Published online on the journal's webpage: **http://journal.ittelkom-pwt.ac.id/index.php/dinda**

Journal of Dinda

Data Science, Information Technology, and Data Analytics

Vol. 4 No. 1 (2024) 8 - 13 E-ISSN: 2809-8064

Random Forest Machine Learning for Spam Email Classification

Rizky Ageng N[1*], Rafdhani Faisal[2], Solahuddin Ihsan[3]

[1,2,3*]Software Engineering, Faculty of Informatics, Institut Teknologi Telkom Purwokerto

[1*]21104045@ittelkom-pwt.ac.id, [2]21104043@ittelkom-pwt.ac.id, [3]21104072@ittelkom-pwt.ac.id

Abstract

This research discusses the crucial role of email as a main element in digital communication, facilitating information transfer and serving as an advertising platform. However, the problem of email spam, which involves sending unsolicited commercial messages, has had negative impacts such as consuming large amounts of resources and disrupting user experience. With its affordable cost and ease of sending messages to thousands of recipients, email spam includes product promotions, pornographic material, viruses and irrelevant content. The impact includes loss of time and damage to the user's computer resources. To address this problem, email services provide advanced spam filters that use email content analysis and machine learning techniques. This research focuses on the use of the Random Forest Classification algorithm as a basis for filtering spam emails. Although Random Forest is known to have strong classification capabilities, the risk of overfitting is a challenge. Therefore, this study adopts the Randomized Search CV method to identify the best parameter combination, ensuring the reliability of the model in dealing with the complexity of diverse email datasets. With this approach, this research contributes to the development of effective solutions to reduce the impact of email spam in digital communications.

Keywords: *Spam Email. Random Forest, Confusion Matrix, ROC-AUC, Randomized Search CV*

1. Introduction

Email (electronic email) is a crucial element used in communicating digitally over the internet. In addition to functioning to transfer information in the form of files, email can also be used as a means for advertising purposes. Electronic messaging is now the top choice in communicating, allowing users to easily send messages simply by being connected to the internet [1].

Spam, also known as unsolicited commercial email or unsolicited excessive email, has created various problems in our daily communication. The negative impact caused by spam includes the use of large resources, such as network bandwidth and storage space. Examples of spam cases include gambling ads and pornographic material [1]. Considering the affordable cost and ease of sending messages to many recipients, some parties use it to send product or service promotions, pornographic materials, viruses, and content that is considered irrelevant to thousands of email users [2].

Spam emails can be a nuisance and cause significant losses, as they can waste time and damaging users' computer resources [3]. To deal with email spam problems, a few email services provide sophisticated spam filters. These filters operate by analyzing the content of an email and comparing it to a list of emails already known as spam. Some filters also adopt machine learning techniques to assess whether an email can be categorized as spam or not [3].

This research focuses on the use of the Random Forest Classification algorithm as the main foundation for filtering spam emails. Random Forest is known to have strong classification capabilities, but there is a risk of overfitting which can reduce the reliability of the model on new data. To overcome this challenge, the study used the Randomized Search CV method. This approach helps identify the best combination of parameters for the model, optimizing the balance between precision and generalization. Thus, Randomized Search CV is a critical step in ensuring the reliability of the Random Forest Classification model in overcoming the complexity of diverse email datasets.

2. Research Methods

The research phase starts from collecting data that is used as a dataset to be included as test data and train data, preprocessing, split data, applying classification to get accuracy [4]. Here's the diagram shown in Figure 1.

Received: 08-01-2024 | Revised: 10-02-2024 | Published: 15-02-2024

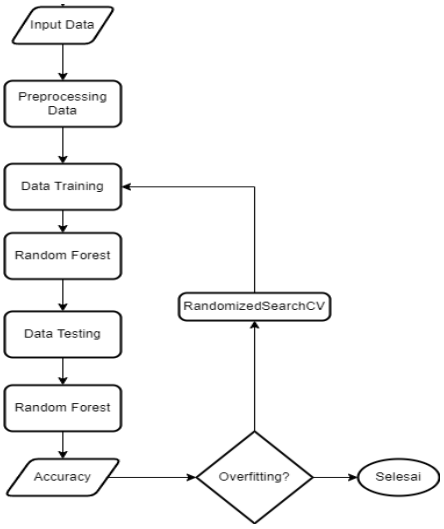

Figure 1. Flowchart Research

2.1. *Collecting and Preprocessing Data*

The dataset consists of 5172 entries representing each email in CSV file format. Each row represents one email with a total of 3002 columns. The first column presents the Name of the email, which is deliberately labeled numerically to maintain the privacy of the email recipient. Meanwhile, the last column is identified as a prediction label, where a value of 1 indicates that the email is spam, while 0 indicates that it is not spam. The remaining 3000 columns include the most common words in all emails, after ignoring non-alphabetic characters and words. For each row, information about the number of occurrences of each word in the email is stored in the corresponding cell. With this approach, all information related to 5172 emails is stored in a structured manner in the CSV data framework, replacing the need to save each email as a separate text file [5]. The following is a table of the division of train data and test data in Tables 1 and 2.

Table 1. Data Training

	the	to	…	ff	dry
2146	3	9	…	0	0
4672	8	3	…	0	0
2152	0	0	…	0	0
4424	33	12	…	1	1
4585	9	4	…	0	0

Table 2. Data testing

	the	to	…	ff	dry
1511	8	3	…	0	0
2212	7	8	…	0	0
3937	10	6	…	3	0
963	1	2	…	0	0
3891	2	2	…	0	0

In the data processing phase, crucial steps are taken to resolve null values in the dataset. This procedure involves identifying and removing null values, so that the data used for training and testing becomes cleaner and more consistent. After processing null values, the next step is the process of dividing data. Data is divided into two main parts: train data and test data. This process ensures that the model can learn from most datasets to then test on never-before-seen data, thus ensuring good generalization. By eliminating null values and dividing the data carefully, this stage of data processing forms a solid basis for the development of accurate and reliable classification models [6].

2.2. *Data Split*

In the context of classification, testing data sets is an important step to evaluate the accuracy and performance of a method. In the distribution of data for training and testing purposes, an 80:20 ratio is used, where 80% is used as training data and 20% as test data. After the data sharing process is complete, the next step is to classify using the method being tested [7].

2.3. *Random Forest*

Random Forest (RF) is an algorithm that uses a binary recursive separation approach to reach the final node in the tree structure, based on the concepts of classification and regression trees [8]. This approach was chosen for its high reputation for accuracy and its ability to cope with small samples as well as feature spaces with high dimensions [9].

Random Forest Classifier is a classification method formed from a set of decision trees. The process involves using the voting results of the trees to obtain the final result in the detection of sarcasm. This method is supported by independent training data and random features, which vary from one feature to another [10]. Here is a visual representation of the decision tree structure that can be seen in Figure 2.

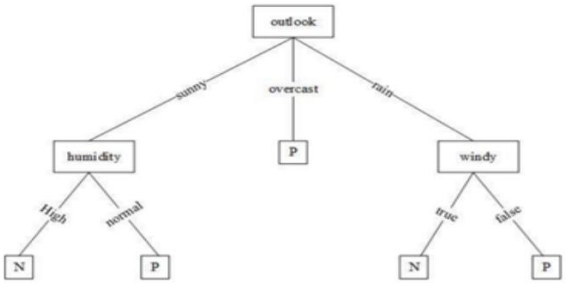

Figure 2. Random Forest Process

The initial step in determining the decision tree involves calculating the value of entropy and information gain [11]. The approach using entropy aims to evaluate the degree of impurity of attributes, while information gain

measures the value of information obtained in separating vertices [12]. Gini calculations have a significant impact on the top node and separator node [13]. The Gini calculation process continues until the final Gini value reaches zero.

2.4. *Accuracy*

Initially, to determine accuracy through the confusion matrix. Confusion Matrix is a method to display the accuracy results of the model that has been created [14]. The Confusion Matrix resumes performance in classifying according to the number of categories classified based on the correct value of a predicted class of objects [15].

Table 3. Confussion Matrix

No	Confusion Matrix	Predicted Positive	Predicted Negative
1	Actual Positive	TP	FN
2	Actual Negative	FP	TN

TP = The amount of positive data clarified correct

TN = The amount of negative data clarified is incorrect

FP = The amount of positive data clarified is incorrect

FN = The amount of negative data clarified correct

After getting the calculation results from the Confusion Matrix, the next step is to create an ROC (Receiver Operating Characteristic) curve based on the relationship between false positives and True Positives [16]. Performance evaluation on the Random Forest algorithm was carried out using Receiver Operating Characteristic (ROC) analysis to measure the level of accuracy, sensitivity, and specificity [17].

Table 4. ROC Criteria

Nilai AUC	Interpretasi
0.90 – 1.00	Excellent Classification
0.80 – 0.90	Good Classification
0.70 – 0.80	Fair Classification
0.60 – 0.70	Poor Classification
0.50 – 0.60	Failure

If accuracy is overfitting, the step taken to overcome the problem is to use RandomizedSearchCV [18]. RandomizedSearchCV is a parameter tuning method that helps identify the best combination of parameters for the model [19]. By utilizing this approach, the ultimate goal is to optimize the balance between precision and model generalization [20]. RandomizedSearchCV effectively helps overcome the risk of overfitting, which can reduce the reliability of the Random Forest Classification model especially in the face of the complexity of diverse email datasets. Thus, the use of RandomizedSearchCV is a critical step to ensure that the Random Forest Classification model is able to provide good performance without being affected by overfitting new data.

```
Train Results

Confusion Matrix for Train :
 [[2782    0]
 [ 736  361]]
Accuracy Score for Train :  0.8102603763856664
ROC AUC for Train :  0.9843855571072535
OOB Score for Train :  0.7963392626965713
```

3. Results and Discussion

The test is carried out by assessing the level of accuracy of the results of spam email identification carried out by the system. From the results of this test, we can determine the parameters that provide the best level of precision using the Random Forest Classifier.

The application of the Random Forest classification model succeeded in achieving an accuracy level of 1.0 and an Area Under the Receiver Operating Characteristic Curve (ROC AUC) value of 1.0 in the training data, as shown in Figure 3. This shows that the model is very effective in identifying and classifying training data with an optimal level of accuracy.

```
++++++++++++++++++++++++++++++++++++++++++++++++++
Train Results

Confusion Matrix for Train :
 [[2782    0]
 [   0 1097]]
Accuracy Score for Train :  1.0
ROC AUC for Train :  1.0
```

Figure 3. Training Result

Meanwhile, in the testing phase using test data, the Random Forest Classifier model still showed excellent performance with an accuracy level of around 0.97 and an ROC AUC value of around 0.99, as seen in Figure 4.

```
++++++++++++++++++++++++++++++++++++++++++++++++++
Test Results

Confusion Matrix for Test :
 [[876  14]
 [ 20 383]]
Accuracy Score for Test :  0.9737045630317092
ROC AUC for Test :  0.9964661109097499
```

Figure 4. Testing Result

The model's ability to make predictions with a high degree of precision, even on never-before-seen data, demonstrates its success in generalizing patterns from training data to test data. However, it is important to note that with accuracy scores and ROC_AUC scores reaching 1.0, there are indications that point to overfitting, which means the model learns too deeply from the training data.

Therefore, as a solution to overcome this problem, we will use the RandomizedSearchCV (Cross-Validation) method to find the best parameters in the Random Forest model, hoping to improve predictions and prevent overfitting, thus ensuring the reliability and credibility of the model in more general situations.

```
Train Results

Confusion Matrix for Train :
 [[2782    0]
 [ 736  361]]
Accuracy Score for Train :  0.8102603763856664
ROC AUC for Train :  0.9843855571072535
OOB Score for Train :  0.7963392626965713
```

Figure 5. Train Result

After successfully overcoming the overfitting problem, the Random Forest model was reimplemented with satisfactory results. At the testing stage with training data, accuracy reached 0.8, indicating the model's ability to classify training data with a good level of precision. In addition, the Area Under the Receiver Operating Characteristic Curve (ROC AUC) value also increased to 0.9, reflecting the quality of the model's predictions against different levels of sensitivity and specificity.

```
++++++++++++++++++++++++++++++++++++++++++++++++++
Test Results

Confusion Matrix for Test :
 [[889   1]
 [307  96]]
Accuracy Score for Test :  0.7617942768754834
ROC AUC for Test :  0.9739231048038587
```

Figure 6. Test Result

When applied to test data, the model showed an accuracy of 0.7, signifying its ability to generalize patterns from the training data to never-before-seen data. In addition, the ROC AUC remained at a value of 0.9, indicating the consistency of the model's performance in maintaining a good comparison between True Positive Rate and False Positive Rate in the test data.

It is important to note that the Out-of-Bag (OOB) Score in Figure 5 reaches a value of 0.7. Although lower than accuracy and ROC AUC on training data, OOB Score values provide a good picture of model performance without requiring additional test data. Overall, these results show that the improved Random Forest model successfully provides predictions with more balanced quality between the training data and the test data in Figure 6.

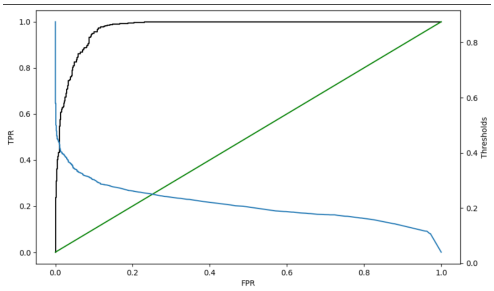

Figure 7. ROC-AUC for Data Test

Figure 7 shows the ROC-AUC (Receiver Operating Characteristic - Area Under the Curve) curve for test data, which is a visual tool for measuring the performance of classification models, especially in terms of sensitivity and specificity. On the ROC-AUC curve, the X-axis represents the False Positive Rate (FPR), which is the proportion of negative classes that are incorrectly classified as positive, while the Y-axis represents the True Positive Rate (TPR), which is the proportion of positive classes that are classified as positive. This curve illustrates how changes in classification thresholds affect the trade-off between False Positives and True Positive Rates.

Analysis of Figure 7 shows that the model has a good ability to distinguish between positive and negative classes in the test data. Area under curve (AUC) is a numerical metric that measures the extent to which the model can distinguish between classes, with a maximum value of 1.0 indicating excellent performance. By looking at this curve, it can be concluded that the tested model is able to make predictions well on the test data, characterized by ROC-AUC that is close to the maximum value.

4. Conclusion

Overall, testing and implementation of the Random Forest Classifier model showed significant results in identifying and classifying spam emails. Although at the training stage the model was able to achieve accuracy and ROC AUC of 1.0, indicating optimal ability to process training data, these results also indicate potential overfitting. The implementation of RandomizedSearchCV successfully overcame, resulting in an accuracy of 0.8 and an ROC AUC of 0.9 on the training data. The results on the test data also showed good performance with an accuracy of 0.7 and an ROC AUC of 0.9, confirming the model's ability to generalize patterns from the training data. Although the OOB Score achieved 0.7 on testing the training data, it gives a good idea of the model's performance. Overall, the optimized Random Forest Classifier model successfully provided quality-balanced predictions between training and test data, confirming its reliability and credibility in

classifying spam emails.

In the conclusion there should be no reference. Conclusions contain the facts obtained, simply answering the problem or purpose of the study (do not constitute any more discussion); State the possibilities of application, implications and speculation accordingly. If needed, provide advice for future research.

References

[1] H. Mukhtar, J. Al Amien, and M. A. Rucyat, "Filtering Spam Email menggunakan Algoritma Naïve Bayes," *J. CoSciTech (Computer Sci. Inf. Technol.*, vol. 3, no. 1, pp. 9–19, 2022.

[2] A. Hidayat, "Aplikasi Teks Mining Untuk Mendeteksi Spam Pada Email Berbasis Naive Bayes," *J. Teknol. Pint.*, vol. 2, no. B, pp. 1–10, 2022.

[3] M. B. Hartono, A. K. Darmawan, and H. Hoiriyah, "Komparasi Deep Learning Dan Traditional Machine Learning Untuk Email Spam Filtering," *J. Minfo Polgan*, vol. 12, no. 1, pp. 636–643, 2023.

[4] H. Iswanto, E. Seniwati, Y. Astuti, and D. Maulina, "Comparison of Algorithms on Machine Learning For Spam Email Classification," *IJISTECH (International J. Inf. Syst. Technol.*, vol. 5, no. 4, p. 446, 2021.

[5] Suryawanshi, Shubhangi, Anurag Goswami, and Pramod Patil. "Email spam detection: an empirical comparative study of different ml and ensemble classifiers." *2019 IEEE 9th International Conference on Advanced Computing (IACC)*. IEEE, 2019.

[6] A. R. Chrismanto, Y. Lukito, and A. Susilo, "Implementasi Distance Weighted K-Nearest Neighbor Untuk Klasifikasi Spam & Non-Spam Pada Komentar Instagram," *J. Edukasi dan Penelit. Inform.*, vol. 6, no. 2, p. 236, 2020.

[7] M. Azhari, Z. Situmorang, and R. Rosnelly, "Perbandingan Akurasi, Recall, dan Presisi Klasifikasi pada Algoritma C4.5, Random Forest, SVM dan Naive Bayes," *J. Media Inform. Budidarma*, vol. 5, no. 2, p. 640, 2021.

[8] Sharfina, Nabilah, and Nur Ghaniaviyanto Ramadhan. "Analisis SMOTE Pada Klasifikasi Hepatitis C Berbasis Random Forest dan Naïve Bayes." *JOINTECS (Journal of Information Technology and Computer Science)* 8.1 (2023): 33-40.

[9] N. L. P. C. Savitri, R. A. Rahman, R. Venyutzky, and N. A. Rakhmawati, "Analisis Klasifikasi Sentimen Terhadap Sekolah Daring pada Twitter Menggunakan Supervised Machine Learning," *J. Tek. Inform. dan Sist. Inf.*, vol. 7, no. 1, pp. 47–58, 2021.

[10] I. Adriansyah, M. D. Mahendra, E. Rasywir, and Y. Pratama, "Perbandingan Metode Random Forest Classifier dan SVM Pada Klasifikasi Kemampuan Level Beradaptasi Pembelajaran Jarak Jauh Siswa," *Bull. Informatics Data Sci.*, vol. 1, no. 2, pp. 98–103, 2022.

[11] F. Diba, "Analisis Random Forest Menggunakan Principal Component Analysis Pada Data Berdimensi Tinggi," *Indones. J. Comput. Sci.*, vol. 12, no. 4, pp. 2152–2160, 2023.

[12] A. P. Mahendra, D. Pradipta, M. R. B. Saputro, and K. Kusrini, "Application of the Decision Tree Method to Forest Fire Detection (Case Study: in Palembang, South Sumatra)," *JTECS J. Sist. Telekomun. Elektron. Sist. Kontrol Power Sist. dan Komput.*, vol. 2, no. 1, p. 75, 2022.

[13] C. M. Sitorus, A. Rizal, and M. Jajuli, "Prediksi Risiko Perjalanan Transportasi Online Dari Data Telematik Menggunakan Algoritma Support Vector Machine," *J. Tek. Inform. dan Sist. Inf.*, vol. 6, no. 2, pp. 254–265, 2020.

[14] M. Y. Aldean, P. Paradise, and N. A. Setya Nugraha, "Analisis Sentimen Masyarakat Terhadap Vaksinasi Covid-19 di Twitter Menggunakan Metode Random Forest Classifier (Studi Kasus: Vaksin Sinovac)," *J. Informatics, Inf. Syst. Softw. Eng. Appl.*, vol. 4, no. 2, pp. 64–72, 2022.

[15] J. Moh, K. Ii, and B. S. Indah, "PERBANDINGAN ALGORITMA DECISION TREE , RANDOM FOREST DAN NAIVE BAYES PADA PREDIKSI PENILAIAN KEPUASAN PENUMPANG," vol. 12, no. 2, pp. 150–159, 2023.

[16] E. C. P. Witjaksana, R. R. Saedudin, and V. P. Widartha, "Perbandingan Akurasi Algoritma Random Forest dan Algoritma Artificial Neural Network untuk Klasifikasi Penyakit Diabetes," *e-Proceeding Eng.*, vol. 8, no. 5, pp. 9765–9772, 2021.

[17] D. Hindarto, "Perbandingan Kinerja Akurasi Klasifikasi K-NN, NB dan DT pada APK Android," *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, vol. 9, no. 1, pp. 486–503, 2022.

[18] R. Halim and H. Bunyamin, "Analisis Penjualan di Cabang Toko Serba Ada dengan Algoritma Machine Learning," vol. 5, no. November, pp. 438–453, 2023.

[19] N. Adila, S. Khasanah, and T. Sutabri, "STRATEGI PERANCANGAN SISTEM AMAVIS DAN," vol. 5, no. 2, pp. 154–166, 2023.

[20] A. Wibisono, "FILTERING SPAM EMAIL MENGGUNAKAN METODE NAIVE BAYES," vol. 3, no. 4, 2023.

[21] W. Van Der Aalst, *Process Mining Data Science In Action*. Springer Heidelberg New York Dordrecht London, 2016.