INCODING: Journal of Informatic and Computer Science Engineering
https://journal.mahesacenter.org/index.php/incoding/index || ISSN xxxx-xxxx (online)
5 (1) 2025: 38-50
DOI: 10.34007/incoding.v5i1.831
Genesis:
Disubmit: 16 April 2025
Direview: 30 April 2025
Diterima: 30 April 2025

Meningkatkan Deteksi Email Phising Melalui Pendekatan SVM yang Dioptimalkan NLP

Enhancing Phishing Email Detection through NLP-Optimized SVM Approach

Rino Nurcahyo Fauzi Tanjung & Sayuti Rahman
Teknik Informatika, Fakultas Teknik, Universitas Medan Area, Indonesia

*Coresponding Email: rinonurcahyofauzitanjung@gmail.com

Abstrak

Serangan email phishing menjadi ancaman serius dalam ekosistem digital karena mampu mengecoh pengguna untuk membocorkan informasi sensitif atau mengakses tautan berbahaya. Penelitian ini bertujuan mengembangkan model klasifikasi email phishing berbasis algoritma Support Vector Machine (SVM) yang dikombinasikan dengan teknik Natural Language Processing (NLP) untuk meningkatkan akurasi deteksi. Proses dimulai dengan tahap tokenisasi, pembersihan teks, dan ekstraksi fitur menggunakan pendekatan TF-IDF, yang selanjutnya digunakan sebagai input ke dalam model klasifikasi. Berbagai kernel SVM, termasuk linear, radial basis function (RBF), dan polynomial diuji melalui metode grid search dengan penyetelan parameter seperti C, gamma, dan degree. Hasil penelitian menunjukkan bahwa SVM dengan kernel polynomial menghasilkan akurasi tertinggi sebesar 97,85%, melampaui algoritma lain seperti Naïve Bayes, Random Forest, dan Logistic Regression. Temuan ini mengindikasikan bahwa integrasi NLP dan SVM dengan tuning parameter yang tepat memberikan solusi efektif dalam mitigasi serangan email phishing. Model ini dapat menjadi fondasi dalam pengembangan sistem keamanan siber yang lebih adaptif dan efisien.
Kata Kunci: Phishing Email; Natural Language Processing; Support Vector Machine; TF-IDF.

Abstract

Phishing email attacks are a serious threat in the digital ecosystem because they can trick users into leaking sensitive information or accessing malicious links. This study aims to develop a phishing email classification model based on the Support Vector Machine (SVM) algorithm combined with Natural Language Processing (NLP) techniques to improve detection accuracy. The process begins with the tokenization, text cleansing, and feature extraction stages using the TF-IDF approach, which is further used as input into the classification model. Various SVM kernels, including linear, radial basis function (RBF), and polynomial, are tested through the grid search method with parameter tuning such as C, gamma, and degree. The results showed that SVMs with polynomial kernels produced the highest accuracy of 97.85%, surpassing other algorithms such as Naïve Bayes, Random Forest, and Logistic Regression. These findings indicate that the integration of NLP and SVM with proper parameter tuning provides an effective solution in mitigating phishing email attacks. This model can be the foundation for the development of a more adaptive and efficient cybersecurity system.
Keywords: *Phishing Email; Natural Language Processing; Support Vector Machine; TF-IDF.*

PENDAHULUAN

Dalam era digital saat ini, email merupakan salah satu sarana komunikasi utama dalam aktivitas pribadi, profesional, maupun bisnis. Namun, peningkatan penggunaan email juga diiringi dengan meningkatnya ancaman keamanan, terutama dalam bentuk serangan phishing [1], [2]. Phishing melalui email telah menjadi salah satu metode serangan siber yang paling umum dan merugikan, karena mampu menipu pengguna untuk mengungkapkan informasi sensitif atau mengklik tautan berbahaya. Studi menunjukkan bahwa email phishing dapat menyebabkan berbagai dampak serius, seperti pencurian data pribadi, penipuan keuangan, penyebaran malware, serangan spear phishing, serta ransomware (Leonov et al. 2021; Simoiu et al. 2020), Bahaya ini menjadi lebih signifikan ketika serangan phishing menyasar lembaga pemerintahan, institusi keuangan, atau organisasi layanan publik, di mana dampaknya dapat bersifat sistemik dan berdampak luas.

Email phishing umumnya menyamar sebagai pesan dari entitas yang sah, dengan tujuan mengelabui pengguna agar melakukan tindakan tertentu, seperti mengklik tautan palsu atau mengunduh lampiran berbahaya. Pengguna yang kurang waspada dan tidak memiliki pemahaman mengenai tanda-tanda penipuan digital cenderung menjadi korban [4]. Oleh karena itu, perlindungan terhadap email phishing bukan hanya persoalan teknis, tetapi juga merupakan kebutuhan mendesak dalam membangun sistem keamanan informasi yang holistik.

Berbagai pendekatan telah dikembangkan untuk mendeteksi dan menangkal email phishing. Salah satu metode dasar adalah dengan verifikasi keabsahan email, misalnya melalui pemeriksaan alamat pengirim, analisis domain pengirim, validasi tautan dan lampiran, serta verifikasi identitas institusi melalui jalur komunikasi resmi [5]. Langkah-langkah tersebut cukup efektif untuk pengguna yang memiliki tingkat kesadaran digital tinggi, tetapi kurang dapat diandalkan untuk pengguna awam yang mudah tertipu oleh format dan bahasa email yang menyerupai entitas resmi.

Dalam beberapa tahun terakhir, pendekatan otomatis berbasis teknologi mulai banyak digunakan, termasuk penerapan algoritma machine learning (ML) untuk klasifikasi email. Algoritma seperti Naïve Bayes, Support Vector Machine (SVM), Decision Tree, Random Forest, dan K-Nearest Neighbors (KNN) telah banyak diteliti dan diimplementasikan dalam deteksi phishing (Gupta, Palwe, dan Keskar 2020; Dinata et al.

2023). Masing-masing algoritma memiliki kelebihan dan keterbatasan, tergantung pada karakteristik dataset, kompleksitas fitur, serta volume data.

Salah satu metode yang menonjol adalah Support Vector Machine (SVM), yang dikenal mampu mengatasi permasalahan data berdimensi tinggi dan tidak seimbang [7], [8]. SVM bekerja dengan mencari hyperplane optimal yang memisahkan kelas data secara maksimal. Dengan memanfaatkan kernel functions seperti Radial Basis Function (RBF) dan polynomial kernel, SVM dapat memproyeksikan data ke ruang berdimensi lebih tinggi, sehingga pola kompleks dalam email phishing dapat dikenali dengan lebih baik (Ding et al. 2021; Gopi et al. 2023; Zhou dan Jetter 2006).

Selain performanya yang baik dalam skenario dengan jumlah fitur besar, SVM juga menunjukkan efisiensi yang tinggi dalam situasi data berjumlah terbatas, yang kerap terjadi pada studi phishing di sektor spesifik. Lebih lanjut, SVM dinilai lebih cepat dalam proses klasifikasi dibandingkan pendekatan deep learning yang membutuhkan waktu pelatihan dan komputasi yang lebih besar [10].

Untuk meningkatkan akurasi klasifikasi teks email, penggunaan Natural Language Processing (NLP) juga telah diintegrasikan. NLP memungkinkan sistem untuk mengekstraksi informasi penting dari konten email seperti kata kunci, entitas bernama, dan frasa bermakna, yang kemudian dapat digunakan sebagai fitur dalam model klasifikasi [11].

Meskipun banyak penelitian telah mengadopsi algoritma machine learning untuk klasifikasi email phishing, beberapa kendala masih ditemukan dalam hal optimalisasi kombinasi fitur dan parameter algoritmik, terutama pada SVM. Sebagian besar studi cenderung menggunakan parameter default atau metode tuning sederhana, tanpa eksplorasi sistematis terhadap pengaruh nilai C, gamma, dan jenis kernel terhadap performa model.

Selain itu, integrasi antara teknik NLP untuk representasi teks optimal dan tuning algoritma SVM belum banyak dikaji secara mendalam dalam konteks klasifikasi email phishing. Padahal, representasi teks yang kaya dapat meningkatkan kualitas input bagi algoritma klasifikasi, dan secara signifikan berdampak pada akurasi serta kemampuan generalisasi model. Kelemahan lainnya adalah masih kurangnya eksplorasi mengenai ketidakseimbangan data antara email phishing dan non-phishing yang dapat

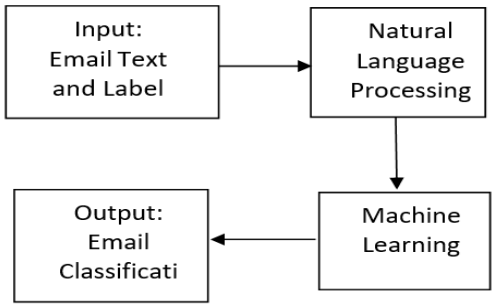
mengakibatkan bias klasifikasi pada model yang dilatih dengan distribusi data tidak proporsional.

Dengan demikian, terdapat kebutuhan akan penelitian yang secara eksplisit mengkaji bagaimana kombinasi fitur berbasis NLP dan penyetelan hiperparameter pada SVM dapat digunakan untuk meningkatkan performa deteksi email phishing, terutama dalam skenario data terbatas dan distribusi yang tidak merata.

Penelitian ini bertujuan untuk mengembangkan model klasifikasi email phishing yang lebih akurat dan efisien dengan memanfaatkan algoritma Support Vector Machine (SVM) yang dikombinasikan dengan teknik Natural Language Processing (NLP) untuk representasi fitur teks. Secara khusus, penelitian ini akan mengevaluasi pengaruh penggunaan berbagai jenis kernel dan nilai parameter seperti C dan gamma terhadap akurasi klasifikasi, serta menguji efektivitas kombinasi fitur linguistik hasil ekstraksi NLP dalam proses kategorisasi email.

METODE PENELITIAN

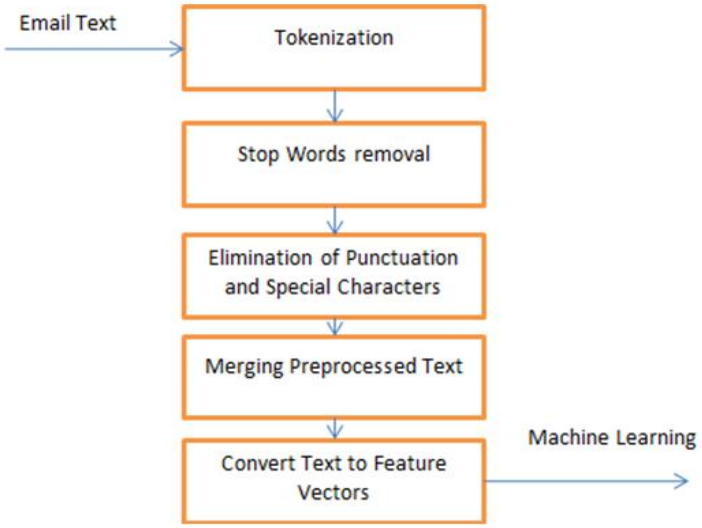
Metode *Natural Language Processing* (NLP) digunakan sebagai pemroses email teks dalam penelitian ini, dan hasilnya dikenali oleh pendekatan pembelajaran mesin. Ada banyak pendekatan untuk menentukan metode terbaik dalam mengklasifikasikan email palsu. Metode-metode ini meliputi SVM, *naive bayes*, dan *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 [12].

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 [13]. SVM bekerja dengan mencari hyperplane optimal yang memisahkan kelas data [14],[15]. 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:

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

Dimana:

$K(x, x')$ 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 [16]. 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 [17].

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 reading.	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

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.

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.

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* [18], *Random Forest* [19], *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 [20]. Namun, perlu adanya pengembangan metode yang lebih cepat dan akurat di masa mendatang.

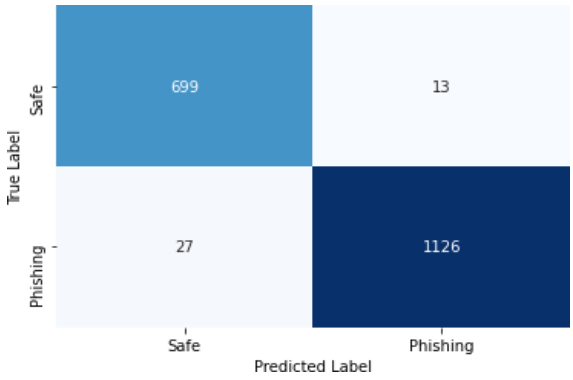
3. Pengaruh NLP pada Klasifikasi

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	0.9573
Logistic Regression	0.9726	0.9667
KNN	0.4761	0.5144
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.

DAFTAR PUSTAKA

[1] J. Lynch, "Identity theft in cyberspace: Crime control methods and their effectiveness in combating phishing attacks," *Berkeley Tech. LJ*, vol. 20, p. 259, 2005.

[2] D. Fatmala Putri and W. Ratna Sari, "Analisis Perlindungan Nasabah BSI Terhadap Kebocoran Data Dalam Menggunakan Digital Banking," *J. Ilm. Ekon. dan Manaj.*, vol. 1, no. 4, pp. 173–181, 2023, [Online]. Available: https://doi.org/10.61722/jiem.v1i4.331

[3] P. Y. Leonov, A. V. Vorobyev, A. A. Ezhova, O. S. Kotelyanets, A. K. Zavalishina, and N. V. Morozov, "The main social engineering techniques aimed at hacking information systems," in *Proceedings* -

2021 Ural Symposium on Biomedical Engineering, Radioelectronics and Information Technology, USBEREIT 2021, 2021, pp. 471–473. doi: 10.1109/USBEREIT51232.2021.9455031.

[4] W. Syahputra, Y. Ananda, and L. A. Siregar, "Perancangan Sistem Keamanan Brankas Bertingkat Menggunakan KTP Elektronik Dan Verifikasi Smartphone," *Semin. Nas. Tek. UISU*, 2022.

[5] G. Chudra *et al.*, "Uji dan analisis kerentanan mahasiswa Universitas X terhadap serangan rekayasa sosial," *J. Inov. Inform. Univ. Pradita*, vol. Volume 7, 2022.

[6] A. Gupta, S. Palwe, and D. Keskar, "Fake Email and Spam Detection: User Feedback with Naives Bayesian Approach," 2020, p. ,pp. 41–47. doi: 10.1007/978-981-15-0790-8_5.

[7] R. Tjut Adek, M. Fikry, and U. Khalil, "News Opinion Classification Application With Support Vector Machine Algorithm Using Framework Codeigniter," *J. Informatics Telecommun. Eng.*, vol. 5, no. 1, pp. 160–166, 2021, doi: 10.31289/jite.v5i1.5189.

[8] D.- Andriansyah and Eka Wulansari Fridayanthie, "Optimization of Support Vector Machine and XGBoost Methods Using Feature Selection to Improve Classification Performance," *J. Informatics Telecommun. Eng.*, vol. 6, no. 2, pp. 484–493, 2023, doi: 10.31289/jite.v6i2.8373.

[9] X. Ding, J. Liu, F. Yang, and J. Cao, "Random radial basis function kernel-based support vector machine," *J. Franklin Inst.*, vol. 358, no. 18, pp. 10121–10140, 2021, doi: 10.1016/j.jfranklin.2021.10.005.

[10] S. Rahman, M. Ramli, F. Arnia, A. Sembiring, and R. Muharar, "Convolutional Neural Network Customization for Parking Occupancy Detection," in *Proceedings of the International Conference on Electrical Engineering and Informatics*, 2020, pp. 1–6. doi: 10.1109/ICELTICs50595.2020.9315509.

[11] E. A. Olivetti *et al.*, "Data-driven materials research enabled by natural language processing and information extraction," *Applied Physics Reviews*, vol. 7, no. 4. 2020. doi: 10.1063/5.0021106.

[12] M. A. Hussain Sujon and H. Mustafa, "Comparative Study of Machine Learning Models on Multiple Breast Cancer Datasets," *Int. J. Adv. Sci. Comput. Eng.*, vol. 5, no. 1, pp. 15–24, 2023, doi: 10.62527/ijasce.5.1.105.

[13] D. S. Pamungkas, S. K. Risandriya, and A. Rahman, "Classification of Finger Movements Using EMG Signals with PSO SVM Algorithm," *Int. J. Adv. Sci. Comput. Eng.*, vol. 4, no. 3, p. pp 210-219, 2022, doi: 10.30630/ijasce.4.3.100.

[14] W. S. Noble, *What is a support vector machine?*, vol. 24, no. 12. 2006. doi: 10.1038/nbt1206-1565.

[15] L. Vanneschi and S. Silva, "Support Vector Machines," in *Natural Computing Series*, 2023, p. pp.207–235. doi: 10.1007/978-3-031-17922-8_10.

[16] G. I. Webb, "Encyclopedia of Machine Learning and Data Science," *Encycl. Mach. Learn. Data Sci.*, no. April, 2020, doi: 10.1007/978-1-4899-7502-7.

[17] R. Yacouby and D. Axman, "Probabilistic Extension of Precision, Recall, and F1 Score for More Thorough Evaluation of Classification Models," 2020, pp. 79–91. doi: 10.18653/v1/2020.eval4nlp-1.9.

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

[19] O. E. Taylor and P. S. Ezekiel, "A Model to Detect Spam Email Using Support Vector Classifier and Random Forest Classifier," 2020.

[20] A. Kadhim Ali, A. Mohsin Abdullah, and S. Fawzi Raheem, "Impact the Classes' Number on the Convolutional Neural Networks Performance for Image Classification," *Int. J. Adv. Sci. Comput. Eng.*, vol. 5, no. 2, pp. 64–69, 2023.