I.J. Information Engineering and Electronic Business, **2022, 6, 27-36**
Published Online on December 8, 2022 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijieeb.2022.06.03

An E-mail Spam Detection using Stacking and Voting Classification Methodologies

Aasha Singh*
KNIT, Sultanpur, India
E-mail: researchcse19@gmail.com
ORCID iD: https://orcid.org/0000-0001-9002-1494
*Corresponding Author

Awadhesh Kumar
KNIT, Sultanpur, India
E-mail: awadhesh@knit.ac.in
ORCID iD: https://orcid.org/0000-0001-9055-2500

Ajay Kumar Bharti
BBDU, Lucknow, India
E-mail: ajay_bharti@hotmail.com
ORCID iD: https://orcid.org/0000-0001-6879-5151

Vaishali Singh
MUIT, Lucknow, India
E-mail: singh.vaishali05@gmail.com
ORCID iD: https://orcid.org/0000-0001-8304-8947

Received: 30 June, 2022; Revised: 05 August, 2022; Accepted: 20 September, 2022; Published: 08 December, 2022

Abstract: Nowadays, we use emails almost in every field; there is not a single day, hour, or minute when emails are not used by people worldwide. Emails can be categorized into two types: ham and spam. Hams are useful emails, while spam is junk or unwanted emails. Spam emails may carry some unwanted, harmful information or viruses with them, which might harm user privacy. Spam mails are used to harm people by wasting their time and energy and stealing valuable information. Due to increasing in spam emails rapidly, spam detection and filtering are the prominent problems that need to be solved. This paper discusses various machine learning models like Naïve Bayes, Support Vector Machine, Decision Tree, Extra Decision Tree, Linear regression., and surveys about these machine learning techniques for email spam detection in terms of their accuracy and precision. In this paper, a comprehensive comparison of these techniques and stacking of different algorithms is also made based on their speed, accuracy, and precision performance.

Index Terms: Email, Spam, SVM, Linear Regression, Stacking, Voting.

1. Introduction

Implementing spam filtering is of the utmost importance to any organization. Spam filtering is a method to keep unwanted, useless mails out of our inbox, which causes unnecessary inconvenience for the one using the mail service. The biggest disadvantage of these spam emails is that they might also carry some bugs or malware that may adversely affect our system's security. Spammers nowadays may also ask for ransoms in exchange for the critical information they have stolen.[13]

1.1 Email spam Filtering

Email spam filtering is the process of classifying a mail as spam or ham (mails that are not spam). Email spam filtering is necessary today as E mail service is in huge demand. It has been used for official purposes in the business and government sector and is popular among the common public for conveying their messages to their friends and

family. As emails are becoming essential to everybody's life, anyone with a good internet connection can receive useless emails. These junk mails divert a person's mind from important and useful mail, which can sometimes lead to harmful situations [14]. These spam emails fill up the memory and might instigate to play on money for the lottery, loans, jobs. These spam mails steal our system's helpful information like our contact details, account details, and security password. With all the harmful impacts on our privacy, email spam filtering is the need of the hour. For this, various machine learning algorithms, such as Naive Bayes, Support vector machine, random forest decision Tree, etc. [15], can be used.

1.2 SMS Spam Filtering

SMS is the service provided on our mobile phones. This service was top-rated among people before the advancement of the internet and its accessibility. SMS is very useful in conveying one's message from one person to another through a mobile where the internet is scarce, or the internet is facing issues in terms of speed. SMS service is handy and does not need any extra platform to fulfill its services. Hence it attracted a large number of spammers. Each day we receive spam messages on our mobile phones regarding some lottery, loans, money making, work from home. [14].

1.3 Social Network Spam Filtering

Online Social Platforms (OSNs) are getting very popular among all generations. As the internet is getting advanced day by day, these social platforms are reaching more audiences. These social platforms are very convenient to use in terms of sharing messages, pictures, and videos. However, with popularity comes inevitable consequences. A large number of users and convenient conditions for connecting with people worldwide have also attracted many spammers that take advantage of the people. According to certain reports, spam messages in the past had limited impact, but now due to large and easy connectivity [21], spammers are creating a large, distributed impact among the common public. Regarding spam filtering, we can say that spam filtering is one of the trending topics in the field of research in machine learning. Identifying which emails are spam or ham is a difficult task. We need to implement different feature extraction and training algorithms to generate an optimized solution.

1.4 Aim and Objective

Email Spam Filtering using ML is that though there are various spam filters already available in the market, there is always a chance of improvement in any system. Each day new findings and research work are being published. The main objective of this work is to develop a methodology that uses the existing technologies but implements something new that gives rise to a more optimized solution to the problem than existing systems are giving. This research paper discusses the various spam filtering algorithms' accuracy on the dataset used. Stacking of different algorithms is also done for an optimal result. Dataset is a CSV file comprising of 5575 mails which are categorized into spam and ham. The algorithm used makes use of this dataset to train the model and predict a new upcoming email if it belongs to ham category or spam category.

2. Literature Review

The research papers that discussed various machine learning algorithms for spam filtering have been thoroughly studied and analyzed.

While studying different papers and articles, some exciting results were found, which helped in comprehension of our proposed methodology for spam detection better the concept and role of machine learning in this advancing field of email spam detection and filter. Research papers regarding various feature extraction algorithms have been paid attention to and analyzed well. Dataset is a CSV file comprising 5575 mails downloaded from Kaggle, categorized into spam and ham. The algorithm uses this dataset to train the model and predict a new upcoming email if it belongs to the ham or spam category. Below is the table showing previous research used for spam detection using supervised machine learning techniques. The table consists of authors with the algorithm they used and the dataset on which research was done. The table also depicts the classifiers' performance in terms of their accuracy [1-12]. A comprehensive comparison of various machine learning classifiers and models has been made to understand. In [22], authors have obtained the results by combining two methods, like Particle Swarm Optimization and Artificial Neural Network algorithms for the purpose of feature selection and for classification and to separate spam used they applied Support Vector Machine algorithm. They have compared their result with K-Means and Data Classification Self Organizing Map methods. In [23], authors provided the performance analysis on various classification methodologies like Hidden Naïve Bayes, Bayesian Logistic Regression, and many more algorithms to find the accuracy, precision, Root Mean Squared Error and others too. The Rotation Forest algorithm gives the better performance in the form of accuracy i.e., 94.2%.

Table 1. Comparisons between different algorithms used for spam filtering

Authors	Dataset	Algorithms	Accuracy
DeBArr and Wechsler [9]	Custom collection	Random forest	95.2%
Arif et al. [1]	Smart home dataset	Bagged model, XG Boost, and generalized linear model with step-wise feature selection	91.8% from a Generalized linear model with step-wise feature selection
Olatunji [3]	Enron dataset	ELM and SVM classifier	94.06 using SVM
Verma and Sofat [6]	Enron dataset	ID3 algorithm and hidden Markov	89%
Zheng et al. [2]	Weibo social network data	SVM	99.5%
Banday and Jan [8]	Real-life dataset	K nearest neighbor, SVM, Na ïve Bayes, and additive regression tree	96.69% using SVM
Rusland et al. [10]	Spam base and spam data	Modified Na ïve Bays with selective features	83% using spam data 88% using spam base
Subasi et al. [5]	UCI dataset	CART, RE tree, NBT, C4.5, and LAD tree	95.1%
Jamil et al. [4]	Health fitness data	SVM, DT, KNN, and LR	92.1 using SVM
Garavand et al. [12]	Standard dataset from UCI educational dataset	deep learning, SVM, and swarm optimization	93% using SVM
Hall zu et al. [11]	Twitter and Facebook dataset	Bayes net, NB, and SVM	90% using SVM
Hizwai et al. [7]	Spam assassin	MLP, Na ïve Bayes, Decision Tree and SVM	99.3% using random forest

3. Comparative Spam Filtering Models

3.1 Machine Learning

Machine learning is one of the most advancing fields in computer science [17]. It has important and valuable applications in artificial intelligence, allowing our system to automatically learn from its environment and improve its functionality without explicit programming. The main motive behind every machine learning algorithm is to develop an automated tool that can access and train data. The main purpose of machine learning is to develop tools that can work automatically without human intervention [15]. Over the decades, researchers have been trying to improve email communication among which email spam detection and filter is one of the major tasks that need attention. Many research papers have been published regarding the issue, but research gaps still need to be covered. Spam detection is one of the topics of curiosity that can fill those research gaps. There are three types of machine learning:

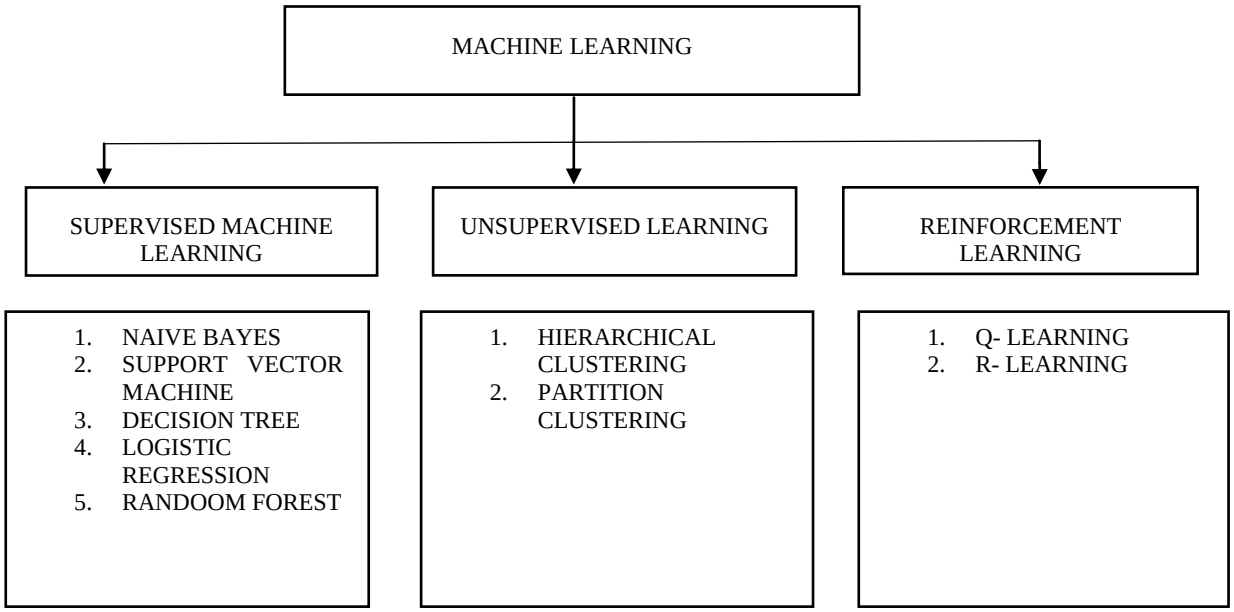

Fig. 1. Various machine learning models

3.1.1 Supervised machine learning

Supervised machine learning is a learning algorithm that needs labelled data [19]. By providing labelled data, our model is trained and predicts future events based on that trained data. In other words, we can say that supervised machine learning models analyse the existing labelled dataset and generate a method to make predictions for upcoming events. This type of learning is used to solve various problems like face recognition, spam classification, object classification, advertisement popularity. Supervised machine learning has various types, including:

- Na ïve Bayes Classifier
- Support Vector Machine
- Random Forest
- Decision Tree
- Logistic Regression

3.1.2 Unsupervised machine learning

Unsupervised machine learning is the algorithm in which the training dataset is not labeled. Unsupervised machine learning algorithms generate a solution for hidden structures inferring a feature from a labeled dataset [15]. Unsupervised learning classifies data by making clusters of similar data from the data used based on the features of data available [18]. Clustering is one of the main features of unsupervised machine learning algorithms, and it is of two types:

- Hierarchical Clustering
- Partition Clustering

3.1.3 Reinforcement learning

Reinforcement learning is a machine learning technique that works on the basis of taking rewards from its environment. In this, an agent works to perceive and interpret information from its environment; positive values are assigned as a reward, and negative values are assigned as penalties [16]. In reinforcement learning, a dataset is not correctly labeled. It is of two types [15]:

- Q Learning
- R learning

3.2 Data Cleaning, Data Preprocessing, and Feature Selection and Extraction

3.2.1 Data cleaning involves the following tasks

- Changing column name
- Encoding column names of spam and ham as 1 and 0, respectively
- Checking null value, if exist then dropping it
- Checking and dropping duplicate values
- Implementing EDA (Exploratory Data Analysis)

3.2.2 Data preprocessing involves the following tasks

- Changing data to lowercase
- Tokenization
- Removing special character
- Removing stop words and punctuation
- Stemming

3.2.3 Feature Selection

Feature selection is the removal of irrelevant data and minimizing the input variables in the developed model. It helps in getting rid of all the noisy data from the dataset used. Feature selection is quite helpful if the message size is enormous and needs to be compressed. The computational complexity and time need to be kept in 5.6 hey Cortana hello search mind for email detection and feature selection minimizes the number of features of a mail which helps increase the model's performance.

3.2.4 Feature Extraction

Feature extraction is the process of creating new features from the existing features, thus, reducing the number of features in the dataset [20]. After the creation of new features, the original features are removed. It is the method by which the most discriminative feature of mail is selected from an arbitrary set of features. Two algorithms are used for feature extraction: TF-IDF (Term Frequency- Inverse Document Frequency) and Count Vectorizer (CV) approach. TF-IDF is basically a text vectorizer that changes a text into a usable vector. It can be divided into subparts: term Frequency (TF) and Document frequency (DF). TF is the calculation of the number of occurrences of a specific term in the dataset,

while DF is the number of documents containing that term. CV follows the method of extracting features or data in the form of vectors and presenting them into a matrix which makes our data highly flexible.

4. Experimental Setup and Results

Initially, a dataset of 5572 emails categorized into two fields as ham and spam was taken. Then we have performed the following tasks on that dataset to check which method gave the best-optimized results.

4.1 Data Cleaning

For cleaning the data, dropping unnamed columns, renaming them to some meaningful name, finding duplicate and missing values, and removing them have been done.

4.2 EDA (Exploratory Data Analysis)

This new dataset has been generated after data cleaning, and EDA helps analyze that dataset. In our experiment, the initial dataset, which was 5572 emails, is now left with 4516 emails.

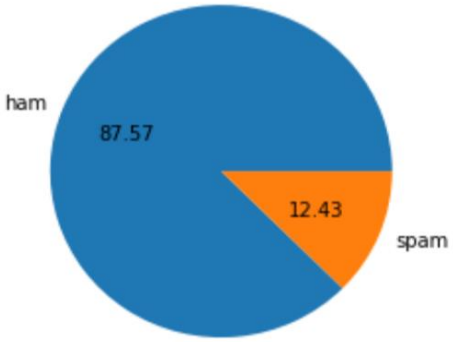

Fig. 2. Ratio of ham and spam mails in the dataset after data cleaning

Then, checking for balance in data is done in which the number of sentences, number of characters, and number of words have been calculated for the dataset.

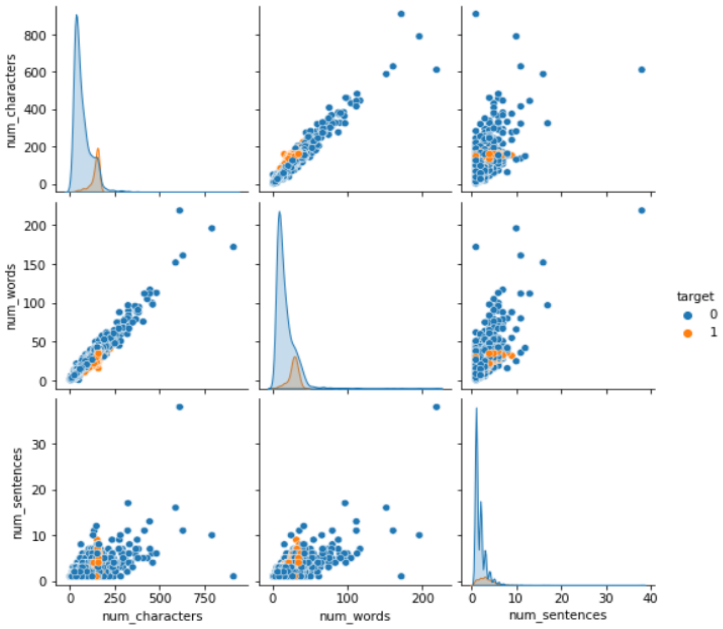

Fig. 3. hart depicting the number of characters, number of words, and number of sentences

4.3 Text Preprocessing

In-text preprocessing, changing dataset of ham and spam mails to lowercase areas, then tokenization of mails is done, which is breaking the sentence into words. After that, the removal of any special characters present is done. After that, removing any stop words and punctuation is done, followed by the stemming of data. Stop words are nothing but a set of very common words used in a language and have no important meaning, while stemming is a machine learning technique in which we can extract the base form of any word by removing affixes from them.

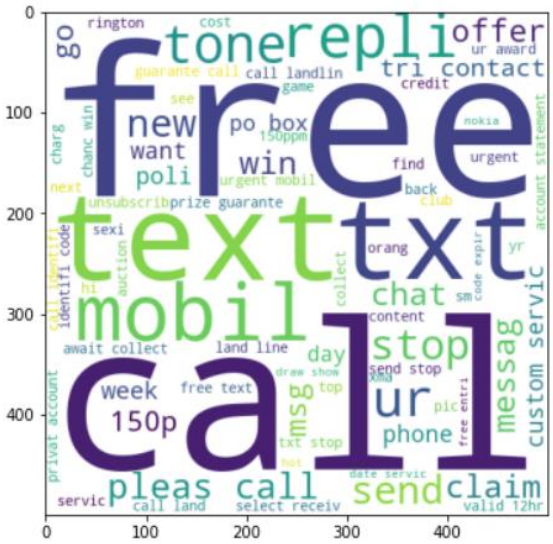

Fig. 4. Depicting common spam words in a mail

Fig. 5. Depicting various words in a ham mail

4.4 Feature Extraction

For feature extraction, two algorithms have been used TFIDF and count vectorizer, in which it was found that the TFIDF algorithm gave better accuracy on the training dataset than the count vectorizer.

4.5 Comparison between different algorithms on the basis of their accuracy

Ten different algorithms were compared for the same dataset, and their accuracy and precision were calculated, the result is shown in the Table 2 given below.

Table 2. Table showing algorithms along with their precision and accuracy

	Algorithm	Accuracy	Precision
0	SVC	0.976744	0.981308
7	ETC	0.975775	0.975775
2	NB	0.971899	1.000000
5	RF	0.971899	1.000000
6	BgC	0.963178	0.963178
4	LR	0.959302	0.938144
8	GBDT	0.948643	0.948643
9	xgb	0.943798	0.943798
3	DT	0.937016	0.829787
1	KN	0.912791	1.000000

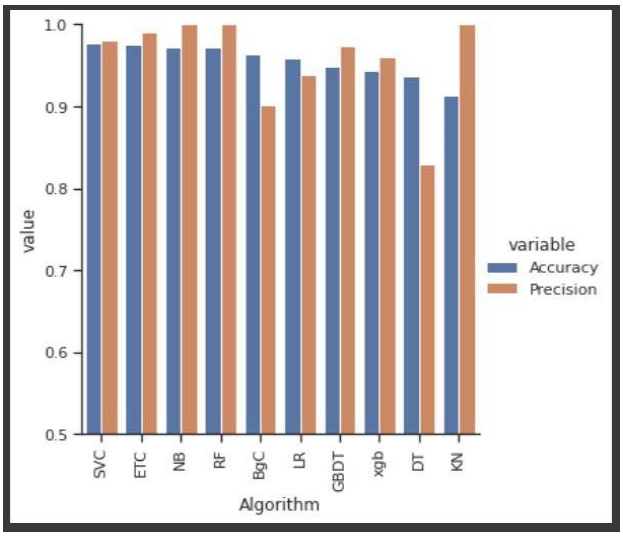

Fig. 6. Chart depicting accuracy and precision of different algorithms

The algorithms used for comparisons include Support Vector Classifier, Extra Tree Classifier, Random Forest, Bagging Classifier, Linear Regression, Decision Tree, Gradient Boosting Classifier, XGB Classifier, K neighbor Classifier, and Naïve Bayes. The comparison concluded that SVC (Support Vector Classifier) gave the best accuracy of 97.67% and precision of 98.13%.

4.6 Ensembling using Stacking and Voting Classifier

Stacking and voting both are ensembling techniques in which various base models are used to improve the model's performance in terms of optimality and robustness.

Three algorithms for Stacking and Voting are used, namely Extra tree Classifier, SVM, and Multinomial Naïve Bayes. We have used these three algorithms for stacking and voting as these algorithms provide the best accuracy, as shown in Table 2 and Fig.-6. As per the experimental results, it was found that stacking gave better accuracy than the voting classifier for the algorithms used for ensembling.

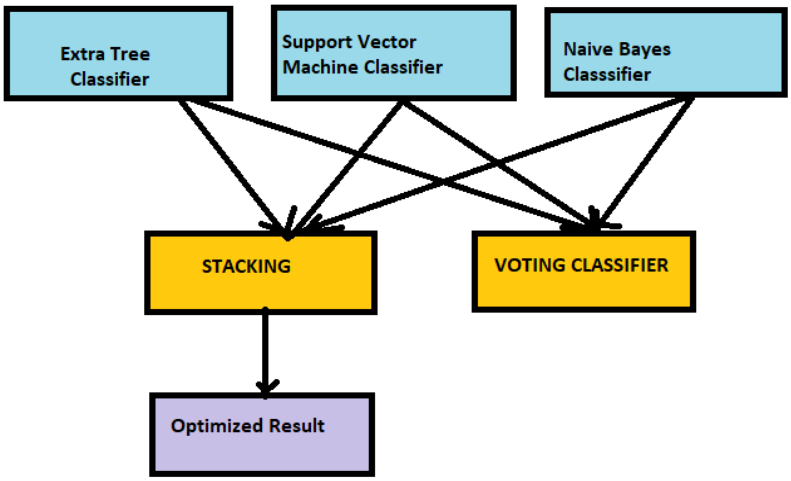

Fig 7. Stacking and voting of different ML classifiers

5. Conclusion

In this study, we have reviewed various machine learning algorithms and what accuracy they give while spam filtering on the given dataset. A detailed review has been provided in this paper how a dataset is trained using a machine learning algorithm and how it predicts the new upcoming mail into spam or ham. This paper also discusses about various machine learning algorithms and their types, processes involved in text pre-processing, feature selection and feature extraction. It has been concluded that TF-IDF algorithm works better than Count vectorizer (CV) approach for feature extraction. A comparative study has been done which machine learning classifier works best for the used dataset. For Stacking and Voting three algorithms namely Extra tree Classifier, SVM and Multinomial Naïve Bayes are used. We have used these three algorithms for stacking and voting as these algorithms are giving the best results as shown in figure-6 and figure-7. During the experiment it is found out that Voting classifier gives the accuracy of 89.24% and precision about 100% while in stacking it is noticed that it gave accuracy of 97.67% and precision of 92.56% in comparison to Rotation Forest algorithm which performs 94.2% accuracy as well as 96.73% accuracy found when SVM algorithm was experimented.

References

[1] M. H. Arif, J. Li, M. Iqbal, and K. Liu, "Sentiment analysis and spam detection in short informal text using learning classifier systems," Soft Computing, vol. 22, no. 21, pp. 7281–7291, 2018.

[2] X. Zheng, X. Zhang, Y. Yu, T. Kechadi, and C. Rong, "ELM- based spammer detection in social networks," The Journal of Supercomputing, vol. 72, no. 8, pp. 2991–3005, 2016.

[3] S. O. Olatunji, "Extreme Learning machines and Support Vector Machines models for email spam detection," in Proceedings of the 2017 IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE), IEEE, Windsor, Canada, April 2017.

[4] F. Jamil, H. K. Kahng, S. Kim, and D. H. Kim, "Towards secure fitness framework based on IoT-enabled blockchain network integrated with machine learning algorithms," Sensors, vol. 21, no. 5, p. 1640, 2021.

[5] A. Subasi, S. Alzahrani, A. Aljuhani, and M. Aljedani, "Comparison of decision tree algorithms for spam E-mail filtering," in Proceedings of the 2018 1st International Conference on Computer Applications & Information Security (ICCAIS), IEEE, Riyadh, Saudi Arabia, April 2018.

[6] M. Verma and S. Sofat, "Techniques to detect spammers in twitter-a survey," International Journal of Computer Applications, vol. 85, no. 10, 2014.

[7] W. Hijawi, H. Faris, J. Alqatawna, A. Z. Ala'M, and I. Aljarah, "Improving email spam detection using content based feature engineering approach," in Proceedings of the Applied Electrical Engineering and Computing Technologies (AEECT), IEEE, IEEE, Aqaba, Jordan, 2017.

[8] M. T. Banday and T. R. Jan, "Effectiveness and limitations of statistical spam filters," 2009, https://arxiv.org/ftp/arxiv/papers/0910/0910.2540.pdf.

[9] D DeBarr and H. Wechsler, "Using social network analysis for spam detection," in Proceedings of the International Conference on Social Computing, Behavioral Modeling, and Prediction, Springer, Ethesda, MD, USA, March 2010.

[10] N. F. Rusland, N. Wahid, S. Kasim, and H. Hafit, "Analysis of Naıve Bayes algorithm for email spam filtering across multiple datasets," in Proceedings of the IOP Conference Series: Materials Science and Engineering, IOP Publishing, Busan, Republic of Korea, 2017.

[11] H. Xu, W. Sun, and A. Javaid, "Efficient spam detection across online social networks," in Proceedings of the 2016 IEEE International Conference on Big Data Analysis (ICBDA), IEEE, Hangzhou, China, March 2016.

[12] M. Zavvar, M. Rezaei, M. Rezaei, and S. Garavand, “Email spam detection using combination of particle swarm optimization and artificial neural network and support vector machine,” International Journal of Modern Education and Computer Science, vol. 8, no. 7, pp. 68–74, 2016.
[13] N. Udayakumar, S. Anandaselvi, and T. Subbulakshmi, “Dynamic malware analysis using machine learning India, algorithm,” in Proceedings of the 2017 International Conference on Intelligent Sustainable Systems (ICISS), IEEE, Palladam, December 2017.
[14] Emmanuel Gbenga Dada, Joseph Stephen Bassi, Haruna Chiroma, Shafi'i Muhammad Abdulhamid, Adebayo Olusola Adetunmbi, Opeyemi Emmanuel Ajibuwa, “Machine learning for email spam filtering: review, approaches and open research problems, Heliyon, Volume 5, Issue 6, 2019, e01802, ISSN 2405-8440,
[15] Naeem Ahmed, Rashid Amin, hamza Aldabbas, Deepika Koundal, Bader Alouffi and Tariq Shah, “Machine Learning Techniques for Spam Detection in Email and IoT platforms, Security and Communication Networks Volume 2022, Article ID 1862888
[16] Stefano Palminteri, Mathias Pessiglione, in International Review of Neurobiology, 2013.
[17] E. Alpaydin, Introduction to Machine Learning, MIT Press, Cambridge, UK, 2020.
[18] Z. Ghahrami, “Unsupervised Machine Learning Springer, Berlin, Germany
[19] S. B. Kotsiantis, I. Zaharakis, and P. Pintelas, “Supervised machine learning: a review of classification techniques, “Emerging artificial intelligence applications in computer engineering, vol. 160, pp. 3–24, 2007
[20] Article on Feature Extraction Techniques by Pier Paolo Ippolito published in Towards Data Science, October10, 2019
[21] J. Dean, “Large scale deep learning,” in Proceedings of the Keynote GPU Technical Conference, San Jose, CA, USA, 2015.
[22] Mohammad Zavvar, Meysam Rezaei, Shole Garavand,"Email Spam Detection Using Combination of Particle Swarm Optimization and Artificial Neural Network and Support Vector Machine", International Journal of Modern Education and Computer Science(IJMECS), Vol.8, No.7, pp.68-74, 2016.DOI: 10.5815/ijmecs.2016.07.08
[23] Shafi’i Muhammad Abdulhamid, Maryam Shuaib, Oluwafemi Osho, Idris Ismaila, John K. Alhassan,"Comparative Analysis of Classification Algorithms for Email Spam Detection", International Journal of Computer Network and Information Security(IJCNIS), Vol.10, No.1, pp.60-67, 2018.DOI: 10.5815/ijcnis.2018.01.07

Authors’ Profiles

Aasha Singh has completed her M.C.A. degree from KNIT Sultanpur in 2011. She is pursuing her Ph.D. from M.U.I.T. Lucknow. She is presently working as an Assistant Professor in the department of Computer Science & Engineering at KNIT Sultanpur. Her research areas are Machine Learning, Software Engineering, Data Mining. She has 05 Years of teaching/industry experience.

Dr. Awadhesh Kumar has completed his B.E. degree from G.B. Pant Engineering College, Pauri (Garhwal) in 1999 and M.Tech. in Computer Science from A.K.T.U. Lucknow, U.P. and Ph.D. from M.N.N.I.T. Allahabad, U.P., India. He is presently working as faculty member in the department of Computer Science & Engineering at KNIT Sultanpur, U.P., India since 2000. His teaching and research interests include Computer Networks, Mobile Ad-Hoc Networks, Wireless Sensor Networks and Machine Learning.

Dr. Ajay Kumar Bharti, Working as Professor, School of Computer Application, Babu Banarasi Das University, Lucknow. He has over 19 years of rich experience in Research, Education and Industry. He worked in numerous premier organizations like Pixellent Solutions, K.N.I.T. Sultanpur, M.I.E.T. Meerut, University of Lucknow, Lucknow, I.E.T. Lucknow and Maharishi University of Information Technology, Lucknow. He has published around 50 of research papers in reputed journals and conference proceedings. He has rich experience in subjects like Operating Systems, DBMS, Data Structures, Computer Graphics, Computer Networks etc. His research interest is in Service Oriented Architecture, Knowledge Based System, e-Governance and Artificial Intelligence.

Dr. Vaishali Singh is currently working as an Assistant Professor in Department of Computer Science, Maharishi University of Information Technology, Lucknow. She has received her Ph. D Degree in the year 2017 from B.B. Ambedkar University, Lucknow, India. Her research interest includes Information Retrieval and Question Answering Systems. She has published the research papers in International Conferences and Journals.

How to cite this paper: Aasha Singh, Awadhesh Kumar, Ajay Kumar Bharti, Vaishali Singh, "An E-mail Spam Detection using Stacking and Voting Classification Methodologies", International Journal of Information Engineering and Electronic Business(IJIEEB), Vol.14, No.6, pp. 27-36, 2022. DOI:10.5815/ijieeb.2022.06.03