E-ISSN : 2988-1986
https://ejournal.warunayama.org/kohesi

Kohesi: Jurnal Multidisiplin Saintek
Volume 5 No 4 Tahun 2024

ANALISIS KINERJA GEMINI 1.5 PRO DALAM VALIDASI DAN DETEKSI EMAIL PHISHING

Moh Sulthan Arief Rahmatullah[1], Andyana Muhandhatul Nabila[2], Atha Rahma Arianti[3]
Departemen Teknologi Informasi, Fakultas Teknologi Elektro dan Informatika Cerdas,
Institut Teknologi Sepuluh Nopember Surabaya
Email: r.sulthanar@gmail.com[1], aa.andyana823@gmail.com[2], atharahmaarianti@gmail.com[3]

Abstrak

Email phishing merupakan ancaman siber yang terus berkembang dan menuntut solusi deteksi yang lebih efektif. Penelitian ini menganalisis kinerja model bahasa besar Gemini 1.5 Pro dalam validasi dan deteksi email phishing. Sebuah dataset yang terdiri dari 200 email berlabel, dengan komposisi seimbang antara 100 email phishing dan 100 email sah, digunakan untuk mengevaluasi kemampuan Gemini 1.5 Pro. Hasil eksperimen menunjukkan performa yang menggembirakan, dengan akurasi mencapai 93% dalam membedakan email phishing dan email sah. Analisis lebih lanjut menggunakan metrik presisi, recall, dan F1-score mengungkap kemampuan model dalam meminimalisir false positive dan false negative. Laporan klasifikasi menunjukkan presisi sempurna (1.00) untuk kategori email phishing, meskipun recall-nya mencapai 0.79, sementara untuk email sah, recall sempurna (1.00) dicapai dengan presisi 0.90. Matriks konfusi memberikan gambaran detail tentang kinerja klasifikasi, dengan 52 email phishing terklasifikasi benar, 14 salah diklasifikasikan sebagai sah, dan seluruh 124 email sah terklasifikasi dengan benar. Temuan ini mengindikasikan potensi Gemini 1.5 Pro sebagai alat yang efektif dalam mendeteksi email phishing dan membuka peluang riset lebih lanjut untuk mengoptimalkan kinerjanya serta mengeksplorasi aplikasinya dalam sistem keamanan siber yang lebih komprehensif.

Kata kunci : Phishing, Ancaman Siber, Gemini 1.5 Pro, Email

Abstract

Phishing emails are a growing cyber threat that demands more effective detection solutions. This study analyzes the performance of the Gemini 1.5 Pro large language model in the validation and detection of phishing emails. A dataset of 200 labeled emails, with a balanced composition of 100 phishing emails and 100 legitimate emails, is used to evaluate the capabilities of Gemini 1.5 Pro. The experimental results showed an encouraging performance, with an accuracy of 93% in distinguishing between phishing and legitimate emails. Further analysis using precision, recall, and F1-score metrics revealed the model's ability to minimize false positives and false negatives. The classification report showed perfect precision (1.00) for the phishing email category, although the recall was

Article History
Received: November 2024
Reviewed: November 2024
Published: November 2024

Plagirism Checker No 234
Prefix DOI : Prefix DOI : 10.8734/Kohesi.v1i2.365

Publish by : Kohesi

0.79, while for legitimate emails, perfect recall (1.00) was achieved with a precision of 0.90. The confusion matrix provides a detailed overview of the classification performance, with 52 phishing emails correctly classified, 14 incorrectly classified as legitimate, and all 124 legitimate emails correctly classified. These findings indicate the potential of Gemini 1.5 Pro as an effective tool in detecting phishing emails and open up further research opportunities to optimize its performance and explore its application in a more comprehensive cybersecurity system.
Keywords : *Phishing, Cyber Threat, Gemini 1.5 Pro, Email*

PENDAHULUAN

Phishing merupakan ancaman siber yang terus berkembang, memanfaatkan email palsu untuk mencuri informasi sensitif seperti kredensial login, data finansial, dan data pribadi. Teknik phishing semakin canggih, menyulitkan pengguna untuk membedakan email palsu dari email yang sah. Konsekuensi dari serangan phishing dapat merugikan individu maupun organisasi, mulai dari kerugian finansial hingga kerusakan reputasi. Oleh karena itu, pengembangan metode deteksi phishing yang efektif dan akurat menjadi krusial.

Penelitian ini ini bertujuan untuk menganalisis kinerja Gemini 1.5 Pro, sebuah model bahasa besar yang dikembangkan oleh Google AI, dalam konteks deteksi email phishing. Gemini 1.5 Pro memiliki kemampuan pemahaman bahasa alami yang kuat dan telah menunjukkan potensi dalam berbagai tugas pemrosesan teks, termasuk klasifikasi teks. Pemanfaatan model bahasa besar seperti Gemini 1.5 Pro untuk deteksi phishing diharapkan dapat meningkatkan akurasi dan efisiensi dalam mengidentifikasi email berbahaya. Dalam penelitian ini, Gemini 1.5 Pro akan dievaluasi kinerjanya dalam membedakan email phishing dari email yang sah. Fokus penelitian ini adalah mengukur kemampuan Gemini 1.5 Pro dalam memahami karakteristik linguistik dan struktural email phishing.

KAJIAN PUSTAKA

Phishing

Phishing adalah metode penipuan yang mengecoh orang untuk melakukan hal-hal yang dapat mengekspos informasi pribadi mereka. Metode ini memanipulasi pengguna untuk mengklik tautan yang mengarahkan mereka ke situs web palsu atau mengunduh perangkat lunak yang dapat membahayakan komputer pribadi. Pada tahun 2006, peretas di Amerika Serikat menggunakan email untuk menciptakan "umpan" yang ditujukan kepada pengguna untuk mendapatkan nama pengguna dan PIN akun American Online. Sejak saat itu, metode phishing semakin berkembang, membuatnya sulit untuk dideteksi sebagai email palsu. [4]

Email Phishing

Dalam beberapa tahun terakhir, banyak orang telah menerima spam dan email phishing. Email phishing dapat meniru struktur email promosi dan iklan dari berbagai perusahaan, yang merusak kredibilitas mereka. Alasan mengapa email pribadi menerima begitu banyak email phishing adalah karena kebanyakan orang menggunakan email pribadi untuk masuk ke berbagai situs web, yang membuat URL email mudah diperoleh dan dijadikan target. Para *spammer*

membeli alamat email secara massal dari penyedia tertentu dan menambahkannya ke dalam daftar pengiriman mereka. [1]

Model LLM (Large Language Model)

Large language models(LLMs) telah merevolusi natural language processing (NLP), secara signifikan meningkatkan akurasi pencarian informasi [2]. Model semacam ini kini mampu menyelesaikan tugas-tugas baru tanpa pelatihan khusus untuk tugas tersebut, biasanya hanya berdasarkan deskripsi teks dalam bahasa alami (natural language). Kemampuan LLM untuk memahami dan menghasilkan teks, serta konten terkait seperti kode program, kini telah mencapai tingkat yang hampir menyerupai manusia.[3]

Pemrosesan Bahasa Alami (NLP)

Rekayasa perangkat lunak, kecerdasan buatan, dan semantik semuanya bergabung dalam bidang Pemrosesan Bahasa Alami (NLP), yang merupakan salah satu disiplin ilmu yang kuat. NLP berfokus pada cara komputer dan bahasa manusia dapat bekerja sama untuk memahami, menafsirkan, dan menghasilkan tulisan atau ucapan yang mirip manusia. Teknik NLP telah berkembang dengan cepat, menjadi sangat penting dalam berbagai aplikasi, terutama dalam meningkatkan interaksi manusia dengan komputer. Kemajuan AI seperti robotika dan arsitektur otak yang merujuk pada kombinasi miliaran koneksi antar neuron di otak, telah berkontribusi pada pertumbuhan NLP. [5]

Keamanan Siber

Keamanan siber adalah upaya untuk melindungi ruang virtual dari serangan siber. Pelanggaran dalam keamanan siber dapat menyebabkan kerugian finansial maupun non-finansial bagi organisasi korban dan kliennya. Oleh karena itu, tujuan keamanan siber adalah melindungi dari pelanggaran-pelanggaran tersebut. [7]

AI dalam Keamanan Siber

AI dapat memperkuat keamanan siber dengan memberikan metodologi baru dalam deteksi, pencegahan, dan respons terhadap ancaman. Dengan menggunakan algoritma *machine learning*, AI dapat menganalisis kumpulan data besar dan dengan cepat menganalisis tren untuk membantu organisasi mengantisipasi metode serangan baru yang mungkin terjadi. Selain itu, alat-alat berbasis AI juga dapat mengotomatisasi perlindungan terhadap ancaman siber dan memperkuat seluruh rencana keamanan dengan menggunakan analisis perilaku lanjutan dan deteksi ancaman prediktif. [6]

METODE PENELITIAN

Penelitian ini mengadopsi pendekatan eksperimental untuk menganalisis kinerja Gemini 1.5 Pro dalam mendeteksi email phishing tanpa melakukan fine-tuning. Alur penelitian meliputi beberapa tahapan, yaitu persiapan data, pemrosesan teks, penggunaan Gemini 1.5 Pro, dan evaluasi kinerja model. Penting untuk dicatat bahwa Gemini 1.5 Pro digunakan as-is tanpa modifikasi atau pelatihan tambahan pada dataset spesifik ini, sehingga penelitian ini mengevaluasi kemampuan bawaan model.

Data:

Data yang digunakan dalam penelitian ini terdiri dari 200 email, yang dibagi secara seimbang menjadi 100 email phishing dan 100 email sah. Pengumpulan email phishing dilakukan melalui beberapa sumber daring terpercaya, termasuk situs web PhishTank dan Open Phish, serta repository publik yang menyediakan sampel email phishing. Sementara itu, email sah dikumpulkan dari kotak masuk email pribadi dengan memastikan tidak adanya informasi sensitif dan privasi yang disertakan. Seluruh email diberi label secara manual, "phishing" untuk email phishing dan "sah" untuk email sah, untuk memfasilitasi proses evaluasi model. Dataset kemudian dibagi menjadi dua bagian: 80% untuk data latih dan 20% untuk data uji. Meskipun Gemini 1.5 Pro tidak difine-tuning, pembagian data ini tetap dilakukan untuk konsistensi metodologi dan memungkinkan perbandingan dengan penelitian di masa mendatang yang mungkin melibatkan fine-tuning.

Proses Validasi dan Evaluasi Kinerja Model (tanpa Fine-tuning):

1. Preprocessing Teks: Tahap preprocessing merupakan langkah krusial dalam mempersiapkan data email sebelum diumpankan ke Gemini 1.5 Pro. Proses ini meliputi beberapa teknik, antara lain: penghapusan karakter khusus (tanda baca, simbol, angka), konversi seluruh teks menjadi huruf kecil untuk penyeragaman, dan penghilangan stopwords (kata-kata umum seperti "yang", "dan", "atau") yang dianggap kurang informatif dalam proses klasifikasi. Preprocessing bertujuan untuk mengurangi noise dalam data dan memberikan input yang lebih bersih ke Gemini 1.5 Pro.
2. Inferensi dengan Gemini 1.5 Pro: Email yang telah melalui tahap preprocessing dikonversi menjadi format yang sesuai dan diumpankan ke Gemini 1.5 Pro untuk inferensi (prediksi) melalui API yang disediakan. Penting untuk digaris bawahi bahwa model tidak dilatih pada data ini, melainkan langsung digunakan untuk memprediksi label ("phishing" atau "sah") berdasarkan pengetahuan yang sudah dimiliki model.
3. Evaluasi Kinerja: Kinerja Gemini 1.5 Pro dalam mendeteksi email phishing tanpa fine-tuning dievaluasi menggunakan berbagai metrik standar, yaitu:
   Akurasi: Mengukur persentase keseluruhan email yang diklasifikasikan dengan benar.
   Presisi: Mengukur ketepatan model dalam mengidentifikasi email phishing.
   Recall: Mengukur kemampuan model dalam menemukan semua email phishing yang ada.
   F1-score: Rata-rata harmonik presisi dan recall.
   Matriks Konfusi: Visualisasi kinerja model dengan menunjukkan jumlah true positive, true negative, false positive, dan false negative.

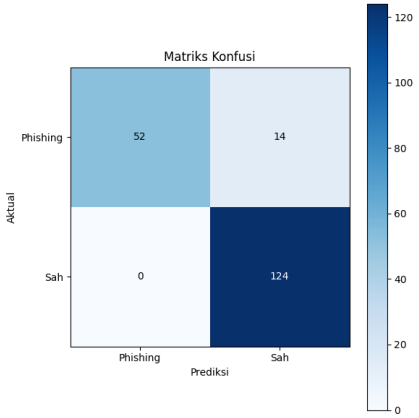
Proses evaluasi dilakukan menggunakan data uji. Hasil evaluasi kemudian dianalisis untuk menentukan efektivitas Gemini 1.5 Pro tanpa fine-tuning dalam mendeteksi email phishing dan memberikan wawasan tentang potensi model untuk tugas klasifikasi teks jenis ini.

HASIL DAN PEMBAHASAN

Penelitian ini mengevaluasi kinerja Gemini 1.5 Pro dalam mendeteksi email phishing tanpa fine-tuning. Hasil evaluasi ditampilkan pada Tabel 1 dan Matriks Konfusi pada Gambar 1.

Tabel 1. Hasil Evaluasi Kinerja Gemini 1.5 Pro

Metrik	Nilai
Akurasi	93%
Presisi	1.00 (Phishing), 0.90 (Sah)
Recall	0.79 (Phishing), 1.00 (Sah)
F1-score	0.88 (Phishing), 0.95 (Sah)



Gambar 1. Matriks Konfusi

Hasil menunjukkan bahwa Gemini 1.5 Pro, tanpa fine-tuning, mencapai akurasi yang tinggi sebesar 93% dalam mengklasifikasikan email phishing dan sah. Hal ini menunjukkan potensi kemampuan generalisasi model yang baik, meskipun tidak dilatih secara spesifik pada dataset ini. Presisi untuk kategori "Sah" mencapai 0.90, menandakan bahwa dari seluruh email yang diprediksi sebagai sah, 90% di antaranya memang benar email sah. Recall sempurna (1.00) untuk kategori "Sah" mengindikasikan bahwa model berhasil mengidentifikasi semua email sah dalam dataset uji.

Namun, terdapat perbedaan kinerja yang cukup signifikan pada kategori "Phishing". Meskipun presisi sempurna (1.00) menunjukkan bahwa semua email yang diprediksi phishing memang benar phishing, recall hanya mencapai 0.79. Ini berarti terdapat 21% email phishing yang tidak terdeteksi oleh model dan salah diklasifikasikan sebagai email sah (false negative). Hal ini terlihat pada matriks konfusi, di mana terdapat 14 email phishing yang salah diklasifikasikan. Analisis lebih lanjut terhadap email-email yang salah klasifikasi ini diperlukan untuk memahami pola kesalahan dan potensi kelemahan model.

Perlu dicatat bahwa penelitian ini tidak melakukan fine-tuning pada Gemini 1.5 Pro. Penelitian sebelumnya yang menggunakan model bahasa besar lain dengan fine-tuning pada dataset spesifik phishing telah melaporkan hasil akurasi yang lebih tinggi. Perbandingan langsung dengan penelitian tersebut agak sulit dilakukan karena perbedaan metodologi dan dataset. Namun, hasil ini tetap menunjukkan bahwa Gemini 1.5 Pro memiliki potensi yang menjanjikan dalam deteksi phishing, bahkan tanpa fine-tuning.

Selanjutnya, analisis kesalahan menunjukkan bahwa kesalahan klasifikasi terutama terjadi pada email phishing yang menggunakan teknik yang lebih canggih, seperti spoofing domain yang sangat mirip atau teknik social engineering yang kompleks. Hal ini

menunjukkan bahwa peningkatan kinerja dapat dicapai melalui fine-tuning model dengan dataset yang lebih besar dan beragam, serta dengan mengembangkan teknik preprocessing yang lebih robust untuk mengidentifikasi ciri-ciri khusus email phishing.

KESIMPULAN

Penelitian ini menyimpulkan bahwa Gemini 1.5 Pro, tanpa fine-tuning, menunjukkan potensi yang menjanjikan dalam mendeteksi email phishing. Model mencapai akurasi 93% dalam mengklasifikasikan email phishing dan sah pada dataset yang digunakan. Meskipun presisi untuk kategori phishing sempurna (1.00), recall yang lebih rendah (0.79) mengindikasikan perlunya peningkatan dalam mendeteksi semua email phishing. Analisis kesalahan menunjukkan bahwa model kesulitan dalam mengidentifikasi email phishing yang menggunakan teknik yang lebih canggih.

Berdasarkan hasil penelitian ini, beberapa saran untuk penelitian selanjutnya adalah:

1. Fine-tuning Gemini 1.5 Pro: Melakukan fine-tuning model dengan dataset email phishing yang lebih besar dan beragam diharapkan dapat meningkatkan kinerja, terutama nilai recall untuk kategori phishing. Dataset yang lebih representatif terhadap berbagai teknik phishing akan membantu model mempelajari pola dan karakteristik email phishing secara lebih komprehensif.
2. Eksplorasi Teknik Preprocessing: Menyelidiki dan mengembangkan teknik preprocessing teks yang lebih maju, seperti penggunaan stemming, lemmatization, atau teknik lain yang dapat mengekstrak fitur yang lebih relevan dari email, dapat meningkatkan akurasi klasifikasi.
3. Analisis Mendalam Kesalahan Klasifikasi: Melakukan analisis lebih lanjut terhadap email-email yang salah klasifikasi oleh model. Analisis ini dapat memberikan pemahaman yang lebih baik tentang kelemahan model dan membantu dalam pengembangan strategi untuk mengatasi kelemahan tersebut.
4. Pengujian pada Dataset yang Berbeda: Menguji kinerja model pada dataset yang berbeda dan independen akan memberikan gambaran yang lebih luas tentang generalisasi dan keandalan model.
5. Integrasi dengan Sistem Keamanan: Mengeksplorasi kemungkinan integrasi Gemini 1.5 Pro dengan sistem keamanan email yang ada. Hal ini dapat membantu dalam membangun sistem deteksi phishing yang lebih robust dan real-time.

DAFTAR PUSTAKA:

[1] A. Chien and P. Khethavath, "Email Feature Classification and Analysis of Phishing Email Detection Using Machine Learning Techniques," 2023 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), Nadi, Fiji, 2023

[2] H. Yu et al., "ERDL: Efficient Retrieval Framework Based on Distillation from Large Language Models," 2024 International Conference on Computational Linguistics and Natural Language Processing (CLNLP), Yinchuan, China, 2024

[3] I. Trummer, "Large Language Models: Principles and Practice," 2024 IEEE 40th International Conference on Data Engineering (ICDE), Utrecht, Netherlands, 2024

[4] P. Saraswat and M. Singh Solanki, "Phishing Detection in E-mails using Machine Learning," 2022 2nd International Conference on Technological Advancements in Computational Sciences (ICTACS), Tashkent, Uzbekistan, 2022

[5] S. Das and D. Das, "Natural Language Processing (NLP) Techniques: Usability in Human-Computer Interactions," 2024 6th International Conference on Natural Language Processing (ICNLP), Xi'an, China, 2024

[6] S. Jawhar, C. E. Kimble, J. R. Miller and Z. Bitar, "Enhancing Cyber Resilience with AI-Powered Cyber Insurance Risk Assessment," 2024 IEEE 14th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 2024

[7] T. M. Mbelli and B. Dwolatzky, "Cyber Security, a Threat to Cyber Banking in South Africa: An Approach to Network and Application Security," 2016 IEEE 3rd International Conference on Cyber Security and Cloud Computing (CSCloud), Beijing, China, 2016