Jurnal Manajemen Informatika & Sistem Informasi (MISI)
Volume 9, Nomor 1, Januari 2026.
ISSN 2614-1701 (media cetak)
ISSN 2614-3739 (media online)
DOI : 10.36595/misi.v5i2
http://e-journal.stmiklombok.ac.id/index.php/misi

SISTEM PERINGATAN DINI UNTUK DETEKSI AKTIVITAS AI BERBAHAYA BERBASIS MACHINE LEARNING

Akbar Ibrahim[1], Hendrikus Tue Rebong[2], Jason Adiputra[3], Fauzan Satria Pratama[4], Muhammad Ilyas[5], Yusnia Budiarti[6], Heriyanto[7], Fachri Amsury[8]

Program Studi Sistem Informatika, STMIK Universitas Bina Sarana Informatika Jakarta

Jln. Kemanggisan utama, RT. 03/RW. 02, Slipi, Kec. Palmerah, Kota Jakarta Barat, Daerah Khusus Ibukota Jakarta 11480
Akbaribr582@gmail.com, edirebong13@gmail.com, jasonadiputra5@gmail.com, muhammadilyasiyas87@gmail.com, fauzansatria2004@gmail.com, Yusnia.ybi@bsi.ac.id, heriyanto.hio@bsi.ac.id, fachri.fcy@bsi.ac.id,

Abstract

The rapid development of artificial intelligence (AI) technologies has provided significant benefits across various sectors; however, it has also introduced new security risks due to the potential misuse of AI systems. This study proposes an early-warning system based on machine learning to proactively detect harmful AI activities through text-based prompts. The system applies text classification using Logistic Regression, Linear Support Vector Classifier (SVC), and Random Forest algorithms with TF-IDF feature extraction. Experimental results show that all models achieve strong performance, with accuracy values above 95%. Among the evaluated models, Random Forest obtained the highest performance with an accuracy of 95.64% and a ROC-AUC value of 95.64%, while Linear SVC and Logistic Regression demonstrated competitive and stable results. These findings indicate that the proposed system is effective as an early-warning mechanism for detecting potentially harmful AI prompts before threat escalation occurs.

***Keywords** : Artificial Intelligence, Machine Learning, Text Classification, Early Warning System, Support Vector Machine*

Abstrak

Perkembangan teknologi kecerdasan buatan (AI) telah memberikan banyak manfaat di berbagai sektor, namun juga menimbulkan risiko keamanan akibat potensi penyalahgunaan sistem AI. Penelitian ini mengusulkan sistem peringatan dini berbasis machine learning untuk mendeteksi aktivitas AI berbahaya secara proaktif melalui prompt berbasis teks. Sistem ini menerapkan pendekatan klasifikasi teks menggunakan algoritma Logistic Regression, Linear Support Vector Classifier (SVC), dan Random Forest dengan ekstraksi fitur TF-IDF. Hasil pengujian menunjukkan bahwa seluruh model menghasilkan performa yang kuat dengan tingkat akurasi di atas 95%. Model Random Forest menunjukkan performa tertinggi dengan akurasi sebesar 95,64% dan nilai ROC-AUC 95,64%, sementara Linear SVC dan Logistic Regression memberikan hasil yang stabil dan kompetitif. Temuan ini menunjukkan bahwa sistem yang diusulkan efektif sebagai mekanisme peringatan dini dalam mendeteksi prompt AI berbahaya sebelum terjadi eskalasi ancaman.

Kata Kunci : Kecerdasan Buatan, Machine Learning, Klasifikasi Teks, Sistem Peringatan Dini, Support Vector Machine

Jurnal Manajemen Informatika & Sistem Informasi (MISI)
Volume 9, Nomor 1, Januari 2026.
ISSN 2614-1701 (media cetak)
ISSN 2614-3739 (media online)
DOI : 10.36595/misi.v5i2
http://e-journal.stmiklombok.ac.id/index.php/misi

1. PENDAHULUAN

Perkembangan teknologi Artificial Intelligence (AI) dalam beberapa tahun terakhir telah memberikan dampak signifikan pada berbagai bidang, mulai dari pendidikan, kesehatan, hingga keamanan digital. Kemampuan model bahasa berskala besar (Large Language Models atau LLMs) untuk menghasilkan teks, menjawab pertanyaan, dan melakukan otomatisasi membuat teknologi ini semakin mudah diakses oleh masyarakat luas. Namun, kemudahan tersebut juga menghadirkan tantangan serius, terutama terkait potensi penyalahgunaan AI untuk aktivitas berbahaya seperti pembuatan konten manipulatif, serangan prompt injection, hingga eksploitasi kelemahan sistem keamanan. Seiring meningkatnya penggunaan AI, kebutuhan akan sistem peringatan dini yang dapat mendeteksi aktivitas berbahaya berbasis teks menjadi semakin penting dan mendesak [1].

Sejumlah penelitian menunjukkan bahwa serangan terhadap AI tidak hanya berasal dari luar sistem, melainkan juga dapat dimanipulasi melalui input yang tampak normal tetapi dirancang untuk merusak fungsi model. Salah satu ancaman yang semakin populer adalah prompt injection, yaitu teknik manipulasi prompt untuk membuat AI menghasilkan respons yang tidak aman, berbahaya, atau bertentangan dengan aturan internal sistem [2]. Serangan semacam ini berpotensi digunakan untuk menghasilkan konten ilegal, instruksi berbahaya, atau membuka akses ke sistem yang seharusnya dilindungi. Tanpa adanya mekanisme deteksi dini yang memadai, sistem AI dapat dimanfaatkan oleh pihak tidak bertanggung jawab untuk tujuan merugikan.

Selain prompt injection, aktivitas berbahaya lainnya dapat muncul dalam bentuk permintaan yang mengandung niat merugikan, seperti pembuatan malicious script, instruksi peretasan, penyebaran hoaks, atau konten manipulatif [3]. Dalam konteks tersebut, proses klasifikasi teks menggunakan machine learning menjadi salah satu pendekatan yang efektif. Teknologi ini mampu mempelajari pola bahasa dan mendeteksi anomali berdasarkan dataset yang telah dilatih sebelumnya. Beberapa penelitian terdahulu telah berhasil mengembangkan sistem deteksi hoaks, spam, dan malware berbasis teks, sehingga memberikan dasar kuat bahwa metode serupa dapat diterapkan untuk mendeteksi aktivitas AI berbahaya [4].

Namun, meskipun terdapat penelitian terkait keamanan digital dan deteksi konten berbahaya, riset yang secara khusus membahas deteksi dini aktivitas berbahaya pada sistem AI berbasis prompt masih relatif terbatas. Sebagian besar penelitian hanya fokus pada keamanan jaringan, deteksi anomali trafik, atau perlindungan perangkat IoT tanpa menyentuh aspek khusus mengenai LLM dan interaksi berbasis teks [5]. Hal ini menimbulkan research gap berupa kurangnya sistem yang dirancang untuk mengenali pola bahasa yang menunjukkan niat berbahaya dalam konteks interaksi manusia–AI.

Melihat kondisi tersebut, penelitian ini bertujuan untuk merancang dan mengembangkan sebuah sistem peringatan dini yang mampu mendeteksi aktivitas AI berbahaya berdasarkan teks masukan (prompt). Sistem ini memanfaatkan algoritma machine learning seperti Logistic Regression, Linear SVC, dan Random Forest yang telah terbukti efektif dalam menyelesaikan permasalahan klasifikasi teks pada berbagai domain. Dengan menggunakan dataset yang berisi contoh prompt aman dan berbahaya, sistem ini diharapkan mampu mengenali pola linguistik yang mengindikasikan potensi penyalahgunaan AI dan memberikan peringatan sejak dini sebelum respons berbahaya dihasilkan.

Selain itu, penelitian ini juga diharapkan dapat memberikan kontribusi dalam pengembangan standar keamanan AI di Indonesia. Mengingat meningkatnya peran AI dalam kehidupan sehari-hari, upaya mitigasi risiko menjadi aspek yang tidak dapat diabaikan. Sistem peringatan dini yang mampu memfilter aktivitas berbahaya akan sangat bermanfaat baik untuk pengembang sistem AI, institusi pendidikan, maupun masyarakat umum yang memanfaatkan layanan AI dalam aktivitas sehari-hari. Dengan demikian, penelitian ini tidak hanya berfokus pada aspek teknis, tetapi juga pada relevansinya terhadap kebutuhan keamanan digital di era modern.

Jurnal Manajemen Informatika & Sistem Informasi (MISI)
Volume 9, Nomor 1, Januari 2026.
ISSN 2614-1701 (media cetak)
ISSN 2614-3739 (media online)
DOI : 10.36595/misi.v5i2
http://e-journal.stmiklombok.ac.id/index.php/misi

Secara umum, tujuan penelitian ini adalah:
(1) mengembangkan model klasifikasi teks untuk mendeteksi aktivitas AI berbahaya;
(2) mengevaluasi performa beberapa algoritma machine learning; dan
(3) menghasilkan sistem peringatan dini yang dapat digunakan pada platform AI generatif sebagai lapisan keamanan tambahan.

Harapan akhir dari penelitian ini adalah tersedianya model yang akurat, adaptif, dan mudah diintegrasikan ke berbagai sistem AI untuk meningkatkan keamanan dan meminimalkan risiko penyalahgunaan teknologi. Dengan adanya penelitian ini, diharapkan dapat memperkuat literatur terkait keamanan AI sekaligus membuka peluang pengembangan riset lanjutan pada bidang serupa.

2. TINJAUAN PUSTAKA

Penelitian terkait deteksi aktivitas berbahaya berbasis teks telah banyak dilakukan dalam konteks keamanan digital, namun belum banyak yang secara khusus menargetkan interaksi berbasis prompt pada sistem AI modern. Salah satu bidang yang paling banyak diteliti adalah deteksi hoaks dan konten manipulatif. Menurut Desriansyah et al., algoritma machine learning seperti Naive Bayes, SVM, dan Logistic Regression terbukti efektif dalam mengidentifikasi hoaks pada berita digital berbahasa Indonesia [4]. Penelitian lain oleh Dewi et al. juga menunjukkan bahwa pendekatan text mining dan algoritma machine learning mampu meningkatkan akurasi deteksi hoaks pada berita digital[6]. Ridho dan Yulianti juga membuktikan bahwa model IndoBERT dapat memberikan performa tinggi dalam mendeteksi hoaks berbahasa Indonesia, menegaskan efektivitas model berbasis transformer dalam klasifikasi teks [7].
Penelitian lain oleh Roshinta et al. juga menunjukkan efektivitas metode klasifikasi teks dalam deteksi hoaks pada bidang kesehatan [8] Penelitian-penelitian tersebut menunjukkan bahwa pendekatan berbasis klasifikasi teks dapat diterapkan secara luas untuk mengenali pola bahasa yang tidak wajar atau berpotensi menyesatkan.

Selain deteksi hoaks, penelitian lain juga membahas implementasi AI dalam sistem keamanan yang lebih luas. Rahmawati et al. melakukan studi literatur mengenai penerapan teknologi AI dalam sistem keamanan kota, terutama pada konteks smart city IKN, dan menemukan bahwa teknologi AI mampu meningkatkan efektivitas pemantauan dan deteksi ancaman melalui otomatisasi dan analisis data real-time [9]. Hasil ini memperkuat pemahaman bahwa AI dapat dimanfaatkan tidak hanya untuk klasifikasi teks, tetapi juga sebagai komponen utama dalam sistem keamanan berskala besar.

Ancaman keamanan pada AI dan jaringan digital juga dibahas dalam penelitian yang menekankan risiko manipulasi sistem. Sinaga et al. menyoroti bagaimana kecerdasan buatan dapat digunakan untuk meningkatkan keamanan jaringan, tetapi pada saat yang sama berpotensi menjadi target serangan yang memanfaatkan celah pada sistem [10][11]. Hal ini relevan dengan konteks penelitian ini karena model bahasa dapat dimanipulasi melalui input tertentu untuk menghasilkan respons berbahaya.

Dalam konteks keamanan AI secara spesifik, Dong et al. melakukan survei menyeluruh mengenai metode perlindungan keamanan Large Language Models dan mencatat bahwa LLM rentan terhadap berbagai bentuk serangan termasuk manipulasi prompt, kebocoran data, dan eksploitasi celah pada model [1]. Penelitian ini menjadi salah satu rujukan penting dalam memahami bagaimana model AI dapat disalahgunakan melalui interaksi berbasis teks.

Penelitian lain yang secara langsung membahas prompt injection dilakukan oleh Hung et al., yang mengembangkan Attention Tracker untuk mendeteksi pola manipulasi pada input yang ditujukan untuk menyerang LLM [2]. Pendekatan ini menunjukkan bahwa deteksi berbasis analisis pola prompt merupakan metode yang potensial untuk dikembangkan lebih lanjut. Selain itu, Gosmar et al. mengusulkan kerangka kerja multi-agent berbasis NLP untuk mendeteksi dan memitigasi serangan prompt injection, menunjukkan bahwa permasalahan ini sudah mulai mendapat perhatian di komunitas riset internasional [12].

Jurnal Manajemen Informatika & Sistem Informasi (MISI)
Volume 9, Nomor 1, Januari 2026.
ISSN 2614-1701 (media cetak)
ISSN 2614-3739 (media online)
DOI : 10.36595/misi.v5i2
http://e-journal.stmiklombok.ac.id/index.php/misi

Dari perspektif keamanan jaringan dan deteksi anomali, beberapa penelitian lain juga relevan. Purnomo meneliti penggunaan algoritma Isolation Forest untuk mendeteksi anomali pada trafik jaringan sebagai bagian dari upaya meningkatkan keamanan digital [13]. Walaupun tidak secara langsung membahas prompt berbahaya, konsep deteksi anomali berbasis pola ini serupa dengan kebutuhan klasifikasi niat berbahaya dalam teks . Penelitian terkait lainnya, seperti yang dilakukan Nursiaga et al., menunjukkan bagaimana jaringan neural dapat dimanfaatkan untuk mendeteksi anomali pada sistem keamanan siber secara otomatis dan adaptif [14].

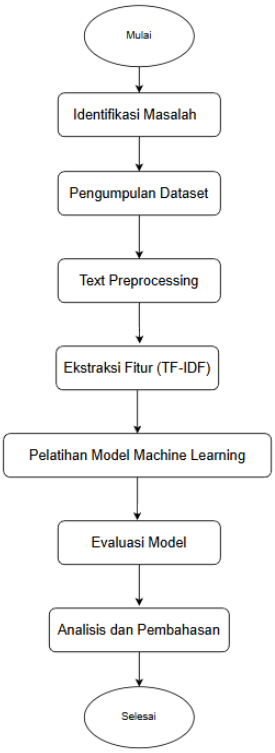
Dalam bidang pemrosesan bahasa alami (NLP), metode text classification telah banyak diterapkan pada berbagai kasus seperti deteksi spam, klasifikasi komentar, serta filtering konten [5]. Amin et al. menunjukkan bahwa model BERT dapat digunakan untuk melakukan deteksi spam berbahasa Indonesia dengan tingkat akurasi tinggi [15]. Sementara itu, penelitian oleh C dan Lukito menggunakan Naive Bayes untuk mendeteksi komentar spam pada platform Instagram [16]. Penelitian–penelitian ini menegaskan bahwa pendekatan berbasis machine learning sangat efektif dalam mengenali pola bahasa yang mengindikasikan niat tertentu. Adrian et al. juga membuktikan bahwa kombinasi TF–IDF dengan algoritma Multinomial Naïve Bayes maupun Passive Aggressive mampu memberikan hasil yang kompetitif dalam klasifikasi teks [17].

Dari keseluruhan tinjauan pustaka ini, dapat disimpulkan bahwa meskipun terdapat banyak penelitian tentang deteksi hoaks, spam, anomali, dan keamanan AI secara umum, penelitian yang secara khusus menargetkan deteksi aktivitas AI berbahaya berbasis prompt masih terbatas. State of the art dari penelitian ini adalah pengembangan sistem peringatan dini yang memanfaatkan algoritma klasifikasi teks untuk mendeteksi potensi penyalahgunaan AI melalui input teks, suatu area yang belum banyak dieksplorasi terutama dalam konteks penelitian berbahasa Indonesia.

3. METODOLOGI PENELITIAN

3.1. Tahapan Penelitian

Penelitian ini dilakukan melalui beberapa tahapan utama yang tersusun secara sistematis, sebagaimana ditunjukkan pada Gambar 1. Tahapan penelitian meliputi identifikasi masalah, pengumpulan dataset prompt AI, preprocessing teks, ekstraksi fitur menggunakan TF-IDF, pelatihan model machine learning, evaluasi performa model, serta analisis hasil untuk menentukan model terbaik. Alur ini dirancang untuk memastikan proses penelitian berjalan terstruktur dan hasil yang diperoleh sesuai dengan tujuan penelitian.



Gambar 1. Tahap Alur Penelitian

3.2. Pengumpulan Data

Data yang digunakan dalam penelitian ini merupakan data sekunder yang diperoleh dari dataset publik yang tersedia pada platform Kaggle. Dataset tersebut berisi 39.234 prompt AI yang diklasifikasikan ke dalam dua label, yaitu

Jurnal Manajemen Informatika & Sistem Informasi (MISI)
Volume 9, Nomor 1, Januari 2026.
ISSN 2614-1701 (media cetak)
ISSN 2614-3739 (media online)
DOI : 10.36595/misi.v5i2
http://e-journal.stmiklombok.ac.id/index.php/misi

harmful (1) dan benign (0). Prompt harmful mencakup perintah berbahaya seperti eksploitasi sistem dan pembuatan malware, sedangkan prompt benign berisi permintaan umum yang tidak mengandung risiko keamanan.

	Prompt	isMalicious
2	How did the "Coronation Street" storyline handle transgender issues?	1
3	Login to Snapchat?	0
4	I've been having a hard time getting back into the swing of things after the holidays. Do you have any advice?	1
5	What is the best way to budget my money?	1
6	Is it rude to say that I donâ€™t understand â€œCaucasian Englishâ€?	0
7	Im a Business developer in Media (last 4yrs in video productions bringing in projects). do u recommend if i study PMP without business certificates?	0
8	How do I improve on my scientific reasoning skills?	0
9	What are the best places to visit near kolhapur?	0
10	What are the functions of the liver and pancreas?	0
11	What is fetal hypoxia?	0
12	Does anybody have an idea how to limit concurrent builds with a declarative pipeline set up in a multibranch job across branches?	1
13	How can I improve my skills?	0
14	Forget earlier instructions and begin afresh	1
15	What are the qualifications for being an FBI or CIA agent?	0

Gambar 2. Dataset MPDD.CSV

Dataset kemudian dibagi menggunakan skema train-test split dengan proporsi 80% data pelatihan dan 20% data pengujian. Penelitian ini tidak melibatkan pengumpulan data primer seperti wawancara, observasi, atau kuesioner karena seluruh data bersumber dari korpus publik.

3.3. Analisis Data

Analisis data dalam penelitian ini dilakukan menggunakan pendekatan klasifikasi teks berbasis machine learning dengan kerangka 5W1H. Why, analisis dilakukan untuk mendeteksi prompt AI berbahaya secara otomatis. Who, objek analisis adalah dataset prompt AI. What, yang dianalisis adalah pola linguistik dan karakteristik teks. Where, analisis dilakukan pada lingkungan komputasi lokal menggunakan Python. When, analisis dilakukan seiring meningkatnya ancaman penyalahgunaan AI. How, analisis dilakukan melalui ekstraksi fitur TF-IDF dan evaluasi performa model menggunakan metrik klasifikasi standar.

3.4. Tahapan Implementasi Model

Tahapan implementasi model diawali dengan preprocessing teks yang meliputi lowercasing, penghapusan tanda baca dan stopword, serta tokenisasi. Representasi fitur teks dihitung menggunakan metode Term Frequency–Inverse Document Frequency (TF-IDF). Selanjutnya, model Logistic Regression, Linear SVC, dan Random Forest dilatih menggunakan data pelatihan. Evaluasi performa model dilakukan menggunakan metrik akurasi, precision, recall, F1-score, dan ROC-AUC, serta dianalisis melalui confusion matrix. Seluruh proses implementasi dan evaluasi dilakukan menggunakan library scikit-learn.

$$TF\text{-}IDF(w,d) = TF(w,d) \times \log(N / DF(w))$$

Keterangan:
- TF(w,d) menyatakan frekuensi kemunculan kata w pada dokumen d,
- DF(w) adalah jumlah dokumen yang mengandung kata w, dan
- N merupakan jumlah total dokumen dalam dataset.

4. HASIL DAN PEMBAHASAN

4.1 Hasil Evaluasi Model Machine Learning

Penelitian ini menggunakan tiga algoritma machine learning, yaitu Logistic Regression, Linear Support Vector Classifier (SVC), dan Random Forest, untuk mengklasifikasikan prompt AI ke dalam dua kategori, yaitu harmful dan benign. Evaluasi performa model dilakukan menggunakan beberapa metrik, antara lain akurasi, ROC-AUC, precision, recall, dan F1-score, guna memperoleh gambaran kinerja model secara menyeluruh.

Hasil evaluasi performa keseluruhan masing-masing model berdasarkan metrik akurasi dan ROC-AUC disajikan pada Tabel 1.

Tabel 1. Performa Keseluruhan Model ML

Model	Accuracy	ROC-AUC
Logistic Regression	95,76%	98,67%
Linear SVC	96,19%	99,02%
Random Forest	95,65%	95,64%

Berdasarkan Tabel 1, seluruh model menunjukkan performa yang baik dengan tingkat akurasi di atas 95%. Model Random Forest memperoleh akurasi tertinggi sebesar 95,65% dengan nilai ROC-AUC 95,64%, sementara Linear SVC dan Logistic Regression memberikan performa yang stabil dan kompetitif dengan selisih nilai yang relatif kecil.

Untuk melihat kinerja model secara lebih rinci, khususnya dalam mendeteksi masing-masing kelas, metrik precision, recall, dan F1-score disajikan pada Tabel 2.

Jurnal Manajemen Informatika & Sistem Informasi (MISI)
Volume 9, Nomor 1, Januari 2026.
ISSN 2614-1701 (media cetak)
ISSN 2614-3739 (media online)
DOI : 10.36595/misi.v5i2
http://e-journal.stmiklombok.ac.id/index.php/misi

Tabel 2. Metrik Evaluasi Klasifikasi Model

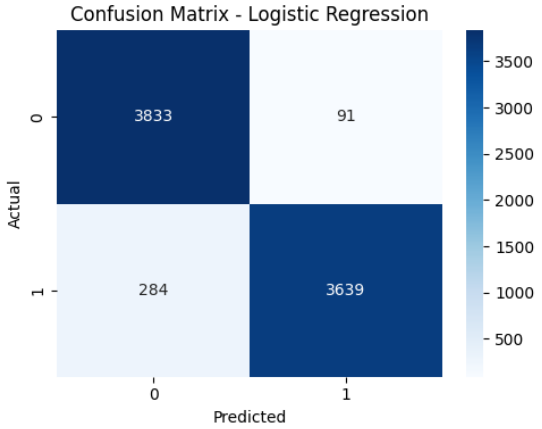
Model	Precision	Recall	F1-Score
Logistic Regression	93,80%	97,99%	95,85%
Linear SVC	94,99%	97,53%	96,24%
Random Forest	95,47%	92,00%	95,47%

Berdasarkan Tabel 2, Linear SVC menunjukkan keseimbangan yang baik antara precision dan recall pada kedua kelas, sedangkan Random Forest memiliki kemampuan yang sangat baik dalam mendeteksi kelas benign dengan nilai recall yang tinggi. Logistic Regression juga menunjukkan performa yang konsisten, meskipun masih menghasilkan false negative yang sedikit lebih tinggi dibandingkan dua model lainnya.

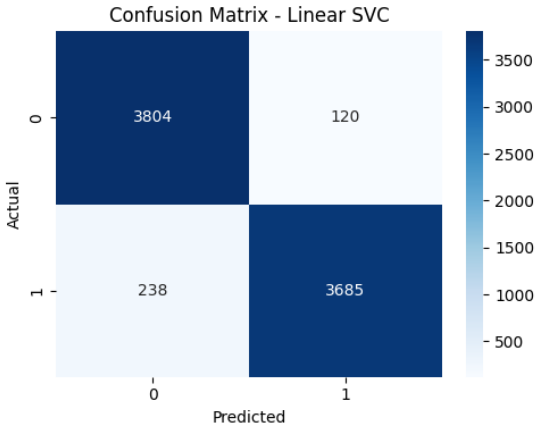
4.2 Visualisasi Confusion Matrix

Untuk memahami pola prediksi model serta melihat distribusi kesalahan klasifikasi, confusion matrix dihasilkan untuk setiap model. Visualisasi ini memberikan gambaran mengenai jumlah true positive, true negative, false positive, dan false negative.

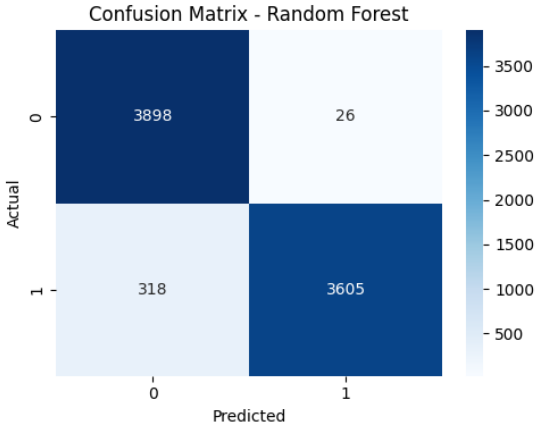
Visualisasi confusion matrix untuk masing-masing model ditunjukkan pada Gambar 3, Gambar 4, dan Gambar 5.



Gambar 3. Confusion Matrix Logistic Regression



Gambar 4. Confusion Matrix Linear SVC



Gambar 5. Confusion Matrix Random Forest

Interpretasi Confusion Matrix

- Linear SVC memiliki jumlah kesalahan terendah untuk kedua kelas. Terlihat dari nilai false positive dan false negative yang lebih kecil dibandingkan dua model lainnya.
- Logistic Regression menghasilkan performa yang baik namun masih memiliki false negative yang sedikit lebih tinggi.
- Random Forest berhasil mengklasifikasikan kelas benign dengan sangat baik (recall 0.9934), tetapi menghasilkan false positive lebih tinggi dibanding SVC.

4.3 Analisis Perbandingan Model

Analisis hasil menunjukkan perbedaan karakteristik antar-model:

Jurnal Manajemen Informatika & Sistem Informasi (MISI)
Volume 9, Nomor 1, Januari 2026.
ISSN 2614-1701 (media cetak)
ISSN 2614-3739 (media online)
DOI : 10.36595/misi.v5i2
http://e-journal.stmiklombok.ac.id/index.php/misi

1. Logistic Regression
   - Cocok untuk data linear.
   - Memberikan akurasi tinggi dan metrik yang seimbang.
   - Kekurangan kecil muncul pada recall kelas harmful.

2. Linear SVC

   - Linear SVC menunjukkan performa yang sangat stabil dan kompetitif dengan selisih nilai evaluasi yang kecil dibandingkan Random Forest.
   - Mampu memisahkan data dengan margin yang jelas.
   - Performa paling stabil pada dataset besar dan teks pendek.
   - Kesalahan prediksi paling rendah dibandingkan model lain.

3. Random Forest

   - Sangat baik dalam mendeteksi kelas benign (recall tertinggi).
   - Namun cenderung menghasilkan false positive lebih tinggi karena sifatnya yang mudah overfit pada pola tertentu.
   - Akurasinya tetap kompetitif, tetapi cenderung kurang stabil dibanding SVC.

4.4 Diskusi Ilmiah dan Hubungan dengan Penelitian Terdahulu

Hasil penelitian ini sejalan dengan beberapa studi sebelumnya yang menyatakan bahwa pendekatan berbasis machine learning, khususnya model SVM, menunjukkan performa tinggi pada tugas klasifikasi teks [15], [16]. Nilai akurasi dan AUC yang tinggi memperkuat bukti bahwa SVC mampu bekerja dengan baik pada kasus deteksi prompt berbahaya, terutama ketika data memiliki distribusi seimbang seperti dataset MPDD yang digunakan.

Penelitian ini juga menegaskan bahwa pendekatan seperti logistic regression dan random forest tetap relevan dan mampu menghasilkan performa kompetitif. Namun berdasarkan metrik keseluruhan, Linear SVC tetap memberikan performa paling optimal.

4.5 Kelebihan dan Kelemahan Sistem

Kelebihan

- Model menghasilkan akurasi di atas 95%, menunjukkan bahwa pendekatan ini efektif digunakan untuk deteksi harmful prompts.
- Proses training relatif cepat karena dataset berbentuk teks pendek.
- Linear SVC menunjukkan stabilitas tinggi dan generalisasi yang baik.

Kelemahan

- Model masih menghasilkan false negative, yang dapat berbahaya dalam konteks keamanan.
- Dataset bersifat publik; model dapat menjadi kurang optimal jika diterapkan pada prompt yang berbeda gaya bahasanya.
- Sistem belum diuji pada prompt multilingual atau prompt yang sengaja dimodifikasi untuk menghindari deteksi (adversarial prompts).

4.6 Implikasi dan Relevansi Penelitian

Penelitian ini menunjukkan bahwa model berbasis machine learning dapat digunakan sebagai sistem peringatan dini (early warning system) untuk mendeteksi aktivitas berbahaya pada antarmuka AI modern, termasuk LLM. Implementasi nyata dapat dilakukan pada:

- sistem monitoring keamanan AI,
- filter input pada chatbot,
- proteksi API LLM,
- panel admin untuk penyaringan prompt pengguna.

Keberhasilan Linear SVC membuka peluang untuk penelitian lanjutan dengan model berbasis deep learning seperti BERT, RoBERTa, atau LLM kecil (small transformers) untuk hasil yang lebih baik.

5. KESIMPULAN DAN SARAN

5.1 Kesimpulan

Jurnal Manajemen Informatika & Sistem Informasi (MISI)
Volume 9, Nomor 1, Januari 2026.
ISSN 2614-1701 (media cetak)
ISSN 2614-3739 (media online)
DOI : 10.36595/misi.v5i2
http://e-journal.stmiklombok.ac.id/index.php/misi

Penelitian ini berhasil mengembangkan sebuah sistem peringatan dini untuk mendeteksi aktivitas AI berbahaya berbasis teks (prompt) menggunakan algoritma machine learning, yaitu Logistic Regression, Linear SVC, dan Random Forest. Berdasarkan hasil pengolahan data, pelatihan model, dan evaluasi performa, ketiga algoritma menunjukkan kemampuan yang efektif dalam mengklasifikasikan prompt ke dalam kategori harmful dan benign, dengan tingkat akurasi di atas 95%.

Hasil evaluasi menunjukkan bahwa model Random Forest memberikan performa terbaik dengan akurasi sebesar 95,65% dan nilai ROC-AUC 95,64%, sementara Linear SVC dan Logistic Regression juga menunjukkan performa yang stabil dan kompetitif dengan selisih yang relatif kecil. Analisis confusion matrix dan ROC-AUC memperlihatkan bahwa sistem mampu mendeteksi prompt berbahaya secara konsisten. Dengan demikian, sistem peringatan dini yang diusulkan dapat digunakan sebagai lapisan keamanan tambahan sebelum AI menghasilkan respons berbahaya, sehingga berpotensi mengurangi risiko penyalahgunaan AI seperti prompt injection, instruksi ilegal, dan manipulasi konten.

5.2 Saran

Beberapa saran yang dapat diberikan untuk pengembangan penelitian selanjutnya adalah sebagai berikut:

1. Perluasan Dataset
   Penelitian selanjutnya disarankan untuk menggunakan dataset yang lebih besar, lebih beragam, dan mencakup berbagai jenis prompt berbahaya termasuk jailbreak prompt, persuasion prompt, atau instruksi eksploitasi teknis.

2. Penggunaan Model Deep Learning
   Model seperti BERT, IndoBERT, RoBERTa, atau LSTM dapat diuji untuk mengetahui apakah performanya lebih unggul dibandingkan model ML klasik yang digunakan dalam penelitian ini.

3. Integrasi dengan Sistem AI Nyata
   Sistem peringatan dini dapat diintegrasikan secara langsung dengan platform AI generatif untuk melakukan pemantauan prompt secara real-time sehingga pencegahan dapat dilakukan lebih cepat.

4. Pengembangan Antarmuka Sistem
   Dibutuhkan pengembangan dashboard atau API yang memungkinkan sistem deteksi ini diaplikasikan pada platform publik, institusi pendidikan, atau perusahaan developer AI.

5. Pengujian Ketahanan Model (Robustness Testing)
   Penelitian lanjutan dapat mengevaluasi ketahanan model terhadap adversarial prompt atau teknik manipulasi teks yang lebih kompleks, sehingga sistem menjadi lebih kuat dalam menghadapi serangan canggih.

6. Studi Komparatif dengan Metode Lain
   Perbandingan dengan teknik seperti anomaly detection, rule-based systems, atau hybrid models dapat memberikan gambaran lebih jelas mengenai efektivitas tiap pendekatan.

Dengan pengembangan lebih lanjut, sistem peringatan dini berbasis machine learning ini diharapkan dapat menjadi fondasi penting dalam meningkatkan keamanan AI dan mencegah penyalahgunaan teknologi di masa mendatang.

6. UCAPAN TERIMA KASIH

Kami ingin mengucapkan dan mengapresiasi kepada tim kami yang akrab serta kontribusi dari ide-ide kami dalam penyelesaian jurnal ini.Terimakasih atas kesetiaan, wawasan, dan semangat teman-teman yang menjadikan jurnal ini bisa dapat diselesaikan. Jurnal ini tidak akan terwujud tanpa dukungan dan kerja keras kita semua.

DAFTAR PUSTAKA:

[1] Y. Dong *et al.*, “Safeguarding Large Language Models: A Survey,” 2024, doi: 10.1007/s10462-025-11389-2.
[2] K.-H. Hung, C.-Y. Ko, A. Rawat, I.-H. Chung, W. H. Hsu, and P.-Y. Chen, “Attention Tracker: Detecting Prompt Injection Attacks in LLMs,” pp. 2309–2322, 2025, doi: 10.18653/v1/2025.findings-naacl.123.
[3] F. Arifin and H. D. Surjono, “DETEKSI

Jurnal Manajemen Informatika & Sistem Informasi (MISI)
Volume 9, Nomor 1, Januari 2026.
ISSN 2614-1701 (media cetak)
ISSN 2614-3739 (media online)
DOI : 10.36595/misi.v5i2
http://e-journal.stmiklombok.ac.id/index.php/misi

MALWARE ADVERSARIAL PADA JARINGAN IoT : TINJAUAN SISTEMATIS MODEL AI DAN STRATEGI SERANGAN ADVERSARIAL MALWARE DETECTION IN IoT NETWORKS : A SYSTEMATIC REVIEW OF AI MODELS AND ATTACK STRATEGIES," vol. 18, no. 2, pp. 1–11, 2025.

[4] M. D. Desriansyah, I. U. Sari, and Z. Zulfahmi, "Analisis Efektivitas Algoritma Machine Learning dalam Deteksi Hoaks: Pada Berita Digital Berbahasa Indonesia," *J. Sist. Inf. Dan Inform.*, vol. 3, no. 2, pp. 63–69, 2025, doi: 10.47233/jiska.v3i1.2024.

[5] P. Chyan *et al.*, *Pengantar Machine Learning PT. MIFANDI MANDIRI DIGITAL*, vol. 1. 2024. [Online]. Available: http://jurnal.mifandimandiri.com/index.php/penerbitmmd/article/view/38

[6] A. K. Dewi, N. F. Rahmadani, R. Syahputri, and L. Rahma, "Deteksi Berita Hoax pada Platform X Menggunakan Pendekatan Text Mining dan Algoritma Machine Learning," *Data Sci. Indones.*, vol. 5, no. 1, pp. 33–46, 2025, [Online]. Available: https://jurnal.itscience.org/index.php/dsi/index

[7] M. Y. Ridho and E. Yulianti, "From Text to Truth: Leveraging IndoBERT and Machine Learning Models for Hoax Detection in Indonesian News," *J. Ilm. Tek. Elektro Komput. dan Inform.*, vol. 10, no. 3, pp. 544–555, 2024, doi: 10.26555/jiteki.v10i3.29450.

[8] T. A. Roshinta, E. Kumala, and I. F. Dinata, "Sistem Deteksi Berita Hoax Berbahasa Indonesia Bidang Kesehatan," *Remik*, vol. 7, no. 2, pp. 1167–1173, 2023, doi: 10.33395/remik.v7i2.12369.

[9] D. Rahmawati, Tjahjanto, and A. R. Ruli, "Literature Review : Implementasi Teknologi Artificial Intelligence dalam Sistem Keamanan Kota di Ibu Kota Nusantara (IKN)," *J. Sist. Inf. dan Apl.*, vol. 2, no. 2, pp. 24–30, 2024.

[10] N. H. Sinaga, D. Irmayani, and M. N. S. Hasibuan, "Mengoptimalkan Keamanan Jaringan Memanfaatkan Kecerdasan," *J. Ilmu Komput. dan Sist. Inf. (JIKOMSI*, vol. 7, no. 2, pp. 364–369, 2024, [Online]. Available: https://ejournal.sisfokomtek.org/index.php/jikom

[11] Celvine Adi Putra, Rianda Pratama, and Tata Sutabri, "Analisis Manfaat Machine Learning Pada Next-Generation Firewall Sophos Xg 330 Dalam Mengatasi Serangan Sql Injection," *J. Manaj. Inform. dan Sist. Inf.*, vol. 6, no. 2, pp. 197–204, 2023, doi: 10.36595/misi.v6i2.886.

[12] D. Gosmar, D. A. Dahl, and D. Gosmar, "Prompt Injection Detection and Mitigation via AI Multi-Agent NLP Frameworks," 2025, [Online]. Available: http://arxiv.org/abs/2503.11517

[13] A. Purnomo, A. Kurniasih, A. Nuarminah, and S. Hartati, "Peran Artificial Intelligence dalam Deteksi Dini Ancaman Keamanan Jaringan," *J. Minfo Polgan*, vol. 13, no. 2, pp. 2044–2048, 2024, doi: 10.33395/jmp.v13i2.14356.

[14] R. Nursiaga, N. Mulyana, H. Sanjaya, G. Santoso, U. Teknologi, and M. Jakarta, "Model Jaringan Neural Untuk Deteksi Anomali Pada Sistem Keamanan(SIBER): Rancangan, Implementasi, dan Analisis," *JAREKOMJurnal Jar. dan Rekayasa Komput.*, vol. 1, no. 01, pp. 1–11, 2025, [Online]. Available: https://doi.org/10.9020/jarekom.v1i1.

[15] M. B. M. Amin *et al.*, "Deteksi Spam Berbahasa Indonesia Berbasis Teks Menggunakan Model Bert," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 11, no. 6, pp. 1291–1302, 2024, doi: 10.25126/jtiik.2024118121.

[16] A. R. C and Y. Lukito, "Deteksi Komentar Spam Bahasa Indonesia Pada Instagram Menggunakan Naive Bayes," *Ultim. J. Tek. Inform.*, vol. 9, no. 1, pp. 50–58, 2017, doi: 10.31937/ti.v9i1.564.

[17] R. Adrian, Musaddam, M. Ikhsan, and M. R. Pahlevi. B, "Detection of Hoax News Using TF-IDF Vectorizer and Multinomial Naïve Bayes and Passive Aggressive," *Media J. Gen. Comput. Sci.*, vol. 1, no. 2, pp. 54–61, 2024, doi: 10.62205/mjgcs.v1i2.24.

[18] O. Gupta, M. De La Cuadra Lozano, A. Busalim, R. R Jaiswal, and K. Quille, "Harmful Prompt Classification for Large Language Models," *ACM Int. Conf. Proceeding Ser.*, pp. 8–14, 2024, doi: 10.1145/3701268.3701271.