A Survey on Email Spam Filtering Techniques

Svati Chaudhari[1], Dr. Mehul P. Barot[2]
[1]Dept of computer engineering
[2]Assistant Professor
[1, 2] L.D.R.P Institute of Technology & Research, KSV University - GANDHINAGAR

Abstract- *E-mail is most using a secure medium for online communication and transferring data or messages through the internet. Different approaches to data filtering exist, to automatically detect and delete unwanted emails. There are several numbers of email spam filtering approaches such as Knowledge-based techniques, Classification techniques, Clustering techniques, Learning-based techniques, machine learning algorithms, Heuristic processes, etc. In this paper, we are discussing the different spam filtering techniques such as naïve Bayes and the decision tree, which are the different parts of machine learning. However, here we try to gift the classification, evaluation, and comparison of various email spam filtering systems and summarize the scenario concerning the accuracy rate of various exiting approaches.*

Keywords- Naïve bayes, corpus, SCA, DBB-RDNN-REL

I. INTRODUCTION

Spam electronic mail is unauthorized and unwanted junk email dispatched out in bulk to an unknown recipient list. Typically unauthorized mail is dispatched for industrial purposes. It can be dispatched in large quantity through botnets, networks of contaminated computers. Email is one of the most important for communication. Private, corporation, corporate, and government contact are all examples of this form of communication. With the rapid increase in email usage, there has also been an increase in SPAM emails. SPAM emails also referred to as spam emails; contain virtually identical messages sent via email to multiple recipients. Apart from being annoying, spam emails can also pose a security threat to the computer system. It is reported that in 2007, spam cost corporations an order of $100 billion. We use text mining to perform automated spam filtering in this project to efficiently use emails. With data-mining classification algorithms, we try to recognize trends to enable us to classify the emails as HAM or SPAM.

1.1 Data Mining

Data mining denotes extracting or "mining" knowledge from large amounts of data. Many individuals give data mining as another word for another widely used word, Knowledge Discovery from Data, or KDD. Knowledge discovery as procedures illustrated in Figure and involves an iterative order of the following steps:

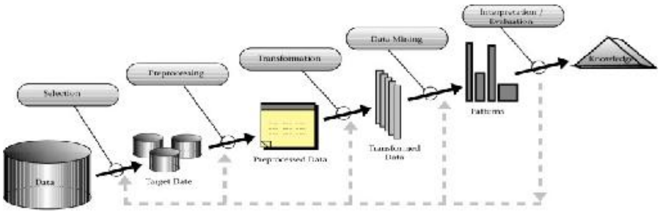


Figure 1.1 KDD process in Data Mining[13]

1. Data cleaning to eliminate noise and unreliable data.
2. Data integration where many data sources may be pooled.
3. Data selection where data applicable to the analysis task are reclaimed from the database.
4. Data transformation where data are transformed or fused into forms suitable for mining by accomplishing summary or aggregation operations.
5. Data mining an indispensable procedure where intelligent approaches are applied to extract data patterns.
6. Pattern evaluation to find the truly interesting patterns signifying knowledge grounded on some interestingness measures.
7. Knowledge presentation where visualization and knowledge representation methods are used to present the mined knowledge to the customer.

Steps 1 to 4 are various ways of preprocessing data, where the data is prepared for mining. The phase of data mining can communicate with the user or a knowledge base. The interesting patterns are presented to the user and may be stored as new knowledge in the knowledge base [2].

Machine Learning

Machine learning is a subfield of artificial intelligence (AI) [14]. The goal of machine learning is generally to understand the data structure and fit the knowledge into models that people can understand and manipulate. In its place, machine learning systems allow computers to train on data inputs and practice statistical analysis to generate values that fall within a specific range. As of this, machine learning enables computers in building models from sample data to mechanize decision-making procedures centered on data inputs. Two of the most commonly accepted approaches to machine learning are supervised learning, which trains algorithms based on human-labeled example input and output data, and unsupervised learning, which gives the algorithm no labeled data to allow it to find structure within its input data.

1. Supervised Learning

In supervised learning, the computer is delivered with sample inputs that are characterized by their preferred outputs. The algorithm can able to "learn" by equating its actual output with the "taught" outputs to detect errors and transform the model accordingly to evaluate this technique. Therefore, supervised learning uses patterns to guess label values on extra unlabeled results. A common use case of supervised learning is to utilize past data to predict statistically probable future happenings.

2. Unsupervised Learning

Since the data in unsupervised learning isn't classified, the learning algorithm is left to find commonalities among the data it's provided. As unlabeled data are ampler than labeled data, machine learning methods that assist unsupervised learning are principally valuable. The objective of unsupervised learning may be as direct as determining hidden patterns inside a dataset, but it may also have an objective of feature learning, which lets the computational machine mechanically learn the representations that are required to categorize raw data.

It is possible to divide learning approaches into linear and nonlinear approaches. Linear methods are easier, whereas the action of nonlinear approaches is more versatile. The approaches can be further categorized as classification- or regression-based methods for supervised learning. Regression-based methods fit the data to a continuous function and thus operate with continuous labels for the data, while classification-based methods aim to classify the data using discrete and categorical labels. For unsupervised learning, the methods are primarily categorized as clustering methods, which group the data into clusters based on underlying similarities [8].

II. LITERATURE REVIEW

In this paper, we discussed various literature searches on Spam Email. Classification is a function of data mining that assigns objects to target categories or classes in a set. For each case in the data, classification aims to accurately predict the target class.

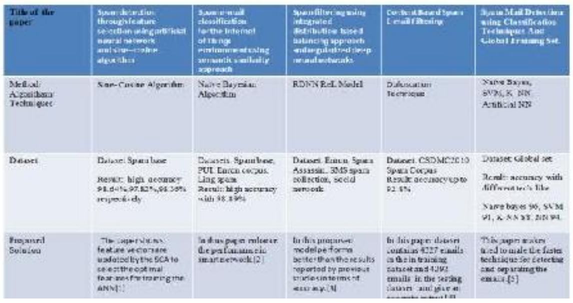
Rozita Talei pashiri, Yaser Rostami, MahramiMohsen have discussed that the SCA reduces Detection errors. To reduce the spam detection error, a feature selection-based method was provided in this paper using the sine–cosine algorithm (SCA). In the proposed process, the SCA updates the feature vectors to pick the best features to train the ANN. It gives the accuracy up to 98.36%.[1]

S. Venkatraman B. Surendiran P. Arun Raj Kumar has discussed in this paper is naive Bayesian algorithm to combat spam emails and enhance the performance in the smart network. To analyze the effectiveness of our approach, the experiments were conducted on benchmark data sets such as Spambase, PU1, Enron corpus, and Lingspam [2].

Aliaksandr Barushka, Petr Hajek has discussed in this paper is about that they compare the performance of the approach with state-of-the-art spam filters and uses several machine-learning algorithms to classify text and develop DBB-RDNN-REL.[3] As demonstrated on four benchmark spam datasets (Enron, Spam Assassin, SMS spam collection, and Social networking), the proposed approach enables capturing more complex features from high-dimensional data by additional layers of neurons. Another benefit of this method is that no additional reduction of dimensionality is required and a changed distribution-based algorithm is used to fix spam dataset imbalance.

Pingchuan Liu, Teng-Sheng Moh has discussed in this paper they proposed a content-based spam email filtering approach that uses keyword-based corpus. The system utilizes corpus based on keywords Built from training datasets for the classification of new Email Incoming Post. To enhance the accuracy of We came up with several different processes from our algorithm, to deal with obfuscated, negligible, or infrequent terms. We some experiments were conducted to test our proposed work. [4]

Vishal Kumar Singh, Shweta Bhardwaj has discussed in this paper all the machine learning and data mining techniques and also test that technique to make the classification able to separate the spam mail. In these training sets, thousands of samples are used to Make the spam mailable to be distinguished by the classifier. Even after this much effort, though, today, junk mail continues. They continue because a new form of spam mail is available every day. Presented. Therefore, even though we are sorted and labeled with old spam mail, new ones are kept He's moving in. [5]



Title of the paper	Spam detection through feature selection using artificial neural network and sine–cosine algorithm	Spam e-mail classification for the Internet of Things environment using semantic similarity approach	Spam filtering using integrated distribution based balancing approach and regularized deep neural networks	Content Based Spam E-mail Filtering	Spam Mail Detection using Classification Techniques And Global Training Set.
Method/ Algorithms/ Techniques	Sine–Cosine Algorithm	Naive Bayesian Algorithm	RDNN ReL Model	Bifurcation Technique	Naive Bayes, SVM, K- NN Artificial NN
Dataset	Dataset: Spam base Result: high accuracy 94.64%,97.82%,98.38% respectively	Datasets: Spambase, PUI, Enron corpus, Ling spam Result: high accuracy with 98.89%	Dataset: Enron, Spam Assassin, SMS spam collection, social network	Dataset: CSDMC2010 Spam Corpus Result: accuracy up to 92.8%	Dataset: Global set Result: accuracy with different tech's like Naive bayes 96, SVM 91, K-NN 88, NN 94.
Proposed Solution	The paper shows feature vectors are updated by the SCA to select the optimal features for training the ANN[1]	In this paper enhance the performance in smart network[2]	In this proposed model performs better than the results reported by previous studies in terms of accuracy.[3]	In this paper dataset contains 4227 emails in the in training dataset and 4293 emails in the testing dataset and give an accurate output[4]	This paper makes tried to make the faster technique for detecting and separating the emails.[5]

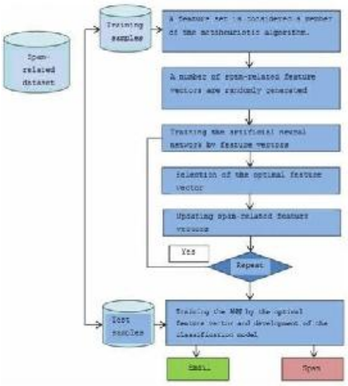
III. **OVERVIEW OF EMAIL SPAM**

My proposed work is on E-MAIL AND SMS SPAM FILTERING APPROACH USING DATA MINING TECHNIQUES and find out fake mail and SMS. So here I use the classification data mining techniques approach on the dataset to increasing accuracy. It can also decrease the error rates from the dataset.

Existing Method And Architecture:

In this method, they were using the metaheuristic algorithm. As an input, several spam-related feature vectors are randomly generated. Train that input data using the artificial neural network by feature vectors. After training the data are trained data divide into the imported emails and spam emails.

The spam detection error relies on varied factors, the most vital of that are ANN inputs and choice of vital options best representing the spam and emails. Using all the options for coaching the ANN will increase the error given that every feature contains a different level of importance. On the other hand, together with all the options will increase each the problem and information size, resulting in a rise within the execution time. A metaheuristic algorithm rule just like the framework can be used for feature choice to cut back or alleviate theseverity of those 2 issues for coaching the ANN. Themetaheuristic algorithmic rule is, in fact, liable for change the feature vectors, whereas the ANN is employed for coaching and leaming.


Spam-related dataset
Training samples
A feature set is considered a member of the metaheuristic algorithm.
A number of spam-related feature vectors are randomly generated
Training the artificial neural network by feature vectors
Selection of the optimal feature vector
Updating spam-related feature vectors
Yes
Repeat
Test samples
Training the ANN by the optimal feature vector and development of the classification model
Email
Spam

Sine-cosine Algorithm:

SCA is a global optimization method that iteratively updates a set of candidate solutions using a mathematical model based on the sine and cosine functions. In the complex problem of nonlinear optimization, this algorithm has shown good efficiency. Seyedali Mirjalili (Mirjalili, 2016) presented the Sine Cosine Algorithm (SCA) as one of the most recent MHAs built in the literature for solving optimization problems. SCA is a mathematical technique for evaluating the best solution by simulating the action of sine and cosine functions (destination). SCA is a mathematical technique for evaluating the best solution by simulating the action of sine and cosine functions (destination).

Artificial neural network

An artificial neural network (ANN) is a computer system designed to simulate how information is analyzed and processed by the human brain. It is the basis of artificial intelligence (AI) and solves problems that, by human or statistical standards, would prove impossible or difficult. An artificial neural network (ANN) is an attempt to replicate the network of neurons that make up a human brain for a computer to learn new things and make decisions in a human-like manner. By programming standard computers to behave as if they are interconnected brain cells, ANNs are formed.

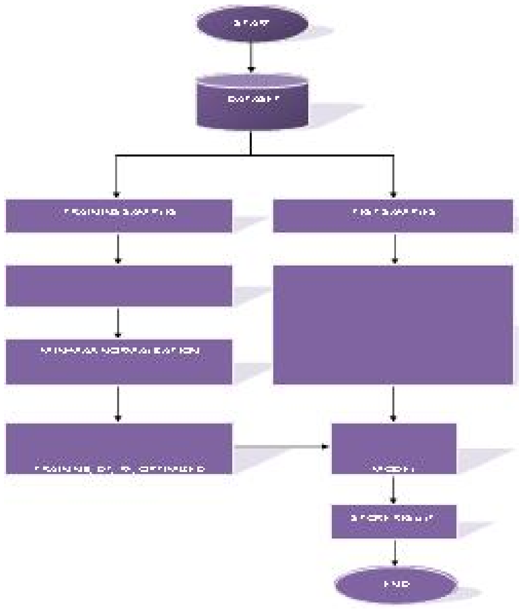
Metaheuristic algorithm

Metaheuristics are higher-level heuristics that control the entire search process such that global optimal solutions can be sought systematically and efficiently. While metaheuristics can not always guarantee that the true global optimal solution is obtained, in many practical problems they can provide very good results. A metaheuristic is a higher-level procedure or heuristic that is used to find, produce, or select a heuristic that can provide a good solution to an optimization problem, particularly when the data is incomplete or the computational power is limited. Metaheuristics sample a subset of solutions that would otherwise be too broad to be entirely enumerated or discussed in any other way. Metaheuristics can be applied to a wide range of problems because they make few assumptions about the optimization problem at hand.

IV. OVERVIEW OF PROPOSED METHOD

proposed method architecture

In my proposed algorithm I'll try to give the higher accuracy from the existing algorithms. In the algorithm, the dataset is divided into two parts training and testing. The training data is 70% and the testing data is 30%. the trained 70%data train the left 30% data and gave the better result. For those results, I used many different methods like naïve Bayes, random forest, decision tree. And also I comparing this method after the implementation.



Algorithm Steps for the method:

STEP:-1 Take Input from Dataset.
STEP:-2 Data-preprocessing from Dataset.
STEP:-3Divide Training (70%) and Testing (30%) data from Dataset.
STEP:-4 Use the SMOT technique on Dataset. STEP:-5 Apply MIN-MAX Normalization on
Dataset.
STEP:-6 Train Model using NB, DT, RF, Optimized XG-BOOSTER Classifier.
STEP:-7 Model Evaluation. STEP:-8 Store Result.

V. CONCLUSION

This survey paper elaborates on different Existing Spam Filtering systems through Machine learning techniques by exploring many ways, concluding the summary of many Spam Filtering techniques, and summarizing the accuracy of various planned approaches concerning many parameters. Moreover, all the existing ways area unit effective for email spam filtering. Some have an effective outcome and a few area units trying to implement another method for increasing their accuracy rate. Tho' all area unit effective however still currently spam filtering system has some lacking that area unit is the major concept for researchers and that they are attempting to generate a next-generation spam filtering method that has the power to think about a sizable amount of multimedia system data and filter the spam email additionally conspicuously.

REFERENCES

[1] Pashiri, Rozita Talaei, Yaser Rostami, and Mohsen Mahrami. "Spam detection through feature selection using artificial neural network and sine–cosine algorithm." *Mathematical Sciences* 14.3 (2020): 193-199.

[2] Venkatraman, S., B. Surendiran, and P. Arun Raj Kumar. "Spam e-mail classification for the internet of things environment using semantic similarity approach." *The Journal of Supercomputing* 76.2 (2020): 756-776.

[3] Barushka, Aliaksandr, and Petr Hajek. "Spam filtering using integrated distribution-based balancing approach and regularized deep neural networks." *Applied Intelligence* 48.10 (2018): 3538-3556.

[4] Liu, Pingchuan, and Teng-Sheng Moh. "Content-based spam e-mail filtering." *2016 International Conference on Collaboration Technologies and Systems (CTS)*. IEEE, 2016.

[5] Singh, Vishal Kumar, and Shweta Bhardwaj. "Spam mail detection using classification techniques and global training set." *Intelligent Computing and Information and Communication*. Springer, Singapore, 2018. 623-632.

[6] Abdulhamid, Shafi'I. Muhammad, et al. "Comparative Analysis of Classification Algorithms for Email Spam Detection." *International Journal of Computer Network & Information Security* 10.1 (2018).

[7] Woitaszek, Matthew, Muhammad Shaaban, and Roy Czernikowski. "Identifying junk electronic mail in Microsoft outlook with a support vector machine." *2003 Symposium on Applications and the Internet, 2003. Proceedings.* IEEE, 2003.

[8] Rathod, Sunil B., and Tareek M. Pattewar. "Content-based spam detection in the email using Bayesian classifier." *2015 International Conference on*

Communications and Signal Processing (ICCSP). IEEE, 2015.

[9] . Ferrara, E.: The history of digital spam. arXiv preprint arXiv :1908.06173 (2019)

[10] Ren, Y., Ji, D.: Learning to detect deceptive opinion spam: a survey. IEEE Access 7, 42934–42945 (2019)

[11] Broadhurst, R., Trivedi, H.: Malware in spam email: trends in the 2016 Australian Spam Intelligence Data. Available at SSRN 3413442 (2018)

[12] 2014 Internet Security Threat Report, Volume 19; Available from:http://www.symantec.com/content/en/us/enterprise/other_resou rces/b-istr_main_report_v19_21291018.en-us.pdf.

[13] The economic of Spam, 2012.Available from: http://pubs.aeaweb.org/doi/pdfplus/10.1257/jep.26.3.87.

[14] A. Qaroush, I. M.khater, M. Washaha, “Identifying spam e-mail based-on statistical header features and sender behavior,” Proceedings of the CUBE International Information Technology Conference, pp. 771–778, 2012.

[15] Parhat Parveen, Prof. Gambhir Halse, “Spam mail detection using classification”, International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 6, June 2016.

[16] Megha Rathi, Vikas Pareek, “Spam mail detection through data mining-A comparative performance analysis”, I.J. Modern Education and Computer Science, May 2013.

[17] R. Malarvizhi, K. Saraswathi, Research scholar, PG & Research, Department of Computer Science, Government Arts College, “Content-Based Spam Filtering and Detection Algorithms- An Efficient Analysis & Comparison”, International Journal of Engineering Trends and Technology, Volume 4, Issue 9, Sep 2013. 5. Vinod Patidar, Divakar Singh, “A Survey on Machine Lear