iJRASET
International Journal For Research in Applied Science and Engineering Technology



# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

**Volume:** 11 **Issue:** III **Month of publication:** March 2023

**DOI:** https://doi.org/10.22214/ijraset.2023.49529

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538*
*Volume 11 Issue III Mar 2023- Available at www.ijraset.com*

# A Comprehensive Overview on Intelligent Spam Email Detection

Jayasree Nallabariki[1], T. Keerthi Chandana[2], D Sai Tejaswi[3], Dr. M Y Babu[4]
*Department of CSE, Vardhaman College of Engineering, Hyderabad, India*

***Abstract:*** ***Spam, usually referred to as unsolicited commercial or bulk e-mail, has recently become a major issue on the internet. Time, storage, and transmission bandwidth are all wasted by spam. Spam email has been a growing issue for years. Nowadays, automatic email filtering appears to be the most successful strategy for preventing spam. Only several years ago most of the spam could be reliably dealt with by blocking e-mails coming from certain addresses or filtering out messages with certain subject lines. Spammers started employing a number of cunning strategies to get beyond filtering techniques, such as utilizing random sender addresses and/or adding random characters to the message subject line's beginning or conclusion. Machine learning techniques now a days are used to automatically filter the spam e-mail in a very successful rate. Machine learning field is a subfield from the broad field of artificial intelligence, this aims to make machines able to learn like human. Understanding, observing, and providing knowledge about a statistical occurrence are all terms used here. In the first place, data collection and representation are typically problem-specific (i.e., for email messages), and in the second place, e-mail feature selection and feature reduction aim to lower the dimensionality (i.e. the number of features).Finally, the e-mail classification phase of the process finds the actual mapping between training set and testing set. Machine Learning approach includes lots of algorithms that can be used in e-mail filtering like Naïve Bayes, K-nearest neighbour, Support VSector Machine, classifiers. In conclusion, we try to summarize the performance results of the few machine learning methods in terms of spam precision and accuracy.***
***Keywords: Naive Bayes, K-nearest neighbour, Support Vector Machine, Spam, Ham.***

## I. INTRODUCTION

Our daily lives now frequently involve the Internet. The firm suffers financial consequences and the recipient user is irritated when the same communication is sent again. In this project, a Spam Mail Detection system is proposed will classify the given email as spam or ham email. Spam filtering mainly focuses on the content of the message. Based on its content, the categorization system assigns a category to the provided email. Feature extraction and selection plays a vital role in the classification. Email data is gathered using the dataset in order to detect spam mail. To obtain the accurate results, data needs to be pre-processed by removing stop words and word tokenization. Using the TF-IDF Vectorizer module, preprocessing of the data is carried out. SVM algorithm is used to detect the given email is spam or harm. Spam, often known as unsolicited commercial bulk emails, has recently grown to be a major problem online. Because the sender of the spam mail is known, it is reported. Such a person collects email addresses from a variety of websites, chat groups, and computer infections. Spam prevents the user from creating full and sensible use of your time, storage capability and network information measure. the massive volume of spam mails flowing through the pc networks have damaging effects on the memory house of email servers, communication information measure, central processing unit power and user time. Over seventy-seven percent of all international email traffic is attributable to the spam email problem, which is growing yearly .Users United Nations agency realises how annoying it is to receive spam emails that they didn't ask for. It has also caused significant losses for a number of users. Various spammers' dishonest tactics and web scams have affected United Nations organisations. In order to persuade people to reveal sensitive personal information like passwords, Bank Verification Numbers (BVN), and Mastercard numbers, organisations send emails purporting to be from respectable businesses.

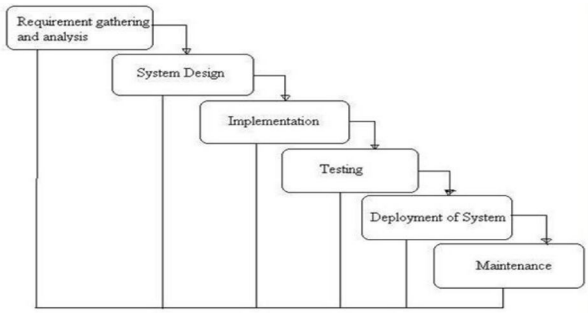
## II. RELATED WORK

To effectively handle the threat expose by email spams, leading email suppliers like Gmail, Yahoo mail and Outlook have utilized the mixture of various machine learning (ML) techniques like Neural Networks in its spam filters. These cc unit} techniques have the capacity to be told and establish spam mails and phishing messages by analyzing many such messages throughout a massive assortment of computers. Since machine learning have the capability to adapt to variable conditions, Gmail and Yahoo mail spam filters do over simply checking junk emails victimization pre-existing rules. They generate new rules themselves supported what they need to learn as they continue in their spam filtering operation.

©IJRASET: All Rights are Reserved 

The machine learning model utilized by Google have currently advanced to the purpose that it will observe and separate spam and phishing emails with regarding ninety nine.9 % accuracy. The implication of this is often that one out of m messages reach evading their email spam filter. Statistics from Google discovered that between 50-70 % of emails that Gmail receives area unit direct mail. Google's detection models have conjointly incorporated tools referred to as Google Safe Browsing for distinctive websites that have malicious URLs. The phishing-detection performance of Google are increased by introduction of a system that delay the delivery of some Gmail messages for a short while to hold out further comprehensive scrutiny of the phishing messages since {they area unit|they're} easier to observe after they are analyzed and put together. The aim of delaying the delivery of a number of these suspicious emails is to conduct a deeper examination whereas a lot of messages arrive in due course of your time and therefore the algorithms are updated in real time. solely regarding zero.05 % of emails are unit plagued by this deliberate delay.

## III. MODEL BUILDING




### A. *Requirement Gathering and Analysis*

It's the first and most important stage of any project because ours is an academic leave for requirements amassing. We followed IEEE journals and gathered a lot of IEEE relegated papers before selecting one titled "individual web revisitation by setting and substance importance input." For the analysis stage, we used references from the paper and conducted a literature review of a few papers to gather all the project's requirements in the stage.

### B. *System Design*

This phase studies the need specifications from the first phase and prepares the system design. System design aids in determining the overall system architecture as well as the hardware and system requirements. In the current phase, the software code is being created.

### C. *Implementation*

The system is first built as small programmes known as units, which are then incorporated into the following phase, with input from the system design. Unit testing is the process of developing and evaluating each unit for functionality.

### D. *Testing*

Integration Testing: Following the testing of each unit created during the implementation phase, the entire system is integrated. Constant software testing is required to check for bugs and mistakes in the developed programme. Testing is carried out to ensure that the client has no issues installing the software.

### E. *Deployment of System*

After functional and non-functional testing is complete, the product is deployed in the client environment or made available for purchase.

### F. *Maintenance*

After installation, there is a maintenance phase where changes are made to the system or a specific component to improve performance or change characteristics. These changes result from either customer-initiated change requests or flaws found when the system is being used in real life. The developed software is regularly maintained and supported for the client.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538*
*Volume 11 Issue III Mar 2023- Available at www.ijraset.com*

## IV. METHODOLOGY

### *A. Proposed System Diagram*

Spam Mail Detection is used to differentiate between spam and ham emails. This method is accomplished by using Support Vector Machine(SVM), KNN, Naive bayes algorithm. Dataset is separated into two sets depending on labels and sent into the algorithm using this way. The system compares all three methods and predicts which one has the highest accuracy. For spam identification, we employ the most accurate algorithm available.

### *B. Datasets*

This is a csv file with associated data from 5172 randomly selected email files, together with the labels for each file's spam or not-spam classification.

Each row in the 5172-row csv file represents a single email. 3002 columns are present. Name of the email is shown in the first column. To safeguard privacy, the name has been specified with numbers rather than the receivers' names. The labels for prediction are in the final column and are 1 for spam and 0 for not spam. After removing the non-alphabetical characters and words, the remaining 3000 columns represent the 3000 most prevalent words across all emails. The corresponding cells for each row contain the number of words in each column of the corresponding email for that row. As a result, rather than storing the data for each of the 5172 emails separately, it is saved in a single compact dataframe.

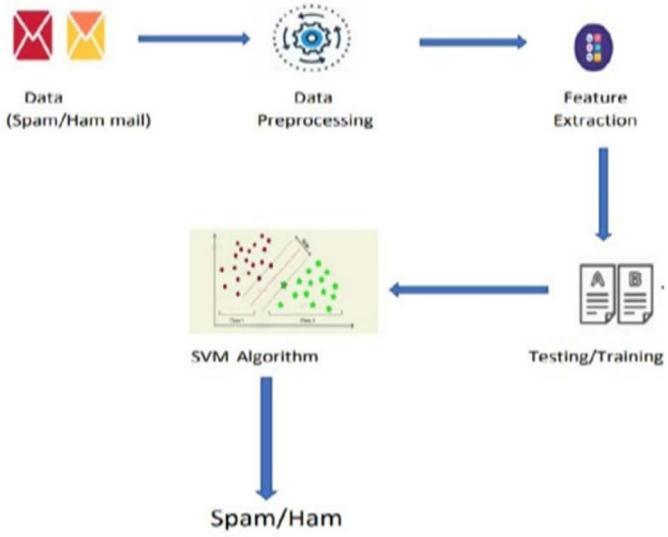



Figure3: Architecture Diagram

### *C. Algorithm Selection*

#### *1) Nearest Neighbour Algorithm*

KNN is a slow supervised learning algorithm; it requires more time to train than other algorithms, which divide training from data into two steps and testing into a third. The foundation of the K Nearest Neighbor algorithm is the neighbouring data points' weights being assigned. For training datasets, K Nearest Neighbor distance is calculated. Currently, classification of each of the K Nearby data points is done using the majority of votes. In KNN, three different lengths must be measured: Euclidian, Manhattan, and Minkowski distances. using the following formula, one may determine which Euclidian will be considered to be the most one. distance.

The steps listed below define the KNN algorithm:

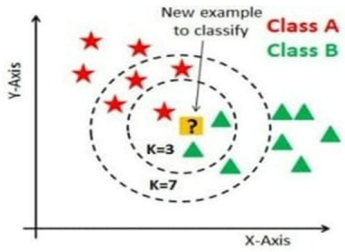
*a)* D stands for the training samples, and k for the number of nearest neighbours.
*b)* For each sample class, create a superclass.
*c)* Each training sample's Euclidian distance should be calculated.
*d)* categorise the sample based on the dominant class in the neighbourhood

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

$$Eucledian\ Distance = D(x, y) = (x_i - y_i)_{2k_i} = 1$$

K=number of cluster

x , y=co-ordinate sample spaces




*2) Naive Bayes Algorithm*

Naïve Bayes classifier is based on Bayes theorem. It has strong independence assumption. It is also known as independent feature model. It assumes the presence or absence of a particular feature of a class is unrelated to the presence or absence of any other feature in the given class. Naïve bayes classifier can be trained in supervised learning setting. It uses the method of maximum similarity. It has been worked in complex real world situation. It requires small amount of training data. It estimates parameters for classification. For each class, only the variance of the variable needs to be determined, not the complete matrix. Inputs are typically high when using naive bayes. It gives ouput in more sophisticated form. The probability of each input attribute is shown from the predictable state. Machine learning and data mining methods are based on naïve bayes classification.

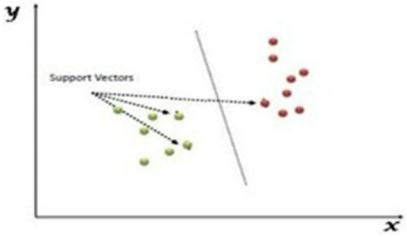
Bayes Theorem

$$P(H|X) = \frac{P(X|H)\ P(H)}{P(X)}$$

Where P(H|X ) is posterior probability of H conditioned on X
P(X|H) is posterior probability of X conditioned on H
P(H)is prior probability of H P(X) is prior probability of X

*3) Support Vector Machine (SVM)*

Support Vector Machine is a very common supervised machine learning method that can function as both a classifier and a predictor (it has a pre-defined target variable). It locates a hyper-plane in the feature space that distinguishes between the classes for classification. An SVM model depicts the training data points as points in the feature space that are mapped in a way that various classes are separated from one another by a maximum achievable margin. In the same space, the test data points are then mapped, and they are categorised according to which side of the margin they fall.




*D. Training and Testing*

Our email spam detector will be trained to identify and classify spam emails using a train-test split method. The train-test split is a method for assessing how well a machine learning algorithm is working. Any supervised learning method can use it for classification or regression. A dataset is split into two independent datasets as part of the operation. The training dataset is the first dataset that is used to fit the model. We give the model's input element for the second dataset, the test dataset. Lastly, we create predictions and contrast them with the actual results.

Train dataset: Used to fit the machine learning model.

Test dataset: used to gauge how well the machine learning model fits the data.
With actual data with known inputs and outputs, we would fit the model. Then, since we lack the target values or expected output for the new cases, we would base our projections on those examples. We will use the information from our sample. using the labels spam and ham, respectively, the csv file contains instances that have been pre-classified as spam and non-spam.We'll employ the train test split() technique from scikit-learn to divide the data into our two datasets.

## V. PERFORMANCE EVALUATION

### *A. Learning Rate*

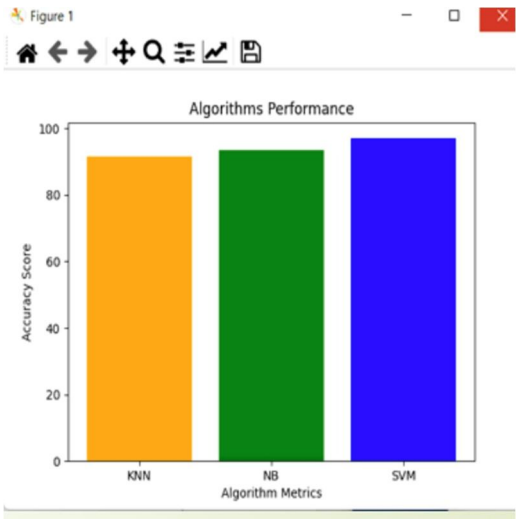
Since last few decades, researchers are trying to make email as a secure medium. Spam filtering is one of the core features to secure email platform. Regarding this several types of research have been progressed reportedly but still there are some untapped potentials. Over time, still now e-mail spam classification is one of the major areas of research to bridge the gaps. Therefore, a large number of researches already have been performed on email spam classification using several techniques to make email more efficient to the users. That's why, this paper tried to arrange the summarized version of various existing Machine Learning approaches. In addition, in order to evaluates the most of the approaches like Naive Bayes , SVM , kNN used reliable and well known dataset for benchmarking performance such as SpamData , The Spam Assassin , The Spambase, Ecml-pkdd 2006 challenge dataset , PU corpora dataset , Enron dataset ,Trec 2005 dataset . Some of these dataset are in a prepared structure e.g. ECML and data accessible in Spambase UCI archive . Among them, some of the classifiers also used novel methods applied in the feature selection for improving classification.

### *B. Result and Outcome*

Data collection: Data from publicly available sources must be gathered in order to train models.
Pre-Processing: Data must be pre-processed in accordance with models in order to improve the model's correctness and provide greater information about the data.
Feature extraction − It is used to extract the features from data to define the attributes in image and text data.
Feature selection − It is used to identify useful attributes to create supervised models.

Results

| | Algorithm | Accuracy | Precision | Recall | F1_Score |
|---|---|---|---|---|---|
| 1 | KNN | 91.50717703349282 | 95.50063371356148 | 69.91525423728814 | 76.12917496129174 |
| 2 | NB | 93.60047846889952 | 96.53272845106935 | 77.33050847457628 | 83.54656108742935 |
| 3 | SVM | 97.00956937799043 | 98.08309915696508 | 89.58382512629244 | 93.247609201344 |

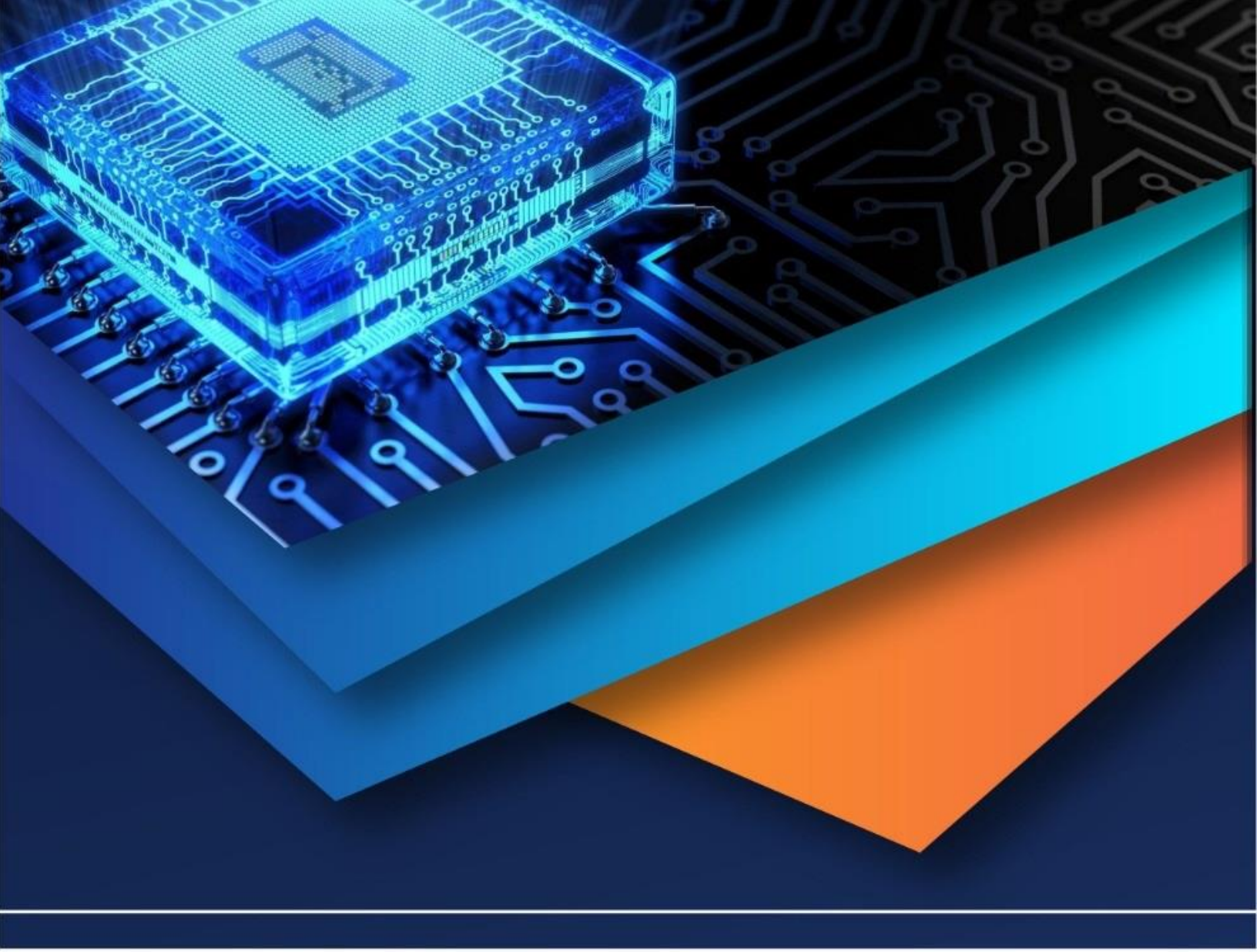


## VI. CONCLUSION

In this study, we have a tendency to reviewed machine learning approaches and their application to the sector of spam filtering. A review of the state of the art algorithms been applied for classification of messages as either spam or ham is provided. The makes an attempt created by totally different researchers to finding the matter of spam through the utilization of machine learning classifiers was mentioned. The evolution of spam messages over the years to evade filters was examined. the essential design of email spam filter and therefore the processes concerned in filtering spam emails were looked into. The paper surveyed a number of the in public accessible datasets and performance metrics that may be wont to live the effectiveness of any spam filter. The challenges of the machine learning algorithms in expeditiously handling the menace of spam was found out and comparative studies of the machine learning technics accessible in literature was done. we have a tendency to additionally disclosed some open analysis issues related to spam filters. In general, the figure and volume of literature we have a tendency to reviewed shows that vital progress are created and can still be created during this field. Having mentioned the open issues in spam filtering, more analysis to reinforce the effectiveness of spam filters got to be done. this may create the event of spam filters to still be an energetic analysis field for academician and business practitioners researching machine learning techniques for effective spam filtering. Our hope is that analysis students can use this paper as a spring board for doing qualitative analysis in spam filtering mistreatment machine learning, deep leaning and deep adversarial learning algorithms. The overall accuracy of the results achieved are 99.9% accuracy on training data and 98.2% on testing data with less false positive rate. It shows that classifiers give better with training data and less compared to testing data. It has also been further observed that the proposed system has the least percent error and hence can be deemed the most accurate method. The future enhancement will be to extend this design to take into account more attributes that could classifies the emails using images and also including different datasets into training the algorithm into producing more accurate results.

## REFERENCES

[1] M. Awad, M. Foqaha Email spam classification using hybrid approach of RBF neural networkand particle swarm optimizationInt. J. Netw. Secur. Appl., 8 (4) (2016)
[2] D.M. Fonseca, O.H. Fazzion, E. Cunha, I. Las-Casas, P.D. Guedes, W. Meira, M. Chaves
[3] Measuring characterizing, and avoiding spam traffic costs IEEE Int. Comp., 99 (2016).
[4] Visited on May 15, 2017 Kaspersky Lab Spam Report (2017 )
[5] 2012 https://www.securelist.com/en/analysis/204792230/Spam_Report_April_2012.
[6] E.M. Bahgat, S. Rady, W. GadAn e-mail filtering approach using classification techniques The1st International Conference on Advanced Intelligent System and Informatics (AISI2015), November 28-30, 2015, Springer International Publishing, BeniSuef, Egypt (2016), pp. 321-331CrossRefView Record in Scopus
[7] N. Bouguila, O. Amayri, A discrete mixture-based kernel for SVMs: application to spam andimage categorization Inf. Process. Manag., 45 (6) (2009), pp. 631-642,
[8] Y. Cao, X. Liao, Y. Li An e-mail filtering approach using neural network InternationalSymposium on Neural Networks, Springer Berlin Heidelberg (2004), pp. 688-694.
[9] F. Fdez-Riverola, E.L. Iglesias, F. Diaz, J.R. Méndez, J.M. CorchadoSpamHunting: an instancebased reasoning system for spam labelling and filtering Decis. Support Syst., 43 (3) (2007), pp.722-736
[10] S. Mason New Law Designed to Limit Amount of Spam in E-Mail (2003) http://www.wral.com/technolog.
[11] I. Stuart, S.H. Cha, C. Tappert A neural network classifier for junk e-mail Document AnalysisSystems VI, Springer Berlin Heidelberg (2004), pp. 442-450.
[12] J. Han, M. Kamber, J. Pei Data Mining: Concepts and Techniques Elsevier (2011)
[13] S.N. Qasem, S.M. Shamsuddin, A.M. Zain Multi-objective hybrid algorithms for radial basisfunction neural network design Knowl. Based Syst., 27 (2012), pp. 475
[14] J.D. Schaffer, D. Whitley, L. Eshelman Combinations of genetic algorithms and neuralnetworks: a survey of the state of the artCombinations of Genetic Algorithms and Neural Networks(1992), pp. 1-37.
[15] E. Elbeltagi, T. Hegazy, D. Grierson Comparison among five evolutionary-based optimization algorithms Adv. Eng. Inf., 19 (2005), pp. 43-53.
[16] L.H. Gomes, C. Cazita, J.M. Almeida, V. Almeida, W.J. Meira Workload models of spam and legitimate e-mails Perform. Eval, 64 (7–8) (2007), pp. 690-714.
[17] C.C. Wang, S.Y. Chen Using header session messages to anti-spamming Comput. Secur., 26 (5) (2007), pp. 381-390.
[18] T.S. Guzella, W.M. Caminhas, A review of machine learning approaches to spam filtering, Expert Syst. Appl., 36 (7) (2009), pp. 10206-10222.
[19] C.P. LuegFrom spam filtering to information retrieval and back: seeking conceptual foundations for spam filtering Proc. Assoc. Inf. Sci. Technol., 42 (1) (2005).
[20] X.L. Wang Learning to classify email: a survey 2005 International Conference on Machine Learning and Cybernetics (Vol. 9, pp. 5716-5719), IEEE (Aug 2005)



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)