Classification Analysis for e-mail Spam using Machine Learning and Feed Forward Neural Network Approaches

Srinivasa Rao Dangeti [1], Dileep Kumar Kadali [2,*], Yesujyothi Yerramsetti [3], Ch Raja Rajeswari [4], D. Venkata Naga Raju [5], Srinath Ravuri [6]

[1,2,*,3,4,5,6] Shri Vishnu Engineering College for Women, Bhimavaram
dileepkumarkadali@gmail.com

Abstract. In the present era, electronic communication plays an essential role in our daily lives. However, this convenience is accompanied by the persistent challenge of email spam, which inundates inboxes and poses a serious cybersecurity threat. Email spam remains a pervasive issue, with conventional spam filters often struggling to adapt to evolving spamming techniques. This paper aims to leverage machine learning advanced techniques to enhance the accuracy and efficiency of email spam classification. By employing state-of-the-art algorithms and models, the goal is to develop a robust and adaptable system capable of effectively identifying and filtering out spam emails. Several machine learning classifiers namely KNN, SVC, DT, NB, RF and Logistic Regression are applied. Later, a deep learning Feed Forward Neural Network model was applied and achieved good accuracy. The experiments' outcome showed that the proposed deep learning gave good accuracy for email spam classification.

Keywords: Email spam, Machine Learning, Deep Learning, Classification, Accuracy etc.

1 Introduction

The ubiquity of email as a primary mode of communication has brought unparalleled convenience, yet it has also given rise to the persistent challenge of email spam. Spam, characterized by unsolicited and often malicious content, not only clutters inboxes but poses significant threats to individuals and organizations alike. Conventional spam filters, relying on predefined rules and heuristics, struggle to keep pace with the ever-evolving tactics employed by spammers. Over the years, spammers have become increasingly sophisticated, employing deceptive tactics to evade traditional filters. From disguised phishing attempts to polymorphic malware, the landscape of email spam is dynamic and complex. As a result, static and rule-based filters find it challenging to discern between legitimate and malicious content effectively. There are several drawbacks with email spam. Email spam may spread phishing and viruses and erode user confidence. As users sort through spam, they lose productivity and take up network bandwidth and storage, costing enterprises more. False positives and negatives in spam screening might miss crucial emails or flag real ones. Spam is worldwide and spammers' strategies change; thus, concerted efforts are needed to reduce it. Spam

K. R. Madhavi et al. (eds.), *Proceedings of the International Conference on Computational Innovations and Emerging Trends (ICCIET 2024)*, Advances in Computer Science Research 112,
https://doi.org/10.2991/978-94-6463-471-6_7

management's environmental impact is another worry. Email spam must be combated by technological advances, user education, and international cooperation. Machine learning, with its ability to learn from data and adapt to changing patterns, presents a promising avenue for addressing the shortcomings of traditional spam filters. By training models on diverse datasets containing both legitimate and spam emails, machine learning algorithms can discern intricate patterns and anomalies, offering a more nuanced approach to email classification. Because of its versatility and capacity to recognize complicated patterns, Machine Learning (ML) is vital for email spam categorization. ML models can automatically learn and update their decision-making processes, making them effective against developing spam methods. These algorithms specialize on spam detection, enhance accuracy by examining varied datasets, and decrease false positives by analyzing word context. ML extraction of key email properties, fast handling of vast and diverse datasets, and real-time detection contribute to a dynamic and evolving spam categorization system. Personalized models provide powerful and specialized defences against the ever-changing email spam environment. In this paper, we applied several ML classifiers and FFNN for email spam classification. The KNN algorithm stands as a robust supervised machine learning technique, adept at handling both classification and regression tasks. Positioned within the realm of instance-based and lazy learning algorithms, KNN doesn't construct an explicit model during training. Instead, its predictive prowess relies on assessing the resemblance between new data points and established examples within the training dataset. The fundamental concept underlying SVC is the identification of a hyperplane that optimally separates data points of different classes. In a two-dimensional space, this hyperplane is a line; in higher dimensions, it becomes a hyperplane. The "support vectors" are the points that are nearest to the decision border and play a crucial role in defining it. DTC is a popular supervised ML technique for classification problems. Iterative dataset partitioning based on the most relevant characteristics at each decision node creates a tree-shaped structure. Decision rules at these nodes homogenize subsets. The method continues until a stopping requirement is met, resulting in leaf nodes providing class labels. Decision trees' interpretable rules and non-parametric nature allow them to capture complex connections. Management of tree depth to minimize overfitting, impurity measurements, and pruning for maximum performance are crucial. Decision trees are useful in credit scoring, medical diagnosis, and customer churn prediction. Bayes' theorem-based probabilistic machine learning method the Naive Bayes Classifier is known for its simplicity and efficiency in classification applications. It excels in text-related tasks like spam filtering and sentiment analysis because it assumes conditional independence across characteristics given the class label. Naive Bayes works even when the independence requirement is violated because of its simplicity and quickness. In email filtering, text categorization, and medical diagnosis, Multinomial, Gaussian, and Bernoulli Naive Bayes are commonly used. Due to its resilience, accuracy, and feature significance insights, the Random Forest Classifier is used in many fields. Its ensemble technique uses many decision trees to make it a powerful machine learning tool for classification and regression. A prominent binary classification method, logistic regression, employs the sigmoid function to transform data into a probability range between 0 and 1. Linear decision boundaries and interpretable coefficients make Logistic Regression useful for spam detection and sickness diagnosis. When interpretability is required and the relationship between attributes and the target variable is linear, the model's computational efficiency and simplicity are excellent. A Feedforward Neural Network, or multilayer perceptron, is a basic deep learning architecture with input, hidden, and output layers. Neurons in each layer execute weighted summations, apply activation functions for non-linearity, and learn hierarchical input data representations via unidirectional flow. These universal function approximators excel in picture classification, natural language processing, and financial forecasts after backpropagation and weight and bias adjustments. Their versatility to capture complex patterns and correlations makes them essential to deep learning models. Feedforward neural networks offer several advantages over conventional machine learning (ML) algorithms, contributing to their widespread adoption in various applications. Feedforward neural networks can automatically learn hierarchical representations of features in the data. Unlike conventional ML algorithms that may struggle with capturing complex relationships, neural networks excel at discerning intricate patterns and extracting relevant features at multiple levels. Neural networks, with their activation functions and multiple layers, can model non-linear relationships effectively. This enables them to handle complex mappings between input and output variables, surpassing the limitations of linear models often employed in conventional ML algorithms.

2 Literature Review

Harsha Dinendra et al. [1] intended to provide a solution based on machine learning that could categorize emails that were not spam and highlight the relevance of these communications. A variety of machine learning models were developed and trained for use in the research by making use of non-spam emails taken from the personal inbox of the first author. Notable examples of algorithms that exhibited significant accuracy included the DT, RF, and deep neural networks. This report summarized the outcomes that were produced as well as provided some insights into the modelling process. To find the best spam classification model, the authors in [2] tested NB, SVC and Random Forests. To maximize model performance, hyperparameters were tweaked. Classifier performance was assessed by accuracy, recall, and F1-score. The research produced a spam classification model tailored for email integration. Using automated spam filtering increased email security and worker efficiency. The research also sought to improve NLP and ML methods for email spam classification. K. Iqbal et al.[3] applied several ML classifiers for spam classification and achieved good results. They applied SVM, KNN, RF and DT algorithms. In [4], word embedding specifically, the pre-trained transformer model BERT was used to classify spam emails. Utilizing attention layers to incorporate text context, the fine-tuned BERT model achieved 99% accuracy and a 98.67% F1 score in spam email detection. The outcomes were examined in comparison to a baseline DNN model that included k-NN and NB classifiers, a BiLSTM layer, and two stacked Dense layers. The model was trained and assessed for robustness and persistence against unknown data using two public datasets. In [5], the authors applied different ML classifiers for email spam detection and achieved a good outcome. They applied three algorithms namely logistic regression, SVM and Naïve Bayes. In [6], the authors suggested using ML models for cybersecurity issues, especially email spam classification.

LSTM, CNN and General Neural Boards (GNB) are applied by authors in [7]. The well-known open-source Spam dataset, categorized and forecasted using this algorithm, was comprised of about 6000 actual email samples. The models were evaluated and trained, and a comprehensive report on the research results was included in the publication. The CNN-LSTM model gave a remarkable prediction score of 98.78% on the spam dataset. The Deep Neural Network (DNN) and Min-hash collaborated to categorize emails into Spam and Ham, boasting an impressive accuracy rate of 98.2% [8]. The authors claimed combination worked as an outstanding system for spam detection and categorization, suggesting its implementation and potential for further enhancements. V. S. Vinitha et al. [9] applied deep learning RNN for email spam detection. As RNN is easily suitable with text datasets, the authors claimed that RNN provided a good detection rate for email spam classification. The RNN with several architectures tested and the model with a good detection rate was retained as the final model. The authors used several variations of RNN including LSTMs and GRUs for email spam classification.

3 Proposed Model

Classification and regression are the key examination regions in managed learning, and this strategy is now and again in building forecast models. It likewise lessens how much thinking is expected to track down suitable autonomous factors.

3.1. Classifiers Using Machine Learning Techniques

Each sort of classifier utilized as an AI strategy can be evaluated as having a few pay and challenges in expectations; for instance, ANN enjoys the benefits of client active and casual to use for the utilization of the results. The enormous advantage of the ANN is to grow an unassuming and open calculation because of AI, which will convey a real arrangement conversely, with customary assessment methods. To put it plainly, we can say that ANN is flexible and is utilized for unique solicitations and issue arrangements. It is a multipurpose classifier [14]. It very well may be utilized for some applications and issues. Notwithstanding, the profits, of ANN likewise have a few downsides; for instance, it will become more slow with immense information that is ready for preparation [10]. Notwithstanding the above benefits of the strategy of the credulous Bayes classifier, we want a slight amount of preparing information, which will be restored into a legitimate order that can be handily applied with a high credit rate [12].

The proposed method for email spam classification is shown in Fig 1. Initially, an email spam dataset from Kaggle was collected. The dataset is verified for data preprocessing. After preprocessing the dataset, it has no missing values. After that, Several Machine Learning classification algorithms namely KNN, SVC, DT, NB, and RF applied to the final dataset. The results with ML models are tabulated. Later, a deep learning Feed Forward Neural Network model was proposed for email spam detection. The performance of all applied ML and DL models compared and best model was used for email spam classification.

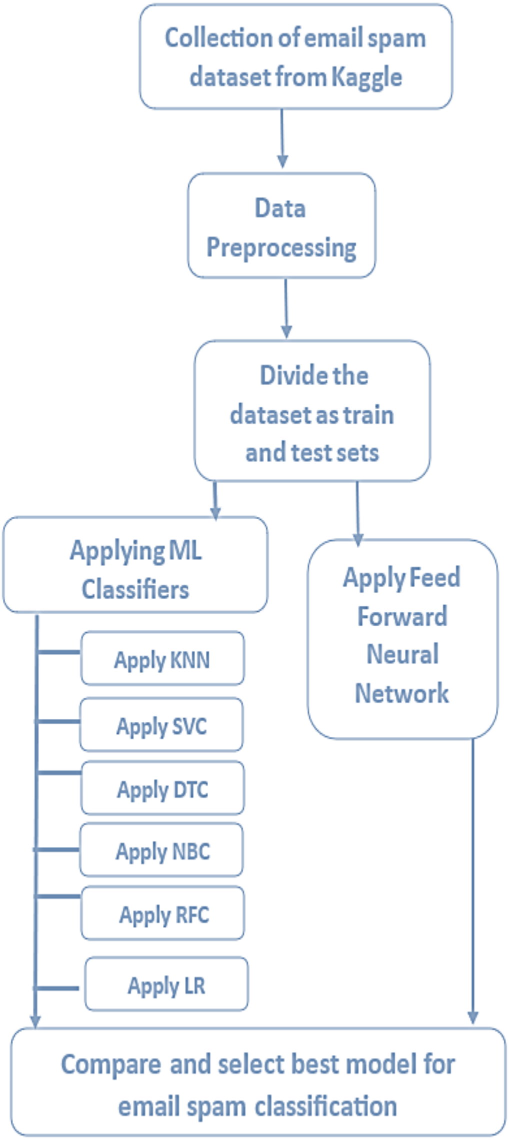
Collection of email spam dataset from Kaggle
Data Preprocessing
Divide the dataset as train and test sets
Applying ML Classifiers
Apply KNN
Apply SVC
Apply DTC
Apply NBC
Apply RFC
Apply LR
Apply Feed Forward Neural Network
Compare and select best model for email spam classification

Fig. 1. Proposed Model

4 Results and Discussion

Data assortment

The dataset is checked for missing values. There are no missing values in the dataset. So, there is no need to apply preprocessing techniques. The dataset contains 3000 features. As the actual dataset is a text dataset, it is already applied with the bag of words model. So, all the words in the actual dataset are 3000, which are represented as columns in the dataset. The attribute values of the dataset are numbers indicating several times the word appeared in the whole email dataset. The number of positive and negative samples in the dataset is shown in Fig.2. There total of 5,172 samples in the dataset, where 3,672 positive samples and 1500 negative samples.

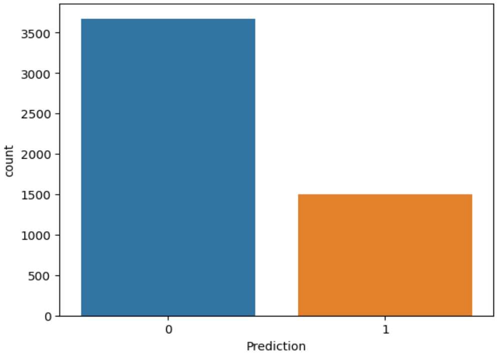

Fig 2. Number of positive and negative samples in the dataset

Applying ML Algorithms classifies applied the dataset. Six methos applied namely K-NN, SVC, DTC, NBC, RFC and LGC. The results achieved with these algorithms are shown in Table 1.

Table 1. Result of ML Algorithms

Alg(Classifier)	Precision	Recall	F1-score	Accuracy
KNN-C	0.83	0.85	0,84	85.7
SV-C	83	85	84	80.3
DT-C	82	69	72	93.04
RF-C	92	91	92	96.9
NB-C	96	96	96	94.02
LR-C	92	94	93	96.32

Fig 3 shows precision, recall and accuracy values for all six algorithms. Fig 4 shows the accuracy values obtained with the proposed six algorithms.

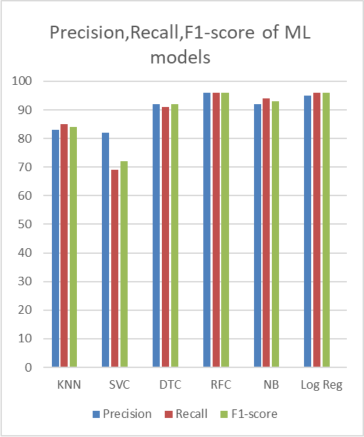

Fig 3. Precision, Recall,f1-score of ML Algorithms

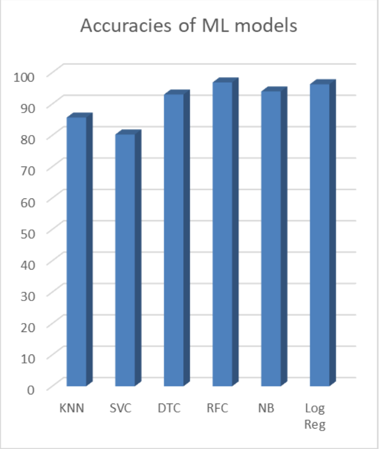

Fig 4. Accuracy of ML Algorithms

From Figure 3, it is observed that, among six applied algorithms, Random Forest gives the highest accuracy with an accuracy value of 96.9%. From the Figure, it is observed that the precision, recall and f1-score recall values are also 96% which is good for the random forest. The next better classifier is logistic regression with an accuracy of 96.32%. The precision, recall and f1-score for logistic regression are 92%, 94% and 93% respectively. Next, Decision Tree and Naïve Bayes also gave reasonable accuracies with values of 93% and 94%. The remaining classifiers' accuracy is less than 90% only. After experimenting with ML classifiers, it is observed that RF and Logistic Regression are the best classifiers for email spam classification. The best accuracy achieved with ML classifiers is 96.9% with RF. To increase the performance, a deep learning FFNN was applied. The FFNN model contains five hidden layers. The neurons in the hidden layers are 1000,700,500,200 and 100. The number epochs in the model is 50. The accuracy achieved with FFNN is 98.3%. The comparison of best ML classifiers (RF, LR) and FFNN is shown in Fig.5.

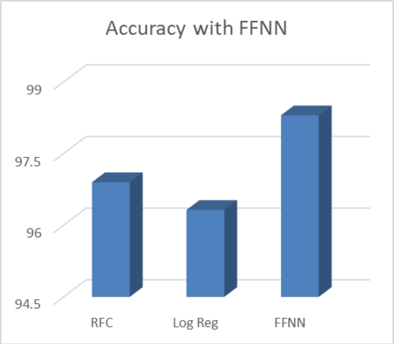

Fig 5. Accuracy Comparison between RF and ANN

5 CONCLUSION

In this paper, an email spam detection model was proposed based on ML and FFNN techniques. A dataset with email spam data was collected from the Kaggle repository. After verifying the cleanliness of the dataset, ML algorithms namely, KNN, SVC, DT, NB, RF and Logistic Regression applied. Among six applied ML models, Random Forest given the highest accuracy of 96.9%. Later, logistic regression gave a good accuracy of 96.4%. Later, a deep learning Feed Forward Neural Network (FFNN) was applied. The accuracy acquired with FFNN was 98.3%, which is higher than Random Forest. The outcomes exposed that the proposed FFNN performed well for email spam classification.

References

1. H. Dinendra et al., “Personalized Classification of Non-Spam Emails Using Machine Learning Techniques,” International Research Conference on Smart Computing and Systems Engineering, Colombo, Sri Lanka, 2022, pp. 171-177.
2. Mrs. Anitha Reddy et al., “Email Spam Detection Using MachineLearning” Journal of Survey in Fisheries Sciences, 10(1) 2658-2664,2023.
3. K.Iqbal et al., “Email classification analysis using machine learning techniques”, Applied Computing and Informatics, May 2022, https://doi.org/10.1108/ACI-01-2022-0012.
4. I. AbdulNabi and Q. Yaseen, “Spam Email Detection Using Deep Learning Techniques,” Procedia Computer Science, vol. 184. Elsevier BV, pp. 853–858, 2021. doi: 10.1016/j.procs.2021.03.107.
5. D. K. Kadali and R. Mohan, “Shortest route analysis for High-Level Slotting using Peer-to-Peer,” in Apple Academic Press eBooks, 2022, pp. 113–122. doi: 10.1201/9781003048367-10.
6. G. V. Sesha Sai Krishna Vineeth et al., “Email Spam: A New Strategy of Screening Spam Emails using Natural Language Processing,” 2023 Third International Conference on Artificial Intelligence and Smart Energy (ICAIS), Coimbatore, India, 2023, pp. 710-715.

7. D. K. Kadali, D. Raju, and P. V. R. Raju, "Cluster query optimization technique using Blockchain," in Cognitive science and technology, 2023, pp. 631–638. doi: 10.1007/978-981-99-2742-5_65.
8. J. M. Muhammad et al., "Classification and Prediction of Spam Emails Based on AI Enabling Models Using Deep and Machine Learning Techniques" International Conference on Emerging Technologies in Electronics, Computing and Communication, 2022, pp. 1-6.
9. D. K. Kadali, "Cluster optimization for similarity process using De-Duplication," Sep. 01, 2016. https://ijsrd.com/Article.php?manuscript=IJSRDV4I60433
10. Madhavi, K. Reddy, K. Suneetha, K. Srujan Raju, Padmavathi Kora, Gudavalli Madhavi, and Suresh Kallam. "Detection of COVID 19 using X-ray Images with Fine-tuned Transfer Learning." Journal of Scientific and Industrial Research (2023): 241-248.
11. D. K. Kadali, R. Mohan, N. Padhy, S. C. Satapathy, N. Salimath, and R. D. Sah, "Machine learning approach for corona virus disease extrapolation: A case study," International Journal of Knowledge-based and Intelligent Engineering Systems, vol. 26, no. 3, pp. 219–227, Dec. 2022, doi: 10.3233/kes-220015.
12. D. K. Kadali, R. Mohan, and M. C. Naik, "Enhancing crime cluster reliability using neutrosophic logic and a Three-Stage model," Journal of Engineering Science and Technology Review, vol. 16, no. 4, pp. 35–40, Jan. 2023, doi: 10.25103/jestr.164.05.
13. A. Lakshmanarao, M.Shashi, "A survey on machine learning for cyber security," International Journal of Scientific and Technology Research Volume 9, Issue 1, Pages 499 – 502, January 2020
14. D. K. Kadali, M. C. Naik, and K. N. Remani, "Estimation of data parameters using cluster optimization," in Lecture notes on data engineering and communications technologies, 2022, pp. 331–342. doi: 10.1007/978-981-19-2600-6_23.
15. Chitteti, Chengamma, and K. Reddy Madhavi. "Taylor African vulture optimization algorithm with hybrid deep convolution neural network for image captioning system." Multimedia Tools and Applications (2024): 1-19.
16. D. K. Kadali, J. Mohan, and Y. Vamsidhar, "Similarity based Query Optimization on Map Reduce using Euler Angle Oriented Approach," https://www.ijser.org/, Jan. 2012,
K. N. Remani, V. S. Naresh, S. Reddi, and D. K. Kadali, "Crime data optimization using neutrosophic logic based game theory," Concurrency and Computation: Practice and Experience, vol. 34, no. 15, Mar. 2022, doi: 10.1002/cpe.6973.
17.Murty, S., Prasad, M., Raja, V., P. Kiran Sree, G. Ramesh Babu and Ch. Phaneendra Varma (2023). A Hybrid Intelligent Cryptography Algorithm for Distributed Big Data Storage in Cloud Computing Security. Lecture Notes in Computer Science, pp.637–648. doi:https://doi.org/10.1007/978-3-031-36402-0_59.
18.Prasad Maddula, Srikanth, P., P. Kiran Sree, P.B.V. Raja Rao and Murty, S. (2023). COVID-19 prediction with Chest X-Ray images using CNN. doi:https://doi.org/10.1109/iitcee57236.2023.10090951.
19.K. Reddy, A. Vinaya Babu, A. Anand Rao, and S. V. N. Raju. "Identification of optimal cluster centroid of multi-variable functions for clustering concept-drift categorical data." In Proceedings of the International Conference on Advances in Computing, Communications and Informatics, pp. 124-128. 2012.
20.Mohan, Sri Aravind Desamsetti, Karunasri Adina, Padma Jyothi Uppalapati, Murty, S. and RajaRao P. B. V (2024). Creating a Protected Virtual Learning Space: A Comprehensive Strategy for Security and User Experience in Online Education. pp.350–361. doi:https://doi.org/10.1007/978-3-031-48888-7_30.
21.https://www.kaggle.com/datasets/balaka18/email-spam-classification-dataset-csv/data