IJFANS INTERNATIONAL JOURNAL OF FOOD AND NUTRITIONAL SCIENCES

ISSN PRINT 2319 1775 Online 2320 7876

 Volume 11, Sp.Iss 5 , 2022

Email Spam Guard Using Machine Learning and Deep Learning Algorithms

I Veda Sai Priya [1*], Y Rohita Lakshmi Vasavi [1*], G Venkata Naga Sai Indu Priya [1*], P Anisha [1*], Dr M Kavitha [1], M Kalyani [2]

[1*] Student, [1] Associate Professor, [2] Assistant Professor

[1*, 1] Department of Computer Science & Engineering, Koneru Lakshmaiah Education Foundation (KLEF), Vaddeswaram, Green fields, Guntur, Andhra Pradesh, India -522302
[2] Department of Mathematics, PACE Institute of Technology and Sciences, Ongole, AP, India
Email: mkavita@kluniversity.in

DOI : 10.48047/IJFANS/11/Sp.Iss5/060

Abstract. With the rapid growth of digital communication, the problem of email spam has become a persistent issue, negatively impacting the user experience and information. This project focuses on leveraging machine learning techniques to effectively detect and filter out email spam, enhancing the efficiency and reliability of email communication. The project begins with a comprehensive preprocessing step that involves cleaning and transforming raw email data into a structured format suitable for analysis. Feature extraction techniques such as word embeddings convert textual content into numerical representations, capturing semantic meanings and contextual information. Machine learning algorithms, including Decision Tree, Random Forest, Extreme Gradient Boosting, and, Deep Learning algorithm including Long Short-Term Memory (LSTM) network are trained and evaluated on a labeled dataset of emails to identify the most suitable algorithm for spam detection. The evaluation metrics encompass accuracy, precision, recall, and F1-score, providing a comprehensive assessment of model performance. The project contributes to the larger goal of creating a seamless and secure digital communication experience by reducing the intrusion of spam emails.

Keywords: Machine Learning, Deep Learning, Email Spam Detection, Feature Selection

1. Introduction

Email, short for electronic mail, is a widely used method of exchanging digital messages over the Internet. Email is widely used for personal, professional, and business communication. Due to its popularity, it became the most common source for spammers to steal user's sensitive information [1].

Email spam, commonly known as spam, refers to unsolicited and often irrelevant or unwanted email messages sent in bulk to a large number of recipients without informing them. It can also be referred to as uninvited or junk email. These can be annoying, time-consuming to deal with, and even harmful. They may contain viruses, malware, advertisements, and phishing links that can steal user's personal information [2].

Email spam detection is the process of identifying and filtering out all such kinds of unwanted emails from reaching the user's inbox. So, it is essential to devise an approach and develop an automatic system to detect spam emails before they are opened ensuring security in this digital era of communication. In this scenario, Machine Learning and Deep learning algorithms are very helpful.

ISSN PRINT 2319 1775 Online 2320 7876
Research paper **© 2012 IJFANS. All Rights Reserved**, Volume 11, Sp.Iss 5 , 2022

ML algorithm Decision Tree, Ensemble method Random Forests, Extreme Gradient Boosting, and Deep learning algorithm- Long Short Term Memory are trained to identify and filter out spam from legitimate messages automatically. The brief description of these algorithms is as follows.

Decision Tree (DT) is a supervised machine learning algorithm used for both regression and classification problems. It creates a structure resembling a tree by dividing the dataset into subsets according to the values of the input features. The leaves of the tree signify the conclusions or forecasts, whereas each node in the tree indicates a decision.

Random Forest (RF) is an ensemble learning method that combines multiple Decision Trees to improve predictive accuracy and reduce overfitting. It creates a collection of Decision Trees by training on random subsets of the data and features. The final prediction is often an average or a voting mechanism of the individual trees.

Extreme Gradient Boosting (XGBoost) is a powerful machine learning algorithm known for its exceptional predictive accuracy and speed. It leverages gradient boosting to optimize the ensemble of decision trees, making it a popular choice in data competitions and a wide range of applications, from regression to classification. XGBoost is recognized for its efficiency, handling large datasets and complex features effectively.

Long Short Term Memory (LSTM) is a type of recurrent neural network (RNN) designed for sequential data, such as time series and natural language processing. It overcomes the vanishing gradient problem in traditional RNNs and can capture long-range dependencies in data. LSTMs use a specialized architecture with memory cells that can store and retrieve information over extended sequences, making them well-suited for tasks that require modelling temporal dependencies.

This approach of filtering spam messages using machine learning and deep learning involves data collection, data pre-processing, feature extraction, model training, and model evaluation techniques.

2. Related Work

Table 1 represents various authors contributions in this research problem.

Table 1 Various authors contributions

Title	Dataset	Methodologies	Observations/Key Findings
Detecting ham and spam emails using feature union and supervised machine learning models, 2023 [1]	Kaggle datasets namely Dataset 1 'Spam or Ham - EMP Week 2 ML HW and Dataset 2 Spam filter are used.	Feature extraction_approach known as Feature union that combines TF-IDF and BoW is used. Various algorithms like Random forest (RF),Gradient Boosting Machine (GBM), Support vector machines (SVM), Gaussian naïve Bayes (GNB, Long short-term memory (LSTM) and Gated Recurrent Unit (GRU) are applied.	Random forest and logistic regression achieve the highest accuracy scores 0.991 and 0.990, respectively.
Comparison of machine learning techniques for spam detection, 2023 [2]	Two Datasets have been used Spam Corpus and Spam base dataset is used	Adaptive Booster, Artificial Neural Network, Bootstrap Aggregating, Decision Table, Decision Tree, J48, K Nearest Neighbor, Linear	Here multiple machine learning classifiers have been implemented for detecting spam emails.we can say that In terms of

ISSN PRINT 2319 1775 Online 2320 7876
Research paper **Volume 11, Sp.Iss 5 , 2022**

		Regression, Logistic Regression, Naïve Bayes, Random Forest, Sequential Minimal Optimization and, Support Vector Machine algorithms are successfully implemented.	accuracy, the Random Forest classifier performs better compared to the rest of the machine learning classifiers. And Naïve Bayes classifier performs poorly in term of accuracy.
Detection of Email Spam using Machine Learning Algorithms: A Comparative Study, 2022 [3]	Gmail Dataset with different volumes of 1000,1500 and 2100 Mails had been used.	support vector machines (SVM), Gaussian naive Bayes (GNB), and logistic regression (LR) methods are used.	various Machine Learning models are applied to the same dataset. The different machine learning models were compared based on accuracy and Precision. Support vector machine results in 98.09% accuracy.
ML Approaches to Detect Email Spam Anomaly, 2022 [4]	Data set called stop-words is used	ML models namely random forest (RF), gradient boosting machine (GBM), support vector machines (SVM), Gaussian naive Bayes are used.	This proposed system tries to recognize a recurrent word group which are used mostly that are classed as spam using machine learning techniques
Spam Email Detection Using Machine Learning and Deep Learning Techniques, 2021 [5]	Spam base and spam data	Naïve bayes, convolutional neural networks, multi layer perceptron, support vector machine are various methodologies used.	SVM (support vector machine) algorithm and compared to the work of others on same dataset with same algorithm and the results are improved
Email Spam Detection using Machine Learning and Neural Networks, 2021[6]	Spam Assassin and non-assassin dataset is used	Logistic Regression, Naive Bayes, Support Vector Machine, Neural Network.	Most Frequent Word Count with Count Vectorization. Both Feature sets are developed using the existing kernel. By the accuracy graphs it can be seen that the artificial neural network has the highest detection rate of whether a file is spam or ham.
Novel email spam detection method using sentiment analysis and personality recognition, 2020 [7]	Datasets named original dataset (CSDMC 2010 dataset), polarity dataset v2.0, validation dataset (TREC 2007) are used.	Sentiment classification is done by adding polarity feature. Personality recognition is done using the Myers–Briggs personality model. Used 10-fold cross-validation technique.	Combining sentiment analysis techniques with personality recognition techniques the best result obtained in Bayesian spam filtering is improved in terms of 99.24% accuracy.
A Spam Email Detection Mechanism for English Language Text Emails Using Deep Learning Approach, 2020 [8]	Enron corpus's dataset	Ontology-based spam filtering methodology is used	It does not capture the misuse of storage and bandwidth resources.

ISSN PRINT 2319 1775 Online 2320 7876
Research paper Volume 11, Sp.Iss 5 , 2022

Detecting Spam Email With Machine Learning Optimized With Bio-Inspired Metaheuristic Algorithms, 2020 [9]	Email datasets namely Ling-Spam dataset, Enron dataset, PUA dataset, PU1, PU2, PU3 datasets, Spam Assassin dataset are used.	Bioinspired algorithms Particle Swarm Optimization(PSO), Genetic Algorithm (GA) are used for optimization	Multinomial Naïve Bayes with Genetic Algorithm performed the best overall. It has the highest accuracy of 98.47%, providing precision of 97.79%, recall of 81.74%, and F1-Score of 87.42% on the split of training size 80% and Testing size 20%.
Efficient Clustering of Emails Into Spam and Ham: The Foundational Study of a Comprehensive Unsupervised Framework, 2020 [10]	A comprehensive novel dataset of 100,000 records of ham and spam emails has been developed and used as the data source.	Spectral and K-means, BIRCH, HDBSCAN and K-modes, FEATURE SCORE,G UNSUPERVISED CLUSTERING ALGORITHMS,HDBSCAN.	The Purity of the clusters produced by the three best performing algorithms in this study, though is extremely good, but some degree of misclassifications is still there. In future we intent to propose the second segment of the framework where algorithms used in this study will be implemented on email body and subject field for clustering purposes.
Email Spam Detection Using Machine Learning Algorithms, 2020 [11]	A spam email data set from Kaggle is used to train	Classic classifiers, support vector machine, k- nearest neighbour	spam detection is proficient of filtering mails giving to the content of the email and not according to the domain names or any other criteria. Therefore, at this it is an only limited body of the email.
Email based Spam Detection, 2020 [12]	Used an email dataset , Enron corpus and CSDMC2010 spam dataset.	Bayes' theorem and Naive Bayes' Classifier-Mean algorithm	it detects unsolicited and unwanted emails and prevents them hence helping in reducing the spam message
Email Spam Detection using integrated approach of Naïve Bayes and Particle Swarm Optimization, 2018 [14]	Dataset obtained from DATAMALL	f Naïve Bayes, KNN algorithm and Reverse DBSCAN algorithm	Using an interleaved water cycle and Simulated Annealing the number of features has decreased to more than 50%.Results show that the minimum number of features (17) was selected using high-level WCA-SA with NB while the best accuracy (95.6%) was achieved with SVM.
Email Spam Detection using integrated approach of Naïve Bayes and Particle Swarm Optimization,2018[14]	Dataset obtained from DATAMALL	Naïve Bayes, KNN algorithm and Reverse DBSCAN algorithm	The major drawback of the concept is that authors have not used any feature extraction technique.

ISSN PRINT 2319 1775 Online 2320 7876
Research paper Volume 11, Sp.Iss 5 , 2022

Email Classification Research Trends: Review and Open Issues, 2017 [15]	Email dataset (dataset containing both spam and non-spam used to train the classifier)	Feature set Analysis, Analysis of text classification techniques, Performance metrics and quantitative analysis	This review only focuses on email classification techniques, dataset analysis, features set analysis, and performance measure analysis due to limited scope of research.
Hybrid Decision Tree and Logistic Regression Classifier for Email Spam Detection, 2016 [16]	Spam base dataset from the UCI machine learning repository is used.	d LRFNT+DT a hybrid DT and LR with FN Threshold for email spam detection	DT drawback
A Comprehensive Study of Email Spam Botnet Detection, 2015 [17]	Two datasets which include two logs, a Hotmail user-login log, and a Hotmail signup log is utilized.	Various email spamming botnet detection techniques are used.	From the experiments, the authors have observed that the main activity concentrates on a small set of IPs particularly four IP addresses that harvested 70% of the email addresses, which ended up receiving 74% of the total spam.
Content Based Spam Detection in Email using Bayesian Classifier, 2015 [18]	Phish tank dataset is used	Spam filtering methods used	Applying the Bayesian Classifier,we experimentally demonstrated that spam mails can be detected with an accuracy of more than 96.46%
Leveraging Social Networks for Effective Spam Filtering, 2014 [19]	data is collected from Facebook	social network based bayesian spam filter, Adaptive Trust Management, False Negative rate and false positive rate, Bayesian are implemented methods.	A Social Network Aided Personalized and effective spam filter (SOAP) is proposed Each node uses SOAP to prevent spam autonomously
Text and Image Based Spam Email Classification using KNN, NaIve Bayes and Reverse DBSCAN Algorithm, 2014 [20]	Enron corpus datasets are used	Trial and error method is used.	Text filtering is that they are time consuming, OCR based detection also has disadvantages

3. Methodology

Figure 1 represents the block diagram of proposed work. We used the emails.csv [21] dataset collected from Kaggle. It has 5695 unique values with zero missing values and zero mismatched data, having a mean of 0.24 and a standard deviation of 0.43. It has two features namely text and spam.text feature is used as an input which is of string format. spam feature is the target attribute which contains binary values where 1 represents spam and 0 represents non-spam (ham).

ISSN PRINT 2319 1775 Online 2320 7876
Research paper **© 2012 IJFANS. All Rights Reserved**, Volume 11, Sp.Iss 5 , 2022

*Data Down Smapling :*Downsampling is a process of reducing the resolution or size of dataset, typically to save space or processing power. It involves decreasing the number of data points while retaining essential information.

Data Preprocessing : Data preprocessing involves cleaning, transforming, and organizing raw data into a format suitable for analysis or machine learning. It includes tasks like handling missing values, scaling features, and encoding categorical data.

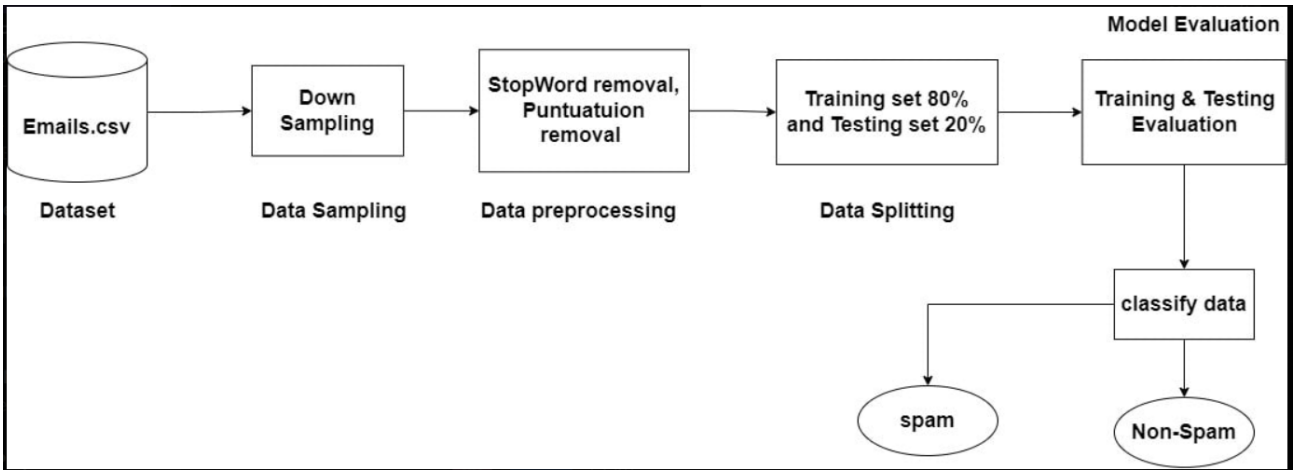

Figure 1 Block diagram of the proposed work

Procedure:

Step1:Import all the necessary libraries namely numpy, pandas, atplotlib.pyplot, seaborn, string, nltk, WordCloud from wordcloud, stopwords from ntlk.corpus,CountVectorizer from sklearn.feature_extraction.text.

Step2:Load the dataset from emails.csv

Step3:Visualize the count of spam and ham messages using a bar plot.And apply downsampling required.

Step4:Remove 'Subject' ,'punctuation' and 'stopwords' from the email text.

Step5:Generate word clouds for non-spam and spam emails to visualize the frequent words in each category.

Step6:Preprocess email text and convert it into numerical features using CountVectorizer.

Step7: Split the dataset into training and testing sets in the ratio 80:20 respectively.

Step8: Apply Classifier

Step8.1:Apply Decision Tree Model

Step8.1.1:Create a Decision Tree classifier by importing DecisionTreeClassifier from sklearn.tree

Step8.1.2:Train the Decision Tree classifier using the training data.

Step8.1.3:Make predictions on the test set using the trained model.

Step8.2:apply random forest model

Step8.2.1: Create a Random Forest classifier by importing RandomForestClassifier from sklearn.ensemble

Step8.2.2:Train the Random Forest classifier using the training data.

ISSN PRINT 2319 1775 Online 2320 7876
Research paper **© 2012 IJFANS. All Rights Reserved**, Volume 11, Sp.Iss 5 , 2022

Step8.2.3:Make predictions on the test set using the trained model.

.Step 8.3:Apply XGBoost Model

Step8.3.1: Create xgboost classifier by importing XGBClassifier from xgboost.

Step8.3.2:Train the XGBoost classifier using the training data.

Step8.3.3:Make predictions on the test set using the trained model.

Step 8.4: Apply LSTM Model

Step8.4.1: import libraries tensor flow, Tokenizer from tensorflow. keras. preprocessing. text, pad_sequences from from tensorflow.keras.preprocessing. Sequence, Early Stopping, ReduceLROnPlateau from keras.callbacks

Step8.4.2: Convert text data to sequences of tokens.

step8.4.3: Pad the sequences to ensure they have the same length (e.g., max_len = 100).

Step8.4.4: Build a Sequential model using TensorFlow/Keras.

Step8.4.5: Add an Embedding layer to learn vector representations of words.

Step 8.4.6: Add an LSTM layer to identify patterns in the text sequences.

Step 8.4.7: Include fully connected (Dense) layers with ReLU activation.

Step 8.4.8: Add an output layer with a sigmoid activation for binary classification.

Step 8.4.9: Set up early stopping (es) and learning rate reduction (lr) callbacks.

Step 8.4.10: Train the model on the training data using fit () with specified parameters, including the number of epochs and batch size.

Step9: Calculate evaluation or performance metrics: Accuracy, Precision, Recall, and F1 Score and compare the results among four implemented methods.

3. Results and Discussion

We have used Jupiter notebook platform to implement Decision Tree, Random Forest, XGBoost, and LSTM models. Accuracy, Precision, Recall, and F1 Score are the performance metrics used here to evaluate this text classification problem spam detection whose results are mentioned in Table 2.

Accuracy: Measures overall correctness of a model by calculating the ratio of correct predictions to the total predictions. It's simple but can be misleading for imbalanced datasets.

Precision: Evaluates how many positive predictions are actually correct, focusing on minimizing false positives, important when the cost of such errors is high.

Recall: Assesses the model's ability to capture all actual positive instances, concentrating on minimizing false negatives, crucial when missing positives is costly.

F1 Score: A balance between precision and recall, combining both metrics into a single value, useful when you need to consider false positives and false negatives in a balanced way.

Table 2 Performance analysis of Machine learning and deep learning models

Model	Accuracy	Precision	Recall	F1 Score
Decision Tree	0.952	0.901	0.91	0.906
Random Forest	0.959	1	0.838	0.912
XG Boost	0.984	0.966	0.972	0.969
LSTM	0.987	0.982	0.993	0.988

ISSN PRINT 2319 1775 Online 2320 7876
Research paper Volume 11, Sp.Iss 5 , 2022

Graphical Representation of performance metrics for proposed methods is shown in figure 2. It represents bargraph that compares accuracy,f1 score,precision and recall among four implemented methods. Figure 3 displays a line graph comparing the performance of four proposed methods using accuracy, F1 score, precision, and recall metrics.

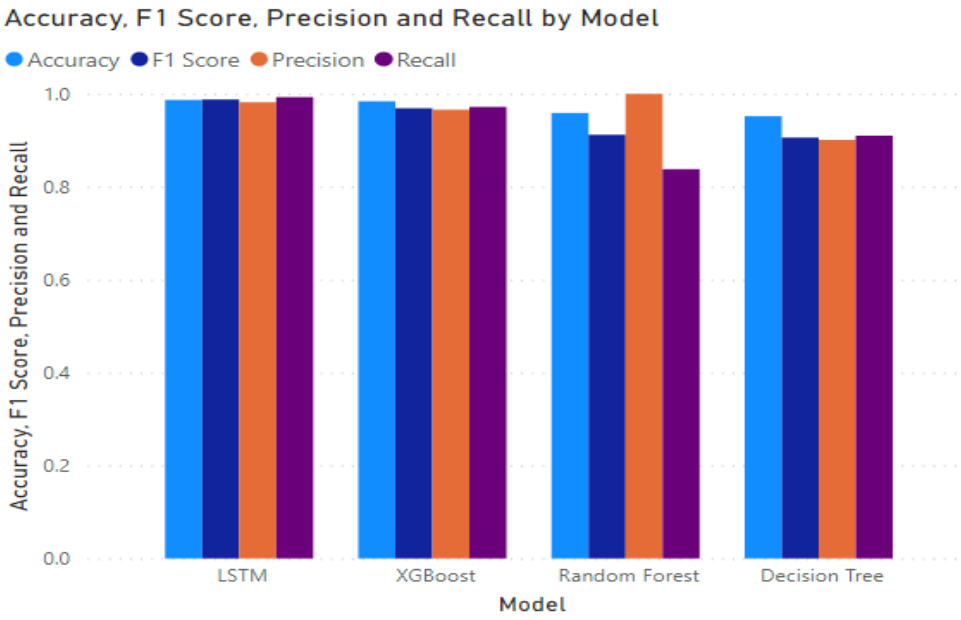

Figure 2 Graphical representation of performance metrics over 4 models

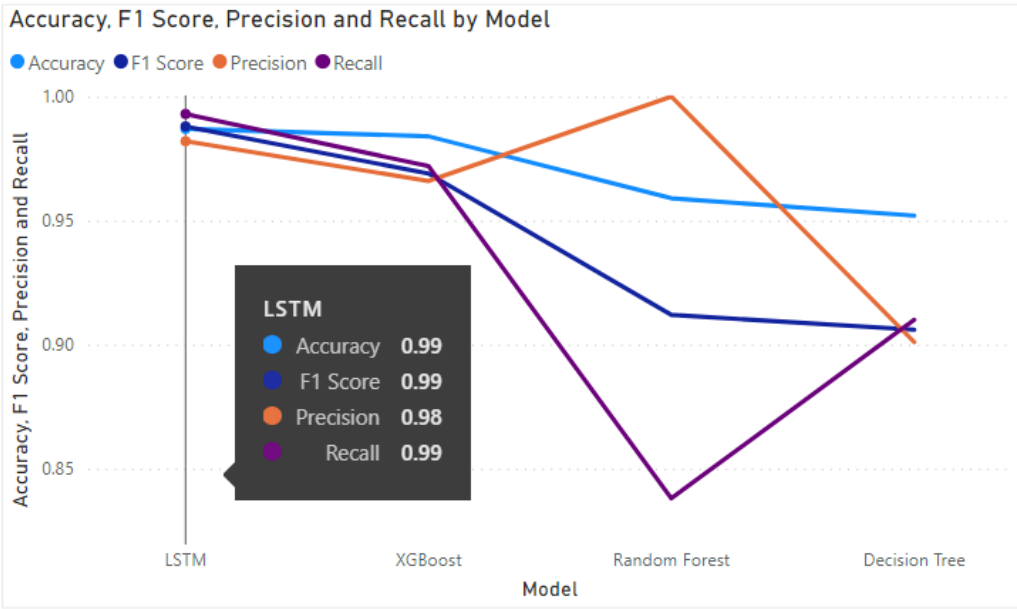

Figure3 Line Graph of performance metrics of 4 models

Among four applied models Long Short-Term Memory (LSTM) network has classified the test data into spam and ham with a highest accuracy of 0.987.

4. Conclusions

Spam classification using machine learning and deep learning algorithms is considered to be more efficient than traditional approaches like whitelisting/blacklisting. This study proposes four algorithms decision tree, random forest, xgboost and LSTM for filtering spam

ISSN PRINT 2319 1775 Online 2320 7876
Research paper **© 2012 IJFANS. All Rights Reserved**, Volume 11, Sp.Iss 5 , 2022

messages and LSTM showed high performance and accuracy. User feedback, such as marking emails as spam or not spam, is used to improve the accuracy of spam filters over time. The goal is to protect users from the annoyance and potential security risks associated with spam while ensuring that legitimate emails reach their intended recipients.

Acknowledgements

Sincere thanks to the management of KL University for providing good infrastructure to do extensive research.

References

1. Ezpeleta, Enaitz, et al. "Novel email spam detection method using sentiment analysis and personality recognition." Logic Journal of the IGPL 28.1 (2020): 83-94.
2. Kaddoura, Sanaa, Omar Alfandi, and Nadia Dahmani. "A spam email detection mechanism for English language text emails using deep learning approach." 2020 IEEE 29th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE). IEEE, 2020.
3. Gibson, Simran, et al. "Detecting spam email with machine learning optimized with bio-inspired metaheuristic algorithms." IEEE Access 8 (2020): 187914-187932.
4. Karim, Asif, et al. "Efficient clustering of emails into spam and ham: The foundational study of a comprehensive unsupervised framework." IEEE Access 8 (2020): 154759-154788.
5. Kumar, Nikhil, and Sanket Sonowal. "Email spam detection using machine learning algorithms." 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA). IEEE, 2020.
6. Sultana, Thashina, et al. "Email based Spam Detection." International Journal of Engineering Research & Technology (IJERT) (2020).
7. Al-Rawashdeh, Ghada, Rabiei Mamat, and Noor Hafhizah Binti Abd Rahim. "Hybrid water cycle optimization algorithm with simulated annealing for spam e-mail detection." IEEE Access 7 (2019): 143721-143734.
8. Agarwal, Kriti, and Tarun Kumar. "Email spam detection using integrated approach of Naïve Bayes and particle swarm optimization." 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS). IEEE, 2018.
9. Mujtaba, Ghulam, et al. "Email classification research trends: review and open issues." IEEE Access 5 (2017): 9044-9064.
10. Wijaya, Adi, and Achmad Bisri. "Hybrid decision tree and logistic regression classifier for email spam detection." 2016 8th International Conference on Information Technology and Electrical Engineering (ICITEE). IEEE, 2016.
11. Khan, Wazir Zada, et al. "A comprehensive study of email spam botnet detection." IEEE Communications Surveys & Tutorials 17.4 (2015): 2271-2295.
12. Rathod, Sunil B., and Tareek M. Pattewar. "Content based spam detection in email using Bayesian classifier." 2015 International Conference on Communications and Signal Processing (ICCSP). IEEE, 2015.
13. Shen, Haiying, and Ze Li. "Leveraging social networks for effective spam filtering." IEEE Transactions on Computers 63.11 (2013): 2743-2759.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper **Volume 11, Sp.Iss 5 , 2022**

14. Harisinghaney, Anirudh, et al. "Text and image based spam email classification using KNN, Naïve Bayes and Reverse DBSCAN algorithm." 2014 International Conference on Reliability Optimization and Information Technology (ICROIT). IEEE, 2014.