Jurnal Teknik dan Science
JTS Vol 5 No. 2 June 2026 | ISSN: 2828-7002 (Print), ISSN: 2828-6871 (online), Hal 88-99

Identifikasi Pesan Penipuan Berkedok Hadiah Digital Menggunakan Machine Learning

Muhammad Shafari Rahmat[1], Syahrizal Andhika[2]
Universitas Gunadarma

Article History
Received : 2026-03-07
Revised : 2026-04-23
Accepted : 2026-05-29
Published : 2026-07-19

Corresponding author*:
shafari@staff.gunadarma.ac.id

Cite This Article:
Muhammad Shafari Rahmat, & Syahrizal Andhika. (2026). Identifikasi Pesan Penipuan Berkedok Hadiah Digital Menggunakan Machine Learning. Jurnal Teknik Dan Science, 5(2), 88–99.

DOI:
https://doi.org/10.56127/jts.v5i2.2948

Abstract: Digital prize scams are increasingly distributed through text messages, social media, and instant messaging platforms using persuasive expressions, suspicious links, and urgent instructions. This study aims to develop a machine learning-based approach for identifying Indonesian-language scam messages disguised as digital prize notifications. The dataset consisted of 6,000 text messages, comprising 3,000 scam messages and 3,000 non-scam messages. The data were manually labeled by two independent annotators and processed through case folding, text cleaning, tokenization, stopword removal, and stemming. Text features were transformed into numerical representations using Term Frequency–Inverse Document Frequency (TF-IDF). Five classification algorithms were evaluated, including Multinomial Naive Bayes, Support Vector Machine, Logistic Regression, Random Forest, and XGBoost. Model performance was assessed using stratified five-fold cross-validation based on accuracy, precision, recall, F1-score, and ROC-AUC. The results showed that XGBoost achieved the best performance, with an accuracy of 97.0%, precision of 97.2%, recall of 96.8%, F1-score of 97.0%, and ROC-AUC of 99.3%. Scam messages were commonly characterized by words related to prizes, winning notifications, free offers, claims, and urgency. These findings indicate that TF-IDF combined with XGBoost can effectively support the automatic detection of digital prize scam messages. Future studies should incorporate URL characteristics, sender metadata, and contextual language models to improve detection performance.

Keywords: digital prize scam, machine learning, text classification, TF-IDF, XGBoost.

PENDAHULUAN

Perkembangan layanan komunikasi digital telah memberikan kemudahan bagi masyarakat dalam menerima informasi, melakukan transaksi, dan berinteraksi dengan berbagai organisasi. Pesan singkat, aplikasi perpesanan, media sosial, dan surat elektronik kini digunakan oleh perusahaan untuk menyampaikan promosi, program loyalitas, voucher, serta informasi hadiah kepada konsumennya. Namun, kemudahan tersebut juga dimanfaatkan oleh pelaku kejahatan untuk menyebarkan pesan penipuan dengan menyamar sebagai perusahaan, bank, platform perdagangan elektronik, penyelenggara dompet digital, maupun lembaga pemerintah. Penipuan melalui pesan digital menjadi persoalan yang penting karena dapat mengakibatkan pencurian data pribadi, pengambilalihan akun, pemasangan perangkat lunak berbahaya, dan kerugian finansial.

Besarnya ancaman penipuan digital di Indonesia dapat dilihat dari data Indonesia Anti-Scam Centre. Sejak mulai beroperasi pada 22 November 2024 hingga 11 November 2025, lembaga tersebut menerima 343.402 laporan penipuan yang berkaitan dengan 563.558 rekening. Nilai kerugian yang dilaporkan mencapai Rp7,8 triliun, sedangkan dana yang berhasil diblokir sebesar Rp386,5 miliar (Otoritas Jasa Keuangan [OJK], 2025a). Data tersebut tidak secara khusus memisahkan penipuan berkedok hadiah digital, tetapi

menunjukkan bahwa penipuan berbasis komunikasi dan transaksi digital telah menimbulkan dampak finansial yang besar di Indonesia.

Salah satu modus yang digunakan pelaku adalah mengirimkan pesan yang menyatakan bahwa penerima memenangkan hadiah, undian, voucher belanja, saldo dompet digital, telepon seluler, kendaraan, atau hadiah uang tunai. Pesan tersebut biasanya mengatasnamakan institusi yang dikenal masyarakat dan disertai instruksi untuk membuka tautan, mengisi formulir, menghubungi nomor tertentu, membayar biaya administrasi, atau memberikan informasi rahasia. OJK menjelaskan bahwa pesan penipuan yang mengatasnamakan bank dapat berisi informasi mengenai hadiah undian atau promosi menarik yang disertai tautan berbahaya. Ketika tautan tersebut dibuka, korban dapat diarahkan menuju halaman palsu yang bertujuan mengambil data pribadi dan informasi keuangan (OJK, 2025b). OJK juga menegaskan bahwa lembaga tersebut tidak pernah meminta pembayaran untuk proses pencairan hadiah atau program giveaway (OJK, 2024).

Pesan penipuan berkedok hadiah digital pada dasarnya memanfaatkan teknik rekayasa sosial. Pelaku tidak selalu menyerang kelemahan teknis suatu sistem, tetapi memengaruhi kondisi psikologis penerima pesan melalui janji keuntungan, rasa penasaran, keterbatasan waktu, dan permintaan tindakan segera. Kata-kata seperti “selamat”, “pemenang”, “hadiah khusus”, “klaim sekarang”, “kesempatan terbatas”, dan “verifikasi data” dapat digunakan untuk membangun kesan bahwa pesan tersebut resmi dan mendesak. Carroll et al. (2022) menunjukkan bahwa pengguna pada umumnya masih mengalami kesulitan dalam membedakan pesan phishing modern dari komunikasi yang sah. Isi pesan yang sesuai dengan konteks atau kebutuhan pengguna dapat semakin meningkatkan tingkat keberhasilan penipuan.

Pencegahan yang hanya mengandalkan kewaspadaan pengguna memiliki keterbatasan karena bentuk pesan penipuan terus berubah. Daftar kata terlarang, pemblokiran nomor, dan blacklist tautan dapat digunakan untuk menyaring pola yang telah dikenali, tetapi kurang efektif ketika pelaku menggunakan nomor baru, domain baru, pemendek tautan, variasi ejaan, serta perubahan susunan kalimat. Goel dan Jain (2018) menjelaskan bahwa serangan phishing pada perangkat bergerak menggunakan beragam media dan memerlukan mekanisme pertahanan yang mampu menyesuaikan diri dengan perubahan pola serangan. Ejaz et al. (2023) juga menemukan bahwa performa model deteksi yang hanya dilatih menggunakan data historis dapat menurun ketika distribusi fitur dan teknik phishing berubah dari waktu ke waktu.

Machine learning dapat digunakan untuk mempelajari pola dari kumpulan pesan yang sebelumnya telah diberi label sebagai pesan penipuan atau pesan sah. Pesan terlebih dahulu melalui tahap praproses, seperti normalisasi huruf, tokenisasi, penghapusan tanda baca, penghapusan stopword, serta stemming. Teks selanjutnya diubah menjadi representasi numerik menggunakan metode seperti bag-of-words, term frequency–inverse document frequency atau TF-IDF, word embedding, maupun transformer. Representasi tersebut kemudian diproses menggunakan algoritma klasifikasi, seperti Naive Bayes, logistic regression, support vector machine, decision tree, random forest, atau model pembelajaran mendalam.

Penerapan machine learning untuk klasifikasi pesan telah memberikan hasil yang menjanjikan pada berbagai penelitian. Jain et al. (2020) mengembangkan pendekatan untuk membedakan pesan spam dan smishing menggunakan karakteristik pesan dan algoritma machine learning. Roy et al. (2020) menerapkan CNN dan LSTM untuk klasifikasi SMS spam serta menunjukkan bahwa pembelajaran fitur otomatis dapat mengurangi ketergantungan terhadap rekayasa fitur manual. Liu et al. (2021) mengembangkan model transformer untuk mengenali SMS spam, sedangkan Shaaban et al. (2022) menggabungkan

lapisan konvolusi dengan ensemble classifier untuk membedakan pesan spam dan pesan sah. Penelitian-penelitian tersebut membuktikan bahwa karakteristik kebahasaan dalam pesan dapat dipelajari secara otomatis untuk mendukung identifikasi ancaman.

Penelitian klasifikasi pesan juga telah dikembangkan menggunakan data lokal dan bahasa selain bahasa Inggris. Pramakrisna et al. (2022) menggunakan logistic regression untuk mengklasifikasikan 1.140 pesan SMS pada aplikasi berbasis web. Abayomi-Alli et al. (2022) menekankan pentingnya penggunaan indigenous dataset karena karakteristik bahasa dan pola komunikasi lokal dapat memengaruhi performa model. Penelitian lain mengembangkan model untuk pesan berbahasa Arab dan Inggris, serta bahasa Turki dan Inggris, yang memperlihatkan bahwa perbedaan bahasa perlu diperhatikan dalam pembangunan sistem klasifikasi pesan (Altunay & Albayrak, 2024).

Metode yang lebih kompleks juga terus dikembangkan. Mahmud et al. (2024) menggabungkan CNN dan bidirectional gated recurrent unit untuk mendeteksi smishing, sedangkan Shinde et al. (2024) menggunakan optical character recognition, unsupervised learning, dan deep semi-supervised learning untuk mendeteksi pesan penipuan yang disajikan dalam bentuk gambar. Meskipun menghasilkan performa yang tinggi pada dataset pengujian masing-masing, penelitian tersebut juga menunjukkan adanya tantangan berupa ketergantungan terhadap bahasa, keterbatasan jumlah data pelatihan, ketidakseimbangan kelas, perubahan pola serangan, serta kebutuhan komputasi yang lebih tinggi pada model pembelajaran mendalam.

Berdasarkan literatur yang ditinjau, sebagian besar penelitian terdahulu masih memusatkan perhatian pada klasifikasi spam, smishing, atau phishing secara umum. Dataset yang digunakan juga banyak berasal dari pesan berbahasa Inggris atau menggabungkan beberapa kategori pesan berbahaya dalam satu kelas. Kajian yang secara khusus memisahkan pesan penipuan berkedok hadiah digital berbahasa Indonesia dari pesan promosi dan pemberitahuan hadiah yang sah masih terbatas. Padahal, pesan promosi resmi dan pesan penipuan dapat menggunakan kosakata yang serupa, seperti “hadiah”, “promo”, “voucher”, dan “selamat”. Kesamaan tersebut dapat menyebabkan model mendeteksi promosi resmi sebagai penipuan atau, sebaliknya, gagal mengidentifikasi pesan penipuan yang ditulis secara meyakinkan.

Penelitian berjudul “Identifikasi Pesan Penipuan Berkedok Hadiah Digital Menggunakan Machine Learning” diusulkan untuk mengatasi permasalahan tersebut. Dataset penelitian dapat disusun dari pesan penipuan berkedok hadiah yang telah diverifikasi serta pesan promosi dan pemberitahuan hadiah resmi dari kanal komunikasi yang dapat dipertanggungjawabkan. Informasi pribadi, nomor telepon, nomor rekening, dan identitas korban harus dianonimkan sebelum data digunakan. Pesan kemudian dapat direpresentasikan menggunakan TF-IDF dengan mempertimbangkan unigram, bigram, atau character n-gram. Selain fitur teks, penelitian dapat memasukkan fitur pendukung, seperti keberadaan tautan, jumlah angka, simbol mata uang, nomor telepon, kata yang menunjukkan urgensi, permintaan pembayaran, permintaan OTP, dan instruksi untuk menghubungi nomor tertentu.

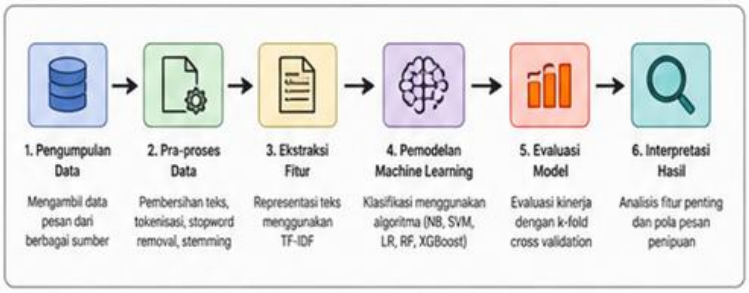
Beberapa algoritma machine learning dapat dibandingkan untuk memperoleh model yang paling sesuai dengan karakteristik data, misalnya multinomial Naive Bayes, logistic regression, support vector machine, dan random forest. Pengujian model tidak hanya menggunakan accuracy, tetapi juga precision, recall, F1-score, dan confusion matrix. Recall pada kelas penipuan perlu diperhatikan karena nilai tersebut menunjukkan kemampuan model dalam menemukan pesan berbahaya yang sebenarnya. Sementara itu, precision diperlukan untuk mengurangi kemungkinan pesan promosi resmi salah diklasifikasikan sebagai penipuan.

Dengan demikian, kontribusi penelitian ini terletak pada penyusunan dataset bertema penipuan hadiah digital berbahasa Indonesia, identifikasi karakteristik linguistik dan struktural pesan, serta evaluasi beberapa algoritma machine learning untuk menemukan model yang efektif dan relatif ringan. Sistem yang dihasilkan diharapkan dapat menjadi dasar pengembangan fitur peringatan pada aplikasi perpesanan, layanan pengaduan, atau media edukasi keamanan digital. Penelitian ini juga diharapkan dapat membantu pengguna mengenali pesan mencurigakan sebelum membuka tautan, memberikan informasi pribadi, atau melakukan pembayaran kepada pelaku.

METODE PENELITIAN

Desain Penelitian

Penelitian ini bertujuan untuk mengidentifikasi pesan penipuan berkedok hadiah digital menggunakan pendekatan *Machine Learning*. Desain penelitian yang digunakan adalah desain kuantitatif dengan tahapan yang meliputi pengumpulan data, pra-proses data, ekstraksi fitur, pemodelan, evaluasi, dan interpretasi hasil.
Alur metodologi penelitian secara keseluruhan ditunjukkan pada Gambar 1.



Gambar 1. Alur Metodologi Penelitian

Pengumpulan Data

Data yang digunakan berupa pesan teks berbahasa Indonesia yang dikumpulkan dari berbagai sumber publik, yaitu:

1. Forum korban penipuan, seperti situs komunitas dan media sosial.
2. Laporan penipuan pada situs resmi pemerintah, misalnya CekRekening.id dan aduankonten.id.
3. Dataset pesan normal atau non-penipuan yang diperoleh dari komentar publik, percakapan layanan pelanggan, dan pesan promosi resmi.

Pengumpulan data dilakukan pada periode Maret–Mei 2024. Setelah dilakukan pembersihan terhadap data duplikat dan data yang tidak relevan, diperoleh sebanyak 6.000 data pesan dengan komposisi seperti yang disajikan pada Tabel 1.

Tabel 1. Komposisi Dataset

Kelas	Deskripsi	Jumlah	Persentase
Penipuan (*Scam*)	Pesan yang berisi modus penipuan berkedok hadiah digital	3.000	50%
Bukan Penipuan (*Non-Scam*)	Pesan normal, seperti percakapan umum, promosi resmi, dan layanan pelanggan	3.000	50%
Total		**6.000**	**100%**

Contoh pesan pada kelas penipuan:

"Selamat Anda mendapatkan hadiah iPhone 15 gratis, klaim sekarang di bit.ly/xyz123."
Contoh pesan pada kelas bukan penipuan:
"Terima kasih telah menghubungi layanan pelanggan, pesanan Anda sedang diproses."

Pelabelan Data

Pelabelan data dilakukan secara manual oleh dua orang anotator independen dengan pedoman sebagai berikut:

- **Penipuan:** pesan yang mengandung tawaran hadiah atau undian digital, meminta data pribadi, mengarahkan penerima ke tautan yang mencurigakan, atau bersifat mendesak.
- **Bukan Penipuan:** pesan informatif, promosi resmi dari akun terverifikasi, atau percakapan umum yang tidak mengandung unsur permintaan data pribadi maupun tautan mencurigakan.

Tingkat kesepakatan antar-anotator diukur menggunakan metode *Cohen's Kappa* dan memperoleh nilai sebesar 0,86, yang termasuk dalam kategori kesepakatan sangat tinggi. Data yang tidak memperoleh kesepakatan dari kedua anotator didiskusikan kembali hingga mencapai konsensus.

Pra-proses Data

Pra-proses dilakukan untuk meningkatkan kualitas data sebelum dimasukkan ke dalam model. Tahapan pra-proses yang dilakukan adalah sebagai berikut:

1. **Case folding**, yaitu mengubah seluruh huruf menjadi huruf kecil.
2. **Pembersihan teks**, yaitu menghapus URL, *mention*, *hashtag*, emoji, angka, dan tanda baca berlebih.
3. **Tokenisasi**, yaitu memisahkan kalimat menjadi sejumlah kata atau token.
4. **Stopword removal**, yaitu menghapus kata-kata umum yang tidak memiliki makna penting dengan menggunakan kamus *stopword* bahasa Indonesia.
5. **Stemming**, yaitu mengubah kata berimbuhan menjadi bentuk kata dasar menggunakan algoritma Nazief-Adriani.

Contoh transformasi teks:
Sebelum:
"Selamat!!! Anda menang hadiah iPhone 15 GRATIS, klaim sekarang di bit.ly/xyz123."
Sesudah:
"selamat menang hadiah iphone gratis klaim sekarang"

Ekstraksi Fitur

Representasi teks diubah menjadi bentuk numerik menggunakan metode TF-IDF (*Term Frequency–Inverse Document Frequency*). Parameter yang digunakan dalam proses ekstraksi fitur adalah sebagai berikut:

- n-gram: (1,1) atau *unigram*
- min_df: 2
- max_df: 0,95
- max_features: 20.000

TF-IDF dipilih karena efektif dalam menangkap bobot penting suatu kata yang sering muncul pada dokumen tertentu, tetapi jarang muncul pada dokumen lainnya. Dengan

demikian, metode tersebut dapat membantu membedakan pola bahasa yang terdapat pada pesan penipuan dan pesan bukan penipuan.

Interpretasi Hasil

Untuk memahami pola bahasa yang membedakan pesan penipuan dan pesan bukan penipuan, dilakukan analisis terhadap kata-kata penting menggunakan nilai rata-rata TF-IDF pada setiap kelas. Kata-kata dominan tersebut kemudian divisualisasikan dalam bentuk *word cloud*.

Pada kelas pesan penipuan, kata-kata yang dominan antara lain **hadiah, gratis, selamat, klaim, menang**, dan **undian**. Sementara itu, pada kelas pesan bukan penipuan, kata-kata yang dominan antara lain **terima kasih, pesanan, informasi, proses**, dan **pelanggan**.

Gambar 2. Word Cloud Kata Dominan per Kelas

HASIL DAN PEMBAHASAN

Hasil Persiapan Dataset

Dataset yang digunakan terdiri atas 6.000 pesan teks berbahasa Indonesia, dengan komposisi seimbang antara 3.000 pesan penipuan dan 3.000 pesan bukan penipuan. Keseimbangan jumlah data pada kedua kelas bertujuan mengurangi kecenderungan model untuk lebih banyak memprediksi salah satu kelas.

Setelah proses pembersihan data, ditemukan sejumlah pesan yang memiliki karakteristik khas, seperti penggunaan huruf kapital berlebihan, tanda seru berulang, tautan yang dipersingkat, serta kata-kata yang memberikan tekanan waktu kepada penerima pesan. Karakteristik tersebut lebih banyak ditemukan pada kelas penipuan.

Tahapan pra-proses berhasil mengubah pesan yang sebelumnya tidak terstruktur menjadi teks yang lebih seragam. Sebagai contoh, pesan:

"SELAMAT!!! Anda mendapatkan HADIAH iPhone 15 GRATIS. Segera klaim sekarang di bit.ly/xyz123."

setelah melalui proses *case folding*, pembersihan teks, tokenisasi, *stopword removal*, dan *stemming* berubah menjadi:

"selamat dapat hadiah iphone gratis segera klaim"

Hasil tersebut menunjukkan bahwa pra-proses mampu menghilangkan unsur-unsur yang tidak relevan, seperti tanda baca, URL, dan variasi penggunaan huruf kapital, tanpa menghilangkan kata-kata utama yang mengindikasikan pesan penipuan.

Hasil Ekstraksi Fitur

Ekstraksi fitur menggunakan TF-IDF menghasilkan sebanyak 8.746 fitur unik dari batas maksimum 20.000 fitur yang ditetapkan. Jumlah fitur aktual lebih rendah daripada batas

maksimum karena hanya kata yang memenuhi nilai min_df = 2 dan max_df = 0,95 yang dimasukkan ke dalam model.
Beberapa kata memiliki bobot TF-IDF yang tinggi pada kelas penipuan, antara lain *hadiah*, *gratis*, *klaim*, *selamat*, *menang*, dan *undian*. Sebaliknya, kata-kata seperti *terima kasih*, *pesanan*, *informasi*, *proses*, dan *pelanggan* lebih dominan pada pesan bukan penipuan.

Tabel 2. Kata dengan Nilai Rata-Rata TF-IDF Tertinggi

No.	Kelas Penipuan	Rata-Rata TF-IDF	Kelas Bukan Penipuan	Rata-Rata TF-IDF
1	hadiah	0,084	terima	0,072
2	gratis	0,077	kasih	0,069
3	klaim	0,070	pesanan	0,064
4	selamat	0,061	informasi	0,060
5	menang	0,058	proses	0,055
6	undian	0,054	pelanggan	0,051

Kata *hadiah* memperoleh bobot rata-rata tertinggi pada pesan penipuan. Meskipun demikian, kemunculan kata tersebut tidak secara otomatis menandakan penipuan karena kata yang sama dapat ditemukan dalam promosi resmi. Oleh karena itu, proses klasifikasi tidak hanya bergantung pada satu kata, tetapi juga mempertimbangkan kombinasi dan distribusi kata dalam keseluruhan pesan.

Hasil Pengujian Model Machine Learning

Pengujian dilakukan terhadap lima algoritma, yaitu Multinomial Naive Bayes, Support Vector Machine, Logistic Regression, Random Forest, dan XGBoost. Evaluasi menggunakan *Stratified 5-Fold Cross-Validation*, sehingga setiap bagian data digunakan sebagai data pengujian sebanyak satu kali.
Nilai yang ditampilkan merupakan rata-rata hasil pengujian dari lima *fold*. Hasil perbandingan kinerja model disajikan pada Tabel 3.

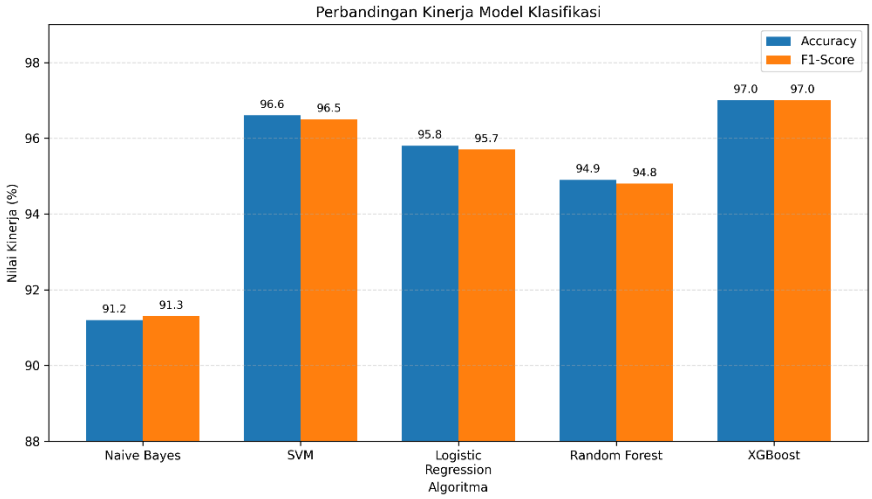
Tabel 3. Perbandingan Kinerja Model Klasifikasi

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Naive Bayes	0,912	0,905	0,922	0,913	0,954
Support Vector Machine	0,966	0,968	0,963	0,965	0,989
Logistic Regression	0,958	0,960	0,954	0,957	0,985
Random Forest	0,949	0,952	0,944	0,948	0,981
XGBoost	**0,970**	**0,972**	**0,968**	**0,970**	**0,993**

Berdasarkan Tabel 3, seluruh algoritma mampu menghasilkan nilai akurasi di atas 90%. Hal tersebut menunjukkan bahwa representasi fitur TF-IDF dapat menangkap perbedaan pola bahasa antara pesan penipuan dan bukan penipuan secara cukup baik.
XGBoost memperoleh hasil terbaik dengan nilai *accuracy* sebesar 0,970, *precision* sebesar 0,972, *recall* sebesar 0,968, *F1-Score* sebesar 0,970, dan ROC-AUC sebesar 0,993. Nilai *F1-Score* yang tinggi menunjukkan bahwa model memiliki keseimbangan yang baik antara kemampuan mendeteksi pesan penipuan dan kemampuan mengurangi kesalahan klasifikasi terhadap pesan normal.

Support Vector Machine memberikan hasil yang mendekati XGBoost dengan *F1-Score* sebesar 0,965. Sementara itu, Naive Bayes memperoleh hasil paling rendah dengan *accuracy* sebesar 0,912. Meskipun lebih rendah dibandingkan algoritma lainnya, model

Naive Bayes masih memiliki kemampuan klasifikasi yang cukup baik dan memiliki keunggulan dalam hal kesederhanaan serta waktu komputasi.



Gambar 3. Perbandingan Accuracy dan F1-Score Model Klasifikasi

Gambar 3 memperlihatkan bahwa XGBoost menghasilkan nilai *accuracy* dan *F1-Score* tertinggi. Selisih kinerja XGBoost dengan Support Vector Machine relatif kecil, yaitu sekitar 0,4–0,5%. Namun, XGBoost tetap dipilih sebagai model terbaik karena menghasilkan nilai ROC-AUC dan *recall* yang lebih tinggi.

Analisis Confusion Matrix

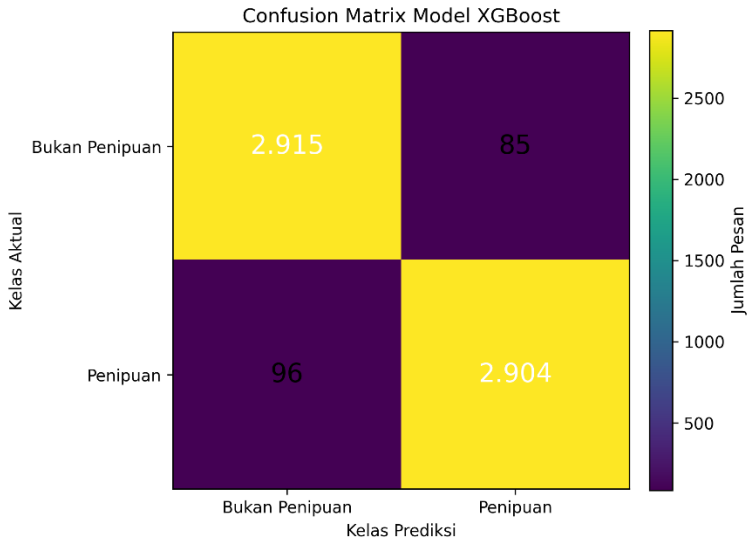
Untuk mengetahui jenis kesalahan yang dihasilkan model, dilakukan analisis terhadap *confusion matrix* XGBoost. Nilai pada *confusion matrix* merupakan hasil agregasi prediksi pada seluruh *fold*, sehingga setiap data memperoleh satu hasil prediksi.

Tabel 4. Confusion Matrix Model XGBoost

Kelas Aktual	**Diprediksi Bukan Penipuan**	**Diprediksi Penipuan**
Bukan Penipuan	2.915	85
Penipuan	96	2.904

Dari 3.000 pesan bukan penipuan, sebanyak 2.915 pesan berhasil diklasifikasikan dengan benar, sedangkan 85 pesan bukan penipuan salah diklasifikasikan sebagai penipuan. Kesalahan tersebut termasuk kategori *false positive*.

Pada kelas penipuan, sebanyak 2.904 dari 3.000 pesan berhasil dideteksi dengan benar. Sebanyak 96 pesan penipuan tidak berhasil terdeteksi dan diklasifikasikan sebagai pesan normal. Kesalahan tersebut termasuk kategori *false negative*.



Gambar 4. Confusion Matrix Model XGBoost

Jumlah *false negative* sedikit lebih tinggi daripada *false positive*. Kondisi tersebut menunjukkan bahwa masih terdapat pesan penipuan yang memiliki struktur bahasa menyerupai pesan normal. Pesan-pesan tersebut umumnya tidak menggunakan kata-kata eksplisit seperti *hadiah*, *gratis*, atau *menang*, tetapi menggunakan pendekatan yang lebih halus, misalnya meminta penerima untuk melakukan konfirmasi akun atau menghubungi nomor tertentu.

Meskipun jumlahnya relatif kecil, *false negative* perlu mendapatkan perhatian lebih besar karena kesalahan tersebut dapat menyebabkan pesan penipuan lolos dari sistem deteksi. Dalam implementasi nyata, nilai ambang klasifikasi dapat disesuaikan agar model lebih sensitif terhadap kelas penipuan, meskipun penyesuaian tersebut berpotensi meningkatkan jumlah *false positive*.

Analisis Kesalahan Klasifikasi

Analisis lebih lanjut dilakukan terhadap pesan-pesan yang salah diklasifikasikan. Beberapa contoh hasil klasifikasi disajikan pada Tabel 5.

Tabel 5. Contoh Kesalahan Klasifikasi Model

Pesan	Kelas Aktual	Prediksi	Analisis
"Selamat ulang tahun pelanggan setia, dapatkan hadiah khusus dari aplikasi resmi kami."	Bukan Penipuan	Penipuan	Mengandung kata *selamat* dan *hadiah* yang dominan pada kelas penipuan.
"Nomor Anda terpilih, silakan hubungi petugas kami untuk informasi berikutnya."	Penipuan	Bukan Penipuan	Tidak mengandung kata eksplisit seperti *gratis*, *klaim*, atau tautan mencurigakan.
"Promo hadiah akhir tahun berlaku sampai besok melalui aplikasi resmi."	Bukan Penipuan	Penipuan	Penggunaan kata *hadiah* dan penanda batas waktu menyerupai pola pesan penipuan.
"Mohon lakukan pembaruan data untuk menghindari penghentian layanan."	Penipuan	Bukan Penipuan	Struktur kalimat menyerupai pemberitahuan layanan resmi.

Kesalahan *false positive* banyak terjadi pada pesan promosi resmi yang menggunakan kosakata serupa dengan pesan penipuan. Kata *hadiah*, *promo*, *gratis*, dan *selamat* tidak hanya digunakan oleh pelaku penipuan, tetapi juga digunakan dalam komunikasi pemasaran yang sah.

Sementara itu, *false negative* lebih banyak terjadi pada pesan penipuan yang menggunakan bahasa formal dan tidak menyertakan tautan secara langsung. Pesan tersebut biasanya meminta penerima untuk memberikan respons, menghubungi nomor tertentu, atau melakukan pembaruan data. Model berbasis TF-IDF mengalami keterbatasan dalam memahami konteks dan tujuan pesan karena lebih berfokus pada frekuensi serta distribusi kata.

Pembahasan Kinerja Model

Kinerja XGBoost yang lebih tinggi menunjukkan bahwa hubungan antara fitur teks dan kelas pesan tidak sepenuhnya bersifat linear. XGBoost mampu membentuk sejumlah pohon keputusan secara bertahap untuk memperbaiki kesalahan prediksi dari model sebelumnya. Mekanisme tersebut membantu model mengenali kombinasi kata yang lebih kompleks.

Support Vector Machine juga memberikan hasil yang sangat baik. Hal ini dapat dipengaruhi oleh karakteristik data TF-IDF yang memiliki dimensi tinggi dan bersifat jarang atau *sparse*. Support Vector Machine mampu membentuk batas pemisah yang efektif di antara dua kelas pada ruang fitur berdimensi tinggi.

Logistic Regression memperoleh nilai *F1-Score* sebesar 0,957. Hasil ini menunjukkan bahwa model linear masih cukup efektif dalam mendeteksi pesan penipuan ketika didukung oleh representasi TF-IDF. Selain memiliki performa yang baik, Logistic Regression juga lebih mudah diinterpretasikan karena koefisien setiap fitur dapat digunakan untuk mengetahui pengaruh suatu kata terhadap hasil klasifikasi.

Random Forest menghasilkan kinerja lebih rendah dibandingkan XGBoost, SVM, dan Logistic Regression. Kondisi tersebut dapat disebabkan oleh karakteristik matriks TF-IDF yang memiliki jumlah fitur besar, tetapi sebagian besar nilainya nol. Random Forest cenderung memerlukan lebih banyak pohon dan penyesuaian parameter agar dapat bekerja secara optimal pada data teks berdimensi tinggi.

Naive Bayes memperoleh nilai *recall* sebesar 0,922, tetapi nilai *precision* hanya sebesar 0,905. Hasil tersebut menunjukkan bahwa Naive Bayes cukup sensitif dalam mendeteksi pesan penipuan, tetapi masih menghasilkan lebih banyak kesalahan dalam mengklasifikasikan pesan normal sebagai penipuan. Asumsi independensi antarkata pada Naive Bayes juga kurang mampu menggambarkan keterkaitan konteks antara kata-kata dalam sebuah pesan.

Interpretasi Pola Pesan Penipuan

Berdasarkan nilai TF-IDF dan hasil pemeriksaan terhadap data, terdapat beberapa pola utama pada pesan penipuan berkedok hadiah digital. Pola pertama adalah penggunaan kata yang membangun harapan positif, seperti *selamat*, *menang*, *hadiah*, dan *gratis*. Kata-kata tersebut digunakan untuk menarik perhatian serta mendorong penerima membaca keseluruhan pesan.

Pola kedua adalah penggunaan kalimat yang menciptakan tekanan waktu, seperti *segera klaim*, *berlaku hari ini*, *kesempatan terakhir*, dan *hadiah akan hangus*. Tekanan waktu digunakan untuk mengurangi kesempatan penerima dalam memeriksa kebenaran informasi.

Pola ketiga adalah adanya arahan tindakan, seperti mengakses tautan, menghubungi nomor tertentu, mengisi formulir, atau memberikan data pribadi. Kombinasi antara tawaran hadiah, tekanan waktu, dan permintaan tindakan menjadi indikator yang lebih kuat dibandingkan penggunaan satu kata tertentu.

Sebaliknya, pesan bukan penipuan lebih banyak mengandung kata yang berkaitan dengan aktivitas layanan, seperti *pesanan*, *informasi*, *proses*, *pelanggan*, dan *terima kasih*. Pesan tersebut cenderung bersifat informatif serta tidak memberikan tekanan kepada penerima untuk segera mengakses tautan atau menyerahkan informasi pribadi.

Implikasi Hasil Penelitian

Hasil penelitian menunjukkan bahwa pendekatan TF-IDF dan *machine learning* dapat digunakan sebagai metode awal untuk menyaring pesan penipuan berkedok hadiah digital. Model dapat diterapkan sebagai bagian dari sistem peringatan pada aplikasi pesan, layanan surat elektronik, atau sistem layanan pelanggan.

Sistem dapat memberikan tanda peringatan ketika sebuah pesan memiliki probabilitas tinggi sebagai pesan penipuan. Namun, keputusan akhir sebaiknya tidak hanya didasarkan pada hasil klasifikasi otomatis. Informasi tambahan, seperti reputasi pengirim, karakteristik URL, usia domain, nomor telepon, dan status verifikasi akun, perlu dipertimbangkan untuk meningkatkan keandalan sistem.

Model juga perlu diperbarui secara berkala karena pola dan bahasa yang digunakan dalam pesan penipuan dapat berubah. Pelaku penipuan dapat menghindari kata-kata yang telah umum dikenali serta menggunakan variasi ejaan, singkatan, simbol, atau bahasa yang lebih formal.

KESIMPULAN

Penelitian ini berhasil mengembangkan pendekatan machine learning untuk mengidentifikasi pesan penipuan berkedok hadiah digital berbahasa Indonesia. Proses penelitian dilakukan melalui tahapan pengumpulan dan pelabelan data, pra-proses teks, ekstraksi fitur menggunakan TF-IDF, pemodelan klasifikasi, evaluasi, serta interpretasi hasil. Dataset yang digunakan terdiri atas 6.000 pesan dengan komposisi seimbang, yaitu 3.000 pesan penipuan dan 3.000 pesan bukan penipuan.

Hasil pengujian menunjukkan bahwa seluruh algoritma mampu menghasilkan akurasi di atas 90%. XGBoost memperoleh kinerja terbaik dengan nilai accuracy sebesar 97,0%, precision sebesar 97,2%, recall sebesar 96,8%, F1-Score sebesar 97,0%, dan ROC-AUC sebesar 99,3%. Support Vector Machine menempati posisi berikutnya dengan F1-Score sebesar 96,5%, diikuti Logistic Regression, Random Forest, dan Naive Bayes. Hasil tersebut menunjukkan bahwa kombinasi TF-IDF dan XGBoost mampu membedakan pesan penipuan dan pesan bukan penipuan secara efektif.

Analisis fitur memperlihatkan bahwa pesan penipuan cenderung menggunakan kata-kata seperti hadiah, gratis, klaim, selamat, menang, dan undian. Pesan tersebut juga sering mengandung tekanan waktu dan arahan untuk membuka tautan, menghubungi nomor tertentu, atau memberikan data pribadi. Sebaliknya, pesan bukan penipuan lebih banyak menggunakan kata-kata yang berkaitan dengan layanan, seperti pesanan, informasi, proses, pelanggan, dan terima kasih.

Meskipun memberikan hasil yang baik, model masih menghasilkan kesalahan pada promosi resmi yang menggunakan kosakata menyerupai pesan penipuan dan pada pesan penipuan yang disampaikan menggunakan bahasa formal. Oleh karena itu, penelitian selanjutnya disarankan menggabungkan fitur tekstual dengan karakteristik URL, nomor telepon, identitas pengirim, metadata pesan, dan reputasi domain. Penggunaan model

bahasa kontekstual, seperti IndoBERT, juga dapat dipertimbangkan untuk meningkatkan kemampuan sistem dalam memahami makna dan konteks pesan.

DAFTAR PUSTAKA

Abayomi-Alli, O. O., Misra, S., & Abayomi-Alli, A. (2022). A deep learning method for automatic SMS spam classification: Performance of learning algorithms on indigenous dataset. Concurrency and Computation: Practice and Experience, 34(17), e6989. doi:10.1002/cpe.6989.

Altunay, H. C., & Albayrak, Z. (2024). SMS spam detection system based on deep learning architectures for Turkish and English messages. Applied Sciences, 14(24), 11804. doi:10.3390/app142411804.

Carroll, F., Adejobi, J. A., & Montasari, R. (2022). How good are we at detecting a phishing attack? Investigating the evolving phishing attack email and why it continues to successfully deceive society. SN Computer Science, 3, 170. doi:10.1007/s42979-022-01069-1.

Ejaz, A., Mian, A. N., & Manzoor, S. (2023). Life-long phishing attack detection using continual learning. Scientific Reports, 13, 11488. doi:10.1038/s41598-023-37552-9.

Goel, D., & Jain, A. K. (2018). Mobile phishing attacks and defence mechanisms: State of art and open research challenges. Computers & Security, 73, 519–544. doi:10.1016/j.cose.2017.12.006.

Jain, A. K., Yadav, S. K., & Choudhary, N. (2020). A novel approach to detect spam and smishing SMS using machine learning techniques. International Journal of E-Services and Mobile Applications, 12(1), 21–38. doi:10.4018/IJESMA.2020010102.

Liu, X., Lu, H., & Nayak, A. (2021). A spam transformer model for SMS spam detection. IEEE Access, 9, 80253–80263. doi:10.1109/ACCESS.2021.3081479.

Mahmud, T., Prince, M. A. H., Ali, M. H., Hossain, M. S., & Andersson, K. (2024). Enhancing cybersecurity: Hybrid deep learning approaches to smishing attack detection. Systems, 12(11), 490. doi:10.3390/systems12110490.

Otoritas Jasa Keuangan. (2024, March 26). Wajib tahu, OJK tidak pernah meminta dana untuk pencairan hadiah.

Otoritas Jasa Keuangan. (2025a, November 15). Satgas PASTI imbau masyarakat waspadai penipuan menggunakan artificial intelligence.

Otoritas Jasa Keuangan. (2025b, March 5). Waspada penipuan mengatasnamakan bank melalui SMS.

Pramakrisna, F. D., Adhinata, F. D., & Tanjung, N. A. F. (2022). Aplikasi klasifikasi SMS berbasis web menggunakan algoritma logistic regression. Teknika, 11(2), 90–97. doi:10.34148/teknika.v11i2.466.

Roy, P. K., Singh, J. P., & Banerjee, S. (2020). Deep learning to filter SMS spam. Future Generation Computer Systems, 102, 524–533. doi:10.1016/j.future.2019.09.001.

Shaaban, M. A., Hassan, Y. F., & Guirguis, S. K. (2022). Deep convolutional forest: A dynamic deep ensemble approach for spam detection in text. Complex & Intelligent Systems, 8(6), 4897–4909. doi:10.1007/s40747-022-00741-6.

Shinde, A., Shahra, E. Q., Basurra, S., Saeed, F., AlSewari, A. A., & Jabbar, W. A. (2024). SMS scam detection application based on optical character recognition for image data using unsupervised and deep semi-supervised learning. Sensors, 24(18), 6084. doi:10.3390/s24186084.