INCODING: Journal of Informatics and Computer Science Engineering
ISSN 2776-432X (Online - Elektronik)
https://journal.mahesacenter.org/index.php/incoding

Date: April 16th, 2025

LETTER OF ACCEPTANCE
Paper Number #831

Dear, Rino Nurcahyo Fauzi Tanjung, Sayuti Rahman

This is to inform you that the manuscript entitled: **"Meningkatkan Deteksi Email Phising Melalui Pendekatan SVM yang Dioptimalkan NLP",** which was sent on **April 16th 2025,** is **ACCEPTED.**

We keep to ensuring a high standard of articles published in the **INCODING: Journal of Informatics and Computer Science Engineering,** and the manuscript that is being sent to you has been submitted after a first selection process based on the agreement of the **Associate Editors.** In general, the standard of manuscripts forwarded to me after the vetting is **good.**

This paper is well organized and follows the manuscript guidelines of the journal to a large extent. The introduction section is good and shows the importance of the study. The literature review is adequate. The outcomes of the study are consistent with the findings. The approach used is praiseworthy. In my opinion, it should be published without **revision again**

Based on the review results, this manuscript is **ACCEPTED**, and **PUBLISHED** in **Mei 2025** for **Volume 3, No. 2, 2025.**

Thank you very much for your contribution. Congratulations on a wonderful job.

Warmest Regards,
Editor In Chief

INCODING
2776-432X (Online - Elektronik)

Agung Suharyanto, S.Sn, M.Si.

Editorial Office:
Mahesa Research Center
Perumahan Griya Nafisa 2, Blok A No 10, Jalan Benteng Hilir
Titi Sewa, RT 06, Dusun XVI Flamboyan,
Kecamatan Percut Sei Tuan, Deli Serdang, 20371
Sumatera Utara, Indonesia
Phone: 082126493527
Email: incodingjournal@gmail.com

INCODING: Journal of Informatic and Computer Science Engineering
https://journal.mahesacenter.org/index.php/incoding/index || ISSN xxxx-xxxx (online)
Vol (Issue) Year: Page
DOI: Genesis:

Meningkatkan Deteksi Email Phising Melalui Pendekatan SVM yang Dioptimalkan NLP

Enhancing Phishing Email Detection through NLP-Optimized SVM Approach

Rino Nurcahyo Fauzi Tanjung[1)], & Sayuti Rahman[2)]
1)Informati, Fakultas Teknik, Universitas Medan Area, Indonesia
Diterima: Maret 2025; Disetujui: Maret 2025; Dipublish: April 25

Coresponding Email: rinonurcahyofauzitanjung@gmail.com

Abstrak

Email phishing merupakan ancaman keamanan siber yang signifikan, dengan risiko mencakup kebocoran data pribadi, penipuan keuangan, dan penyebaran malware. Studi ini bertujuan untuk meningkatkan deteksi email phishing dengan pendekatan machine learning, mengoptimalkan teknik ekstraksi fitur, serta memilih algoritma klasifikasi yang paling efektif. Penelitian ini menggunakan Dataset Deteksi Email Phishing, yang berisi berbagai teks email yang telah dikategorikan sebagai aman atau phishing. Teknik *Natural Language Processing* (NLP), khususnya *Term Frequency-Inverse Document Frequency* (TF-IDF), diterapkan untuk mengubah teks menjadi vektor numerik guna meningkatkan representasi fitur. Model klasifikasi utama yang digunakan adalah *Support Vector Machine* (SVM) dengan kernel polynomial, yang dibandingkan dengan algoritma lain seperti *Random Forest, Naïve Bayes, Logistic Regression*, dan *K-Nearest Neighbors* (KNN). Evaluasi dilakukan menggunakan metrik akurasi, *precision, recall*, dan *F1-score*. Hasil eksperimen menunjukkan bahwa SVM dengan kernel polinomial dan TF-IDF memberikan akurasi tertinggi sebesar 97,85%. Teknik NLP seperti tokenisasi dan penghapusan kata henti juga berkontribusi terhadap peningkatan akurasi klasifikasi. Studi ini menunjukkan bahwa kombinasi NLP dan SVM secara efektif meningkatkan deteksi email phishing. Penelitian selanjutnya dapat mengeksplorasi integrasi model pembelajaran mendalam dan teknik NLP lanjutan untuk meningkatkan akurasi serta efisiensi sistem deteksi phishing.

Abstract

Email phishing merupakan ancaman keamanan siber yang signifikan, dengan risiko mencakup kebocoran data pribadi, penipuan keuangan, dan penyebaran malware. Studi ini bertujuan untuk meningkatkan deteksi email phishing dengan pendekatan machine learning, mengoptimalkan teknik ekstraksi fitur, serta memilih algoritma klasifikasi yang paling efektif. Penelitian ini menggunakan Dataset Deteksi Email Phishing, yang berisi berbagai teks email yang telah dikategorikan sebagai aman atau phishing. Teknik Natural Language Processing (NLP), khususnya Term Frekuensi-Invers Dokumen Frekuensi (TF-IDF), diterapkan untuk mengubah teks menjadi vektor numerik guna meningkatkan fitur representasi. Model klasifikasi utama yang digunakan adalah Support Vector Machine (SVM) dengan kernel polinomial, yang dibandingkan dengan algoritma lain seperti Random Forest, Naïve Bayes, Logistic Regression, dan K-Nearest Neighbors (KNN). Evaluasi dilakukan menggunakan metrik akurasi, presisi, recall, dan F1-score. Hasil eksperimen menunjukkan bahwa SVM dengan kernel polinomial dan TF-IDF memberikan akurasi tertinggi sebesar 97,86%. Teknik NLP seperti tokenisasi dan penghapusan kata henti juga berkontribusi terhadap peningkatan akurasi klasifikasi. Studi ini menunjukkan bahwa kombinasi NLP dan SVM secara efektif meningkatkan deteksi email phishing. Penelitian selanjutnya dapat mengeksplorasi integrasi model pembelajaran mendalam dan teknik NLP lanjutan untuk meningkatkan akurasi serta efisiensi sistem

Document Accepted 3/7/25

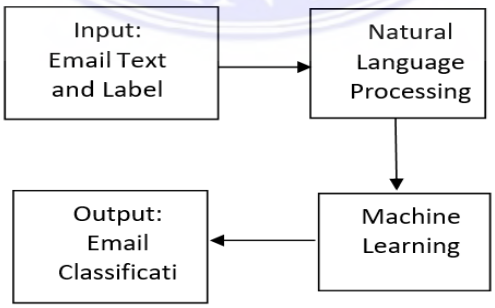
PENDAHULUAN

Email phishing sering kali mengakibatkan berbagai masalah, termasuk kehilangan data pribadi, penipuan keuangan, distribusi malware, serangan phishing, spam, penipuan spear phishing, dan bahaya ransomware (Leonov et al. 2021),(Simoiu et al. 2020). Email palsu sering kali berupaya mengumpulkan informasi sensitif, menipu pengguna agar mentransfer uang, menginfeksi perangkat dengan malware, atau mencuri informasi pribadi. Untuk melindungi diri dari ancaman email palsu, pengguna harus waspada terhadap sinyal penipuan, memverifikasi keabsahan email, dan menerapkan sistem keamanan yang solid (Chudra et al. 2022). Ada berbagai metode untuk memverifikasi keabsahan email dan melindungi diri dari email palsu (Salloum et al. 2022). Memverifikasi alamat pengirim sambal mencari tanda-tanda penipuan, memverifikasi domain email untuk memastikan keaslian, pengecekan link dan lampiran sebelum mengklik atau membukanya, verifikasi identitas instansi terkait melalui jalur komunikasi resmi, penggunaan system keamanan yang handal, dan peningkatan pengetahuan taktik merupakan metode yang umum dilakukan (Syahputra, Ananda, dan Siregar 2022), (Martin dan Swapna 2022). Serta Teknik penipuan email. Keamanan penipuan dapat dikurangi dengan menggunakan metode-metode ini. Metode *machine learning* digunakan dalam beberapa algoritma klasifikasi email, termasuk *Naïve Bayes* (Gupta, Palwe, dan Keskar 2020) , *Support Vector Machine* (SVM) , *Decision Tree, Random Forest*, dan K-Nearest Neighbors (KNN), SVM mencari hyperplane pemisah antara kelas, Decision Tree menggunakan struktur pohon keputusan, Random Forest menggabungkan beberapa pohon keputusan, dan K -NN mengklasifikasikan email berdasarkan tetangga terdekat dalam ruang fitur (Dinata et al. 2023). SVM merupakan metode klasifikasi email yang populer karena memiliki berbagai kelebihan. Pertama-tama, SVM dapat menangani data yang kompleks dan non-linier dalam email dengan memanfaatkan kernel seperti RBF, (Ding et al. 2021),(Gopi et al. 2023). dan kernel *polynomial* (Zhou dan Jetter 2006). Hal ini memungkinkan SVM untuk memproyeksikan data ke dimensi yang lebih tinggi dan secara efektif menemukan pola dan hubungan yang kompleks dalam email. Lebih jauh lagi, karena ide margin terbesar mendorong pemisahan berbagai kelas, SVM dapat menangani data dengan sejumlah besar fitur tanpa kehilangan efisiensi. SVM juga bagus dalam menangani masalah ketidakseimbangan data, seperti ukuran

sampel yang sangat besar dalam email untuk kelas phishing dan aman. Terakhir, SVM memberikan kinerja yang baik dengan ukuran kumpulan data yang relatif kecil, sehingga cocok untuk klasifikasi email yang umumnya melibatkan kumpulan data berukuran terbatas. SVM juga memiliki proses yang lebih cepat dibandingkan *Deep Learning* (Sayuti Rahman et al. 2020),(S. Rahman et al. 2022). Selain itu, kami menggunakan algoritma Natural Language Processing (NLP) untuk mengekstrak aspek-aspek penting dari teks, seperti kata kunci, nama entitas, dan frasa penting. Hal ini berkontribusi pada terciptanya representasi teks yang lebih optimal dalam proses kategorisasi (Olivetti et al. 2020). Penelitian ini membangun fondasi yang kuat untuk memperluas temuannya ke berbagai domain dalam ranah keamanan siber dan klasifikasi teks. Melalui perbandingan sistematis hasil SVM menggunakan berbagai metode dan penerapan penyetelan hiperparameter untuk setiap kernel, termasuk penyetelan halus parameter seperti nilai C dan gamma, tujuan kami adalah mengoptimalkan kinerja model. Sasaran utamanya adalah menentukan kombinasi fitur dan NLP yang paling efektif, dengan antisipasi bahwa hal ini akan meningkatkan akurasi kategorisasi email phishing secara substansial.

METODE PENELITIAN

Metode *Natural Language Processing* (NLP) digunakan sebagai pemroses email teks dalam penelitian ini, dan hasilnya dikenali oleh pendekatan pembelajaran mesin. Ada banyak pendekatan untuk menentukan metode terbaik dalam mengklasifikasikan email palsu. Metode-metode ini meliputi SVM, *naive bayes*, dan *random forest*. langkah-langkah proses penelitian ditunjukkan pada gambar di bawah ini.

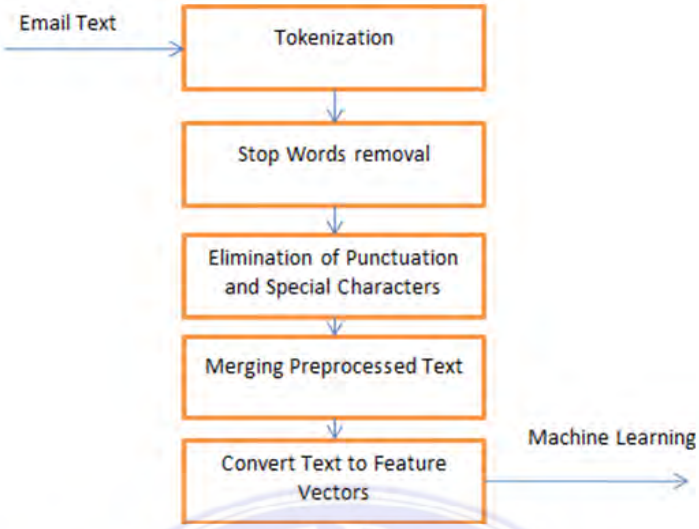


Gambar 1. Proses *Clasification Email*

Natural Language Processing (NLP)

Metode NLP dalam penelitian ini terdiri dari banyak fase kunci yang mengubah teks menjadi atribut yang dapat diklasifikasikan menggunakan *Machine Learning.*

Berikut ini adalah Langkah-langkah proses NLP dalam penelitian ini.



Gambar 2. Proses NLP

Seperti yang terlihat pada Gambar 2, teks email ditokenisasi, yaitu Proses pemrosesan teks dimulai dengan tokenisasi, yaitu membagi teks menjadi unit kecil seperti kata. Selanjutnya, *stop words* (kata umum yang tidak memiliki makna signifikan, seperti "the", "is", "and") dihapus, bersama dengan tanda baca dan karakter khusus. Setelah pembersihan, teks yang telah diproses digabungkan kembali menjadi satu string. Teks ini kemudian dikonversi menjadi vektor fitur menggunakan *Term Frequency-Inverse Document Frequency* (TF-IDF) atau *Bag of Words* (BoW). Nilai vektor fitur ini digunakan sebagai input dalam model *machine learning* untuk klasifikasi (Hussain Sujon dan Mustafa 2023).

Machine Learning

Metode machine learning yang digunakan dalam penelitian ini adalah SVM, Random Forest, dan Naive Bayes. Kami mencoba membandingkan metode-metode ini untuk menemukan akurasi terbaik untuk klasifikasi email phishing.

Support Vector Machine (SVM)

mengenai materi dan metode yang digunakan dalam penelitian, desain percobaan, teknik pengambilan sampel, variabel yang diukur, serta metode analisis data. Support Vector Machine (SVM) digunakan sebagai algoritma utama untuk klasifikasi email phishing, dengan berbagai kernel seperti linier, Radial Basis Function (RBF), dan polynomial (Pamungkas, Risandriya, dan Rahman 2022). SVM bekerja dengan mencari hyperplane optimal yang memisahkan kelas data (Noble 2006),(Vanneschi dan Silva 2023). Pemilihan parameter seperti gamma dan D dalam

kernel SVM memengaruhi kinerja model dan dapat dioptimalkan menggunakan validasi silang atau pencarian grid. Kernel linier digunakan untuk pemisahan data yang dapat diklasifikasikan secara linier, sedangkan kernel RBF menangani hubungan kompleks dan non-linier. Kernel polinomial memungkinkan transformasi ke ruang fitur berdimensi lebih tinggi untuk mengatasi data yang tidak terpisahkan secara linier. Model dievaluasi menggunakan metrik akurasi, *precision, recall*, dan *F1-score*. Pendekatan ini bertujuan untuk meningkatkan efektivitas deteksi phishing dengan teknik *Natural Language Processing* (NLP) dan optimalisasi model SVM.

Fungsi keputusan dalam SVM dengan kernel linier memiliki rumus dasar berikut:

$$f(x) = w^T x + b \quad (1)$$

Di mana:

f (x) adalah fungsi keputusan yang memprediksi kelas sampel x.

w adalah vector normal pada hiperbidang pembagi.

T adalah operasi transposisi pada vector w.

X adalah vector fitur sampel yang akan diprediksi.

Kernel RBF (Radial Basis Function) memiliki pengaruh yang signifikan terhadap algoritma SVM. Kernel RBF memungkinkan SVM untuk memodelkan hubungan yang kompleks dan non-linier antara fitur dalam.

Rumus dasar untuk fungsi kernel RBF adalah sebagi berikut:

$$K(x, x') = exp(-gamma ||x - x'||^2) \quad (2)$$

Di mana :

$K(x, x')$ adalah representasi kernel RBF dari dua vector fitur x dan x'

dalam fungsi kernel RBF, parameter gamma mengatur sejauh mana pengaruh satu titik data meluas ketitik data lainnya. $||x - x'||^2$ adalah jarak kuadrat antara dua vector fitur x dan x'.

Kernel polinomial dalam SVM digunakan untuk mentransformasikan data dari ruang fitur asli ke ruang berdimensi lebih tinggi, memungkinkan pemisahan kelas yang tidak dapat dipisahkan secara linear. Kernel ini efektif dalam menangani pola kompleks dan menentukan batas keputusan yang lebih fleksibel.

Naive Bayes (BN)

Naïve Bayes adalah algoritma klasifikasi berbasis Bayes' Theorem yang mengasumsikan bahwa setiap fitur bersifat independen, meskipun asumsi ini tidak

selalu benar (Webb 2020). Dalam klasifikasi email, Naïve Bayes menghitung probabilitas suatu email termasuk dalam kategori tertentu (misalnya, phishing atau aman) berdasarkan frekuensi kemunculan kata atau frasa dalam teks. Algoritma ini tetap efektif meskipun dengan asumsi yang sederhana, karena cepat dalam pelatihan dan prediksi serta berkinerja baik dalam banyak kasus klasifikasi teks.

Evaluasi Kinerja

Evaluasi kinerja model sangat krusial dalam pengembangan algoritma *machine learning*, termasuk *Support Vector Machine* (SVM). Evaluasi ini mengukur kemampuan model dalam menggeneralisasi dan memprediksi data baru secara akurat. Dalam studi ini, kinerja model dianalisis menggunakan metrik akurasi, presisi, *recall*, dan skor F1 untuk menilai efektivitas deteksi email phishing (Yacouby dan Axman 2020).

Dataset

Dataset yang digunakan bersifat publik dan dapat diunduh melalui situs Kaggle. Dataset ini berisi informasi mengenai isi teks email serta kategorinya, yang dapat dimanfaatkan untuk mengidentifikasi email phishing melalui analisis teks dan klasifikasi berbasis machine learning. Dengan menerapkan teknik analisis teks dan machine learning, dataset ini membantu meningkatkan kemampuan dalam mendeteksi email phishing.

Tabel 1. Contoh Dataset

No	Email Text	Email Type
1	On Sun, Aug 11, 2002 at 11:17:47AM +0100, wintermute mentioned: > > The impression I get from readin...	Safe Email
2	entourage, stockmogul newsletter ralph velez ,genex pharmaceutical , inc . (otcbb : genx) biotec...	Phishing Email
3	we owe you lots of money dear applicant, after further review upon receiving your application your ...	Phishing Email

Seperti yang ditunjukkan pada Tabel 1, dataset berisi teks email dan jenis email. Teks email adalah pesan yang diterima melalui email sedangkan jenis email adalah hasil pelabelan email, yaitu email aman atau email penipuan. Dataset dipisahkan menjadi data pelatihan dan data uji, dengan data pelatihan sebesar 90% dan data uji sebesar 10%.

Tabel 2. Junlah Dataset

Document Accepted 3/7/25
Access From (repository.uma.ac.id)3/7/25

Labels	Amount
Phishing Email	7.328
Safe Email	11.322
Total	18.650

Tabel 2 menunjukkan jumlah kumpulan data adalah 18.650 dengan 7.328 email phishing dan 11.322 email aman.

HASIL DAN PEMBAHASAN

Percobaan dilakukan dengan menggunakan kumpulan data Deteksi Email Pishing. Beberapa strategi percobaan digunakan, termasuk membagi kumpulan data ke dalam pengaturan yang berbeda, memeriksa efek konversi teks menjadi vektor fitur, membandingkan berbagai metode klasifikasi, dan menilai manfaat NLP dalam klasifikasi email phishing. Hasil percobaan ini diharapkan dapat mengungkap teknik terbaik untuk meningkatkan akurasi klasifikasi, yang kemudian dapat diterapkan dalam upaya deteksi email phishing.

1. Hasil Uji Dataset

Kumpulan data dibagi menjadi beberapa bagian untuk pelatihan dan pengujian. Eksperimen ini melibatkan perubahan susunan data pelatihan, dengan 0,1 hingga 0,9 dari keseluruhan kumpulan data berfungsi sebagai data pelatihan dan data yang tersisa berfungsi sebagai data pengujian. Untuk mengubah nilai string menjadi representasi vektor, NLP digunakan, diikuti oleh pendekatan Bag of Words (BoW). Selanjutnya, vektor ini diklasifikasikan menggunakan teknik SVM. Kernel linier dalam SVM digunakan dalam eksperimen awal. Hasil akurasi klasifikasi yang diperoleh berdasarkan variasi komposisi data pelatihan dan pengujian, dengan penerapan BoW dan SVM kernel linier, didokumentasikan.

Tabel 3. Klasifikasi SVM dan BOW

Train Size	Accuracy Train	Accuracy Test	Precision	Recall	F1-Score
0.1	0.9892	0.9310	0.9328	0.9310	0.9313
0.2	0.9890	0.9428	0.9442	0.9428	0.9431
0.3	0.9883	0.9443	0.9455	0.9443	0.9446
0.4	0.9880	0.9521	0.9526	0.9521	0.9522
0.5	0.9887	0.9574	0.9580	0.9574	0.9575
0.6	0.9882	0.9591	0.9596	0.9591	0.9592
0.7	0.9885	0.9614	0.9617	0.9614	0.9614
0.8	0.9887	0.9619	0.9624	0.9619	0.9620
0.9	**0.9890**	**0.9640**	**0.9644**	**0.9640**	**0.9641**

Tabel 3 menggambarkan hasil klasifikasi SVM menggunakan representasi Bag of Words

(BoW) pada berbagai ukuran data training. Terbukti bahwa penggunaan 90% data set untuk training dan 10% untuk pengukuran menghasilkan akurasi tertinggi pada kedua tahap. Semakin banyak data yang digunakan untuk melatih model, semakin baik model tersebut dalam mengklasifikasikan teks. Secara spesifik, akurasi pada training set adalah 96,40%, sedangkan pada testing set adalah 98,90%. Selain itu, precision, recall, dan skor F1 untuk konfigurasi ini masing-masing adalah 96,44%,96,40%,dan 96,41%. Langkah selanjutnya adalah mereplikasi percobaan menggunakan skema data yang sama. Namun, terdapat perubahan pada proses transformasi teks menjadi vektor pada percobaan ini, yaitu menggunakan pendekatan TF-IDF.

Tabel 4. Hasil SVM Linear dengan TF-IDF

Train Size	Accuracy Train	Accuracy Test	Precision	Recall	F1-Score
0.1	0.9871	0.9588	0.9589	0.9588	0.9588
0.2	0.9879	0.9668	0.9671	0.9668	0.9669
0.3	0.9871	0.9692	0.9695	0.9692	0.9692
0.4	0.9864	0.9714	0.9716	0.9714	0.9714
0.5	0.9873	0.9736	0.9738	0.9736	0.9736
0.6	0.9869	0.9772	0.9773	0.9761	0.9772
0.7	0.9870	0.9756	0.9758	0.9756	0.9757
0.8	0.9874	0.9772	0.9773	0.9772	0.9772
0.9	**0.9877**	**0.9780**	**0.9781**	**0.9780**	**0.9780**

Seperti yang ditunjukkan pada Tabel 4, hasil terbaik untuk Pengujian Akurasi, Presisi, Recall, dan Skor F-1 diperoleh ketika proporsi data pelatihan sebesar 90% dan proporsi data pengujian sebesar 10% dari total dataset digunakan. Hasil pengujian ini menunjukkan bahwa skema data pelatihan dan pengujian sebesar 90%-10% adalah ideal, dan metode ini akan digunakan sebagai referensi untuk eksperimen selanjutnya.

1. Hasil Perbandingan Kernel SVM

Pengujian kemudian difokuskan pada perbandingan berbagai kernel yang digunakan dalam SVM dengan data yang telah diubah menjadi vektor fitur. BoW dan TF-IDF adalah dua pendekatan untuk mengubah data string menjadi vektor fitur. Kernel SVM RBF, Polinomial, dan Linear sedang diselidiki. Dalam percobaan ini, kami menjalankan sejumlah kombinasi nilai untuk menemukan pengaturan terbaik bagi setiap kernel. Kami menguji nilai gamma 0, 1, dan 10 dalam kernel RBF, dengan gamma 1 menghasilkan hasil terbaik. Berbagai kombinasi nilai diuji dalam kernel Polinomial, dan parameter optimal ditemukan sebagai d=2, =2, dan C=100.Parameter C mengontrol keseimbangan antara margin dan kesalahan klasifikasi dalam model SVM. Rincian hasil uji coba ini tercantum dalam.

Document Accepted 3/7/25

Tabel 5. SVM Akurasi dengan Vektor Fitur

Kernels	Accuracy BoW	Accuracy TF-IDF
RBF	0.6621	0.9774
Polynomial	0.9727	**0.9785**
Linier	0.9640	**0.9780**

Seperti yang ditunjukkan pada Tabel 5, metode konversi fitur TF-IDF menghasilkan akurasi yang lebih tinggi daripada menggunakan BoW. SVM dengan kernel polinomial adalah yang terbaik dalam mengklasifikasikan email phishing. Akurasi klasifikasi email dengan SVM dan kernel *polynomial* menghasilkan akurasi 97,27% untuk fitur BoW dan 97,85% untuk fitur TF-IDF yang mengungguli kernel lainnya.

2. Perbandingan dengan Metode lain

Berdasarkan hasil penelitian sebelumnya, pendekatan konversi fitur yang digunakan adalah TF-IDF, dan dataset dipisahkan menjadi 90% pelatihan dan pengujian 10%. Metode konversi fitur dan skema distribusi himpunan data akan digunakan sebagai kerangka kerja untuk menguji berbagai algoritma kategorisasi yang telah dievaluasi oleh para peneliti sebelumnya. Pengujian ini dilakukan dengan menggunakan himpunan data yang sama seperti yang digunakan dalam penelitian ini *Naive Bayesian* (Nagaraj et al. 2023), *Random Forest* (Taylor dan Ezekiel 2020), *Logistic Regression*, K-NN, dan SVM termasuk di antara metode kategorisasi yang diteliti. Temuan dari rangkaian pengujian ini dirinci.

Tabel 6. Perbandingan beberapa Metode

Methods	Accuracy Test
Naive Bayesian	0.9046
Random Forest	0.9576
SVM RBF	0.9699
Logistic Regression	0.9705
KNN	0.5153
SVM Polynomial + NLP	**0.9785**

Berdasarkan Tabel 6, penggunaan metode SVM dengan kernel Polinomial, bersama dengan pendekatan NLP dan TF-IDF, menghasilkan hasil akurasi tertinggi jika dibandingkan dengan berbagai metode lain yang dipelajari. Hasilnya, dapat disimpulkan bahwa SVM dengan kernel polynomial merupakan pilihan terbaik untuk mengklasifikasikan email phishing. Kami membandingkan berbagai metode machine learning yang ditetapkan oleh akademisi sebelumnya sebagai yang tercanggih dengan akurasi maksimum 97,05%. Penelitian ini memiliki tingkat akurasi yang lebih tinggi dari penelitian sebelumnya yaitu sebesar 97,85%.

Selain itu, penelitian ini tidak membahas penggunaan deep learning karena membutuhkan komputasi yang besar (Kadhim Ali, Mohsin Abdullah, dan Fawzi Raheem 2023). Namun, perlu adanya pengembangan metode yang lebih cepat dan akurat di masa mendatang.

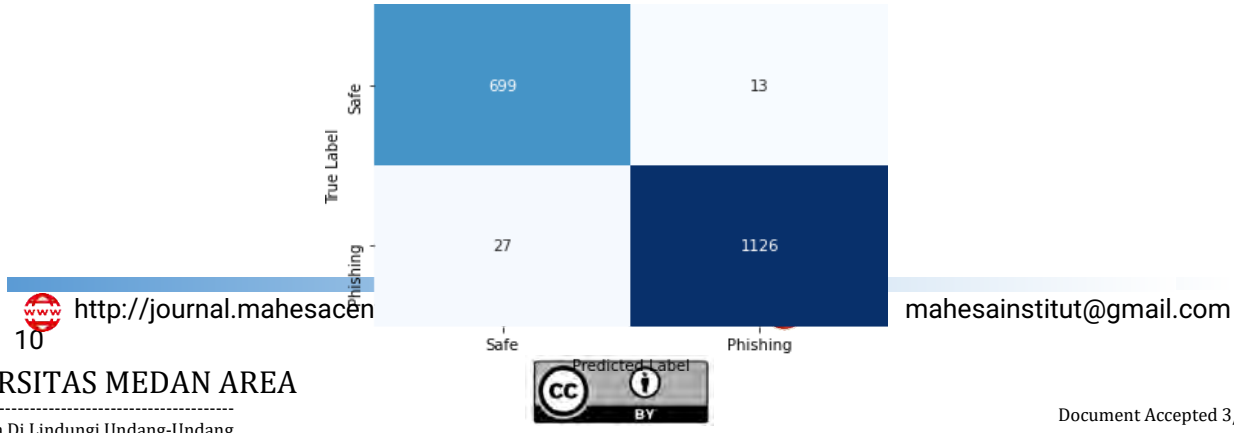
3. Pengaruh NLP pada Klasifikasi

Pemrosesan Bahasa Alami (NLP) sangat penting untuk mengekstraksi fitur tekstual. Kami meneliti efek penggunaan NLP dan ketiadaan NLP pada proses klasifikasi email untuk meneliti dampak penggunaan NLP. Tokenisasi, penghapusan kata-kata umum (stopword), penghapusan tanda baca dan karakter khusus, penggabungan teks yang telah diproses sebelumnya, dan transformasi teks menjadi vektor fitur adalah semua proses dalam proses NLP. Jika NLP tidak ada, proses ini hanya terdiri dari transformasi teks aktual menjadi vektor fitur menggunakan pendekatan TF-IDF. Berikut ini adalah perbandingan jumlah akurasi yang dihasilkan dengan menggunakan NLP dibandingkan dengan tidak menggunakan NLP.

Tabel 7. Pengaruh NLP terhadap Akurasi

Methods	With NLP	Without NLP
Naïve Bayesian	0.9324	0.9046
Random Forest	0.9641	0.9576
Logistic Regression	0.9716	0.9705
KNN	0.4718	0.5153
SVM Polynomial	0.9785	0.9753

Berdasarkan data pada Tabel 7, kita dapat menyimpulkan bahwa penggunaan NLP meningkatkan akurasi. Empat dari lima pengujian yang menggunakan berbagai metode klasifikasi menunjukkan peningkatan kinerja saat NLP digunakan. Berdasarkan hasil eksperimen sebelumnya, SVM dengan kernel Polinomial yang menggunakan NLP dan metode konversi TF-IDF tampaknya menjadi cara yang lebih efektif dalam mengklasifikasikan email phishing. Matriks untuk SVM dengan kernel polynomial dirinci dalam.



Gambar 3. SVM *Polynomial Confusion Matrix*

Seperti yang diilustrasikan pada Gambar 3, 1126 dari 1153 Email yang diidentifikasi sebagai phishing diklasifikasikan dengan benar, sementara 27 email lainnya diklasifikasikan secara salah. Dalam kategori email aman, 699 dari 712 email diklasifikasikan dengan benar, sementara 13 email diklasifikasikan secara salah.

SIMPULAN

Studi ini berfokus pada penilaian dan peningkatan klasifikasi email phishing menggunakan metode SVM dengan berbagai pengaturan dan pendekatan, serta beberapa metodologi tambahan. Temuan studi menunjukkan metode yang sangat baik untuk mendeteksi email phishing dengan mengubah teks email menjadi fitur yang dapat digunakan dalam proses klasifikasi menggunakan NLP. Pekerjaan ini memperoleh akurasi yang tinggi dalam mengkategorikan email phishing dengan mengubah teks email menjadi bentuk vektor fitur menggunakan pendekatan TF-IDF dan menerapkan SVM dengan kernel polinomial, memperoleh akurasi 97,85% dalam kumpulan data yang dievaluasi. Hasil ini menunjukkan manfaat penggunaan SVM dengan kernel polynomial berdasarkan karakteristik teks yang diproses dengan NLP dalam konteks email klasifikasi yang diharapkan dapat mengatasi masalah penipuan email dan meningkatkan keamanan dalam pengiriman informasi melalui email.

UCAPAN TERIMAKASIH (Optional)

Saya mengucapkan terima kasih kepada Universitas Medan Area beserta seluruh jajaran akademik yang telah memberikan bimbingan dan fasilitas dalam proses penelitian ini. Ucapan terima kasih juga disampaikan kepada dosen pembimbing Bapak Dr.Sayuti Rahman ST, M.Kom yang telah memberikan arahan dan masukan berharga dalam penyusunan jurnal ini. Tak lupa, penghargaan yang setinggi-tingginya diberikan kepada keluarga, rekan-rekan, serta semua pihak yang telah memberikan dukungan moral dan teknis dalam berbagai tahap penelitian ini. Semoga penelitian ini dapat memberikan manfaat bagi pengembangan ilmu pengetahuan dan masyarakat luas.

DAFTAR PUSTAKA

Document Accepted 3/7/25

Chudra, Glenny, Theresia Herlina Rochadiani, Handri Santoso, Tommi Poltak Mario, Universitas Pradita, Scientia Bussines Park, Kelapa Dua, dan Tangerang Regency. 2022. "Uji dan analisis kerentanan mahasiswa Universitas X terhadap serangan rekayasa sosial." *Jurnal Inovasi Informatika Universitas Pradita* Volume 7.

Dinata, Rozzi Kesuma, Rizal Tjut Adek, Novia Hasdyna, dan Sujacka Retno. 2023. "K-Nearest Neighbor Classifier Optimization Using Purity." In *AIP Conference Proceedings*, American Institute of Physics Inc. doi:10.1063/5.0117058.

Ding, Xiaojian, Jian Liu, Fan Yang, dan Jie Cao. 2021. "Random radial basis function kernel-based support vector machine." *Journal of the Franklin Institute* 358(18): 10121–10140. doi:10.1016/j.jfranklin.2021.10.005.

Gopi, Arepalli Peda, R. Naga Sravana Jyothi, V. Lakshman Narayana, dan K. Satya Sandeep. 2023. "Classification of tweets data based on polarity using improved RBF kernel of SVM." *International Journal of Information Technology (Singapore)* 15(2): 965–980. doi:10.1007/s41870-019-00409-4.

Gupta, Ayushi, Sushila Palwe, dan Devyani Keskar. 2020. "Fake Email and Spam Detection: User Feedback with Naives Bayesian Approach." In , ,pp. 41–47. doi:10.1007/978-981-15-0790-8_5.

Hussain Sujon, Md. Arman, dan Hossen Mustafa. 2023. "Comparative Study of Machine Learning Models on Multiple Breast Cancer Datasets." *International Journal of Advanced Science Computing and Engineering* 5(1): 15–24. doi:10.62527/ijasce.5.1.105.

Kadhim Ali, Amna, Abdulhussein Mohsin Abdullah, dan Sabreen Fawzi Raheem. 2023. "Impact the Classes' Number on the Convolutional Neural Networks Performance for Image Classification." *International Journal of Advanced Science Computing and Engineering* 5(2): 64–69.

Leonov, Pavel Y., Alexander V. Vorobyev, Anastasia A. Ezhova, Oksana S. Kotelyanets, Aleksandra K. Zavalishina, dan Nikolay V. Morozov. 2021. "The main social engineering techniques aimed at hacking information systems." In *Proceedings - 2021 Ural Symposium on Biomedical Engineering, Radioelectronics and Information Technology, USBEREIT 2021*, , 471–473. doi:10.1109/USBEREIT51232.2021.9455031.

Martin, R. John, dan S. L. Swapna. 2022. "A Machine Learning Framework for Epileptic Seizure Detection by Analyzing EEG Signals." *International Journal of Computing and Digital Systems* 11(1). doi:10.12785/ijcds/1101112.

Nagaraj, P., V. Muneeswaran, G. Shyam Sundar Reddy, V. Bharath Kumar, B. Madhan Mohan, dan Sriram Kumar. 2023. "Automatic Email Spam Classification Using Naïve Bayes." In *2023 International Conference on Computer Communication and Informatics, ICCCI 2023*, , 1–5. doi:10.1109/ICCCI56745.2023.10128233.

Noble, William S. 2006. 24 Nature Biotechnology *What is a support vector machine?* doi:10.1038/nbt1206-1565.

Olivetti, Elsa A., Jacqueline M. Cole, Edward Kim, Olga Kononova, Gerbrand Ceder, Thomas Yong Jin Han, dan Anna M. Hiszpanski. 2020. "Data-driven materials research enabled by natural language processing and information extraction." *Applied Physics Reviews* 7(4). doi:10.1063/5.0021106.

Pamungkas, Daniel Sutopo, Sumantri K Risandriya, dan Adam Rahman. 2022. "Classification of Finger Movements Using EMG Signals with PSO SVM Algorithm." *International Journal of Advanced Science Computing and*

Engineering 4(3): pp 210-219. doi:10.30630/ijasce.4.3.100.

Rahman, S., M. Ramli, F. Arnia, R. Muharar, M. Ikhwan, dan S. Munzir. 2022. "Enhancement of convolutional neural network for urban environment parking space classification." *Global Journal of Environmental Science and Management* 8(3): 315–326. doi:10.22034/gjesm.2022.03.02.

Rahman, Sayuti, Marwan Ramli, Fitri Arnia, Arnes Sembiring, dan Rusdha Muharar. 2020. "Convolutional Neural Network Customization for Parking Occupancy Detection." In *Proceedings of the International Conference on Electrical Engineering and Informatics*, , 1–6. doi:10.1109/ICELTICs50595.2020.9315509.

Salloum, Said, Tarek Gaber, Sunil Vadera, dan Khaled Shaalan. 2022. "A Systematic Literature Review on Phishing Email Detection Using Natural Language Processing Techniques." *IEEE Access* 10. doi:10.1109/ACCESS.2022.3183083.

Simoiu, Camelia, Ali Zand, Kurt Thomas, dan Elie Bursztein. 2020. "Who is targeted by email-based phishing and malware?: Measuring factors that differentiate risk." In *Proceedings of the ACM SIGCOMM Internet Measurement Conference, IMC*, , 567–576. doi:10.1145/3419394.3423617.

Syahputra, Wahyu, Yussa Ananda, dan Lisa Andriana Siregar. 2022. "Perancangan Sistem Keamanan Brankas Bertingkat Menggunakan KTP Elektronik Dan Verifikasi Smartphone." *Seminar Nasional Teknik (SEMNASTEK) UISU*.

Taylor, O E, dan P S Ezekiel. 2020. 6 International Journal of Computer Science and Mathematical Theory E *A Model to Detect Spam Email Using Support Vector Classifier and Random Forest Classifier*.

Vanneschi, Leonardo, dan Sara Silva. 2023. "Support Vector Machines." In *Natural Computing Series*, , pp.207–235. doi:10.1007/978-3-031-17922-8_10.

Webb, Geoffrey I. 2020. "Encyclopedia of Machine Learning and Data Science." *Encyclopedia of Machine Learning and Data Science* (April). doi:10.1007/978-1-4899-7502-7.

Yacouby, Reda, dan Dustin Axman. 2020. "Probabilistic Extension of Precision, Recall, and F1 Score for More Thorough Evaluation of Classification Models." In , 79– 91. doi:10.18653/v1/2020.eval4nlp-1.9.

Zhou, Ding Xuan, dan Kurt Jetter. 2006. "Approximation with polynomial kernels and SVM classifiers." *Advances in Computational Mathematics* 25(1–3): 323–344. doi:10.1007/s10444-004-7206-2.

Document Accepted 3/7/25