

Analisis Indikasi Phishing Pada Pesan Email Menggunakan Metode SVM

Alfin Nur Fadli¹, Jagad Satria Dewa Arta²

^{1,2}Teknik Informatika, Fakultas Teknik dan Ilmu Komputer, Universitas Nusantara PGRI Kediri

E-mail: ¹0070.alfinnur@gmail.com, ²dewarta807@gmail.com

Abstrak –Serangan phishing melalui layanan surat elektronik (email) merupakan salah satu ancaman keamanan siber terbesar saat ini yang bertujuan untuk mencuri informasi sensitif pengguna. Volume serangan yang terus meningkat menuntut adanya sistem deteksi otomatis yang akurat dan efisien. Penelitian ini bertujuan untuk membangun model deteksi phishing pada pesan email menggunakan algoritma Machine Learning, yaitu Support Vector Machine (SVM). Dataset yang digunakan adalah Phishing Validation Emails yang terdiri dari email kategori aman (ham) dan phishing. Tahapan penelitian meliputi pra-pemrosesan teks (preprocessing) seperti cleaning, stopword removal, dan stemming, serta ekstraksi fitur menggunakan metode Term Frequency-Inverse Document Frequency (TF-IDF) untuk mengonversi teks menjadi vektor numerik. Data dibagi menjadi data latih dan data uji untuk mengevaluasi kinerja model. Hasil eksperimen menunjukkan bahwa metode SVM dengan kernel Linear mampu memisahkan kelas email dengan sangat efektif. Berdasarkan pengujian menggunakan confusion matrix, model yang diusulkan berhasil mencapai performa sempurna dengan nilai Akurasi, Presisi, Recall, dan F1-Score masing-masing sebesar 100%. Hasil ini mengindikasikan bahwa kombinasi SVM dan TF-IDF sangat andal dalam mengidentifikasi pola serangan phishing pada dataset yang digunakan.

Kata Kunci —Email, Keamanan Siber, Phishing, Support Vector Machine (SVM), TF-IDF.

1. PENDAHULUAN

Perkembangan teknologi informasi saat ini berlangsung sangat pesat. Hal ini tentunya memberikan keuntungan besar karena teknologi mampu mempermudah berbagai aktivitas manusia. Sektor komunikasi digital berkembang signifikan seiring dengan kemajuan teknologi internet, yang memungkinkan masyarakat untuk bertukar informasi tanpa batasan ruang dan waktu. Namun, dampak negatif yang muncul dari kemajuan ini adalah meningkatnya kejahatan siber, salah satunya adalah penipuan melalui internet. Salah satu bentuk ancaman yang paling dominan adalah phishing, yaitu upaya mengeksploitasi pengguna internet melalui pesan palsu untuk mencuri informasi sensitif seperti kredensial akun atau data finansial. Data menunjukkan bahwa pesan spam dan penipuan digital terus berevolusi, sehingga metode penyaringan konvensional seringkali tidak lagi efektif menahan laju serangan ini [4]. Oleh karena itu, diperlukan sistem klasifikasi otomatis untuk memastikan bahwa email yang masuk ke kotak pengguna aman dari indikasi phishing yang dapat merugikan berbagai pihak.

Machine Learning merupakan cabang dari kecerdasan buatan (AI) di mana komputer memiliki kemampuan untuk belajar dari data secara mandiri tanpa harus diprogram secara eksplisit untuk setiap aturannya. Dalam konteks klasifikasi teks, Support Vector Machine (SVM) dikenal sebagai salah satu algoritma *supervised learning* yang sangat andal. SVM bekerja dengan mencari *hyperplane* atau garis pemisah terbaik yang memisahkan kelas data (email phishing dan legitimate) dengan margin terbesar. Penelitian komparatif menunjukkan bahwa SVM memiliki performa yang lebih stabil dan akurasi yang kompetitif dibandingkan algoritma lain seperti Naïve Bayes, serta efektif dalam memisahkan *hyperplane* pada dataset berdimensi tinggi [1].[3].[6]. Hal ini menjadikan SVM pilihan metode yang tepat untuk memberikan kinerja klasifikasi yang optimal.

Untuk memproses data teks email agar dapat dipahami oleh mesin, diperlukan metode ekstraksi fitur. Term Frequency-Inverse Document Frequency (TF-IDF) adalah metode pembobotan yang digunakan untuk mengonversi teks menjadi vektor numerik dengan menghitung relevansi kata dalam dokumen. Studi terbaru mengindikasikan bahwa penggunaan TF-IDF bersama SVM masih sangat efisien dan relevan dibandingkan metode yang lebih kompleks, terutama jika didukung dengan tahapan *preprocessing* berbasis Natural Language Processing (NLP) seperti *stopword removal* dan *stemming* yang terbukti meningkatkan akurasi [2].[5].[6]. Kinerja model ini kemudian diukur menggunakan *confusion matrix* untuk mengetahui nilai akurasi, presisi, dan recall.

Tujuan dilakukannya penelitian ini adalah membangun model deteksi dini yang mampu mengklasifikasikan pesan email yang terindikasi phishing menggunakan metode SVM, sehingga risiko pencurian data pengguna dapat diminimalisir. Pemilihan algoritma SVM didasarkan pada keberhasilan penelitian-penelitian terdahulu dalam domain serupa. Dalam penelitian yang dilakukan membahas mengenai prediksi email phishing menggunakan SVM dengan pembagian data latih dan uji sebesar 70:30 [1]. Penelitian tersebut menunjukkan bahwa SVM mampu mengenali pola serangan pada *body email* dengan sangat baik setelah melalui proses *text mining*. Selain itu, penelitian lain juga memperkuat hal ini dengan melakukan analisis komparatif yang membuktikan bahwa SVM unggul dalam membedakan email sah dan berbahaya dibandingkan algoritma *tree-based* lainnya [5]. Berdasarkan landasan tersebut, penelitian ini akan menerapkan metode SVM untuk

menganalisis indikasi phishing pada pesan email guna mendapatkan akurasi klasifikasi yang tinggi, Dengan demikian, penelitian ini tidak hanya memberikan kontribusi akademik, tetapi juga solusi praktis yang dapat digunakan untuk meningkatkan keamanan komunikasi digital di dunia nyata.

2. METODE PENELITIAN

Dataset yang akan digunakan pada penelitian ini yaitu merupakan dataset email yang didapatkan dari repositori data penelitian University of Twente (research.utwente.nl) dengan judul Phishing Validation Emails Dataset. Dengan total data yang digunakan yaitu sebanyak [2000] data email yang terdiri dari [1000] data safe email (ham) dan [1000] data phishing email. Untuk visualisasi dataset diberikan pada Gambar [1].

	Email Text	Email Type
0	Dear Jordan, your subscription has been successfully renewed.	Safe Email
1	Dear Casey, thank you for your purchase.	Safe Email
2	Congratulations! You've won a \$3000 gift card.	Phishing Email
3	You have a new secure message from your bank.	Phishing Email
4	Your package delivery is pending. Please provide your personal information to confirm delivery.	Phishing Email

Gambar 1. Dataset penelitian

Dari total [2000] data teks tersebut, akan dibagi menjadi 80% data pelatihan dan 20% data pengujian. Yang dimana data pelatihan digunakan agar model dapat berlatih mengenali pola dari data. Sedangkan, data pengujian berguna untuk melakukan proses evaluasi performa model dalam melakukan proses klasifikasi.

Support Vector Machine merupakan algoritma yang akan mencari hyperplane dimana hyperplane tersebut merupakan hyperplane yang paling optimal dan memiliki amplitudo tertinggi di setiap kelasnya untuk membagi data ke dalam kelas-kelas yang ada. Untuk rumus SVM diberikan pada poin [1],[2] dan [3].

$$f(x) = \omega \cdot x + b \quad \dots (1)$$

Dimana:

- ω = nilai vector bobot (untuk menentukan orientasi arah dari hyperplane)
- x = nilai fitur yang digunakan dalam bentuk vector
- b = nilai bias (nilai ruang jarak hyperplane menuju titik asal (0,0) pada lingkup ruang fitur)

$$\text{Margin} = \frac{2}{|\omega|} \quad \dots (2)$$

Dimana $|\omega|$ = nilai Panjang dari vector bobot pada SVM.

$$y_i(\omega \cdot x_i + b) \geq 1 \quad \dots (3)$$

Dimana dengan memaksimalkan margin dengan batasan setidaknya jarak 1 dari hyperplane dengan nilai y_i merupakan nilai target pada data yang diberikan dan x_i merupakan nilai fitur dari data.

Vectorizer TF-IDF (Term Frequency-Inverse Document Frequency) adalah metode yang digunakan untuk melakukan proses konversi teks ke format vektor dan juga digunakan untuk melakukan proses penghitungan kemunculan kata-kata penting dalam dokumen dengan menggunakan metode statistik. Rumus TF-IDF diberikan pada poin [6]. Pada proses ini kita akan mencari nilai TF yang digunakan untuk mengukur frekuensi kemunculan kata dalam dokumen. Rumus TF diberikan pada poin [4]. Kemudian, nilai IDF digunakan untuk menentukan kepentingan kata secara keseluruhan dalam keseluruhan isi dokumen. Rumus IDF diberikan pada poin [5].

$$TF(t, d) = \frac{\text{jumlah muncul kata } t \text{ dalam dokumen } d}{\text{total kata dalam dokumen } d} \quad \dots (4)$$

$$IDF(t, D) = \log \left(\frac{\text{total dokumen dalam korpus } D}{\text{jumlah dokumen yang mengandung kata } t + 1} \right) + 1 \quad \dots (5)$$

$$TF-IDF = TF(t, d) \times IDF(t, D) \quad \dots (6)$$

Confusion matrix adalah matriks yang dapat digunakan untuk menghitung performa suatu model menggunakan nilai presisi, akurasi, recall, serta nilai f1-score. Dengan demikian kita dapat mengetahui kualitas model yang dibangun untuk dapat melakukan proses klasifikasi. Pada poin [7] disajikan rumus perhitungan keakuratan prediksi model. Poin [8] menunjukkan rumus yang menentukan performa model dalam membuat prediksi akurat untuk setiap data atau kelas yang ada. Sedangkan poin [9] menyajikan rumus untuk mencari nilai

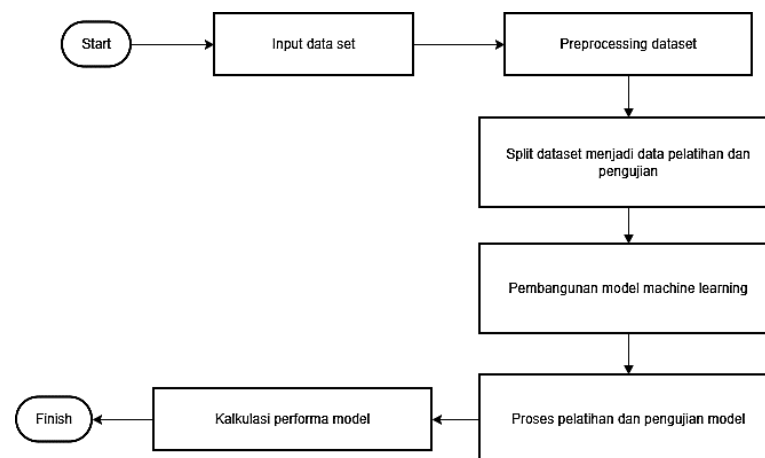
keselarasan atau keseimbangan antara nilai presisi dan nilai recall yang diperoleh sebelumnya. Pemilihan rasio pembagian data latih dan uji sebesar 80:20 didasarkan pada prinsip *hold-out validation*, di mana data latih digunakan untuk membangun model sementara data uji digunakan untuk mengukur kemampuan generalisasi model terhadap data baru.

$$\text{Presisi} = \frac{\text{Pos Benar}}{\text{Pos Benar} + \text{Pos Salah}} \quad \dots (7)$$

$$\text{Recall} = \frac{\text{Pos Benar}}{\text{Pos Benar} + \text{Neg Salah}} \quad \dots (8)$$

$$F1\text{-Score} = 2 \times \frac{\text{Presisi} \times \text{Recall}}{\text{Presisi} + \text{Recall}} \quad \dots (9)$$

Pada penelitian ini, akan menggunakan IDE Jupyter Notebook dan Bahasa pemrograman Python untuk melakukan proses klasifikasi email phishing. Untuk alur proses klasifikasi diberikan pada Gambar [2].



Gambar 2. Preprocessing

Gambar 2 menunjukkan alur proses klasifikasi email phishing menggunakan algoritma Support Vector Machine (SVM). Proses diawali dengan pengumpulan dataset email yang kemudian melalui tahap preprocessing, meliputi pembersihan teks, tokenisasi, penghapusan stopwords, dan pembobotan kata menggunakan metode TF-IDF. Tahapan preprocessing ini bertujuan untuk mengurangi noise pada data teks serta meningkatkan kualitas fitur yang digunakan dalam proses klasifikasi [2], [6].

Selanjutnya, data hasil ekstraksi fitur dibagi menjadi data latih dan data uji dengan rasio 80:20. Data latih digunakan untuk membangun model SVM, sedangkan data uji digunakan untuk mengevaluasi kinerja model. Pada tahap pelatihan, SVM bekerja dengan mencari hyperplane optimal yang mampu memisahkan kelas email aman dan email phishing dengan margin maksimum. Proses ini memungkinkan SVM untuk menangani data berdimensi tinggi secara efektif, khususnya pada data teks berbasis TF-IDF [3], [9].

Hasil akhir dari proses ini berupa prediksi kelas email yang kemudian dievaluasi menggunakan confusion matrix dan classification report. Evaluasi tersebut memberikan gambaran menyeluruh mengenai performa model, termasuk nilai akurasi, presisi, recall, dan F1-score. Alur yang ditampilkan pada Gambar 2 memperlihatkan bahwa setiap tahapan saling berkaitan dan berkontribusi dalam menghasilkan sistem deteksi phishing yang akurat dan andal [1], [5].

3. HASIL DAN PEMBAHASAN

Pada penelitian ini, proses pengujian akan menggunakan IDE Jupyter Notebook dan Bahasa pemrograman Python untuk dapat melakukan proses klasifikasi indikasi phishing pada pesan email. Setelah data teks email disiapkan dan diproses melalui tahapan preprocessing serta ekstraksi fitur, maka selanjutnya akan melakukan proses pelatihan model. Setelah model dilakukan proses pelatihan untuk dapat mengenali pola dari data, maka model tersebut dilakukan proses pengujian dengan menggunakan data pengujian agar dapat diketahui performa model yang dibangun. Untuk hasil pengujian untuk proses klasifikasi email phishing pada penelitian ini mendapatkan akurasi sebesar 100% dan dengan nilai-nilai performa jika dihitung dengan confusion matrix yang sudah didapatkan diberikan pada Tabel [1].

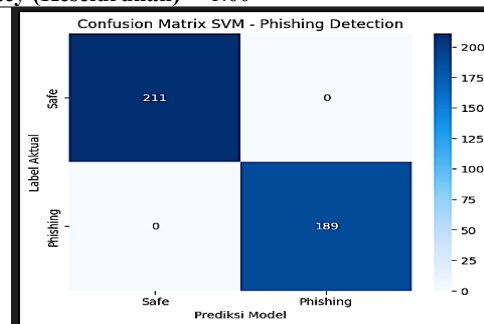
Hasil pengujian menunjukkan bahwa model SVM yang dibangun mampu mencapai tingkat akurasi sebesar 100% dalam mengklasifikasikan email aman dan email phishing. Nilai performa model secara rinci ditampilkan pada Tabel [1], memperlihatkan nilai presisi, recall, dan F1-score masing-masing kelas bernilai 1.00. Hasil ini mengindikasikan bahwa model tidak mengalami kesalahan klasifikasi pada data pengujian.

Nilai presisi dan recall yang sama-sama mencapai 1.00 pada kedua kelas menunjukkan bahwa model tidak bersifat bias terhadap salah satu kelas tertentu. Kondisi ini sangat penting dalam konteks keamanan siber, karena kesalahan dalam mengklasifikasikan email phishing sebagai email aman dapat menyebabkan kebocoran data sensitif pengguna. Sebaliknya, kesalahan dalam mengklasifikasikan email aman sebagai phishing dapat mengganggu aktivitas komunikasi pengguna. Oleh karena itu, keseimbangan antara nilai presisi dan recall menjadi indikator utama bahwa model yang dibangun layak untuk diterapkan dalam sistem deteksi phishing secara nyata [3], [5].

Meskipun hasil yang diperoleh menunjukkan performa yang sangat baik, penelitian ini masih memiliki keterbatasan pada penggunaan dataset yang bersifat statis. Dataset yang digunakan belum sepenuhnya merepresentasikan variasi serangan phishing terbaru yang terus berkembang, sehingga perlu diuji lebih lanjut menggunakan dataset yang lebih beragam atau bersumber dari email real-time agar dapat mengukur kemampuan generalisasi model terhadap variasi serangan phishing yang berkembang [5], [8].

Tabel 1. Nilai Confusion Matrix

Metrik	Safe Email	Phishing Email
Precision	1.00	1.00
Recall	1.00	1.00
F1-score	1.00	1.00
Accuracy (Keseluruhan)	1.00	—



Gambar 3. Confusion Matrix

Arti Dari Angka terlihat pada Gambar [3].

1. True Negative (TN) = 211
Sebanyak 211 email aman berhasil diklasifikasikan dengan benar sebagai *safe email*.
2. False Positive (FP) = 0
Tidak ada *email aman* yang salah diklasifikasikan sebagai *phishing*.
3. False Negative (FN) = 0
Tidak ada *email phishing* yang salah diklasifikasikan sebagai *safe email*.
4. True Positive (TP) = 189
Sebanyak 189 email phishing diklasifikasikan dengan benar oleh model.

Pada Tabel [1] menunjukkan hasil klasifikasi dengan menggunakan algoritma SVM. [cite_start] dapat terlihat dari Tabel [1] bahwa model yang dibangun memiliki nilai keakuratan prediksi yang sangat baik yaitu sebesar 100%, nilai ketepatan prediksi pada semua kelas yang sangat baik juga yaitu sebesar 100% dan nilai harmonik yang baik antara presisi dan recall yaitu sebesar 100%. Dari nilai-nilai tersebut, maka dapat dilihat bahwa model yang dibangun dapat melakukan proses klasifikasi dengan sangat baik dan dapat menghasilkan keakuratan prediksi yang sempurna sehingga dapat dengan baik membantu dalam melakukan proses klasifikasi email phishing.

Hasil pengujian menunjukkan bahwa metode SVM dengan kernel linear memberikan hasil akurasi sempurna sebesar 100%. Hal ini menunjukkan bahwa fitur yang diekstraksi menggunakan TF-IDF memiliki kemampuan yang sangat baik untuk membedakan karakteristik email phishing dan email aman. Nilai presisi dan recall yang sama-sama mencapai 1.00 menunjukkan bahwa model tidak hanya mampu mengklasifikasikan dengan benar, tetapi juga tidak melakukan kesalahan klasifikasi terhadap kelas mana pun [2], [3].

Jika dibandingkan dengan hasil penelitian lain yang menggunakan algoritma *Naïve Bayes* dan *Decision Tree*, metode SVM menunjukkan tingkat kestabilan yang lebih baik terhadap variasi data dan *outlier*. Kinerja tinggi ini disebabkan karena SVM bekerja dengan mencari *hyperplane optimal* yang memaksimalkan jarak antar

kelas sehingga menghasilkan pemisahan yang tegas antara email phishing dan non-phishing [4], [9]. Selain itu, penggunaan kernel linear juga mempercepat waktu komputasi karena kompleksitasnya lebih rendah dibandingkan kernel non-linear seperti RBF atau polinomial.

4. SIMPULAN DAN SARAN

Setelah dilakukan proses pelatihan dan pengujian dengan menggunakan algoritma Support Vector Machine (SVM) untuk melakukan proses klasifikasi email phishing, maka dapat disimpulkan bahwa metode SVM dapat dengan baik melakukan proses klasifikasi data terutama email phishing. Hasil akurasi yang maksimal menunjukkan bahwa metode ini efektif mengenali pola teks berbahaya, sehingga dapat membantu masyarakat untuk dapat tidak tertipu karena phishing lagi.

Pada penelitian selanjutnya, diharapkan untuk dapat melakukan proses klasifikasi dengan model lainnya seperti Naïve Bayes Classifier, Random Forest, KNN, Decision Tree, dan lainnya sehingga dapat dilakukan proses perbandingan metode mana yang terbaik untuk melakukan proses klasifikasi email phishing dengan variasi dataset yang lebih beragam.

Penelitian ini diharapkan tidak hanya memberikan kontribusi dalam ranah akademik, tetapi juga menjadi solusi praktis dalam upaya pencegahan kejahatan siber. Pengembangan sistem deteksi phishing berbasis SVM dapat dijadikan referensi bagi lembaga pendidikan, perbankan, maupun perusahaan teknologi untuk meningkatkan kesadaran keamanan informasi di tingkat pengguna. Selain itu, penelitian lanjutan dapat menguji efektivitas model ini menggunakan dataset real-time dari layanan email populer untuk menilai ketahanan sistem terhadap variasi serangan phishing terbaru

DAFTAR PUSTAKA

- [1] Chaerul Umam, L. Budi Handoko (2024). “PREDIKSI EMAIL PHISING MENGGUNAKAN SUPPORT VECTOR MACHINE”. <https://proceeding.unindra.ac.id/index.php/semnasristek/article/view/7138>.
- [2] Rino Nurcahyo Fauzi Tanjung, Sayuti Rahman (2025). “Meningkatkan Deteksi Email Phising Melalui Pendekatan SVM yang Dioptimalkan NLP”
<https://repositori.uma.ac.id/jspui/handle/123456789/27660>
- [3] Mui Sunjaya (2025). “Analisis Komperatif Akurasi Deteksi Phising Pada Jaringan Komputer Dengan Metode Naïve Bayes Dan Support Vector Machine (SVM)”.
https://www.researchgate.net/publication/396595263_Analisis_Komperatif_Akurasi_Deteksi_Phising_Pada_Jaringan_Komputer_Dengan_Metode_Naive_Bayes_Dan_Support_Vector_Machine_SVM
- [4] SITI AISYAH (2020). “KLASIFIKASI SPAM PADA EMAIL MENGGUNAKAN METODE SUPPORT VECTOR MACHINE DAN DETEKSI ANOMALY”
https://repository.unsri.ac.id/32024/3/RAMA_56201_09011181621024_0003047905_0216068101_01_front_ref.pdf
- [5] Raweia S MohamedAli , Razan Abdulhammed (2025). “Comparative Analysis of Machine Learning Algorithms for Phishing Email Detection”
<https://iasj.rdd.edu.iq/journals/uploads/2025/10/07/c8ee8afbba932d46ec2c63e56442d7b6.pdf>
- [6] Arar Al Tawil, Laiali Almazaydeh, Doaa Qawasmeh, Baraah Qawasmeh (2024). “Comparative Analysis of Machine Learning Algorithms for Email Phishing Detection Using TF-IDF, Word2Vec, and BERT” .
https://www.researchgate.net/publication/385775640_Comparative_Analysis_of_Machine_Learning_Algorithms_for_Email_Phishing_Detection_Using_TF-IDF_Word2Vec_and_BERT.
- [7] Herliyani Hasanah, Nurmalitasari Nurmalitasari, “Perbandingan Tingkat Akurasi Algoritma Support Vector Machines (SVM) dan C45 dalam Prediksi Penyakit Jantung” <https://proceeding.unpkediri.ac.id/index.php/stains/article/view/2819>.
- [8] Ahmad Assril Karim, Muhammad Ary Prasetyo, Muhammad Rohid Saputro, Perbandingan Metode Random Forest, K-Nearest Neighbor, dan SVM Dalam Prediksi Akurasi Pertandingan Liga Italia. <https://proceeding.unpkediri.ac.id/index.php/stains/article/view/2877>.
- [9] Fajar P. Wibowo, “Optimasi Support Vector Machine untuk Filtering Email Spam,” *Jurnal Teknologi dan Sains (JTS)*, vol. 4, no. 2, pp. 45–52, 2024.
- [10] Ahmad R. Nugraha, “Hybrid Feature Extraction untuk Deteksi Phishing Email Menggunakan TF-IDF dan Word2Vec,” *International Journal of Cyber Security and Digital Forensics*, vol. 13, no. 2, pp. 84–93, 2025.