conditional probability and classifies the email as spam and non-spam based on the email content. To further optimize the parameters of the Naïve Bayes approach Particle Swarm Optimization is used, which results in increased the accuracy of the whole classification process. Correlation Based Feature selection is used for feature selection i.e. to select useful feature words from the email. The evaluation of the experiment is done on the basis of f-measure, precision, accuracy and recall. By evaluating the results, we can say that the integrated concept results in increased accuracy and precision than the individual Naïve Bayes approach. In future any other optimization algorithm can be used with Naïve Bayes algorithm. Also, any other ML approach can be used instead of NB approach.

REFERENCES

[1] Zhang, L., Zhu, J., Yao, T.: An evaluation of statistical spam filtering tech-niques. ACM Transactions on Asian Language Information Processing (TALIP) 3(4) (2004) 243–269.

[2] Androutsopoulos I., J. Koutsias, K. V. Chandrinos, G. Paliouras, and C. D. Spyropoulos, "An analysis of naive theorem antispam filtering", In: eleventh European Conference on Machine Learning, pp.9-17, Barcelona, Spain, 2000.

[3] Parsopoulos, Konstantinos E., ed. Particle swarm optimization and intelligence: advances and applications: advances an application. IGI global, 2010.

[4] KritiAgarwal, Tarun Kumar. "Spam Detection using integrated approach of naïve bayes and particle swarm optimization." In Proceeding of the second International Conference on Intelligent Computing and Control System (ICICC 2018) IEEE Xplore 2018.

[5] Harisinghaney, Anirudh, Aman Dixit, Saurabh Gupta, and Anuja Arora. "Text and image based mostly spam email classification using KNN, Naïve Bayes and Reverse DBSCAN algorithmic rule." In improvement, Reliabilty, 2014 International Conference on, pp. 153-155.

[6] 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.

[7] Renuka, Karthika D., and P. Visalakshi. "Latent semantic Indexing primarily based SVM Model for Email Spam Classification." (2014).

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

[9] 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.

[10] 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.

[11] 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.