International Journal of Computer Applications Technology and Research
Volume 11–Issue 03, 66-71, 2022, ISSN:-2319–8656
DOI:10.7753/IJCATR1103.1003

An Email Spam Filtering Model Using Ensemble of Machine Learning Techniques

Aju Omojokun Gabriel
Department of Computer Science
Adekunle Ajasin University
Akungba-Akoko, Nigeria

Adedeji Ayomiposi Joy
Department of Computer Science
Adekunle Ajasin University
Akungba-Akoko, Nigeria

Abstract: The growth of spam emails is on the increase responsible for larger portions of the global email traffics. Aside the annoyance and the time wasted sifting through the unwanted messages; spam emails can also cause immeasurable harms through malicious software capable of damaging systems and compromising confidential information. The risks of filtering spam emails is that sometimes, legitimate mails are marked as spam, yet the results of not filtering spam are the constant flood of spam clogs on networks that adversely impacts users inboxes while draining valuable resources on the networks such as bandwidth and storage capacity, productivity loss and interfere with the expedient delivery of legitimate emails. Several researchers had worked on the design of models for spam email filtering using different techniques, however the detection accuracy of these models have also become subject of discussions. This study developed spam email filtering model using Ensemble of Decision Tree, Support Vector Machine and Multilayer Perceptron (DT-SVM-MLP) technique as a solution approach to solving issues of low spam emails detection accuracy. The ensemble model was trained using forward propagation training technique and the performance was evaluated using five performance metrics of Accuracy, False Positive (FP) Rate, Precision, Recall and F-Measure.

Keywords: Spam Email, Email Filtering, Ensemble Machine Learning, Forward Propagation Training, Performance Metrics.

1. INTRODUCTION

The internet has become an integral part of everyday life and electronic mail (email) has become a powerful and indispensable tool for information exchange. It is one of the most commonly used features over communication networks that may contain texts, files, images, or other attachments. Email messages are sent through email servers and uses multiple protocols within the Transmission Control Protocol/Internet Protocol (TCP/IP) suite which allows users to send and receive messages anywhere in the world because the access mobility to email system is independent of physical locations.

The email is significant for many kinds of group connection and is being widely used by many people; individuals and organizations for both official and personal correspondence (Naem et al, 2018). In the 1990s, there was an increase use of email facilities as more companies and institutions joined the Internet system, as the significant advances made in telecommunication technologies, couple with the reduced costs of computers and telecommunication devices made the internet system more accessible. Email allows users to send and receive messages anywhere with an email address, the system can also be accessed from anywhere in the world and can deliver messages instantaneously. Because the mobile access to email is neither attached to a physical location nor restricted to a fixed place, rather the mobility of email allows people to work and communicate from anywhere. Due to these factors, email communication is used over other modes of communication because it is economical, flexible and reasonable (Palival et al., 2018).

Today, e-mail has become an efficient, rapid and cheap means of communication. Likewise, the dramatic growth in the spread of unwanted email messages, otherwise known as Spams cannot be overemphasised. One of the fast rising and costly problems linked with the internet today is the spam email which are predominantly mercantile and mostly have attractive links to famous websites that lead to meddlesome sites (Naem et al, (2018).

In recent times, unwanted commercial bulk emails have become a huge problem on the email systems. In April 2021, it was estimated that 89.35% of all emails were accounted as spam mails and 482.65billion daily spam mails were sent globally (Palmote et al., 2021). The huge volume of spam mails flowing through the internet networks have destructive effects on the memory space of email servers, communication bandwidth, central processing unit, power consumption and user time (Dada et al., 2019). Aside the cost of spam mails on the internet networks infrastructures, it has also been reported that the spread in spam mails has resulted to untold financial loss for many internet users who have fallen victim of internet scams and other fraudulent practices of spammers who send emails, pretending to be from a reputable source with the intention to persuade individuals to disclose sensitive personal information like passwords, Bank Verification Number (BVN) and credit card numbers. The cost of spam mails to companies worldwide in 2019 was estimated to be US$260 billion (Palmote et al., 2021).

The risk in filtering spam is that sometimes, legitimate mails may be rejected or marked as spam, however, the risks of not filtering spam are the constant flood of spam clogs on networks which adversely impacts the users inboxes, drain valuable network resources such as bandwidth and storage capacity, productivity loss and interfere with the expedient delivery of legitimate mails (Mallampati, 2019). Different machine learning algorithms have been used in the development of email filtering techniques to solve the problem of spam emails wreaking havoc on email users. These machine learning algorithms have been successfully applied to classify emails into either spam or non-spam. These algorithms include Logistic Model Tree Induction, Decision

Tree, Artificial Immune System, Support Vector Machine, and Artificial Neural Networks (Dada et al., (2019).
These algorithms have been giving varying accuracy in the filtering process and accuracy rates has become a point of research. This paper went further in increasing the performance of these machine learning algorithms by developing an ensemble algorithm that combines the three Support Vector Machine (SVM), Decision Tree (DT) and Multilayer Perceptron (MLP) algorithms to form an optimal model.

2. LITERATURE REVIEW

The number of spam email has increased for several reasons such as advertisements, multi-level marketing, chain letters, political emails, stock market advice, among others. Email Filtering have usually relied on keyword patterns, to be more efficient and prevent the danger of accidental removal of ham messages which are called Ham or allowed messages. These patterns need to be checked with each user's received emails. However, detailed setting of such patterns needs time and proficiency which are unfortunately not always available (Takhmiri and Haroonabadi, 2016).

In restricting spam email, several methods and spam filtering algorithms have been developed using machine learning techniques such as Naive Bayes, Support Vector Machine, K-Nearest Neighbor, Bayes Additive Regression, KNN Tree, Decision Tree and rules. Chan et al., (2010), the authors combined the Best Stepwise feature selection with a classifier of Euclidean nearest neighbor and created a Naïve Euclidean approach to develop email filtering system. Each email was represented in D-dimensional Euclidean space. Using SpamBase from the UCI repository, and a 10-fold cross validation, they achieved an accuracy of 82.31% compared to 60.6% for the Zero rule.

Rathi et al., (2013) proposed a data mining technique approach for finding the best classifier for email classification. They analyzed various data mining technique for measuring the performance of several classifiers through “with feature selection algorithm” and “without feature selection algorithm”. After selecting the Best feature selection algorithm, they considered the selected algorithm for their feature selection purpose. They experimented their data using Naïve Bayes, Support vector machine, J48, Random Forest and Random Tree algorithms. The dataset used consists of 58 attributes and 4601 instances. Bhat et al., (2014) proposed some community-based topological features to learn improved classification models for identifying spammers in online social networks. However, the results only spanned over single classifiers. Mahmoud et al., (2014) The proposed a combined Naïve Bayes, Clonal selection and Negative selection algorithms filtering technique that consists of four phases of Training phase, Classification phase, Optimization phase and Testing phase to classify the email messages. The worked used 2,500 spam messages and 2,500 non-spam messages to train the system.

Rusland et al., (2017) performed email spam filtering analysis using Naïve Bayes algorithm on two datasets which are evaluated based on the accuracy, recall, precision and F-measure metrics. The Naïve Bayes algorithm as a probability-based classifier counts the frequency and combination of values in a dataset. The work performed through three phases such as pre-processing, Feature Selection, and implementation. Abdulhamid et al., (2018) studied the analysis based on the classification of algorithms and their efficiencies. For this study various methodologies considered and their efficiencies were measured in terms of basic metrics. Any function collection or efficiency improve approach was used to provide a holistic view of the efficiency of classification techniques. Study shows that there are a variety of classification techniques that are more reliable if better investigated by way of selecting features. Of all the various methodologies utilized, Rotation Forest is the most reliable classifier of 94.2 percent.

Agarwal and Kumar (2018) proposed a combined methodology of machine learning techniques such as the NB algorithm and optimization algorithm namely, the PSO algorithm for identification of spam emails. NB algorithm is mainly utilized for classification of the obtained emails into two categories such as spam or non-spam. PSO algorithm is utilized for the optimization parameters that are of the NB algorithm. The implementation of this algorithm was made with the aid of the popular dataset of Ling spam evaluated the efficiency based on the popular metrics. PSO outperforms relative to individual NB approaches based on the validated findings. Palival et al., (2018) presented an email spam filtering model using ID3 Decision Tree based Algorithm. ID3 is a non-incremental algorithm used to build a decision tree from a fixed set of observations. The resulting tree is then used to classify test observations and each observation is represented by features or attributes and a class to which it belongs. ID3 uses information gain measure to select decision node. Enron dataset was used for training as well as testing the filter system. The Enron dataset contains emails of both types stored in plain text format with 3672 legitimate (ham) emails and 1500 spam emails.

Dada et al., (2019) analyzed the core principles, attempts, performance, and spam filtering study patterns. The latest study investigates the implementations of machine learning environments to the leading ISPs, including Gmail, Yahoo, and Outlook spam filters, to the spam processing e-mail process. There has been debate about the general approach of spam filtering and the efforts of different researchers to tackle spam using machine learning techniques. The study contrasts the advantages and disadvantages of the existing methodologies of machine learning and brings new problems with spam filter growth. The study suggested broad and strong opposing education as the strategies for managing spam e-mail risks to cope successfully with the potential.

Mallampati et al., (2019) presented a kernel function SVM approach to build a spam detection. System, when the support vector machine algorithm analyzes a single mail then it returns a 0 else it returns a 1. The authors considered the dataset from the UC Irvine Machine Learning Repository for spam emails. Olatunji (2019) proposed a model based on support vector machines that are suggested for spam identification when carefully searching for optimized parameters for better results. Experimental findings indicate that all earlier models on the same common dataset used in this work succeeded the model suggested. 95.87 and 94.06% accuracy for preparation is reached and collections of testing respectively

3. METHODOLOGY

The email spam filtering method is designed to separate the spam (unwanted emails) from the non-spam (wanted emails). The recent spam mail classification is mostly handled by machine learning (ML) algorithms intending to differentiate between spam and non-spam messages, the machine learning

algorithms achieve this by using an automatic and adaptive technique, rather than depending on hand-coded rules that are susceptible to the continuously evolving features and varying characteristics of spam messages. Machine learning techniques have the capacity to obtain information from a set of messages provided, and then use the acquired information to classify new messages that it just received.

This research study used machine learning ensemble technique for the email classification. The machine learning ensemble technique combines the Decision Tree (DT), Multilayer Perceptron (MLP) and Support Vector Machine (SVM) base models in order to produce one optimal predictive model. The study used Python programming language (version-3.9) and the Jupyter notebook editor. The activities involved in the chosen methodology to ensure the success of the study include the: Data Collection, Data Pre-processing, Ensemble Model Design, Model Training and Testing, and Model Performance Evaluation.

3.1 Data Collection

The Enron dataset is used for training as well as testing the model. The Enron dataset contains emails of both types stored in plain text format. The Enron directory contains 3672 legitimate (ham) emails and 1500 spam emails, the first attribute contains the subject and body of the email, and the last attribute of the dataset is the nominal attribute, which consists of the value 0's and 1's to represent whether a mail is spam or not. The dataset is divided into a ratio of 80:20 wherein the 80% data is used for training the model and the remaining 20% is used for testing the accuracy of the developed model.

3.2 Data Preprocessing

Pre-processing is a very crucial step in spam email filtering techniques. There are three steps involved in the pre-processing: tokenization, stop word removal and stemming. The initial step consists of the process called tokenization. In the process, all of the unnecessary word, the punctuations and the symbols are removed from the sentences. The strings that are left is split up into various tokens. The next step is stop word removal. Stop-words are basically nothing but unnecessary and non-informative words, e.g. 'a', 'an', 'the', and 'is', among others that doesn't add any sense and information to the message. In the second step all such words which carry no information are removed. English language has around 300-400 stop words.

The last step is the stemming which is the reduction of inflection in words and bringing it to their root form is known as stemming. The root word can just be a canonical form of the original word. Word-stemming is a term used to describe a process of converting words to their morphological base forms, mainly eliminating plurals, tenses, prefixes and suffixes. Stemming is closely related to lemmatization which while reducing a word considers the part of speech and the context of the word.

3.3 Ensemble Model Design

The model design for this study consists of three machine learning techniques algorithms ensembled to form a more optical model (DT – MLP – SVM model).The Decision Tree generates the output as a binary tree like structure called a decision tree, in which each branch node represents a choice between a number of alternatives, and each leaf node represents a classification or decision. MLP networks are general- purpose, flexible, nonlinear models consisting of a number of units organized into multiple layers. The complexity of the MLP network can be changed by varying the number of layers and the number of units in each layer. Given enough hidden units and enough data, it has been shown that MLPs can approximate virtually any function to any desired accuracy. In other words, MLPs are universal approximators. SVM considers data as points in space mapped in a way such that the difference between the closest data points is maximum.

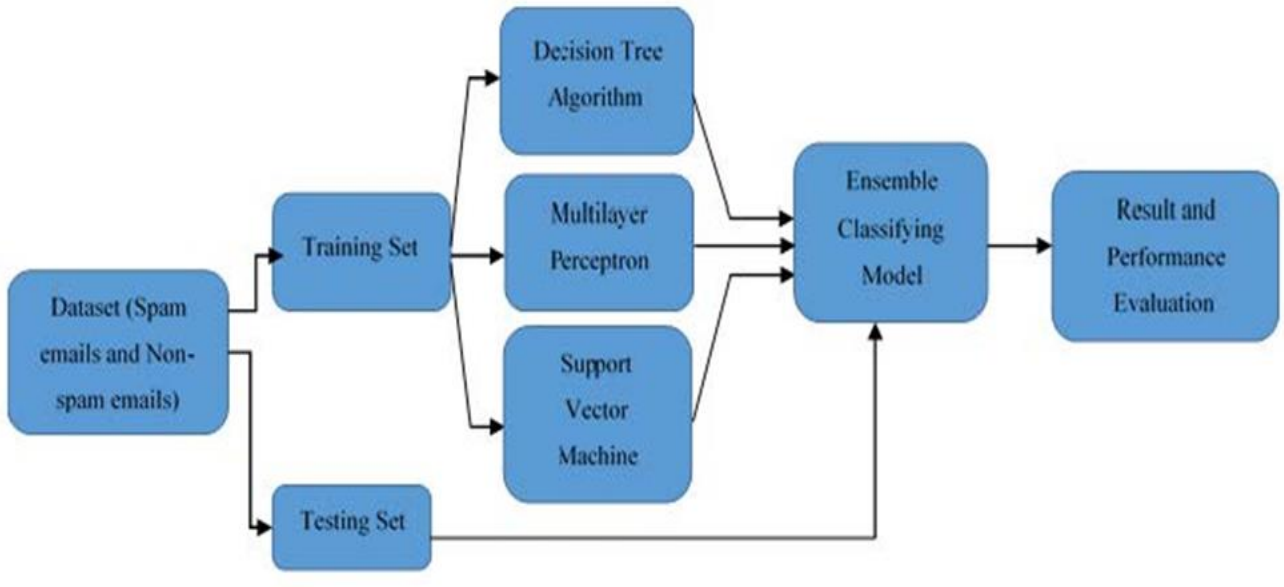

Figure 3.1: The Ensemble Model Architecture

International Journal of Computer Applications Technology and Research
Volume 11–Issue 03, 66-71, 2022, ISSN:-2319–8656
DOI:10.7753/IJCATR1103.1003

In this study, an ensemble Decision Tree – Multilayer Perceptron – Support Vector Machine (DT- MLP -SVM) model is developed to form a more optimal spam email filtering model by taking the advantages of each base models into consideration. The DT- MLP -SVM) model algorithm is as shown in algorithm 1. The outputs of the ensemble model (DT-SVM-MLP) were compared with that of the base models.

Algorithm 3.1: The Study Algorithm for SVM – DT – MLP Model

(1) Input: training data $D = \{ x_i, y_i \}_{i=1}^{m}$.

(2) Output: Ensemble classifier H

/* learn based-level classifiers*/

(3) **for** t = 1 to T **do**

(4) learnt h_t based on D

(5) **end for**

/* construct new data set for classification*/

(6) **for** i = 1 to m **do**

(7) $Dh = \{x_i', y_i\}$, where $x_i' = \{h_1(x_i), \ldots\ldots\ldots\ldots h_T(x_i)\}$

(8) **end for**

/* learn a meta-classifier*/

(9) learn H based on D_h

(10) return

3.4 Model Training and Testing

The forward propagation technique is used in this study for the training of the network. In the forward propagation, the input data is fed in the forward direction through the network. Each layer accepts the input data, processes it as per the activation function and passes to the successive layer. The dataset is divided into a ratio of 80:20 wherein the 80% data is used for training the model and the remaining 20% is used for testing the accuracy of the developed model. The model testing involves explicit checks for the behaviours that the model exhibits. Testing the model performance in terms of accuracy and other metrics on which the model is evaluated.

3.5 Model Performance Evaluation

The performance evaluation is done by measuring the percentage of spam detected and how many misclassifications are done by a particular technique and the ensemble model. The results obtained are then compared on the basis of the performance of each of the techniques (Sharaff, 2019). The ensemble model is evaluated using the five-performance metrics: Accuracy; FP rate; Precision; Recall and F-Measure.

The model resulted into a confusion matrix which consists of four parts: True Positive (TP); True Negative (TN); False Positive (FP) and False Negative (FN). These values are used to determine model performances.

4. FINDINGS AND DISCUSSION

4.1 The Baseline Models

The training dataset (spam and legitimate message) was generated from the mails. The class labels are designated as spam to represent spam and ham to represent legitimate emails. The machine learning techniques (Algorithms): Decision Tree, Multilayer Perceptron and Support Vector Machine Algorithms were used for training the data on the Jupyter notebook environment. The performance of the trained models was evaluated using 10-fold cross validation for its predictive accuracy. Predictive accuracy is used as a performance measure for email spam classification. The prediction accuracy is measured as the ratio of number of correctly classified instances in the test dataset and the total number of test cases. The outputs process for the base models and the ensemble model are shown in Tables 4.1 and 4.2.

Table 4.1: Results of the Base Models Performance Evaluation

Techniques	Accuracy (in %)	Precision	Recall	F-Measure	Support
Decision Tree Classifier	96.74	0.98	1.00	0.99	1139
Multilayer Perceptron	97.15	0.98	0.99	0.98	1139
Support Vector Machine	98.35	0.94	0.96	0.95	1139

Accuracy, Precision, Recall and F-Measure Metrics: From the values obtained for the base models as shown in table 4.1. The SVM model has the best performance in terms of the models' accuracy, however, in term of precision, Recall and F-Measure, the SVM performed below the other two base models. Followed by the Multilayer Perceptron while the Decision Tree had the lowest value in term of the models' accuracy. In term of Precision, Recall and F-Measure; Decision Tree performed better than the Multilayer Perceptron. F-Measure is dependent on Precision and Recall..

4.2 The Developed Ensemble Model

The results of the ensemble model with the base models are shown in table 4.2. The main principle behind the technique is that the combined knowledge of multiple models, in this case, the Decision Tree, Multilayer Perceptron and Support Vector Machine Algorithms can performance better and give a more accurate results as compared to a single model considered for same task.

Table 4.2: Results of the Ensemble Model Performance Evaluation

Techniques	Accuracy (in %)	Precision	Recall	F-Measure	Support
Decision Tree Classifier	96.74	0.98	1.0	0.99	1139
Multilayer Perceptron	97.15	0.98	0.99	0.98	1139
Support Vector Machine	98.35	0.94	0.96	0.95	1139
**Ensemble Model	**99.86**	**0.99**	**1.0**	**0.99**	**1139**

Figure 4.1 shows the performance of precision, recall and f-measure for the different models and Figure 4.2 shows the accuracy performance results of the models.

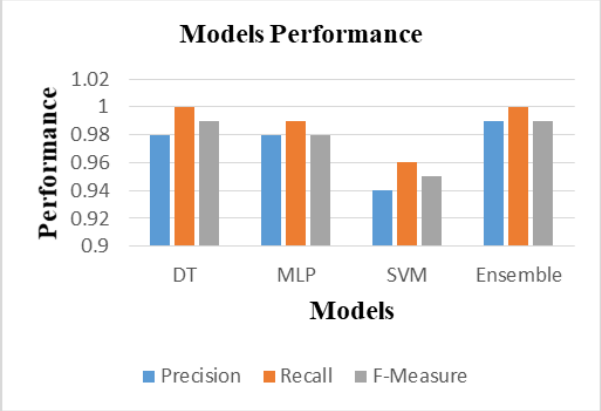

Figure 4.1: Models Precision, Recall & F-Measure Results

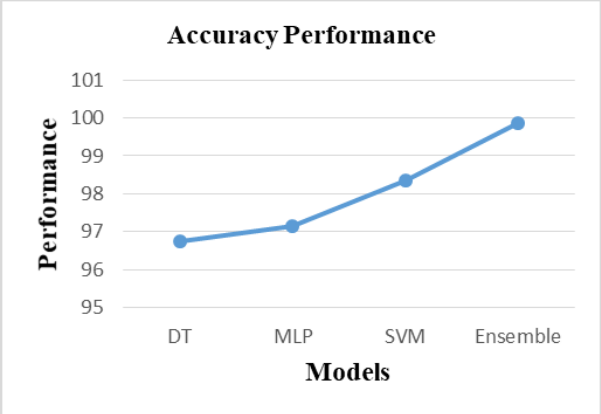

Figure 4.2: Models Accuracy Results

The developed ensemble model was evaluated using all the performance evaluation metrics used for the base models and the model gave an overall high accuracy of 99.86%. This resulted in a more promising approach of email spam filtering technique with more consistent and accurate results.

5. CONCLUSION

Email spam filtering is challenging but a highly desirable task. Different machine learning techniques have been used in different literatures for filtering genuine messages from spam messages. An ensemble model combining the three machine learning techniques of Decision Tree, Multilayer Perceptron and Support Vector Machine Algorithms is used and measured with the chosen performance evaluation metrics to observe the effectiveness and accuracy of each base techniques. Though a slight change was observed in the performance of the base models, this deviation indicates that the performance of a technique depends on the data used more than the algorithm. However, the developed ensemble model performed better than each of the base models. It resulted in a more promising approach of email spam filtering technique producing more consistent and accurate results.

6. REFERENCES

[1] Abdulhamid M. S, Shuaib M, Osho O, Ismaila I, and Alhassan J K, Comparative Analysis of Classification Algorithms for Email Spam Detection 2018, *Inter. J. Comp. Net. Inf. Sec.*, Vol. 10, pp. 60–67.

[2] Agarwal K and Kumar, T. (2018). Approach Of Naïve Bayes and Particle Swarm Optimization. *2018 Sec. Int. Conf. on Intel. Comp. and Cont. Sys.* pp. 685–90.

[3] Bhat, S. Y., Abulaish, M and Mirza, A. A. (2014). Spammer Classification Using Ensemble Methods over Structural Social Network Features. *Proceedings - 2014 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology - Workshops, WI-IAT 2014, 2*, 454–458.

[4] Chan, T. Y., Jie Ji, and Qiangfu Zhao. (2010). Learning to Detect Spam: Naive-Euclidean Approach. *International Journal of Signal Processing*. Issue 1 (2010), pp 31-38.

[5] Christina, V. (2010). Email Spam Filtering using Supervised Machine Learning Techniques. *International Journal on Computer Science and Engineering Vol. 02, No. 09. pp. 3126-3129*

[6] Cortez, P. (2010). Spam Email Filtering Using Network-Level Properties. *July*. https://doi.org/10.1007/978-3-642-14400-4

[7] Dada, E. G and Joseph, S. B. (2018). Random Forests Machine Learning Technique for Email Spam Filtering. Heliyon, Vol. *9, No.* 1. pp. 29–36.

[8] Dada E G, Bassi J S, Chiroma H, Abdulhamid S M, Adetunmbi A O, and Ajibuwa O. E. (2019). Machine Learning for Email Spam Filtering: Review, Approaches and Open Research Problems *Heliyon*, Vol. **5, No.** 6. https://doi.org/10.1016/j.heliyon.2019.e01802

[9] Delany, S. J., Cunningham, P., Tsymbal, A and Coyle, L. (2005). A Case-Based Technique for Tracking Concept Drift in Spam Filtering. *Knowledge-Based Systems*, Vol. *18 No.* 4–5. pp. 187–195. https://doi.org/10.1016/j.knosys.2004.10.002

[10] Enron mail dataset, "http://www2.aueb.gr/users/ion/data/enron-spam (Accessed: 24 March, 2021)

[11] Jantan, A., Ghanem, W. A. H. M and Ghaleb, S. A. A. (2017). Using Modified Bat Algorithm to Train Neural

Networks for Spam Detection. *Journal of Theoretical and Applied Information Technology*, Vol. *95, No.* 24. pp. 6788–6799.

[12] Kaur, H and Sharma, A. (2016). Improved Email Spam Classification Method using Integrated Particle Swarm Optimization and Decision Tree. *2016 2nd International Conference on Next Generation Computing Technologies (NGCT-2016) Dehradun, India* 14-16 October 2016.

[13] Mahmoud, T. M., El-hafeez, T. A and Khairy, M. (2014). *An Efficient Three-phase Email Spam Filtering Technique An Efficient Three-phase Email Spam Filtering Technique*. *Journal of Theoretical and Applied Information Technology*, Vol. *62, No.* 8. pp. 1742–1751.

[14] Mallampati, D., Shekar, K. C and Ravikanth, K. (2019). Supervised Machine Learning Classifier for Email Spam Filtering. (Issue January). Springer, Singapore. https://doi.org/10.1007/978-981-13-7082-3

[15] Naem, A. A., Ghali, N. I and Saleh, A. A. (2018). Antlion Optimization and Boosting Classifier for Spam Email Detection. *Future Computing and Informatics Journal*, Vol. *3, No.* 2. pp.436–442. https://doi.org/10.1016/j.fcij.2018.11.006.

[16] Olatunji S O, Improved email spam detection model based on support vector machines 2019, *Neu. Comp.and App*., **31**, pp. 691–99.

[17] Palimote, J., Anireh, V.I.E and Nwiabu, N. D. (2021). International Journal of Advanced Research in Computer and Communication Engineering. Vol. 10, Issue 4. pp. 17-24

[18] Palival, D., Printer, K., Devre, R and Lemos, N. (2018). Email Spam Filtering Using Decision Tree Algorithm. International Journal of Scientific and Engineering Research, Vol. 9, Issue 3. pp. 40-42.

[19] Park, I., Sharman, R., Rao, H and Upadhyaya, S. (2007). The Effect Of Spam And Privacy Concerns on E-Mail Users' Behaviour. *J. Info. Syst. Security*, Vol. *3, No.* 1. pp. 40–62.

[20] Takhmiri, H and Haroonabadi, A. (2016). Identifying Valid Email Spam Emails Using Decision Tree. International Journal of Computer Applications Technology and Research. Volume 5, Issue 2, pp. 61 - 65.