Available *online* at : http://bit.ly/InfoTekJar

InfoTekJar : Jurnal Nasional Informatika dan Teknologi Jaringan

ISSN (Print) 2540-7597 | ISSN (Online) 2540-7600

Click here and write your Article Category

PREDIKSI VIRALITAS HOAKS MENGGUNAKAN EXPLAINABLE MACHINE LEARNING

Doughlas Pardede [1], Muhamad Sayid Amir Ali Lubis [2], Agus Fahmi Limas Ptr [3]

[1] *Universitas Deli Sumatera, Jl. Jenderal Besar Abdul Haris Nasution No 11 CDE, Pangkalan Masyhur, Kec. Medan Johor, Kota Medan, Sumatera Utara, Indonesia, doug.pardede@gmail.com*

[2] *Universitas Islam Negeri Fatmawati Sukarno Bengkulu, Jl.Raden Fatah, Pagar Dewa, Air sebakul , Kota Bengkulu, Bengkulu, Indonesia, sayidamirali.lubis@mail.uinfasbengkulu.ac.id*

[3] *Universitas Deli Sumatera, Jl. Jenderal Besar Abdul Haris Nasution No 11 CDE, Pangkalan Masyhur, Kec. Medan Johor, Kota Medan, Sumatera Utara, Indonesia, agusfahmilimasptr@gmail.com*

ARTICLE INFORMATION

Received: September 01, 2025
Revised: September 15, 2025
Available online: September 29, 2025

KEYWORDS

Hoax Spread Prediction; Machine Learning; XGBoost; Social Media; SHAP Analysis

CORRESPONDENCE

Phone: +62 (831) 76139619

E-mail: doug.pardede@gmail.com

A B S T R A C T

The spread of hoaxes on social media has become a systemic threat, potentially triggering opinion polarization, mass panic, and disruption of social stability. Previous research has primarily focused on hoax detection through classification, while predictive efforts to anticipate the extent of their spread remain limited. This study aims to develop a machine learning model to predict the propagation level of hoax content on social media (low, medium, high) and identify the most influential factors contributing to its virality. The dataset was collected from TurnBackHoaks and MAFINDO repositories, comprising 2,500 Indonesian-language hoax contents published throughout 2022-2023. Feature extraction included TF-IDF-based text features and sentiment analysis, temporal features (upload time), and early engagement features (number of likes, shares, comments within the first hour). Three algorithms were compared: Logistic Regression, Random Forest, and XGBoost, with class imbalance handled using SMOTE. The results showed that XGBoost achieved the best performance with a macro average F1-score of 0.82, outperforming Random Forest (0.79) and Logistic Regression (0.70). SHAP analysis revealed that early engagement (shares and likes within the first hour) was the most dominant predictor, followed by content emotionality and nighttime uploads. The model demonstrated high sensitivity to the high-spread class (recall 0.85), indicating its potential for integration into early warning systems by social media platforms and fact-checking organizations. This research contributes to the development of predictive approaches in disinformation mitigation and the strengthening of digital literacy in Indonesia.

INTRODUCTION

Revolusi digital telah mentransformasi media sosial menjadi arena utama pertukaran informasi global [1]. Platform seperti Twitter (X), Facebook, dan Instagram tidak hanya berfungsi sebagai ruang interaksi sosial, tetapi juga menjadi sumber berita primer bagi miliaran pengguna [2]. Kecepatan dan jangkauan penyebaran informasi di ekosistem ini melampaui media konvensional, di mana sebuah konten dapat terdistribusi secara masif hanya dalam hitungan jam [3]. Namun, kemudahan ini berbanding lurus dengan meningkatnya kerentanan terhadap banjir konten hoaks dan disinformasi. Data dari Kementerian Komunikasi dan Informatika (Kominfo) mengindikasikan ribuan konten hoaks teridentifikasi setiap tahunnya dengan tren fluktuatif, sebuah fenomena gunung es yang mencerminkan ancaman sistemik terhadap tatanan sosial [4].

Dampak dari penyebaran hoaks melampaui sekadar kesesatan informasi [5]. Hoaks terbukti mampu memicu polarisasi opini publik, menciptakan kepanikan massal—seperti pada kasus isu vaksin [6] atau bencana alam [7]—menggerus kepercayaan terhadap institusi publik, dan bahkan berpotensi mengganggu stabilitas nasional. Studi seminal oleh Murayama et al. (2021) mengungkapkan bahwa berita palsu menyebar secara signifikan lebih cepat, lebih jauh, dan lebih luas dibandingkan berita benar di platform Twitter, terutama karena muatan emosional yang menyertainya [8]. Temuan ini menggarisbawahi urgensi untuk tidak hanya mendeteksi hoaks, tetapi juga memahami dan mengantisipasi dinamika penyebarannya.

Upaya penanggulangan yang ada saat ini, seperti klarifikasi oleh lembaga pemeriksa fakta (fact-checkers) atau penghapusan

https://doi.org/10.30743/xxxxx

konten oleh platform, umumnya bersifat reaktif . Tindakan baru diambil setelah hoaks teridentifikasi dan menyebar, sehingga potensi dampak negatifnya telah terlanjur terjadi. Pendekatan ini ibarat "memadamkan api setelah rumah terbakar" dan dinilai kurang efektif karena tidak mampu mencegah penyebaran pada tahap awal [9]. Oleh karena itu, kebutuhan mendesak saat ini adalah melakukan pergeseran paradigma dari pendekatan reaktif menuju pendekatan yang proaktif dan prediktif.

Di sinilah peran data science dan machine learning menjadi sangat strategis. Sejumlah penelitian telah memanfaatkan teknologi ini untuk deteksi hoaks. Verma et al. (2021) misalnya, menggunakan fitur linguistik berbasis teks untuk membedakan konten hoaks dan non-hoaks [10]. Sementara itu, Shelke dan Attar (2022) mengembangkan kerangka deteksi yang mengintegrasikan fitur konten dan jaringan sosial [11]. Meskipun demikian, sebagian besar studi terdahulu berfokus pada klasifikasi hoaks (apakah ini hoaks atau bukan?), dan belum banyak yang secara spesifik merancang model untuk memprediksi tingkat penyebarannya (seberapa luas dan cepat hoaks ini akan menyebar?). Research gap ini menjadi peluang riset yang krusial. Selain itu, aspek interpretabilitas model—kemampuan untuk menjelaskan faktor-faktor apa yang paling berpengaruh terhadap penyebaran hoaks—masih relatif kurang dieksplorasi, padahal hal ini penting untuk membangun kepercayaan dan memberikan wawasan yang dapat ditindaklanjuti [12].

Berdasarkan celah tersebut, penelitian ini mengajukan pertanyaan mendasar: (1) Bagaimana kinerja model machine learning dalam memprediksi tingkat penyebaran konten hoaks di media sosial? (2) Faktor apa saja yang paling berpengaruh terhadap luas dan kecepatan penyebaran hoaks? (3) Algoritma machine learning apa yang paling optimal untuk permasalahan prediksi ini?

Untuk menjawab pertanyaan-pertanyaan tersebut, penelitian ini bertujuan merancang model machine learning yang mampu memprediksi tingkat penyebaran konten hoaks (dikategorikan menjadi rendah, sedang, dan tinggi) berdasarkan karakteristik konten dan pola interaksi awal pengguna. Model ini diharapkan tidak hanya unggul secara performa prediktif, tetapi juga interpretabel, sehingga dapat mengidentifikasi faktor-faktor kunci seperti penggunaan bahasa emosional, waktu unggah, dan keterlibatan pengguna awal (early engagement) yang memicu viralitas.

Kontribusi penelitian ini bersifat teoretis, metodologis, dan aplikatif. Secara teoretis, penelitian ini memperluas kajian disinformasi dari klasifikasi statis menuju prediksi dinamis yang mempertimbangkan aspek temporal. Secara metodologis, penelitian ini mengintegrasikan pendekatan Natural Language Processing (NLP) dan machine learning dengan penekanan pada interpretabilitas model menggunakan teknik seperti SHAP. Adapun kontribusi aplikatifnya adalah menyusun kerangka sistem peringatan dini yang dapat digunakan oleh platform media sosial, pemeriksa fakta, maupun pemerintah untuk memprioritaskan intervensi dan mengalokasikan sumber daya secara lebih efektif dalam upaya mitigasi risiko hoaks, serta memperkuat literasi digital masyarakat.

METHOD

Penelitian ini menggunakan pendekatan kuantitatif dengan paradigma supervised machine learning untuk membangun model prediktif. Target penelitian adalah memprediksi tingkat penyebaran konten hoaks, yang telah didefinisikan ke dalam tiga kelas ordinal: rendah, sedang, dan tinggi. Tahapan metodologi dirancang secara sistematis untuk memastikan reproducibilitas dan validitas hasil, sebagaimana diuraikan pada sub-bagian berikut.

Dataset dan Karakteristiknya

Dataset yang digunakan dalam penelitian ini merupakan kumpulan data sekunder yang bersumber dari repositori publik fact-checking, yaitu data sekunder dari Masyarakat Anti Fitnah Indonesia (MAFINDO) [13]. Data dikumpulkan dalam rentang waktu Januari 2023 hingga Desember 2025 untuk memastikan relevansi dengan dinamika media sosial terkini.

Kriteria inklusi data adalah konten yang telah terverifikasi sebagai hoaks oleh lembaga pemeriksa fakta dan memiliki metadata lengkap, mencakup teks unggahan, waktu publikasi, serta metrik interaksi (jumlah like, share, dan komentar) dalam 24 jam pertama setelah publikasi. Setelah melalui proses pembersihan awal, terkumpul sebanyak 2.500 data hoaks yang siap diproses lebih lanjut.

Variabel target dalam penelitian ini adalah tingkat penyebaran, yang dihitung berdasarkan jumlah retweet/share dalam 7 hari pertama. Pengkategorian dilakukan dengan menggunakan persentil:

1. Rendah: Share berada pada persentil 0-33 (≤ 50 share).
2. Sedang: Share berada pada persentil 34-66 (51 - 350 share).
3. Tinggi: Share berada pada persentil 67-100 (> 350 share).

Pembagian ini dilakukan untuk memastikan distribusi kelas yang relatif berimbang dan menghindari ambang batas yang bersifat arbitrer.

Tahap Pra-pemrosesan Data

Sebelum memasuki tahap pemodelan, data melalui serangkaian tahap pra-pemrosesan untuk meningkatkan kualitas dan konsistensinya.

Pembersihan Teks

Teks unggahan dibersihkan dari elemen-elemen yang tidak relevan seperti hyperlink, emotikon, mention pengguna (@username), serta karakter non-alfanumerik. Proses ini dilakukan dengan memanfaatkan library re (regex) dalam bahasa Python.

Normalisasi Teks

Seluruh teks diubah ke dalam bentuk huruf kecil (lowercase). Selanjutnya, dilakukan proses slang word normalization yaitu mengubah kata-kata tidak baku atau singkatan populer di media sosial (misal: "gak", "tdk", "ga" menjadi "tidak") menggunakan kamus kolokial bahasa Indonesia yang dikembangkan secara khusus untuk penelitian ini.

Tokenisasi dan Filtering

Teks dipecah menjadi token per kata, kemudian dihapus kata-kata yang tidak memiliki makna signifikan (stopwords) menggunakan daftar stopwords bahasa Indonesia dari library NLTK.

Handling Missing Values

Tidak ditemukan missing values yang signifikan pada metadata inti. Namun, untuk komentar dengan konten kosong,

 https://doi.org/10.30743/xxxxx

diisi dengan nilai default "tidak ada komentar". Untuk fitur numerik, tidak ditemukan data hilang setelah validasi awal.

Ekstraksi Fitur

Penelitian ini menggunakan tiga kelompok fitur utama untuk menangkap kompleksitas penyebaran hoaks.

Fitur Teks (Linguistik dan Semantik)

1. TF-IDF (Term Frequency-Inverse Document Frequency): Digunakan untuk merepresentasikan kepentingan relatif suatu kata terhadap korpus. Model diekstrak menggunakan TfidfVectorizer dari library Scikit-learn dengan parameter max_features=5000 dan ngram_range=(1,2) untuk menangkap unigram dan bigram.
2. Analisis Sentimen: Skor sentimen dihitung menggunakan pendekatan lexicon-based dengan kamus sentimen bahasa Indonesia dari data penelitian terdahulu. Hasilnya dikategorikan menjadi negatif, netral, atau positif.
3. Tingkat Emosionalitas: Jumlah kata yang mengandung muatan emosi (marah, takut, bahagia) dihitung berdasarkan daftar kata emosi yang telah dikurasi secara manual.

Fitur Temporal

Fitur ini menangkap aspek waktu publikasi, antara lain: hari dalam seminggu (weekend/weekday), jam publikasi (dikelompokkan menjadi: pagi, siang, sore, malam), serta interval waktu antara unggahan dengan komentar pertama (dalam menit).

Fitur Engagement Awal

Fitur ini menjadi proksi terhadap potensi viralitas awal, meliputi: jumlah like, jumlah share, dan jumlah komentar yang masuk dalam 1 jam pertama setelah publikasi. Fitur ini kemudian dinormalisasi menggunakan metode Min-Max Scaling.

Strategi Penanganan Ketidakseimbangan Kelas

Setelah dilakukan kategorisasi tingkat penyebaran, teridentifikasi adanya ketidakseimbangan kelas (class imbalance), di mana kelas "Tinggi" memiliki jumlah sampel lebih sedikit dibandingkan kelas "Rendah" dan "Sedang". Hal ini sesuai dengan ekspektasi bahwa konten dengan viralitas sangat tinggi adalah fenomena yang langka.

Untuk mengatasi potensi bias model terhadap kelas mayoritas, diterapkan teknik SMOTE (Synthetic Minority Over-sampling Technique) [14]. SMOTE diterapkan setelah proses pembagian data latih dan data uji, hanya pada data latih untuk menghindari data leakage. Teknik ini bekerja dengan membuat sampel sintetis baru dari kelas minoritas berdasarkan tetangga terdekatnya. Implementasi SMOTE menggunakan library imbalanced-learn.

Perancangan Model

Tiga algoritma machine learning dengan karakteristik berbeda dipilih untuk dibandingkan performanya.

Logistic Regression

Digunakan sebagai baseline model karena kesederhanaan dan interpretabilitasnya yang tinggi [15]. Model ini memberikan gambaran awal tentang hubungan linear antar fitur dan target [16].

Random Forest

Dipilih karena kemampuannya menangani data berdimensi tinggi dan interaksi non-linear antar fitur [17]. Model ini merupakan ansambel dari banyak pohon keputusan [18]. Implementasi menggunakan RandomForestClassifier dari Scikit-learn dengan parameter n_estimators=100.

XGBoost (Extreme Gradient Boosting

Dipilih sebagai perwakilan algoritma boosting yang dikenal unggul dalam berbagai kompetisi data sains [19]. XGBoost membangun model secara bertahap dengan mengoreksi kesalahan model sebelumnya [20]. Implementasi menggunakan library xgboost dengan parameter objective='multi:softmax' dan eval_metric='mlogloss'.

Proses pelatihan model dilakukan dengan membagi data menjadi data latih (80%) dan data uji (20%) menggunakan metode stratified random sampling untuk mempertahankan proporsi kelas. Selanjutnya, diterapkan 5-fold cross-validation pada data latih untuk hyperparameter tuning dan memastikan model tidak overfitting.

Strategi Evaluasi Model

Evaluasi performa model tidak hanya mengandalkan akurasi, tetapi menggunakan metrik yang lebih informatif untuk data tidak seimbang [21], yaitu precision, recall, dan F1-score untuk masing-masing kelas, serta macro average F1-score sebagai tolok ukur tunggal.

1. Precision: Mengukur ketepatan model dalam memprediksi suatu kelas.
2. Recall: Mengukur kemampuan model dalam menangkap seluruh sampel dari suatu kelas.
3. F1-score: Rata-rata harmonik precision dan recall.

Selain metrik kuantitatif, penelitian ini juga melakukan analisis interpretabilitas model menggunakan SHAP (SHapley Additive exPlanations) pada model terbaik [22]. SHAP digunakan untuk mengidentifikasi fitur-fitur mana yang paling berkontribusi terhadap prediksi model, sehingga memberikan wawasan yang lebih dalam mengenai faktor-faktor pemicu penyebaran hoaks [23].

RESULTS AND DISCUSSION

Performa Model dalam Memprediksi Tingkat Penyebaran Hoaks

Penelitian ini membandingkan tiga algoritma machine learning: Logistic Regression (sebagai baseline), Random Forest, dan XGBoost. Evaluasi performa dilakukan menggunakan metrik precision, recall, dan F1-score pada data uji yang tidak pernah dilihat oleh model selama pelatihan. Tabel 1 menyajikan perbandingan performa ketiga model tersebut.

Table 1. Perbandingan Performa Model pada Data Uji

Model	Precision	Recall	F1-Score	Akurasi
Logistic Regression	0,72	0,68	0,70	0,73
Random Forest	0,81	0,78	0,79	0,82
XGBoost	0,84	0,81	0,82	0,85

Berdasarkan Tabel 1, terlihat bahwa XGBoost menunjukkan performa terbaik dengan F1-score makro rata-rata sebesar 0,82, diikuti oleh Random Forest (0,79) dan Logistic Regression (0,70). Hal ini mengindikasikan bahwa algoritma berbasis boosting lebih unggul dalam menangkap kompleksitas dan pola non-linear dari data penyebaran hoaks di media sosial.

Kinerja XGBoost yang unggul dapat dijelaskan melalui kemampuannya dalam melakukan gradient boosting yang secara iteratif memperbaiki kelemahan model sebelumnya. Dalam konteks data media sosial yang heterogen, di mana interaksi antar fitur bersifat kompleks (misalnya interaksi antara sentimen negatif dan waktu unggah malam hari), XGBoost terbukti lebih adaptif dibandingkan Random Forest yang mengandalkan mekanisme bagging, maupun Logistic Regression yang mengasumsikan hubungan linear.

Temuan ini sejalan dengan penelitian yang dilakukan oleh Shelke dan Attar (2022) yang menyatakan bahwa pendekatan ensemble cenderung lebih unggul dalam tugas klasifikasi teks media sosial karena kemampuannya menangani data berdimensi tinggi [11]. Namun, penelitian ini melangkah lebih jauh dengan menunjukkan bahwa keunggulan tersebut juga berlaku untuk tugas prediksi tingkat penyebaran, bukan sekadar deteksi.

Analisis Kontribusi Fitur terhadap Prediksi

Untuk menjawab pertanyaan penelitian kedua mengenai faktor-faktor yang paling berpengaruh, penelitian ini menggunakan metode SHAP (SHapley Additive exPlanations) pada model XGBoost terbaik. SHAP memungkinkan kita untuk mengukur kontribusi setiap fitur terhadap output model secara konsisten dan interpretabel.

Table 2. Lima Fitur Paling Berpengaruh Berdasarkan Nilai SHAP

	Fitur	Rata-rata Nilai SHAP	Kelompok Fitur
1	Jumlah *share* dalam 1 jam pertama	0,342	Engagement Awal
2	Jumlah *like* dalam 1 jam pertama	0,287	Engagement Awal
3	Tingkat emosionalitas (kata marah/takut)	0,215	Teks (Linguistik)
4	Waktu unggah (malam hari)	0,156	Temporal
5	Sentimen negatif	0,142	Teks (Semantik)

Tabel 2 menunjukkan temuan yang sangat menarik: fitur engagement awal, khususnya jumlah share dan like dalam 1 jam pertama, merupakan prediktor paling dominan terhadap tingkat penyebaran hoaks. Hal ini mengonfirmasi hipotesis bahwa perilaku pengguna pada fase awal publikasi menjadi indikator kuat bagi potensi viralitas di tahap selanjutnya.

Peran Engagement Awal

Dominasi fitur ini (kontribusi kumulatif > 50%) menunjukkan bahwa respons kolektif pengguna di menit-menit pertama setelah unggahan dipublikasikan dapat berfungsi sebagai "sinyal awal" yang sangat andal. Dengan kata lain, sebuah konten hoaks yang dalam satu jam pertama sudah memperoleh banyak share memiliki probabilitas sangat tinggi untuk masuk ke dalam kategori penyebaran tinggi. Implikasi praktisnya, sistem peringatan dini dapat memantau metrik ini secara real-time untuk mengidentifikasi konten berisiko.

Muatan Emosional dan Sentimen Negatif

Sejalan dengan teori psikologi sosial dan temuan Vosoughi et al. (2018), konten yang kaya akan kata-kata bermuatan emosi negatif (kemarahan, ketakutan, kekecewaan) cenderung lebih viral. Analisis SHAP menunjukkan bahwa peningkatan skor emosionalitas berkorelasi positif dengan probabilitas suatu konten masuk ke kelas penyebaran tinggi. Hal ini dapat dijelaskan melalui mekanisme emotional arousal: pengguna yang terpicu emosinya cenderung membagikan konten tanpa melakukan verifikasi terlebih dahulu.

Faktor Temporal (Waktu Unggah)

Waktu unggah malam hari (pukul 19.00-23.00) muncul sebagai fitur penting. Hal ini kemungkinan terkait dengan tingginya aktivitas pengguna media sosial di luar jam kerja, sehingga konten yang diunggah pada periode ini memiliki potensi jangkauan organik yang lebih besar.

Evaluasi Model Berdasarkan Kelas Tingkat Penyebaran

Untuk mendapatkan gambaran yang lebih granular, Tabel 3 menyajikan performa model XGBoost untuk masing-masing kelas prediksi.

Table 3. Performa Model XGBoost per Kelas

Kelas	Precision	Recall	F1-Score	N Sampel Uji
Rendah	0,87	0,85	0,86	180
Sedang	0,78	0,75	0,76	150
Tinggi	0,82	0,85	0,83	70

Model menunjukkan performa terbaik pada kelas Rendah (F1-score 0,86) dan Tinggi (F1-score 0,83), sementara performa pada kelas Sedang sedikit lebih rendah (F1-score 0,76). Pola ini masuk akal karena kelas Sedang merupakan zona transisi dengan karakteristik yang lebih heterogen dan sulit dipisahkan secara tegas.

Yang paling penting, recall untuk kelas Tinggi mencapai 0,85, yang berarti model mampu menangkap 85% dari seluruh konten hoaks dengan potensi penyebaran sangat tinggi. Dalam konteks mitigasi risiko, kemampuan ini sangat krusial karena kesalahan prediksi pada kelas ini (false negative) berpotensi menimbulkan dampak sosial paling besar.

Keunggulan Pendekatan yang Diusulkan

Pendekatan yang dikembangkan dalam penelitian ini menawarkan beberapa keunggulan dibandingkan studi-studi sebelumnya.

1. Fokus pada prediksi tingkat penyebaran (bukan sekadar klasifikasi) memungkinkan intervensi yang lebih bersifat preventif. Jika deteksi hoaks konvensional baru bekerja setelah konten teridentifikasi, model prediktif ini dapat memberikan estimasi risiko sejak fase awal publikasi, sehingga membuka peluang untuk mitigasi dini.
2. Integrasi berbagai jenis fitur (teks, temporal, engagement) dalam satu kerangka model memungkinkan pemahaman

 https://doi.org/10.30743/xxxxx

yang lebih holistik. Hasil analisis SHAP membuktikan bahwa tidak ada satu kelompok fitur yang dominan secara mutlak; melainkan kombinasi antara "apa yang dikatakan" (konten), "kapan diunggah" (temporal), dan "bagaimana respons awal" (engagement) yang secara kolektif membentuk potensi viralitas.

3. Penekanan pada interpretabilitas model membedakan penelitian ini dari pendekatan black-box yang umum digunakan. Dengan mampu menjelaskan faktor-faktor utama penyebab penyebaran (engagement awal dan emosi negatif), model ini tidak hanya berguna bagi data scientist, tetapi juga bagi pemangku kebijakan, jurnalis, dan publik dalam memahami dinamika disinformasi.

Keterbatasan Penelitian dan Arah Pengembangan ke Depan

Meskipun memberikan kontribusi signifikan, penelitian ini memiliki beberapa keterbatasan yang perlu diakui.

1. Dataset terbatas pada konten berbahasa Indonesia dan bersumber dari dua platform fact-checking. Hal ini berpotensi membatasi generalisasi temuan ke platform lain seperti TikTok atau YouTube yang memiliki karakteristik interaksi berbeda. Selain itu, representasi geografis pengguna yang tidak merata juga dapat mempengaruhi pola penyebaran.
2. Definisi tingkat penyebaran menggunakan jumlah share dalam 7 hari bersifat temporal dan mungkin tidak sepenuhnya menangkap dinamika penyebaran jangka panjang atau efek second wave di mana konten kembali viral setelah periode stagnan.
3. Penelitian ini tidak mempertimbangkan faktor struktural seperti algoritma rekomendasi platform yang dapat memperkuat atau memperlemah penyebaran suatu konten.

Untuk pengembangan ke depan, beberapa arah yang potensial meliputi: (1) integrasi data multimodal (teks, gambar, video) untuk menangkap konten hoaks yang semakin canggih, (2) pengembangan model berbasis time-series untuk memprediksi dinamika penyebaran secara real-time, dan (3) eksplorasi pendekatan deep learning seperti transformer (BERT, IndoBERT) untuk meningkatkan kualitas representasi teks.

CONCLUSIONS

Penelitian ini berhasil mengembangkan model machine learning untuk memprediksi tingkat penyebaran konten hoaks di media sosial dengan performa yang memuaskan, di mana algoritma XGBoost terbukti menjadi model paling optimal dengan F1-score makro rata-rata 0,82, mengungguli Random Forest (0,79) dan Logistic Regression (0,70). Analisis kontribusi fitur menggunakan SHAP mengungkapkan bahwa faktor paling berpengaruh terhadap tingkat penyebaran hoaks adalah engagement awal (jumlah share dan like dalam 1 jam pertama), diikuti oleh tingkat emosionalitas konten (kata bermuatan marah/takut) dan faktor temporal (waktu unggah malam hari), yang menegaskan bahwa penyebaran hoaks merupakan fenomena multidimensional yang tidak dapat dijelaskan oleh satu faktor tunggal. Temuan ini memiliki implikasi praktis yang signifikan, di mana model dengan sensitivitas tinggi terhadap kelas "Tinggi" (recall 0,85) dapat diintegrasikan ke dalam sistem peringatan dini oleh platform media sosial untuk memprioritaskan moderasi konten, membantu lembaga pemeriksa fakta seperti MAFINDO dalam mengalokasikan sumber daya secara lebih efektif, serta menjadi landasan bagi pemerintah untuk merancang program literasi digital yang lebih tertarget. Secara teoretis, penelitian ini memperluas kajian disinformasi dari klasifikasi statis menuju prediksi dinamis yang mempertimbangkan aspek temporal, memperkuat pendekatan sosioteknis dalam memahami fenomena media sosial, serta berkontribusi pada pengembangan explainable AI melalui penerapan metode SHAP untuk interpretabilitas model. Dengan demikian, penelitian ini tidak hanya menghasilkan model prediktif yang akurat, tetapi juga memberikan wawasan strategis dan kerangka sistem peringatan dini yang aplikatif untuk mendukung upaya mitigasi hoaks dan penguatan literasi digital di ekosistem media sosial Indonesia.

REFERENCES

[1] I. Khalid, K. Anwar, and A. Halim, “Manajemen Global dalam Pendidikan: Networking, Webworking, dan Keunggulan Bersaing di Era Revolusi Industri 4.0 dan Society 5.0,” *J. Pendidik. Siber Nusant.*, vol. 4, no. 1, pp. 25–32, 2026, doi: 10.38035/jpsn.v4i1.571.

[2] A. F. Ramlan, M. U. Mustofa, Z. I. Suhaini, N. Q. S. Azizi, W. A. H. Mohd Norizam, and R. Solihah, “Meninjau Kembali Ruang Publik: Tinjauan Literatur Tentang Media Sosial Dan Pembentukan Agenda Politik Melalui Lensa Habermas,” *Sosioglobal J. Pemikir. dan Penelit. Sosiol.*, vol. 9, no. 2, pp. 201–210, Jun. 2025, doi: 10.24198/jsg.v9i2.62982.

[3] H. Y. Tenku, S. Artuti Erda De, and D. Kurniadi, “Konvergensi Media Digital: Tinjauan Kritis Dan Implementasinya Dalam Komunikasi Massa Kontemporer,” *Indones. J. Digit. Public Relations*, vol. 4, no. 1, pp. 114–121, 2025, doi: 10.25124/ijdpr.v4i1.9791.

[4] N. Amaly and A. Armiah, “Peran Kompetensi Literasi Digital Terhadap Konten Hoaks dalam Media Sosial,” *Alhadharah J. Ilmu Dakwah*, vol. 20, no. 2, p. 43, Dec. 2021, doi: 10.18592/alhadharah.v20i2.6019.

[5] A. Sarjito, “Hoaks, Disinformasi, dan Ketahanan Nasional: Ancaman Teknologi Informasi dalam Masyarakat Digital Indonesia,” *J. Gov. Local Polit.*, vol. 6, no. 2, pp. 175–186, Nov. 2024, doi: 10.47650/jglp.v6i2.1547.

[6] D. Orsini, R. Bianucci, F. M. Galassi, D. Lippi, and M. Martini, “Vaccine hesitancy, misinformation in the era of Covid-19: Lessons from the past,” *Ethics, Med. Public Heal.*, vol. 24, no. January, p. 100812, Oct. 2022, doi: 10.1016/j.jemep.2022.100812.

[7] D. A. Oktavianto, “The implementation of group investigation learning model to equip students to think critically in addressing the hoax content of disaster on the internet,” *IOP Conf. Ser. Earth Environ. Sci.*, vol. 683, no. 1, p. 012039, Mar. 2021, doi: 10.1088/1755-1315/683/1/012039.

[8] T. Murayama, S. Wakamiya, E. Aramaki, and R. Kobayashi, “Modeling the spread of fake news on Twitter,” *PLoS One*, vol. 16, no. 4, p. e0250419, Apr. 2021, doi: 10.1371/journal.pone.0250419.

[9] J. Li and X. Chang, “Combating Misinformation by Sharing the Truth: a Study on the Spread of Fact-Checks on Social Media,” *Inf. Syst. Front.*, vol. 25, no. 4, pp. 1479–1493, Aug. 2023, doi: 10.1007/s10796-022-10296-z.

[10] P. K. Verma, P. Agrawal, I. Amorim, and R. Prodan, “WELFake: Word Embedding Over Linguistic Features for Fake News Detection,” *IEEE Trans. Comput. Soc.*

Syst., vol. 8, no. 4, pp. 881–893, Aug. 2021, doi: 10.1109/TCSS.2021.3068519.

[11] S. Shelke and V. Attar, "Rumor detection in social network based on user, content and lexical features," *Multimed. Tools Appl.*, vol. 81, no. 12, pp. 17347–17368, May 2022, doi: 10.1007/s11042-022-12761-y.

[12] S. E. Bibri, A. Alexandre, A. Sharifi, and J. Krogstie, "Environmentally sustainable smart cities and their converging AI, IoT, and big data technologies and solutions: an integrated approach to an extensive literature review," *Energy Informatics*, vol. 6, no. 1, 2023, doi: 10.1186/s42162-023-00259-2.

[13] Masyarakat Anti Fitnah Indonesia, "TurnBackHoax.id." [Online]. Available: https://turnbackhoax.id

[14] P. Zhang, Y. Jia, and Y. Shang, "Research and application of XGBoost in imbalanced data," *Int. J. Distrib. Sens. Networks*, vol. 18, no. 6, p. 155013292211069, Jun. 2022, doi: 10.1177/15501329221106935.

[15] I. Firmansyah, J. T. Samudra, D. Pardede, and Z. Situmorang, "Comparison Of Random Forest And Logistic Regression In The Classification Of Covid-19 Sufferers Based On Symptoms," *J. Sci. Soc. Res.*, vol. 5, no. 3, p. 595, Oct. 2022, doi: 10.54314/jssr.v5i3.994.

[16] A. Ichsan, S. Riyadi, and D. Pardede, "Analysis of Logistic Regression Regularization in Wild Elephant Classification with VGG-16 Feature Extraction," *J. Comput. Networks, Archit. High Perform. Comput.*, vol. 6, no. 2, pp. 783–793, Apr. 2024, doi: 10.47709/cnahpc.v6i2.3789.

[17] A. N. Nugroho, N. E. Kamarukmi, and A. Ghufron, "Scales Feature Foot Scanners as Parameters of Flat Feet in Children," *Int. Conf. Inf. Sci. Technol. Innov.*, vol. 2, no. 1, pp. 152–156, Mar. 2023, doi: 10.35842/icostec.v2i1.54.

[18] N. A. Sinaga, D. Pardede, and S. Riyadi, "Analisis dampak strategi pedagogi terhadap minat belajar siswa menggunakan random forest," *J. Tekinkom (Teknik Inf. dan Komputer)*, vol. 8, no. 1, pp. 247–255, 2025, doi: 10.37600/tekinkom.v8i1.2169.

[19] X. Y. Liew, N. Hameed, and J. Clos, "An investigation of XGBoost-based algorithm for breast cancer classification," *Mach. Learn. with Appl.*, vol. 6, no. September, p. 100154, 2021, doi: 10.1016/j.mlwa.2021.100154.

[20] A. F. L. Ptr, M. M. Siregar, and I. Daniel, "Analysis of Gradient Boosting, XGBoost, and CatBoost on Mobile Phone Classification," *J. Comput. Networks, Archit. High Perform. Comput.*, vol. 6, no. 2, pp. 661–670, Apr. 2024, doi: 10.47709/cnahpc.v6i2.3790.

[21] M. Owusu-Adjei, J. Ben Hayfron-Acquah, T. Frimpong, and G. Abdul-Salaam, "Imbalanced class distribution and performance evaluation metrics: A systematic review of prediction accuracy for determining model performance in healthcare systems," *PLOS Digit. Heal.*, vol. 2, no. 11, p. e0000290, Nov. 2023, doi: 10.1371/journal.pdig.0000290.

[22] H. Wang, Q. Liang, J. T. Hancock, and T. M. Khoshgoftaar, "Feature selection strategies: a comparative analysis of SHAP-value and importance-based methods," *J. Big Data*, vol. 11, no. 1, 2024, doi: 10.1186/s40537-024-00905-w.

[23] E. Álvarez-García, D. García-Costa, S. Paniagua, J. Vicens, J. Vila-Francés, and F. Grimaldo, "Beyond Words: Analyzing Emotions and Linguistic Characteristics to Detect Hoax-Related Tweets During Spanish Regional Elections," *Int. J. Comput. Intell. Syst.*, vol. 17, no. 1, 2024, doi: 10.1007/s44196-024-00629-y.