P-ISSN: 2963-4008
E-ISSN: 2963-430X

OPTIMASI DETEKSI SPAM EMAIL DENGAN *RANDOM FOREST* DAN *RANDOM SEARCH*

Doni Gunawan[1], Riffa Haviani Laluma[2]
[1, 2] Program Studi Teknik Informatika, Universitas Sangga Buana

[1] korespondensi: donigunawanusbypkp@gmail.com

ABSTRACT

Email spam *detection is a crucial component in digital information security and filtering systems. To improve classification accuracy, this study utilizes the Random Forest Classifier algorithm, known for its stability and reliability in handling Datasets with numerous features. The research applies hyperparameter tuning using Randomized Random Search Cross-Validation to optimize the model's performance by identifying the best combination of parameters. Random Search Cross-Validation enables an efficient parameter search without the computational burden of testing all combinations, as is done in Grid Search Cross-Validation. Evaluation results show that the tuned model outperforms the default configuration, particularly in terms of accuracy and precision in detecting* spam*. This system can assist both individual users and organizations in managing email more effectively and securely.*

Keywords: Hyperparameter Tuning, Random Forest, Random Search Cross-Validation, Spam *Detection, Email*

ABSTRAK

Deteksi spam *pada email merupakan komponen penting dalam sistem keamanan dan penyaringan informasi digital. Untuk meningkatkan akurasi dalam klasifikasi* spam*, digunakan algoritma Random Forest Classifier yang dikenal karena stabilitas dan keandalannya dalam memproses data dengan banyak fitur. Penelitian ini menerapkan teknik hyperparameter tuning menggunakan metode Random Search Cross-Validation untuk mengoptimalkan kinerja model dengan menemukan kombinasi parameter terbaik. Random Search Cross-Validation memungkinkan proses pencarian parameter yang efisien tanpa harus menguji seluruh kemungkinan seperti pada Grid Search Cross-Validation. Hasil evaluasi menunjukkan bahwa gabungan antara model Random Forest dan Hyperparameter-Tuning menggunakan metode Random Search Cross-Validation memiliki performa yang lebih tinggi dibandingkan model default, terutama dalam hal akurasi dan presisi deteksi* spam*. Sistem ini dapat membantu pengguna maupun organisasi dalam mengelola email secara lebih efektif dan aman.*

Kata Kunci: Hyperparameter-Tuning, Random Forest, Randomized Search Classifier, Deteksi Spam*, Email*

PENDAHULUAN

Spam email merupakan pesan elektronik yang dikirim secara massal tanpa izin penerima, biasanya berisi iklan, promosi, maupun tautan berbahaya yang berpotensi digunakan untuk *phishing* atau penyebaran *malware* (1). Keberadaan spam tidak hanya mengganggu komunikasi digital, tetapi juga menimbulkan ancaman serius terhadap keamanan data dan privasi pengguna (1). Peningkatan jumlah spam sejalan dengan penggunaan email yang masif di sektor personal, akademis, maupun bisnis, sehingga deteksi spam menjadi komponen penting dalam sistem keamanan informasi (2) .

Metode konvensional berbasis aturan (*rule-based*) seringkali gagal menghadapi pola spam yang dinamis. Oleh karena itu, pendekatan berbasis *machine learning*, khususnya klasifikasi teks, dipandang lebih efektif karena mampu mempelajari pola dari jumlah data yang besar dan terus berkembang (3). Salah satu algoritma yang banyak digunakan adalah *Random Forest Classifier*,

sebuah metode ensemble learning berbasis pohon keputusan yang dikenal tahan terhadap *overfitting* dan efektif dalam mengolah data berdimensi tinggi (1).

Kinerja model *machine learning* sangat dipengaruhi oleh pemilihan *hyperparameter*. Tanpa pengaturan parameter yang tepat, model dapat mengalami *underfitting* atau *overfitting,* yang berdampak pada rendahnya akurasi sistem (4). Untuk itu, diperlukan teknik optimasi *hyperparameter* yang efisien. *Random Search Cross-Validation* menjadi salah satu metode yang populer karena dapat mengeksplorasi ruang parameter secara acak tanpa harus menguji seluruh kemungkinan kombinasi sebagaimana dilakukan *Grid Search* (5).

Beberapa penelitian sebelumnya menunjukkan bahwa integrasi *Random Forest* dengan teknik optimasi *hyperparameter* dapat meningkatkan performa sistem deteksi spam. Misalnya, penelitian sebelumnya mencapai akurasi 98,2% dengan kombinasi *Continuous Bag-of-Words dan Random Forest* (6). Sementara itu, penelitian sebelumnya membuktikan bahwa *Random Forest* unggul dibandingkan algoritma lain seperti *Naïve Bayes dan SVM* dalam mendeteksi spam, dengan akurasi mencapai 94,2% (7).

Berdasarkan latar belakang tersebut, penelitian ini berfokus pada pengembangan sistem deteksi spam berbasis algoritma *Random Forest Classifier* yang dioptimalkan melalui *Random Search Cross-Validation*. Tujuan utamanya adalah meningkatkan akurasi, presisi, dan *recall* dalam klasifikasi email spam, sekaligus menghasilkan sistem yang efisien dan adaptif terhadap pola spam yang terus berkembang (6).

METODE

Penelitian ini menggunakan pendekatan kuantitatif eksperimental dengan fokus pada pengembangan dan evaluasi sistem deteksi spam berbasis *machine learning*. Tahapan metode penelitian dapat dijelaskan sebagai berikut:

1. *Dataset*

Dataset yang digunakan adalah Enron Email *Dataset*, salah satu *Dataset* publik yang sering dipakai dalam penelitian deteksi spam karena berisi kombinasi pesan ham (*legitimate email*) dan spam. Setelah dilakukan proses *preprocessing*, digunakan 39.282 email dengan pembagian data latih 80% dan data uji 20% (8).

2. *Preprocessing Data*

Tahapan pra-pemrosesan dilakukan untuk membersihkan teks email sebelum dikonversi ke bentuk numerik. Proses yang dilakukan meliputi:

a. Tokenisasi (*tokenization*) dan penghapusan kata umum (*stopword removal*) menggunakan NLTK.
b. Normalisasi teks dengan mengubah semua huruf menjadi *lowercase*.
c. Penghapusan karakter non-alfabet, angka, tanda baca, serta nilai kosong (*missing values*).

d. Langkah ini bertujuan untuk mengurangi noise sehingga model dapat mengenali pola dengan lebih akurat (9).

3. Ekstraksi Fitur

Ekstraksi fitur dilakukan dengan *Term Frequency–Inverse Document Frequency* (TF-IDF), yang mengubah teks menjadi representasi numerik berbasis bobot kata.

$$[TF(t,d) = count(t,d)/(\sum_{t^{\prime}} count(t^{\prime},d))][IDF(t) = \log N/(1+n_t)][TFIDF(t,d) = TF(t,d)x\ IDF(t)] \quad \text{..... (1)}$$

dengan (N) adalah jumlah dokumen, dan (n_t) adalah jumlah dokumen yang mengandung kata (t).

4. Model Klasifikasi

Algoritma *Random Forest Classifier* dipilih karena mampu menangani data berdimensi tinggi dan mengurangi risiko *overfitting* (10). Model ini membangun sejumlah pohon keputusan, kemudian hasil prediksi akhir ditentukan berdasarkan *majority voting* antar pohon:

$$[TF(t,d) = \frac{count(t,d)}{\sum_{t'} count(t',d)}][IDF(t) = log\ \frac{N}{1+n_t}][TFIDF(t,d) = TF(t,d)x\ IDF(t)] \quad \text{..... (2)}$$

5. *Hyperparameter Tuning*

Optimasi dilakukan menggunakan *RandomizedSearchCV* yang memilih kombinasi parameter secara acak dari ruang pencarian yang ditentukan, kemudian mengevaluasi kinerja model melalui *k-fold cross-validation* (4). Parameter yang disetel antara lain:

a. jumlah pohon ((*n_estimators*))
b. kedalaman maksimum pohon ((*max_depth*))
c. jumlah fitur yang digunakan pada setiap split ((*max_features*))
d. ukuran sampel minimum per node

Metode ini dipilih karena lebih efisien dibandingkan *Grid Search* dalam ruang parameter besar (6).

6. Evaluasi Model

Evaluasi performa model dilakukan menggunakan metrik standar klasifikasi, yaitu:

$$[Accuray = \frac{TP+TN}{TP+TN+FP+FN}][Precision = \frac{TP}{TP+FP}][Recall = \frac{TP}{TP+FN}][F1 = \frac{2\ x\ Precision\ x\ Recall}{Precision+Recall}] \quad \text{..... (3)}$$

Selain itu, performa juga diukur dengan *Area Under the Curve* (AUC) dari kurva *Receiver Operating Characteristic* (ROC), untuk menilai keseimbangan antara *true positive rate* dan *false positive rate* (5).

HASIL DAN PEMBAHASAN

Hasil *Preprocessing Data*

Dataset Enron yang digunakan awalnya berjumlah 829.210 email, terdiri dari 407.140 *ham* dan 378.561 spam. Setelah tahap *cleaning* dan *filtering* (penghapusan *missing values* dan karakter non-English), jumlah data berkurang menjadi 785.701 email. Selanjutnya diambil 5% sampel untuk efisiensi komputasi sehingga diperoleh 39.282 email, dengan distribusi seimbang antara *ham* dan spam. *Dataset* dibagi menjadi data latih (80%, 31.425 email) dan data uji (20%, 7.857 email). Proses *preprocessing* berupa tokenisasi, penghapusan *stopword*,

normalisasi teks, dan transformasi fitur menggunakan TF-IDF (8).

Eksperimen Model Model *Random Forest Classifier* dilatih menggunakan data hasil ekstraksi TF-IDF dengan konfigurasi awal (*default parameters*). Kemudian dilakukan *hyperparameter tuning* menggunakan *RandomizedSearchCV* dengan ruang pencarian meliputi jumlah pohon ((*n_estimators*)), kedalaman pohon ((*max_depth*)), jumlah fitur per split ((*max_features*)), dan ukuran sampel minimum.

Evaluasi Performa Hasil evaluasi model ditampilkan pada Tabel 1.

Tabel 1: Perbandingan hasil klasifikasi antara model default dan model dengan *Random Search* Cross-Validation

Model	Akurasi (%)	Presisi (%)	*Recall* (%)	F1-score (%)	AUC
Random Forest (default)	94.3	93.7	92.8	93.2	0.95
Random Forest (tuned)	97.8	97.5	97.1	97.3	0.98

Hasil menunjukkan bahwa penerapan *RandomizedSearchCV* memberikan peningkatan yang signifikan pada semua metrik. Model yang telah dioptimasi mencapai akurasi 97,8%, lebih tinggi dibandingkan konfigurasi default sebesar 94,3%. Peningkatan juga terlihat pada nilai AUC, dari 0,95 menjadi 0,98, menunjukkan kemampuan model dalam membedakan spam dan *ham* semakin baik.

Analisis dan Diskusi

Kinerja tinggi dari algoritma *Random Forest* dipengaruhi oleh kemampuannya menggabungkan hasil dari banyak pohon keputusan (*bagging*), sehingga mengurangi risiko *overfitting* (1). Hasil ini sejalan dengan penelitian sebelumnya yang menemukan bahwa *Random Forest* unggul dibandingkan algoritma lain dalam klasifikasi spam dengan akurasi mencapai 94,2% (11).

Optimasi melalui *RandomizedSearchCV* terbukti efektif dalam menemukan konfigurasi parameter terbaik tanpa harus mengeksplorasi seluruh ruang parameter. Hal ini konsisten dengan temuan sebelumnya yang menyatakan bahwa pencarian acak lebih efisien dibandingkan pencarian grid, khususnya pada ruang parameter yang besar (12).

Dari sisi penerapan praktis, sistem deteksi spam yang dikembangkan dapat mengurangi beban pengguna dan organisasi dalam memilah email, meningkatkan keamanan informasi, serta menekan risiko serangan *phishing* dan *malware*. Selain itu, penggunaan teknik TF-IDF memungkinkan model menangkap representasi numerik dari kata-kata kunci yang sering muncul pada spam, sehingga pola dapat diidentifikasi lebih akurat (13) .

SIMPULAN

Penelitian ini membuktikan bahwa penerapan algoritma *Random Forest Classifier* yang dioptimalkan melalui teknik *Random Search Cross-Validation* mampu meningkatkan efektivitas sistem deteksi spam pada email. Model yang telah dituning menunjukkan peningkatan signifikan pada metrik akurasi, presisi, *recall*, dan AUC dibandingkan dengan konfigurasi default.

Dengan demikian, integrasi *Random Forest* dan *RandomizedSearchCV* dapat menjadi pendekatan yang efisien dan andal untuk mengembangkan sistem klasifikasi spam yang adaptif terhadap pola spam yang dinamis. Sistem ini berpotensi diterapkan dalam berbagai layanan email untuk meningkatkan keamanan komunikasi digital, mengurangi risiko serangan *phishing*, serta meningkatkan kenyamanan pengguna.

DAFTAR PUSTAKA

1. S. T. Ibrahim, O. B. Adjunct Lecturer, and O. H. Part Time Lecturer, "Spam Email Detection Scheme Based On *Random Forest* Algorithm," LAUJCI, 2023. (Online). Available: Www.Laujci.Lautech.Edu.Ng

2. M. A. Bouke, A. Abdullah, M. T. Abdullah, S. A. Zaid, H. El Atigh, And S. H. Alshatebi, "A Lightweight *Machine learning*-Based Email Spam Detection Model Using Word Frequency Pattern," *Journal Of Information Technology And Computing*, Vol. 4, No. 1, Pp. 15–28, Jun. 2023, Doi: 10.48185/Jitc.V4i1.653.

3. C. Beaman Craigbeaman, "Anomaly Detection In Emails Using *Machine learning* And Header Information."

4. P. Probst, M. Wright, And A.-L. Boulesteix, "Hyperparameters And Tuning Strategies For *Random Forest*," Feb. 2019, Doi: 10.1002/Widm.1301.

5. R. Ageng, R. Faisal, And S. Ihsan, "*Random Forest Machine learning* For Spam Email Classification," *Journal Of Dinda Data Science, Information Technology, And Data Analytics*, Vol. 4, No. 1, Pp. 8–13, 2024, (Online). Available: Http://Journal.Ittelkom-Pwt.Ac.Id/Index.Php/Dinda

6. T. O. Omotehinwa And D. O. Oyewola, "Hyperparameter Optimization Of Ensemble Models For Spam Email Detection," *Applied Sciences (Switzerland)*, Vol. 13, No. 3, Feb. 2023, Doi: 10.3390/App13031971.

7. M. Sahami, S. Dumais, D. Heckerman, And E. Horvitz, "A Bayesian Approach To Filtering Junk E-Mail," 1998. (Online). Available: Www.Aaai.Org

8. B. Klimt And Y. Yang, "The Enron Corpus: A New *Dataset* For Email Classification Research." (Online). Available: Http://Www-2.Cs.Cmu.Edu/~Enron/.

9. N. Al-Shanableh Mazen, S. Alzyoud, And E. Nashnush, "Enhancing Email Spam Detection Through Ensemble Enhancing Email Spam Detection Through Ensemble *Machine learning*: A Comprehensive Evaluation Of *Machine learning*: A Comprehensive Evaluation Of Model Integration And Performance Model Integration And Performance Part Of The Management Information Systems Commons." (Online). Available: Https://Scholarworks.Lib.Csusb.Edu/Ciima

10. A. Kosmopoulos, G. Paliouras, and I. Androutsopoulos, "Adaptive Spam Filtering Using Only Naive Bayes Text Classifiers." (Online). Available: http://www.aueb.gr/users/ion/

11. M. A. Ghani and A. Subekti, "Email Spam Filtering Dengan Algoritma

Random Forest," *IJCIT (Indonesian Journal on Computer and Information Technology*, vol. 3, no. 2, pp. 216–221, 2018.

12. J. Bergstra, J. B. Ca, and Y. B. Ca, "*Random Search* for Hyper-Parameter Optimization Yoshua Bengio," 2012. (Online). Available: http://scikit-learn.sourceforge.net.

13. G. M. Sai, K. 1#, K. Eswar, T. 2#, D. Harshavardhan, and R. 3#, "Study of SPAM Email Detection."