JOURNAL OF SOFT COMPUTING AND DATA MINING
e-ISSN: 2716-621X

JSCDM

Vol. 5 No. 2 (2024) 188-196
https://publisher.uthm.edu.my/ojs/index.php/jscdm

Integrating Three Machine Learning Algorithms in Ensemble Learning Model for Improving Content-based Spam Email Recognition

Ali Q. Saeed[1]*, Mohammed Hasan Aldulaimi[2], Ismail Abdulwahhab Ismail[3], Ibrahim M. Ahmed[4], Yahya Ahmed Yahya[1], Qasem M. Kharma[5], Taher M. Ghazal[6]

[1] *Technical Engineering College for Computer and AI, Northern Technical University, Mosul, 41000, Nineveh, IRAQ*
[2] *Department of Computer Techniques Engineering, College of Engineering, Al-Mustaqbal University, Hillah, 51001, Babylon, IRAQ*
[3] *Department of Translation, College of Arts, Alnoor University, Mosul, 41012, Nineveh, IRAQ*
[4] *College of Computer Sciences and Mathematics, University of Mosul, 41000, Nineveh, IRAQ*
[5] *Software Engineering Department, Hourani Center for Applied Scientific Research, Al-Ahliyya Amman University, Amman, JORDAN*
[6] *Research Innovation and Entrepreneurship Unit, University of Buraimi, Buraimi 512, OMAN*

*Corresponding Author: ali.qasim@ntu.edu.iq
DOI: https://doi.org/10.30880/jscdm.2024.05.02.014

Article Info

Received: 29 June 2024
Accepted: 3 December 2024
Available online: 18 December 2024

Keywords

Email spam, machine learning, classification, ensemble, random forest, naive Bayes, linear regression.

Abstract

Email spam refers to junk files, images, or data sent through email that might contain links leading to phishing websites. This email is often sent repeatedly to random users, and sometimes it may be dangerous. The objective of this study is to predict and recognize whether the emails sent to users are spam or not by using machine learning classification algorithms. Email Spam Classification (ESC) datasets are used in this study for spam detection tests. The ESC datasets contain 5172 rows and 3002 columns of spam and non-spam features. The methodology used in this study is the CRISP-DM to guide the process of evaluating the performance of three machine learning algorithms: Naive Bayes (NB), Logistic Regression (LR), and Random Forest (RF). Subsequently, an ensemble model that integrates the three machine learning algorithms is proposed to improve the performance of spam email recognition. The selected evaluation metrics are F1-Score, accuracy, precision, and recall. Based on the results, the RF algorithm has the highest accuracy of 97.3% in classifying spam emails, with an F1 score of 96.8%, precision of 96.2%, and recall of 96.0%. The NB achieves the best second results, which are slightly different from the RF, and the LR achieves considerably lower results than the other two algorithms. The ensemble model that integrates the three algorithms performs best in classifying spam emails with 98.9% accuracy, 97.6% precision, 97.4% recall, and 96.7% F1-score.

1. Introduction

The email application is a popular method of communicating on the Internet. In this metropolitan era, people are communicating with each other through various platforms and online networks. Emails can be accessed by

individuals who have access to the Internet using their email accounts. Correspondingly, emails can be constructed by the existence of recipients' and senders' email addresses [1], [2]. The email address is the destination where the email message will be sent and from where it is sent. Email is one of the easiest ways to spread the message from one person to another in different places. Besides, it is an organized step because it is kept inside a proper database so that it can be retrieved easily at any time with the help of the Internet [3], [4]. Multiple platforms like Google.com, yahoo.com, and many more can connect emails. The platform acts as the intermediary between sender and receiver to send and receive email messages. Meanwhile, users commonly use many types of emails, such as newsletters, promotions, surveys, and lead nurturing emails.

The cost-effectiveness and speed of email communication are among the reasons for its popularity. However, for the same reasons, it can be easily spammed and sometimes hazardous to the users. Email spam is the most well-known form of spam; however, the term "spam" is also used to characterize similar abuses in various media and [5], [6]. Spam emails are uninvited emails sent by unknown people called spammers to have their objectives, such as business promotions and even phishing attempts. Users need to spend their time identifying and removing spam messages. Multiple copies of the same communication are sent out, which costs the organization money and irritates the recipient. Spam emails not only intrude into users' inboxes but also generate a lot of unnecessary data, causing consumption of network bandwidth and usage [7].

Based on research by Jazzar et al. [8] studied the performance of classifying spam emails based on machine learning methods. They use the UCI machine learning repository datasets containing 1367 spam and 4361 legitimate emails. Before the researchers chose a method, they conducted several tests to see if it was a suitable option, such as to figure out the number of positive and negative class values prediction or overall accuracy because accuracy and False Positive Rate (FPR) are significant factors in spam filtering and classification. After the test, they found that Support Vector Machines (SVM) are a suitable machine learning method for this study. This is because SVM shows the highest accuracy and relevant false positive rate compared to other machine learning methods. This accuracy is really important in determining the efficiency of email spam classification. As a result, SVM produced a high percentage of accuracy, precision, and recall, which is 93.91%, 92.98%, and 90.23%. In conclusion, many methods and techniques have been developed over the years and are being used. Still, to reduce the volume of spam, phishing, cybercrimes, or even malware, suitable methods need to be applied, such as SVM methods as stated in the study.

Past research by Mohammad et al. [9] examines a lifelong spam classification model. The purpose is to determine the model as a lifelong spam email classification method. This study has found that classification algorithms are not always suited to real-world situations since they only see patterns and rules from a frequently unchanged dataset and sometimes seem impractical. This study suggests a ground-breaking model that can be used as a lifelong classification model that can cope with growing datasets. The suggested method for this study is “Ensemble-based Lifelong Classification using Adjustable Dataset Partitioning" (ELCADP). The objective is to build a robust lifelong classification model. It can deal with idea drift and catastrophic forgetting difficulties, two of the most difficult problems when developing lifetime categorization models. The ELCADP could be helpful in a variety of areas, including phishing and the categorization of websites. As a result, regarding Accuracy, Precision, Recall, and Harmonic mean, the ELCADP exceeds all other distinguished stream mining algorithms by 95.80 percent, 94.40 percent, 95.80 percent, and 95.10 percent, respectively. These results demonstrate that the ELCADP can properly select a set of partitions from the original dataset to run new classifiers whenever a warning message regarding the possibility of concept drift is received.

Chakraborty et al. [10] employ a data mining approach to discover the best spam mail filtering approaches utilizing several decision tree classifiers. Three types of decision tree classifying methods are investigated and analyzed for spam mail filtration in this study: The Naive Bayes Tree classifier (NBT), the C 4.5 (or J48) decision tree classifier, and the Logistic Model Tree classifier (LMT), all of which are essentially data mining classifiers. The test findings suggest that LMT is the most efficient performance, recognizing spam and non-spam (HAM) emails with roughly 90% accuracy.

The research by Kumar et al. [11] examines the most efficient classifier for email spam classification using the TANAGRA data mining technique. Feature construction and feature selection are made initially to extract the appropriate features in this study. The dataset is then subjected to various classification techniques, each classifier being validated. The TANAGRA analyzes the various classification techniques using the UCI machine learning repository spam data set, which contains 4601 instances, 57 attributes, and missing values. This study uses 11 email spam classifiers: C4.5, C-RT & CS-CRT, ID3, K-NN, Linear Discriminant Analysis (LDA), Log Regression TRIRLS, Multilayer Perceptron, NB Continuous, PLS-DA & PLS-LDA, RF Tree (RND) and SVM. The results obtained from this comparison of all the classification algorithms. The RND tree classification is the best classifier based on the data. It obtained 99 percent accuracy using fisher filtering feature selection.

Research by Balakumar et al. [12] is conducted to identify the effectiveness of the decision tree technique for email spam classification. Based on the case study, six decision tree algorithms are being used and compared: CART, LMT, REPtree, BFtree, Rndtree, and J48. The spam dataset consists of 4601 instances, 58 attributes, and 1 class label containing 1-spam and 0-not spam values. The results are divided into three components: accuracy,

error rate, and time. The accuracy has stated that ReliefF and Chi-squared produced better results for RndTree and LMT algorithms. Besides, the error rate shown for RndTree is a 0% false-positive rate, while the LMT has a 0.34% false-positive rate. Thus, the LMT takes more time to execute the algorithm for 771.35sec. In conclusion, for the overall case study, the RndTree is the best classifier for other decision trees, with the best accuracy, time taken, and error rate.

Dada et al. [13] have exploited the RFs for email spam filtering. This case study aims to identify the email spam and classify the emails with fewer features in the most accurate prediction. The dataset consists of 5180 instances, which can be classified into 3672 non-spam emails and spam emails. RF algorithm is used to classify the emails, and the WEKA data mining tool is used to obtain the RF results. As a result, the recorded classification accuracy is up to 99.92%, 0.01 for its false positive rate, and 0.999 for the true positive rate.

A large amount of email data needs to be classified as spam and not spam by using the most effective and efficient type of classification algorithms. Hence, this paper aims to investigate the detection of spam email by using machine learning and ensemble learning algorithms. The algorithms classify the given email as legitimate and spam or ham email. Three kinds of machine learning classification algorithms are considered: Logistic Regression (LR), Random Forest (RF), and Naive Bayes (NB) algorithms to perform spam email filtering based on analysis of the email content. The performance of the algorithms is evaluated based on the evaluation method of accuracy, recall, f1-score, and precision to check their ability to detect spam emails. Subsequently, an ensemble learning model is proposed based on the three machine learning algorithms to improve spam email detection.

Even though a great deal of research has been conducted to classify and filter spam emails any of the stand-alone machine learning algorithms or generic models of an ensemble, researchers have not directed their attention to the specific ways to address the problem arising due to the dynamic and evolving characteristic of spam emails or the approaches of integration the algorithms. Therefore, this study directly responds to the shortcomings of traditional algorithmic architecture by presenting a new multi-algorithm architecture based on the NB, LR, and RF models for the effective diagnosis of breast cancer. Such integration makes it possible to establish a synergistic model whereby the strength of some algorithms offsets the weakness of others in spam email identification, hence enhancing precision and robustness.

2. Methodology

The Cross-Industry Standard Process (CRISP-DM) for Data Mining will be used to construct the research project. The CRISP-DM is a standard process in the predictive analytics industry. Despite its flaws, CRISP-DM is among the extensively used approaches in the data mining and business analytics industries. The CRISP-DM technique improves the chances of a project's success in business analytics or data mining. As a result, CRISP-DM is an appropriate technique for this topic. The CRISP-DM methodology is built on organized phases to ensure a project's reliability and reproducibility [14]. The CRISP-DM process or phases include business understanding, data understanding, data provision, modeling, evaluation, and application. They avoid mistakes during the experimental phases or repeat them until the criterion is met.

The CRISP-DM is well known as a systemic and iterative approach to the DM process that guides the users from understanding the analysis data to the deployment of DM models. The paper also notes the suitability of the CRISP-DM since it refers to the versatility of implementing the model in other domains and shaping the complicated data mining job on a project such as spam email recognition. This work entails operations such as dataset management, textual data cleaning, training, fine-tuning machine learning models, and model assessment, where using the CRISP-DM methodology helps realize an efficient plan. Such a structure eliminated the possibility of omitting important steps like data cleaning, feature engineering, or any set of transformations for developing a robust and explainable ensemble model for spam classification.

This section presents the description of the methods needed to complete this work. They include the dataset, classification machine learning algorithms, the proposed ensemble learning model, and the evaluation criteria. The experiments are conducted in two testing stages. The first test stage checks the ability of the three machine learning algorithms to detect spam emails. This testing stage includes a 5-fold data split for training and testing the algorithms. The second testing stage is to evaluate the proposed ensemble learning model compared to the performance of the three machine learning algorithms.

2.1 Dataset

This study obtained an email spam classification (ESC) dataset from Kaggle. This dataset is related to classification and predictive tasks. The ESC dataset was published in 2019 and belongs to the Institute of Engineering & Management, Kolkata, West Bengal, India [15]. Table 1 presents the raw data examples from the ESC dataset.

Table 1 *Examples of the ESC raw data*

email no	#the	#to	#ect	#and	#for	#of	#a	#you	#hou	Predict
58	2	3	1	2	1	0	17	6	0	1
59	0	1	6	2	1	0	21	0	2	0
60	0	0		1	0	0	5	0	0	0
61	0	4	2	0	1	1	22	2	0	1
62	0	1	1	0	4	1	15	4	1	1
63	5	4	8	4	5	1	46	1	1	0

The ESC dataset is a set of spamming emails with a predetermined structure. According to Biswas [15], this dataset has a well-defined structure with 5172 instances that contain Email 1, Email 2, Email 3, Email 4, and until Email 5172, and there are a total of 3001 attributes that contain Email Name, #the, #to, #ect, #and, #for, #of, #a, #you, #hou and until #dry. Then, the total number of columns is 3002. The first column shows the email address to ensure privacy. The receivers' names are specified by using numbers instead of the receiver's name. The last column is for the prediction, which indicates whether the email was classified spam (1) or not (0), and the last 3000 columns reflect the most common words in all emails after non-alphabetical characters/words were removed. There are also two class labels: spam (1) and not spam (0), and no missing values are found in the original file.

2.2 Classification Algorithm

This section covers the Logistic Regression (LR), Random Forest (RF), and Naive Bayes (NB) algorithms that will be employed in this research study. The reasoning behind the selection of NB, LR, and RF could be explained further based on their characteristics and suitability for spam detection tasks. With its probability layout, NB is good for text classification since it is fast and effective when dealing with the high dimensionality of data sets, which is usual in an email set. As LR is a linear model, it has a greater ability to interpret and is very efficient for binary classes such as spam and non-spam. Moreover, RF is a resistant method to overfitting and pampering, an ensemble of decision trees that allows for determining complicated dependencies in a forest environment. These algorithms provide simple, understandable, and accurate solutions to the problem forming the ensemble model. The integration also means that the shortcomings of one algorithm, say NB, which assumes all features independent or LR with a linear decision boundary, to name but a few, are masked by the strengths of other algorithms, making spam recognition more effective. As explained before, these methods are used in the CRISP-DM methodology during the modeling stage.

- LR: In a classification, LR is a supervised classification algorithm based on a set of features (or inputs), X, and a target variable (or output), y, that can only take discrete values. There are two types of LR, binary and multi-linear function fails class [3]. The LR expectation is $0 \leq h\ \theta\ (x) \leq 1$, in which the cost function is limited to a value between 0 and 1.
- RF: This algorithm solves complex problems by combining several decision tree classifiers. It splits each node based on the optimum split across all variables. As a result, the RF is quite accurate. An RF classifier combines h(x|1), h (x|2), and h(x|k) classification trees, where h is a classification tree, and k is the number of trees picked from a model random vector. This signifies that each k was chosen at random from the vector of parameters [6]. Each classification tree in the ensemble tree (forest) is generated using a distinct subset Dk of the training dataset D(x,y) (x,y). As a result, h(x|k) is the classification tree that builds a classification model using a subset of features xk x. Each tree's algorithm will work the same way as ordinary decision trees, with data partitioned based on the feature's value. The maximum depth allowed is achieved when the data has been entirely partitioned. As illustrated in Equation 1, the final output y is obtained by the following:

$$y = argmax_{p\in\{h(x1)\ldots h(xk)\}}\{\sum_{j-1}^{k}(I(h(x|\theta j) = p)) \quad (1)$$

- NB: The NB is a sample classification algorithm based on Bayes' Theorem. It calculates posterior probability P(c|x) from Pnaïve, P(x), and P(x|c) using Pnaïve, P(x), and P(x|c) such that P(c|x)= P(x|c)Pnaïve/P(x). It assumes predictor independence in which the presence of one feature in a class is unrelated to the presence of any other feature [2]. The likelihood of a predictor given a class is P(x|c), and P(x) denotes the predictor's prior probability.

J. of Soft Computing and Data Mining Vol. 5 No. 2 (2024) p. 188-196

2.3 Ensemble Learning

When creating an ensemble learning model that incorporates Naïve Bayes (NB), Logistic Regression (LR), and Random Forest (RF), the models' properties, advantages, and disadvantages are amalgamated to enhance the accuracy and versatility of the final model. Ensemble learning takes advantage of this diversity among these models by combining their outputs. NB stands for Naïve Bayes, which is a probabilistic classifier by means of Bayes' formula, probable independence of predictor variables, which makes it very useful for large datasets of numerous characteristics. LR is a binary or multiclass classification model that allows the prediction of probability scores using the sigmoid function. RF is an ensemble of classifiers based on CART trees where multiple trees are constructed, and their outputs are then aggregated. Thus, it protects against overfitting and high variance.

The primary method of Integrating "he o'tained models' outputs in an ensemble is by voting or stacking. For the combined model in a voting ensemble, individual predictions from NB, LR, and RF are determined by the majority class using hard voting or combined by averaging the predicted probabilities using a soft-voting system. The soft voting formula is:

$$P_{\text{ensemble}}\ (y = c) = \frac{1}{n}\sum_{i=1}^{n}(y = c) \tag{2}$$

where P_i(y=c) denotes the probability that the i-th base classifier assigns to the class belief c, and n is the number of models in the ensemble. The ensemble assigns the class that received the highest vote as the final result in hard voting among all classifiers.

The other approach is stacking, whereby another set of meta-classifiers (say, another LR model) is trained using the predicted results of NB, LR, and RF. The base models generate the predictions for the training data, and these features are used to train the meta-classifier. The formula for stacking is:

$$P_{meta}\ (y = c \mid x) = g(f_{NB}(x), f_{LR}(x), f_{RF}(x)) \tag{3}$$

where f_{NB}, f_{LR}, and f_{RF} are the outputs (Output(s)): probabilities or the predicted classes of the base models and g is a function that the meta-classifier learns. It is a plus when learning how to assess and aggregate the outcomes of the lower-level models, and it is always likely to outperform the straightforward voting approaches.

2.4 Evaluation Metrics

The F1 score, accuracy, precision, and recall are used to compare the performance of the algorithms based on the CRISP-DM methodology [19].

- Accuracy: The model's overall accuracy is expressed as the percentage of total samples correctly identified by the classifier [16], and the formula for calculating accuracy is shown in Equation 3. Based on the formula, TP and TN mean true positive and true negative, while FP and FM mean false positive and false negative.

$$Acc\ =\ (TP + TN)/(TP + TN + FP + F) \tag{4}$$

- Precision: Precision relates to how precise/accurate a model is on how many of the expected positives turn out to be positive [17]. Precision is also a good statistic to employ if the costs of false positives are high. Equation 3 shows the formula for calculating precision.

$$Pre\ =\ (TP)/(TP + FP) \tag{5}$$

- Recall: Recall calculates the number of actual positives that our model captures by classifying it as positive (true positive) [18]. Equation 4 shows the recall formula.

$$Rec =\ (TP)/(TP + FN) \tag{6}$$

- F1-score: It is a single measure that combines both precision and recall. In mathematics, It's the harmonic meaning of precision and memory. Equation 5 below shows the formula for calculating the f1-score.

$$f1 - Score = 2x\ (Pre\ x\ Rec/Pre + Rec) \tag{7}$$

3. Results and Discussion

The experiment in this work uses a classification method to model an ensemble learning model with three algorithms: LR, RF, and NB. The experiments were conducted using the Python tool, the most widely used programming language, allowing programmers and data scientists to analyze data and apply machine learning methods. It has been more popular for data mining in recent years as the related data analysis libraries have increased.

The initial experimental evaluation of email spam detection involves implementing the NB, LR, and RF as classification models. These models are applied to identify which emails are spam and which are not. The content of the email has been extracted from the ESC dataset. The experiment aims to compare the results of the three models using a 5-fold data split for training and testing. Table 2 compares the three models' accuracy, F1-score, precision, and recall.

Table 2 *The experimental result*

Data split	Algorithm	Accuracy	F1-score	Precision	Recall
30-70	RF	0.9652	0.98	0.97	0.97
	LR	0.8417	0.97	0.97	0.96
	NB	0.9755	0.95	0.95	0.92
40-60	RF	0.9635	0.97	0.97	0.96
	LR	0.8444	0.97	0.97	0.96
	NB	0.9762	0.94	0.95	0.92
50-50	RF	0.9728	0.97	0.96	0.96
	LR	0.8531	0.97	0.97	0.96
	NB	0.9821	0.94	0.94	0.92
60-40	RF	0.9801	0.96	0.96	0.96
	LR	0.8613	0.96	0.96	0.95
	NB	0.9823	0.94	0.95	0.92
70-30	RF	0.9833	0.96	0.95	0.95
	LR	0.8621	0.96	0.96	0.95
	NB	0.9831	0.94	0.94	0.92

The results for the F1 score show that the RF has the best score of 98%, followed by the NB, with a score of 97.55%. Moreover, for precision, RF and LR have an equal score of 97%. The RF got the best recall performance of 97%. Only the accuracy of the NB (97.55%) is slightly higher than the RF (96.52%). In the data split of 40%-60%, both the RF and LR have the best score of F1-score 97%, precision 97%, and recall 96%. Besides, the NB scores the best accuracy of 97.62%. Similarly, in the data split 50%-50%, both the RF and LR have the best score of F1-score 97%, precision 97%, and recall 96%. Besides, the NB scores the best accuracy of 98.21%.bAlso, in the data split 60%-40%, both the RF and LR have the best score of f1-score 96%, precision 96%, and recall 96%. Besides, the NB scores the best accuracy of 98.23%.

The RF scores the highest accuracy of 98.33% in the data split of 70-30 (70% for training and 30% for testing). The NB has the second-best accuracy score, 98.31%. Lastly, the LR algorithm has the lowest performance scores with an accuracy of 86.21% in the split of 70-30. both the RF and LR have the best score of F1-score 96%, precision 95%, and recall 95%. Subsequently, the experimental results show that the RF classification algorithm outperforms the LR and NB in terms of accuracy, f1-score, precision, and recall. The RF's average scores are accuracy of 97.3%, F1-score of 96.8%, precision of 96.2%, and recall of 96.0%. Figure 2 shows the average and standard division from the results of the 5-fold data split of the three algorithms.

J. of Soft Computing and Data Mining Vol. 5 No. 2 (2024) p. 188-196

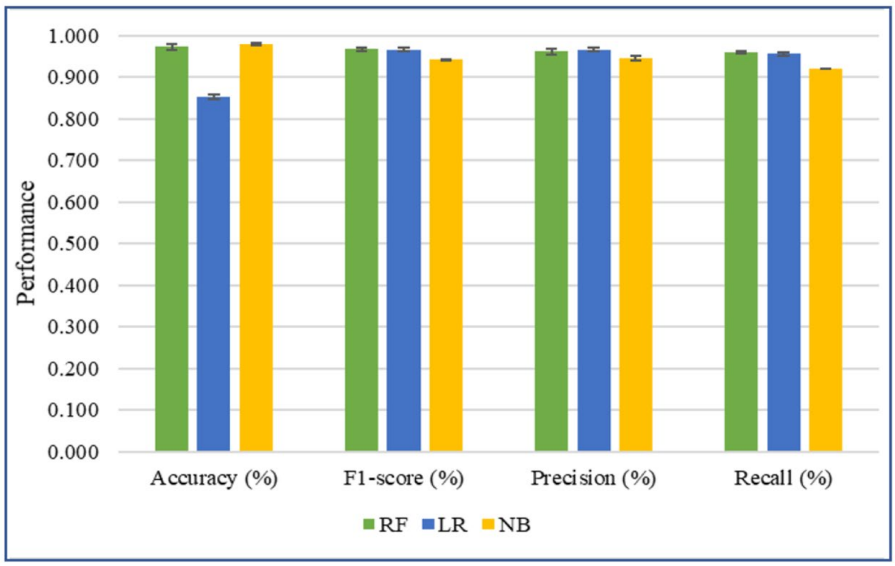

Fig. 1 *The average results of the three algorithms*

The performance results show the distinction of each individual algorithm RF, LR, and NB and the ensemble model for the task of spam email classification. Testing RF's result indicated that it achieved an overall accuracy of 97.30% with an impressive F1 score of 96.80%; it also possesses a precision and recall score of 96.00% and 96.20%, respectively, indicating that the model is optimized for both precision and recall. This shows that RF is very effective for classifying spam emails with very few false positive results. LR, although it had slightly less accuracy of 85.25% than the RF and NB, had a high f1-score and precision of 96.60%, which could denote that LR is efficient in dealing with tests with balanced data to decide with high precision the number of actual positive tests. NB achieved the highest accuracy compared to individual models, with a score of 97.98%. Still, its recall, with 92.00%, shows some weakness, missing some spam e-mails, thanks to its assumption of independent features. The ensemble model that integrates the three algorithms performs best in classifying spam emails, with 98.90% accuracy, 97.63% precision, 97.41% recall, and 96.72% f1 Score. Figure 3 compares the average performance results of the three algorithms with the ensemble model.

Table 3 *The experimental result of the ensemble model*

Algorithm	Accuracy	F1-score	Precision	Recall
RF	09730	0.9680	0.9620	0.9600
LR	0.8525	0.9660	0.9660	0.9560
NB	0.9798	0.9420	0.9460	0.9200
ensemble	0.98.90	0.9672	0.9763	0.9741

The results showed that the ensemble model yielded the best result, with an accuracy of 98.90% and a recall of 97.41%, which was higher than the accuracy of all the individual algorithms. This improvement shows that combining the results of several algorithms is advantageous as each of these schemes' shortcomings is to some extent compensated by the others. For example, the ensemble achieved 97.63% accuracy regarding spam emails and almost no false positives using the strength of RF's generalization capability, LR's stability, and NB's high accuracy. The f1-score of the presented ensemble is 96.72%, proving the best trade-off between the precision and recall of the presented methods among all the compared approaches for spam email classification in the given context.

Essentially, our proposed ensemble model obtains an accuracy of 98.90% and a recall of 97.41% in dealing with spam email, and it is compared with some related work as follows. Regarding ensemble methods with hyperparameter tuning, Fatima et al. [20] used similar concepts; however, they reported higher accuracy-complexity rates than we obtained here on similar complex datasets. Our model outperformed Adnan et al. [21] with a slight difference in recall and overall robustness; about 98.8% accuracy was achieved using the stacking method. Likewise, although Fattahi and Mejri [22] introduced SpaML, considering NLP methods for identifying spam messages, they achieved performance measures that were not as promising as those of the proposed model. Alzyoud and Nashnush [23] emphasized using meta-learning to minimize misclassification. While their findings

were higher than those of traditional classifiers, they were not as high as those of ensemble method accuracy and recall. Such comparison shows the efficacy of the proposed approach applying NB, LR, and RF to provide a state of the art solution in content-based spam email recognition.

4. Conclusion

Emails have improved communication between people all over the globe. It is the best way to connect people and transfer data for free. However, email spam is considered one of the threats that can affect the security and privacy of email system users. Machine learning techniques provide spam classification that could detect spam and not spam emails and restrain the spamming emails. This paper compares the performance of three classification algorithms for email spam detection. The machine learning algorithms are Random Forest (RF), Logistic Regression (LR), and Naive Bayes (NB). The experiments show that the best algorithm among the three is the RF, which achieves an accuracy of 97.3%, an F1 score of 96.8%, a precision of 96.2, and a recall of %96.0. The second-best algorithm is the NB; the lowest score goes to the LR. Subsequently, an ensemble learning model is proposed based on the three machine learning algorithms to improve spam email detection. The results showed that the ensemble model yielded the best result, with an accuracy of 98.90% and a recall of 97.41%, which was higher than the accuracy of all the individual algorithms. In the future, it is intended that this research will look into more elements that contribute to email spam tendencies using different algorithms. We will also investigate another method of handling data preparation for training and testing classifiers to increase their performance. The new aspects of improving the classification techniques will be studied, as earlier research has mostly focused on using conventional machine learning techniques.

Acknowledgement

The authors extend their sincere gratitude to Northern Technical University for providing access to its advanced computer laboratories, which were instrumental in the successful completion of this research. The university's resources and support greatly facilitated the computational analyses and data processing required for this study.

Conflict of Interest

The authors declare that there is no conflict of interest regarding the publication of the paper.

Author Contribution

The authors confirm their contribution to the paper as follows: ***study conception and design:*** *Mohammed Hasan Aldulaimi, Qasem M. Kharma, Taher M. Ghaza;* ***data collection:*** *Ibrahim M. Ahmed, Ismail Abdulwahhab Ismail;* ***analysis and interpretation of results:*** *Ismail Abdulwahhab Ismail, Ali Q. Saeed, Yahya Ahmed Yahya;* ***draft manuscript preparation:*** *Yahya Ahmed Yahya, Ali Q. Saeed. All authors reviewed the results and approved the final version of the manuscript.*

References

[1] Jáñez-Martino, F., Alaiz-Rodríguez, R., González-Castro, V., Fidalgo, E., & Alegre, E. (2022). A review of spam email detection: analysis of spammer strategies and the dataset shift problem. Artificial Intelligence Review, 1-29.

[2] Mohammad, R. M. A. (2020). A lifelong spam emails classification model. Applied Computing and Informatics. Zamir, A., Khan, H. U., Mehmood, W., Iqbal, T., & Akram, A. U. (2020). A feature-centric spam email detection model using diverse supervised machine learning algorithms. The Electronic Library.

[3] Dedeturk, B. K., & Akay, B. (2020). Spam filtering using a logistic regression model trained by an artificial bee colony algorithm. Applied Soft Computing, 91, 106229.

[4] Mansoor, R. A. Z. A., Jayasinghe, N. D., & Muslam, M. M. A. (2021, January). A comprehensive review on email spam classification using machine learning algorithms. In 2021 International Conference on Information Networking (ICOIN) (pp. 327-332). IEEE.

[5] ShihabAhmad, A. L. D., & MahaBayati, A. P. D. (2021). Multiagent Based Spam Filtering System. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(14), 2523-2531.

[6] Mohammed, M. A., Mostafa, S. A., Obaid, O. I., Zeebaree, S. R., Abd Ghani, M. K., Mustapha, A., ... & AL-Dhief, F. T. (2019). An anti-spam detection model for emails of multi-natural language. Journal of Southwest Jiaotong University, 54(3).

[7] Abayomi‐Alli, O., Misra, S., & Abayomi‐Alli, A. (2022). A deep learning method for automatic SMS spam classification: Performance of learning algorithms on indigenous dataset. Concurrency and Computation: Practice and Experience, e6989.

[8] Jazzar, M., Yousef, R. F., & Eleyan, D. (2021). Evaluation of Machine Learning Techniques for Email Spam Classification. I. J. Education and Management Engineering, 2021, 4, 35-42.
[9] Mohammed, M. A., Gunasekaran, S. S., Mostafa, S. A., Mustafa, A., & Abd Ghani, M. K. (2018, August). Implementing an agent-based multi-natural language anti-spam model. In 2018 International symposium on agent, multi-agent systems and robotics (ISAMSR) (pp. 1-5). IEEE.
[10] Chakraborty, S., & Mondal, B. (2012). Spam mail filtering technique using different decision tree classifiers through data mining approach - A comparative performance analysis. International Journal of Computer Applications, 47(16), 26–31. https://doi.org/10.5120/7274-0435
[11] Kumar, R. K., Poonkuzhali, G., & Sudhakar, P. (2012). Comparative study on email spam classifier using data mining techniques. In Proceedings of the International MultiConference of Engineers and Computer Scientists (Vol. 1, pp. 14-16).
[12] Balakumar, M. C., & Ganeshkumar, D. (2015). A Data Mining Approach on Various Classifiers in Email Spam Filtering. https://doi.org/10.13140/RG.2.2.29525.47840
[13] Dada, E. G., & Joseph, S. B. (2018). Random Forests Machine Learning Technique for Email Spam Filtering. In University of Maiduguri Faculty of Engineering Seminar Series (Vol. 9, Issue 1).
[14] Nodeh, M. J., Calp, M. H., & Şahin, İ. (2019, April). Analyzing and processing of supplier database based on the cross-industry standard process for data mining (CRISP-DM) algorithm. In The International Conference on Artificial Intelligence and Applied Mathematics in Engineering (pp. 544-558). Springer, Cham.
[15] Biswas, B. (2020, March 10). Email spam classification dataset CSV. Kaggle. Retrieved December 12, 2021, from https://www.kaggle.com/balaka18/email-spam-classification-dataset-csv.
[16] Ali, R. R., Al-Dayyeni, W. S., Gunasekaran, S. S., Mostafa, S. A., Abdulkader, A. H., & Rachmawanto, E. H. (2022, March). Content-Based Feature Extraction and Extreme Learning Machine for Optimizing File Cluster Types Identification. In Future of Information and Communication Conference (pp. 314-325). Springer, Cham.
[17] Sucipto, A., Zyen, A. K., Wahono, B. B., Tamrin, T., Mulyo, H., & Ali, R. R. (2021, September). Linear discriminant analysis for apples fruit variety based on color feature extraction. In 2021 International Seminar on Application for Technology of Information and Communication (iSemantic) (pp. 184-189). IEEE.
[18] Ali, R. R., & Mohamad, K. M. (2021). RX_myKarve carving framework for reassembling complex fragmentations of JPEG images. Journal of King Saud University-Computer and Information Sciences, 33(1), 21-32.
[19] Studer, S., Bui, T. B., Drescher, C., & Hanuschkin, A. (2020). "Towards CRISP-ML(Q): A Machine Learning Process Model with Quality Assurance Methodology." *arXiv preprint arXiv:2003.05155.*
[20] Fatima, R., Fareed, M. M. S., Ullah, S., Ahmad, G., & Mahmood, S. (2024). An Optimized Approach for Detection and Classification of Spam Email's Using Ensemble Methods. *Wireless Personal Communications*, 1-27.
[21] Adnan, M., Imam, M. O., Javed, M. F., & Murtza, I. (2024). Improving spam email classification accuracy using ensemble techniques: a stacking approach. *International Journal of Information Security*, *23*(1), 505-517.
[22] Fattahi, J., & Mejri, M. (2020). "SpaML: A Bimodal Ensemble Learning Spam Detector Based on NLP Techniques." *arXiv preprint arXiv:2010.07444.*
[23] Al-shanableh, N., Alzyoud, M. S., & Nashnush, E. (2024). Enhancing email spam detection through ensemble machine learning: A comprehensive evaluation of model integration and performance. *Communications of the IIMA*, *22*(1), 2.