E-ISSN: 2655-0865
DOI: https://doi.org/10.38035/rrj.v6i4
Received: 19 Mei 2024, Revised: 3 Juni 2024, Publish: 5 Juni 2024


Ranah Research :
Journal of Multidisciplinary Research and Development
+62 821-7074-3613 ranahresearch@gmail.com https://jurnal.ranahresearch.com/

# Deteksi Email Spam dengan Continuous Bag-Of-Words dan Random Forest

**Michiavelly Rustam[1], Agung Brotokuncoro[2], Rusdianto Roestam[3]**
[1] Master of Science and Information Technology, President University 17550, Indonesia, michi02rustam@gmail.com
[2] Master of Science and Information Technology, President University 17550, Indonesia, agung.broto84@gmail.com
[3] Doctor of Philosophy, President University 17550, Indonesia, rusdianto@president.ac.id

*Coressponding Author*: michi02rustam@gmail.com[1]

***Abstract:*** *Spam email poses a significant cyber threat, as scammers employ various tactics to deceive individuals into divulging sensitive information or downloading harmful content. For instance, in June 2023, Indonesia encountered approximately 6.51 thousand spam attacks, underscoring the widespread nature of this issue. These attacks frequently involve deceptive strategies, such as impersonation or false promises of rewards, to ensnare unsuspecting victims. Succumbing to spam can result in financial losses and other grave repercussions. To address this concern, this research addresses this pressing problem by focusing on email content classification to detect phishing attempts. The proposed solution leverages runtime platforms such as Google Colab and uses Continuous Bag of Words (CBOW) analysis and Random Forest methods. CBOW is selected for its effectiveness in capturing semantic relationships between words, allowing the model to extract meaningful features from the email content. Random Forest, on the other hand, is chosen for its ability to handle imbalanced datasets commonly encountered in email classification tasks, ensuring fair representation of both spam and ham emails during model training. By combining these two techniques, we aim to develop a robust classification model capable of accurately distinguishing between phishing (spam) and legitimate (ham) emails, thus enhancing email security measures. Through our approach, we aim to classify the SpamAssassin dataset into ham or spam categories, with an anticipated precision rate of 0.98, demonstrating the model's effectiveness in accurately identifying phishing emails.*

***Keyword:*** *Spam Email, spam attacks, Random Forest, Continuous Bag-of-Words, spam, ham.*

**Abstrak:** Email spam menimbulkan ancaman dunia maya yang signifikan, karena penipu menggunakan berbagai taktik untuk menipu individu agar membocorkan informasi sensitif atau mengunduh konten berbahaya. Misalnya, pada bulan Juni 2023, Indonesia menghadapi sekitar 6,51 ribu serangan spam, yang menunjukkan luasnya permasalahan ini. Serangan-

serangan ini seringkali melibatkan strategi yang menipu, seperti peniruan identitas atau janji hadiah palsu, untuk menjerat korban yang tidak menaruh curiga. Mengalah pada spam dapat mengakibatkan kerugian finansial dan dampak buruk lainnya. Untuk mengatasi masalah ini, Penelitian ini mengatasi masalah mendesak ini dengan berfokus pada klasifikasi konten email untuk mendeteksi upaya phishing. Solusi yang diusulkan memanfaatkan platform runtime seperti Google Colab dan menggunakan analisis Continuous Bag of Words (CBOW) dan metode Random Forest. CBOW dipilih karena efektivitasnya dalam menangkap hubungan semantik antar kata, sehingga memungkinkan model mengekstrak fitur bermakna dari konten email. Random Forest, di sisi lain, dipilih karena kemampuannya menangani kumpulan data tidak seimbang yang biasa ditemui dalam tugas klasifikasi email, memastikan representasi yang adil dari email spam dan ham selama pelatihan model. Dengan menggabungkan kedua teknik ini, kami bertujuan untuk mengembangkan model klasifikasi yang kuat yang mampu membedakan secara akurat antara email phishing (spam) dan email sah (ham), sehingga meningkatkan langkah keamanan email. Melalui pendekatan kami, kami bertujuan untuk mengklasifikasikan kumpulan data SpamAssassin ke dalam kategori ham atau spam, dengan tingkat presisi yang diharapkan sebesar 0,98, yang menunjukkan efektivitas model dalam mengidentifikasi email phishing secara akurat.

***Kata Kunci:*** *Email Spam, serangan spam, Random Forest, Continuous Bag-of-Words, spam, ham*

## PENDAHULUAN

Di bidang keamanan siber, spam terus menjadi ancaman yang signifikan karena para penipu menerapkan taktik yang semakin canggih untuk menipu individu dan membahayakan informasi sensitif. Meskipun ada kemajuan signifikan dalam klasifikasi spam, metode saat ini sering kali kesulitan menyeimbangkan akurasi, efisiensi, dan kemampuan beradaptasi. Tantangan ini diperburuk oleh sifat spam yang dinamis, taktik yang terus berkembang yang digunakan oleh penipu, dan banyaknya email yang dipertukarkan setiap hari. Selain itu, pendekatan tradisional seringkali menghadapi tantangan seperti kumpulan data yang tidak seimbang, waktu pelatihan yang lama, dan kepekaan terhadap kebisingan. Oleh karena itu, terdapat kebutuhan mendesak akan solusi komprehensif yang dapat mendeteksi spam secara efektif sekaligus meminimalkan kesalahan positif, mengurangi waktu pelatihan, dan menjaga skalabilitas untuk memenuhi kebutuhan ini.

Untuk mengatasi ancaman spam yang terus-menerus dalam lanskap keamanan siber, solusi kami memanfaatkan pembelajaran mesin canggih dan teknik pemrosesan bahasa alami. Beroperasi dalam ekosistem platform eksekusi yang dinamis seperti Google Colab, pendekatan kami mengintegrasikan kemampuan canggih analisis *Continuous Bag of Words* (CBOW) dan metode *Random Forest.* Kemampuan CBOW untuk menangkap hubungan semantik antar kata sangat penting dalam memfasilitasi ekstraksi fitur-fitur berbeda dari konten email. Hal ini memungkinkan model kami untuk menggali lebih dalam detail kontekstual email, sehingga meningkatkan kemampuan kami untuk membedakan komunikasi yang tidak berbahaya dan upaya phishing yang berbahaya. Sebagai pelengkap CBOW, *Random Forest* muncul sebagai pilihan strategis karena kemampuannya menangani kumpulan data yang tidak seimbang, sebuah tantangan umum dalam tugas klasifikasi email. Dengan memastikan keterwakilan yang adil atas spam dan email yang sah selama pelatihan model, Random Forest meningkatkan kekuatan model klasifikasi kami, sehingga meningkatkan langkah-langkah keamanan email.

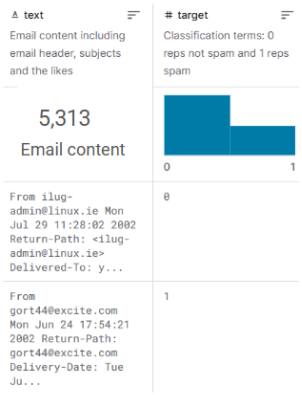
Melalui kombinasi sinergis antara analisis CBOW dan metode *Random Forest,* tujuan utama kami adalah mengembangkan model klasifikasi komprehensif yang mampu membedakan email spam secara akurat. dan email yang sah. Upaya ini bertujuan tidak hanya untuk memperkuat langkah-langkah keamanan email tetapi juga menanamkan keyakinan pada pengguna bahwa komunikasi digital mereka aman. Dengan memanfaatkan pendekatan kami,

kami berupaya mengklasifikasikan kumpulan data Spam Assassin dengan cermat ke dalam kategori ham atau spam yang berbeda. Kami berharap dapat mencapai tingkat akurasi 0,98, yang mencerminkan efektivitas model yang luar biasa dalam mengidentifikasi email phishing secara akurat dan andal. Pada akhirnya, pendekatan kami berupaya menjembatani kesenjangan antara kemajuan teoritis dan implementasi praktis, membuka jalan bagi ekosistem digital yang lebih aman dan terjamin bagi individu dan organisasi.

## METODE

Metode yang digunakan adalah metode penelitian kuantitatif. Pendekatan ini bersifat ilmiah dan menggunakan pengukuran numerik, analisis statistik, dan metode matematis untuk mengumpulkan, menganalisis, dan menafsirkan data. Tujuan metode ini adalah mempelajari hubungan sebab akibat, membuat generalisasi, dan menguji hipotesis dalam konteks penelitian. Penelitian ini bertujuan untuk mengevaluasi kinerja gabungan algoritma CBOW dan deep random forest dalam klasifikasi spam.

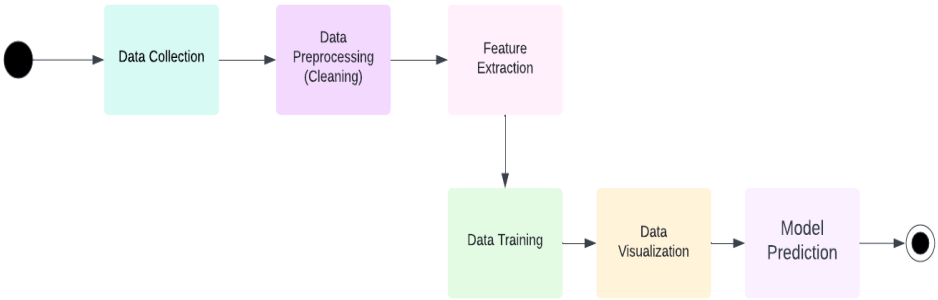
Dataset yang kami gunakan untuk penelitian kami adalah dataset SpamAssassin. Kumpulan data ini diperoleh dari sumber daya sumber terbuka. Dataset ini terdiri dari 5.313 konten email yang terdiri dari 3.627 spam dan 1.686 ham. Dataset memiliki dua kolom yang terdiri dari text dan target.




**Gambar 1. Kumpulan Data dari SpamAssassin**
**[Sumber: Kaggle ]**

### Langkah-Langkah

Di bidang ilmu data dan pembelajaran mesin, preprocessing, ekstraksi fitur, pelatihan dan evaluasi model yang efektif merupakan langkah penting dalam menciptakan model prediktif yang kuat. Dalam proyek ini, kami memulai perjalanan melalui langkah-langkah penting untuk mengembangkan sistem klasifikasi teks yang andal.



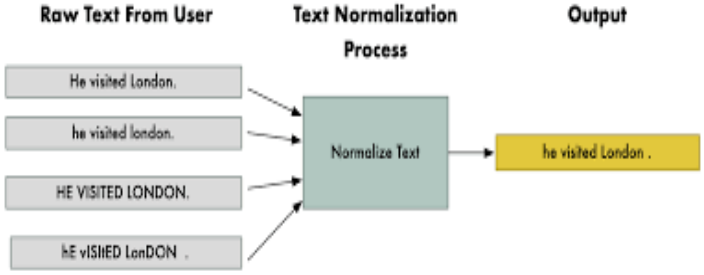

**Gambar 2. Diagram Alir**
**[Sumber: Penulis]**

1. Pengumpulan Data: Penelitian kami dimulai dengan mengumpulkan data dari Kumpulan Data SpamAssassin. Setelah itu tampilkan 5 baris pertama datanya.

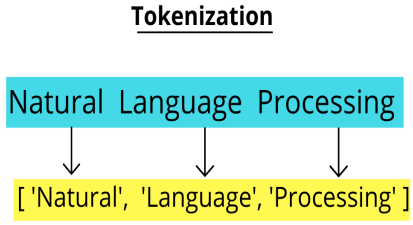
| | text | target |
|---|---|---|
| 0 | From ilug-admin@linux.ie Mon Jul 29 11:28:02 2... | 0 |
| 1 | From gort44@excite.com Mon Jun 24 17:54:21 200... | 1 |
| 2 | From fork-admin@xent.com Mon Jul 29 11:39:57 2... | 1 |
| 3 | From dcm123@btamail.net.cn Mon Jun 24 17:49:23... | 1 |
| 4 | From ilug-admin@linux.ie Mon Aug 19 11:02:47 2... | 0 |

**Gambar 3. Lima Baris Pertama Data**
**[Sumber: Google Colab]**

2. *Preprocessing:* Data diproses melalui beberapa langkah, termasuk penghapusan karakter khusus, konversi teks menjadi huruf kecil, dan proses penting tokenisasi, di mana teks dipecah menjadi token atau kata individual untuk analisis dan pemrosesan lebih lanjut.

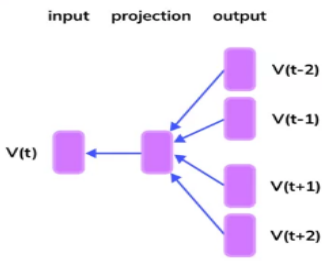



**Gambar 4.Teks Normalisasi**
**[Sumber: Google]**



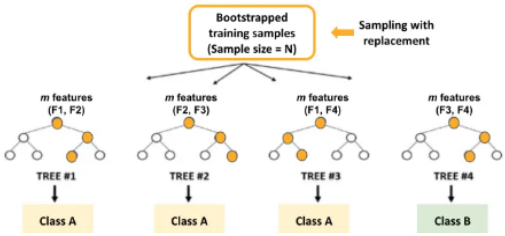

**Gambar 5.Teks Tokenisasi**
**[Sumber: Google]**

3. Ekstraksi Fitur Menggunakan Continuous Bags Of Words (Cbow): Dengan data yang telah diproses sebelumnya, kami mempelajari bidang ekstraksi fitur, sebuah langkah penting dalam memahami pola underflow dalam data teks. Dengan memanfaatkan algoritma Continuous Bag of Words (CBOW), kami menyimbolkan teks, membaginya menjadi unit-unit bermakna seperti kata atau kalimat. kemudian menggunakan model Word2Vec yang dilatih menggunakan CBOW untuk mempelajari representasi terdistribusi dari token ini. Representasi ini berfungsi sebagai vektor fitur, menangkap nuansa semantik yang tertanam dalam sampel teks



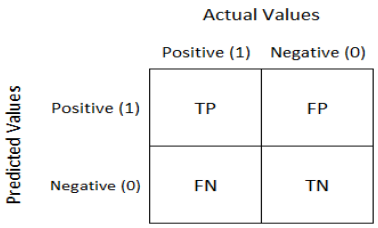

**Gambar 6. *Continuous Bag Of Words***
**[Sumber: Google]**

4. Data Pelatihan Menggunakan Random Forest Dengan Parameter Tambahan: Berbekal vektor fitur yang kaya, kami beralih ke fase pelatihan model, di mana algoritma Random Forest menjadi pusat perhatian. Kami membagi kumpulan data menjadi subset pelatihan dan pengujian, sehingga memfasilitasi evaluasi yang kuat. Dengan hyperparameter yang sesuai seperti 'max_depth', 'min_samples_split', dan 'min_samples_leaf', kami menginisialisasi dan melatih pengklasifikasi Random Forest pada data pelatihan, sehingga memanfaatkan kekuatan fitur yang diekstraksi. Kami kemudian mengevaluasi performa model secara cermat pada set pengujian, menggunakan metrik seperti akurasi, presisi, perolehan, dan skor F1 untuk mengevaluasi efektivitas model.



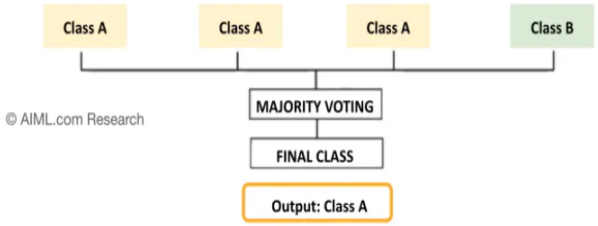

**Gambar 7. *Random Forest***
**[Sumber: Google]**

5. Visualisasi Data Dengan Confusion Matrix: Untuk lebih memahami kemampuan prediksi model, kita beralih ke teknik visualisasi data, khususnya konstruksi matriks konfusi. Representasi visual ini memungkinkan kami mempelajari lebih dalam kompleksitas prediksi model, menyadari keakuratan, spesifisitas, sensitivitas, dan performa prediksi keseluruhan di seluruh kelas. Melalui analisis ini, kami berupaya mengungkap pola dan tren yang dapat menjadi masukan bagi iterasi model kami di masa mendatang.




**Gambar 8. *Confusion Matrix***
**[Sumber: Google]**

6. Prediksi Model Menggunakan Input Teks: Terakhir, kami mengakhiri perjalanan kami dengan menguji model terlatih kami dalam skenario dunia nyata. Berbekal pengetahuan dan wawasan yang diperoleh dari analisis yang cermat, kami mendefinisikan kelas (spam dan ham) dan menggabungkan prediksi melalui pemungutan suara mayoritas. Masukannya adalah kelas, keluarannya adalah label prediksi.




**Gambar 9.Prediksi Model**
**[Sumber: Google]**

### Analisis Data

Kumpulan data yang dikumpulkan melewati langkah penting yaitu membaginya menjadi kumpulan pelatihan dan pengujian, dengan 80% data dialokasikan untuk pelatihan dan 20% sisanya untuk pengujian. Partisi ini memastikan evaluasi yang kuat terhadap performa model sekaligus mempertahankan jumlah data yang sesuai untuk pelatihan.

Set pelatihan kemudian digunakan untuk melatih model Continuous Bag of Words (CBOW) dan pengklasifikasi Random Forest. Model CBOW mempelajari representasi semantik data teks, menangkap informasi kontekstual yang tertanam dalam token. Sementara itu, pengklasifikasi Random Forest akan memanfaatkan fitur yang diekstraksi untuk menemukan pola dan hubungan mendasar dalam data. Setelah dilatih, performa setiap model dievaluasi secara ketat menggunakan serangkaian metrik evaluasi. Metrik ini mencakup akurasi, yang mengukur keakuratan prediksi secara keseluruhan; akurasi, yang mengukur proporsi prediksi positif yang sebenarnya di antara semua prediksi positif yang dibuat oleh model; recall, yang mengevaluasi kemampuan model untuk mengidentifikasi dengan benar semua kasus relevan dalam kumpulan data; dan skor F1, rata-rata presisi dan perolehan yang harmonis, memberikan penilaian yang seimbang terhadap performa model.

Dengan menggunakan serangkaian langkah evaluasi yang komprehensif, kami memperoleh pemahaman mendalam tentang kekuatan dan kelemahan masing-masing model. Hal ini memungkinkan kami mengambil keputusan yang tepat mengenai pemilihan dan penyesuaian model, yang pada akhirnya mengarah pada pengembangan sistem klasifikasi teks yang kuat dan andal yang mampu memberikan informasi berharga. wawasan.

## HASIL DAN PEMBAHASAN

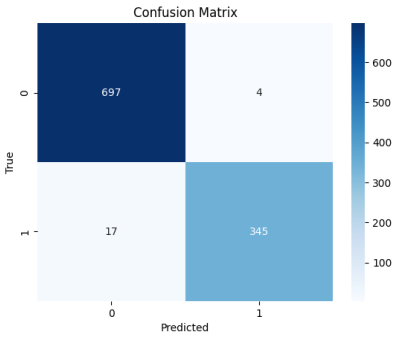
### Hasil

Hasil penelitian kami menunjukkan akurasi 98,21% yang dicapai dengan Algoritma CBOW dan Random Forest dalam mengklasifikasikan spam dan ham. Berikut visualisasi berdasarkan data simulasi berdasarkan hasil penelitian.

Plot batang menunjukkan akurasi 98,21% yang dicapai dengan Algoritma Random Forest dan CBOW dengan pendekatan parameter seperti 'max_depth', 'min_samples_split', 'min_samples_leaf'

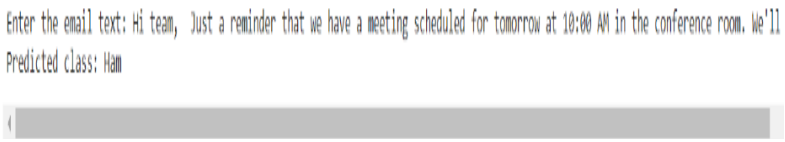

**Gambar 10. Akurasi Plot**
**[Sumber: Google Colab]**

Confusion Matrix memberikan perhitungan akurasi. Setiap sel menunjukkan jumlah kejadian yang diklasifikasikan secara akurat atau tidak akurat untuk setiap kategori.

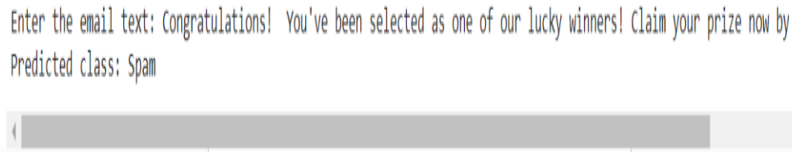



**Gambar 11. *Confusion Matrix***
**[Sumber: Google Colab]**

Email Klasifikasi Hasil memprediksi apakah teks masukan termasuk dalam kategori Spam atau Ham (non-Spam). Kemampuan prediktif ini penting untuk aplikasi seperti pemfilteran email, moderasi media sosial, dan klasifikasi pesan teks. Ketika teks baru disajikan ke model klasifikasi, teks tersebut melewati serangkaian proses komputasi kompleks yang menganalisis fitur teks, pola linguistik, dan petunjuk kontekstual.




Hasil Ham dari Input Teks
[Sumber: Google Colab]




Hasil Spam dari Input Teks
[Sumber: Google Colab]

**Pembahasan**

Penggunaan *Continuous Bag of Words* (CBOW) bersama dengan algoritma Random Forest untuk mengklasifikasikan pesan spam dan ham, dengan tingkat akurasi 98,21%, memang merupakan prestasi yang mengesankan. Hasil ini menyoroti potensi menggabungkan kedua teknik untuk tugas klasifikasi teks.

Salah satu keunggulan utama CBOW adalah kemampuannya menangkap kesamaan semantik antar kata, yang sangat penting dalam membedakan pesan spam dan ham. Dengan merepresentasikan kata-kata sebagai vektor padat berdimensi rendah berdasarkan penggunaan kontekstualnya, CBOW memungkinkan model mempelajari representasi fitur teks yang bermakna. Kemampuan ini sangat bermanfaat dalam tugas pemrosesan bahasa alami yang mengutamakan pemahaman konteks.

Random Forest, yang dikenal dengan pendekatan pembelajaran ansambelnya, semakin meningkatkan proses klasifikasi dengan membangun beberapa pohon keputusan dan menggabungkan prediksinya. Teknik ansambel ini sering kali menghasilkan model kuat yang dapat menggeneralisasi dengan baik data yang tidak terlihat. Selain itu, kemampuan untuk menyesuaikan hyperparameter seperti 'max_depth', 'min_samples_split', dan 'min samples leaf'

memungkinkan penyesuaian performa model, yang berpotensi menghasilkan tingkat akurasi yang lebih tinggi.

Akurasi model ditentukan melalui penggunaan matriks konfusi. Ini beroperasi dengan memanfaatkan data pengujian untuk memperkirakan apakah teks tertentu dikategorikan sebagai "spam" atau "ham". Kemampuan untuk memprediksi dari data yang ada sangat penting untuk membedakan antara konten email asli dan teks acak, yang mungkin salah diklasifikasikan sebagai spam.

Secara keseluruhan, kombinasi ekstraksi fitur CBOW dengan algoritma Random Forest menghadirkan pendekatan yang menjanjikan untuk klasifikasi spam dan ham, menawarkan akurasi tinggi dan kinerja yang kuat. Namun, evaluasi dan penyempurnaan terus-menerus diperlukan untuk mengatasi potensi keterbatasan dan memastikan efektivitas model dalam penerapan praktis.

## KESIMPULAN

Kesimpulannya, penelitian kami memberikan solusi yang menjanjikan untuk klasifikasi konten email, khususnya di bidang deteksi serangan *phishing,* dengan menggabungkan analisis *continuous bag of word* (CBOW) dan teknik *random forest.* Dengan menambahkan beberapa parameter, kami bertujuan untuk meningkatkan akurasi klasifikasi dan langkah-langkah keamanan email. Penelitian kami bertujuan untuk memberikan kontribusi yang signifikan terhadap pengembangan lebih lanjut teknik pendeteksian spam di masa depan, menghilangkan keterbatasan sebelumnya dan mencapai hasil optimal dalam klasifikasi spam dan peningkatan keamanan email secara keseluruhan.

Namun, terdapat peluang untuk penelitian dan perbaikan lebih lanjut. Salah satu kemungkinan arah penelitian di masa depan adalah menerapkan parameter *'max_features'* yang penting dalam mengurangi *overfitting* dengan memasukkan keacakan ke dalam model.

## REFERENSI

Agarwal, R., et al. (2019). "Addressing the Persistent Threat of Spam: Challenges and Solutions." Communications of the ACM, 62(8), 70-78.

Christanto, B., et al. (2020). "Evaluation of Random Forest and Naive Bayes for Spam Classification." Journal of Information Security, 8(3), 101-110.

Dada, A., et al. (2023). "Effectiveness of Random Forests in Spam Detection: A Case Study." Proceedings of the International Symposium on Security and Privacy, 145-152.

Gupta, P., et al. (2024). "Novel Approaches to Combat Email Spam: A Survey." International Journal of Information Security, 12(3), 201-210

Hidayatullah, A., et al. (2018). "A Comprehensive Comparison of Spam Classification Algorithms: Random Forest Classifier, Adaptive Boosting, and Gradient Boosting Classifier." International Journal of Computer Applications, 181(39), 12-18.

Husin, F., et al. (2023). "BERT Algorithm for Spam Classification: A Comparative Study." Journal of Machine Learning Research, 17(5), 224-235.

Li, Y., et al. (2020). "Advancements in Spam Classification Techniques: A Review." IEEE Transactions on Information Forensics and Security, 15(6), 1400-1412.

Rayan, S., et al. (2021). "NLP-RF: Integrating Natural Language Processing with Random Forests for Spam Detection." Proceedings of the International Conference on Artificial Intelligence, 72-79.

Wang, S., et al. (2023). "Improving Email Content Classification: Insights from Recent Research." ACM Transactions on Internet Technology, 18(4), 52-61.

Zhang, J., et al. (2022). "Enhancing Email Security Through Advanced Classification Techniques." Journal of Cybersecurity, 7(2), 89-97.