



Analisis Sentimen Konten Deepfake pada Tiktok untuk Mengukur Literasi Media Baru di Indonesia

Agnes Melliana Eviyanti¹, Amri Sinuraya²

Universitas Persada Indonesia Y.A.I¹, Universitas Kristen Indonesia²
Jl. Pangeran Diponegoro No.74¹, Jl. Diponegoro No.84-86², RT.2/RW.6, Kenari, Kec. Senen,
Kota Jakarta Pusat, Daerah Khusus Ibukota Jakarta 10430¹

e-mail : agnesmelliana.eviyanti@upi-yai.ac.id¹, amrisinuraya@gmail.com²

Abstract

The advancement of artificial intelligence (AI) has facilitated the emergence of deepfake content, radically transforming the landscape of misinformation and blurring the boundaries between reality and digital fabrication. This phenomenon presents a significant challenge to new media literacy amidst the high penetration of the TikTok platform. This study aims to fill the research gap by computationally mapping public responses while evaluating societal media literacy levels in responding to such manipulations. A quantitative approach was applied using Natural Language Processing (NLP) and the IndoBERT model to classify 385 text data samples from TikTok. The model demonstrated robust performance, achieving an overall accuracy of 85.71% and a macro F1-Score of 83.34%. Classification results indicate that the polarity distribution is dominated by neutral (225 data) and positive sentiments (95 data), whereas negative (critical) sentiment accounts for only 65 data points. These empirical findings confirm the existence of a digital paradox, where high technology adoption clashes with low algorithmic awareness, leading the public to normalize visual manipulation as a form of entertainment without critical filtering. This study recommends reorienting the national digital literacy curriculum toward AI literacy to enhance public information resilience.

Keywords: Deepfake, Digital Paradox, IndoBERT, Media Literacy, Misinformation, Sentiment Analysis, TikTok

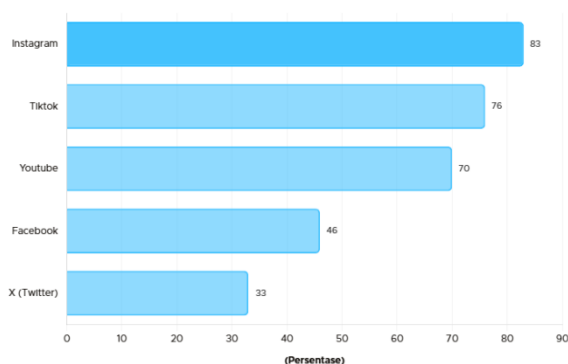
Abstrak

Perkembangan teknologi kecerdasan buatan (AI) telah memfasilitasi kemunculan konten *deepfake*, yang secara radikal mengubah lanskap misinformasi dan mengaburkan batasan antara realitas dan fabrikasi digital. Fenomena ini menghadirkan tantangan signifikan terhadap literasi media baru di tengah tingginya penetrasi platform TikTok. Penelitian ini bertujuan mengisi kekosongan literatur dengan memetakan respons publik secara komputasional sekaligus mengevaluasi tingkat literasi media masyarakat dalam merespons manipulasi tersebut. Pendekatan kuantitatif diterapkan menggunakan *Natural Language Processing* (NLP) dan model IndoBERT untuk mengklasifikasikan 385 sampel data teks dari TikTok. Kinerja model terbukti andal dengan mencatatkan tingkat akurasi keseluruhan sebesar 85,71% dan nilai rata-rata *F1-Score* 83,34%. Hasil klasifikasi menunjukkan distribusi polaritas didominasi oleh sentimen netral (225 data) dan positif (95 data), sementara sentimen negatif (kritis) hanya mencakup 65 data. Temuan empiris ini mengonfirmasi eksistensi paradoks digital, di mana tingginya adopsi teknologi berbenturan dengan rendahnya kesadaran algoritmik, sehingga masyarakat cenderung menormalisasi manipulasi visual sebagai bentuk hiburan tanpa filter kritis. Penelitian ini merekomendasikan reorientasi kurikulum literasi digital nasional menuju literasi kecerdasan buatan (*AI literacy*) guna meningkatkan ketahanan informasi publik.

Kata Kunci : Deepfake, Paradoks Digital, IndoBERT, Literasi Media, Misinformasi, Analisis Sentimen, TikTok

PENDAHULUAN

Perkembangan teknologi kecerdasan buatan (Artificial Intelligence) telah memicu krisis literasi digital yang belum pernah terjadi sebelumnya. Secara historis, literasi media secara fundamental didefinisikan sebagai kemampuan kognitif kritis masyarakat untuk mengakses, menganalisis, mengevaluasi, dan mengkreasi pesan di berbagai medium (Buckingham, 2007; Hobbs, 2010). Namun, paradigma klasik ini menghadapi tantangan besar ketika manipulasi realitas melalui rekayasa visual dan auditori semakin mengaburkan batas antara fakta dan fabrikasi (Nugroho & Wulandari, 2023). Transformasi ini melahirkan gelombang baru misinformasi (*misinformation*) yang jauh lebih kompleks dan sulit dideteksi oleh kognisi manusia awam (Pratama & Yulianti, 2023; Hidayat & Wijaya, 2024). Salah satu manifestasi paling disruptif dari disrupsi teknologi ini adalah deepfake, yakni sintesis media berbasis deep learning yang mampu memanipulasi citra manusia dengan tingkat realisme yang sangat meyakinkan (Fernandes & Fatma, 2025). Dalam konteks penyebaran informasi, deepfake telah mengubah lanskap misinformasi dari sekadar narasi teks palsu menjadi rekayasa realitas visual yang secara psikologis lebih mudah memanipulasi kepercayaan publik (Setiawan et al., 2024). Di tengah masifnya penetrasi media sosial berbasis video pendek, konten deepfake menemukan ekosistem yang ideal untuk menyebar secara eksponensial. Algoritma rekomendasi pada platform seperti TikTok yang berpusat pada pengguna dan sangat dipersonalisasi—tidak hanya memicu lonjakan interaksi visual, tetapi juga secara tersembunyi berdampak pada fungsi kognitif, memori, dan kontrol impulsif penggunaanya (Wicaksono et al., 2024). Agresivitas distribusi algoritmik ini menciptakan fenomena hiper-realitas, di mana masyarakat menjadi sangat rentan mengonsumsi informasi manipulatif tanpa adanya filter kritis maupun kemampuan verifikasi yang memadai.



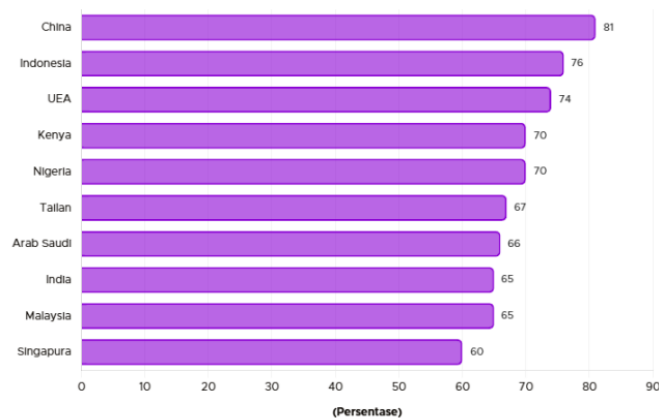
Sumber: Jakpat

GoodStats

Gambar 1.1 Grafik Media Sosial Paling Sering Digunakan Gen Z

Sumber: <https://data.goodstats.id/>

Pemilihan TikTok sebagai objek penelitian didasarkan pada tingkat penetrasi platform yang sangat masif di Indonesia. Mengacu pada data penetrasi media sosial (Gambar 1.1), TikTok menempati urutan kedua sebagai platform yang paling banyak diakses oleh Generasi Z (76%). Tingginya angka keterlibatan ini sayangnya berbenturan dengan fakta mengenai kerentanan publik terhadap kredibilitas informasi. Merujuk pada Gambar 1.2, tercatat bahwa 76% publik di Indonesia memiliki tingkat kepercayaan yang sangat tinggi terhadap media. Kondisi sosiologis yang kontradiktif inilah yang melahirkan fenomena paradoks digital (*digital paradox*); yakni suatu keadaan di mana masifnya adopsi teknologi mutakhir dan tingginya konektivitas masyarakat justru berbanding terbalik dengan kecakapan literasi digital mereka, sehingga menciptakan ruang gema yang subur bagi amplifikasi misinformasi (Nugroho & Wulandari, 2023; Setiawan et al., 2024).



Sumber: Edelman Trust Institute

GoodStats

Gambar 1.2 76% Grafik tingkat kepercayaan terhadap media

Sumber: <https://data.goodstats.id/>

Meskipun ancaman manipulasi realitas dan penyebaran misinformasi ini semakin persisten, diskursus akademik mengenai deepfake saat ini masih sangat terfragmentasi dan didominasi oleh pendekatan teknis komputasional semata. Tinjauan literatur menunjukkan bahwa mayoritas penelitian berfokus eksklusif pada pengembangan metode forensik pendeteksian piksel menggunakan arsitektur seperti *Convolutional Neural Network* (CNN), *Generative Adversarial Networks* (GAN), maupun *Long Short-Term Memory* (LSTM) (Fernandes & Fatma, 2025; Patrick et al., 2024). Di sisi lain, implementasi *Natural Language Processing* (NLP) dan analisis sentimen di ruang digital Indonesia umumnya masih dikerahkan secara terbatas untuk evaluasi komersial atau politik praktis. Hal ini terlihat pada pemanfaatan NLP untuk sentimen layanan perbankan (Cahya & Ichwani, 2026), persepsi publik terhadap kinerja awal pemerintahan (Ihalauw et al., 2025), kategorisasi intensi audiens *live streaming* (Maharani & Ihsan, 2026), hingga evaluasi kepuasan pengguna terhadap aplikasi komersial (Maulana et al., 2025).

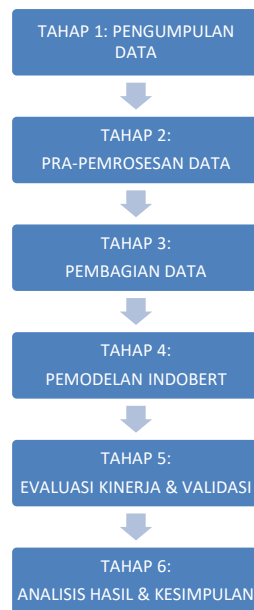
Berdasarkan pemetaan literatur tersebut, penelitian ini secara eksplisit mengidentifikasi kekosongan (*research gap*) yang esensial, belum adanya kajian yang mengintegrasikan ketangguhan model bahasa komputasional (*Language Model*) dengan instrumen pengukuran sosiologis untuk mengevaluasi literasi media baru (*new media literacy*) masyarakat dalam merespons ancaman misinformasi berbasis deepfake. Studi-studi terdahulu hanya memosisikan deepfake sebagai objek deteksi mesin, namun luput mengevaluasi bagaimana secara psikologis dan sosiologis masyarakat merespons keberadaan manipulasi visual tersebut di platform algoritmik seperti TikTok. Berangkat dari *research gap* tersebut, penelitian ini menawarkan kebaruan (*novelty*) dengan menggeser paradigma analisis dari sekadar "deteksi forensik komputasi" menuju "pengukuran ketahanan informasi publik". Mengingat kapabilitas arsitektur komputasi modern yang telah terbukti andal dalam memetakan dinamika sentimen dari teks yang tidak terstruktur (Maharani & Ihsan, 2026; Maulana et al., 2025), penelitian ini memanfaatkan model prapelatihan *Bidirectional Encoder Representations from Transformers* (IndoBERT) (Koto et al., 2020). Melalui integrasi antara NLP dan kajian literasi media, penelitian ini bertujuan memetakan respons masyarakat terhadap fenomena deepfake di TikTok guna membongkar eksistensi paradoks digital di Indonesia, di mana tingginya adopsi teknologi berbenturan dengan rendahnya kesadaran algoritmik. Hasil analisis ini diorientasikan untuk menjadi landasan strategis dan komprehensif bagi rekayasa ulang kurikulum literasi digital nasional dalam menghadapi krisis manipulasi realitas dan misinformasi di masa depan.

METODOLOGI PENELITIAN

Desain Penelitian

Penelitian ini dirancang menggunakan pendekatan kuantitatif eksplanatori yang diintegrasikan dengan metode eksperimental penambangan data (data mining) dan Natural Language Processing (NLP). Pendekatan kuantitatif dikerahkan untuk mengukur fenomena sosial secara objektif yakni respons dan tingkat literasi digital masyarakat terhadap konten deepfake melalui parameter statistik. Sementara itu,

pendekatan eksperimental data mining menggunakan arsitektur deep learning diterapkan guna mengotomatisasi ekstraksi dan klasifikasi informasi dari teks berskala besar. Hal ini bertujuan untuk meminimalisasi bias subjektivitas kognitif peneliti dalam menilai polaritas sentimen. Secara keseluruhan, alur arsitektur sistem pada penelitian ini dapat divisualisasikan melalui bagan berikut:



Gambar 1.3 Diagram Alur Metodologi Analisis Sentimen IndoBERT
Sumber: Olahan data peneliti

Populasi, Sampel, dan Teknik Pengumpulan Data

Objek material dalam penelitian ini adalah jejak digital masyarakat berupa teks deskripsi video dan komentar pada platform media sosial TikTok yang merespons fenomena *deepfake*. Pengumpulan data sekunder diekstraksi melalui teknik *web scraping* dengan memanfaatkan *Application Programming Interface* (API) komputasi awan spesifik untuk platform TikTok. Data yang ditarik mencakup variabel tekstual (komentar) serta metrik interaksi (jumlah tayangan, suka, dan pembagian).

Mengingat populasi percakapan di TikTok bersifat tidak terhingga (*unknown population*), penarikan sampel dilakukan melalui teknik *purposive sampling* berdasarkan parameter kueri pencarian spesifik, yakni #deepfake, #AI, dan #faceswap. Pemilihan ketiga tagar ini didasarkan pada popularitas dan tingkat relevansinya yang tinggi terhadap diskursus sintesis media pada saat data dikumpulkan. Namun demikian, pendekatan ini secara sadar memiliki potensi bias seleksi (*selection bias*). Penggunaan tagar spesifik berisiko membatasi ruang sampel hanya pada kelompok pengguna yang telah memiliki intensi atau paparan awal terhadap terminologi teknologi, sementara percakapan mengenai manipulasi visual yang menggunakan bahasa sehari-hari atau tanpa tagar organik (*untagged conversations*) tidak ikut terjaring dalam populasi sampel. Guna memastikan signifikansi dan representasi yang solid secara statistik dari ruang sampel yang ditarik tersebut, ukuran sampel ditetapkan sebanyak 385 baris data tekstual.

Pra-pemrosesan Data

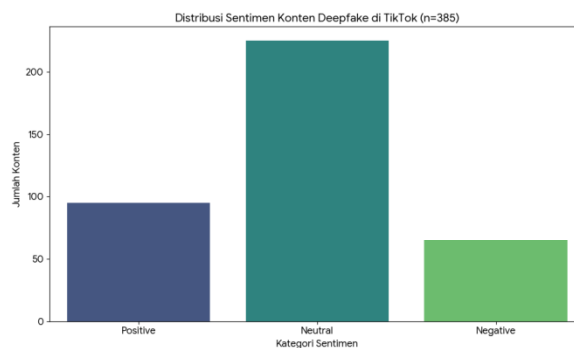
Data tekstual yang diekstraksi dari media sosial umumnya bersifat tidak terstruktur dan mengandung tingkat derau (*noise*) yang tinggi. Agar data dapat dipahami oleh model komputasi matematis, dilakukan tahapan pra-pemrosesan yang ketat secara sekuensial:

1. **Case Folding:** Penyeragaman seluruh karakter alfabet menjadi huruf kecil (*lowercase*) agar mesin tidak membedakan bobot kata berdasarkan kapitalisasi.

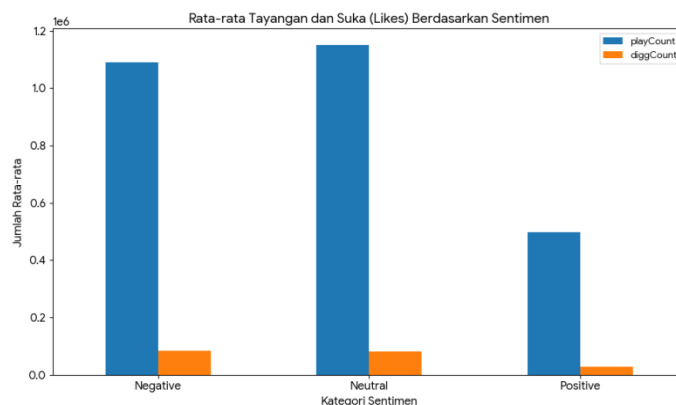
2. **Cleansing:** Pembersihan teks dari elemen-elemen yang tidak memiliki nilai semantik yang relevan, meliputi penghapusan tautan URL, sebutan pengguna (*mentions*), tagar (*hashtags*), emotikon, karakter angka, serta tanda baca.
3. **Tokenization:** Pemecahan struktur kalimat yang panjang menjadi unit kata tunggal (*token*) agar algoritma dapat menganalisis frekuensi dan konteks kata secara individual.
4. **Stopword Removal:** Eliminasi kata-kata hubung atau kata umum yang sering muncul namun tidak membawa makna sentimen spesifik, menggunakan leksikon bahasa Indonesia.
5. **Stemming:** Transformasi kata-kata berimbuhan ke bentuk dasarnya (kata pudar) guna menyeragamkan variasi linguistik dalam korpus data.
6. **Pelabelan Manual (*Ground Truth*):** Teks yang telah bersih kemudian diklasifikasikan oleh anotator manusia ke dalam kelas polaritas sentimen positif, negatif, atau netral, yang akan berfungsi sebagai kebenaran dasar (*ground truth*) bagi mesin.

Pembagian Data

Dataset yang telah melalui tahap pra-pemrosesan dan pelabelan selanjutnya didistribusikan ke dalam dua subset komputasi yang terpisah. Penelitian ini menerapkan rasio pembagian standar 80:20, di mana 80% data (308 sampel) dialokasikan sebagai data latih (*training data*) untuk membangun kecerdasan pola model, dan 20% sisanya (77 sampel) dialokasikan sebagai data uji (*testing data*) yang sengaja disembunyikan dari mesin untuk menguji performa klasifikasi di akhir tahap. Pembagian ini dilakukan secara proporsional menggunakan teknik stratifikasi (*stratified random sampling*). Pendekatan ini esensial untuk menjamin bahwa distribusi kelas sentimen (positif, negatif, dan netral) tetap seimbang pada setiap subset, sehingga mencegah model mengalami bias akibat dominasi kelas mayoritas.



Gambar 2.1 Distribusi Sentimen Konten Deepfake di Tiktok
Sumber: Hasil Olah Data



Gambar 2.2 Rata-Rata Tayangan dan Suka (like) berdasarkan Sentimen
Sumber: Hasil Olah Data

Skenario Pemodelan BERT

Penelitian ini mengimplementasikan algoritma klasifikasi *Bidirectional Encoder Representations from Transformers* (BERT), khususnya model pra-latih IndoBERT yang arsitekturnya telah dioptimalkan secara leksikal untuk memahami korpus bahasa Indonesia yang kompleks.

Dalam skenario pemodelan ini, teks dari data latih dikonversi menjadi representasi token *input_ids* dan *attention_mask* agar dapat diinterpretasikan oleh arsitektur Transformer. Model IndoBERT kemudian dilatih ulang (*fine-tuning*) dengan memodifikasi lapisan terakhirnya (lapisan klasifikasi) melalui beberapa siklus iterasi pembelajaran (*epoch*). Proses optimisasi pembaharuan bobot (*weights*) model dilakukan menggunakan algoritma *optimizer* AdamW, diiringi oleh perhitungan *Cross-Entropy Loss* untuk mengukur dan menekan tingkat kesalahan klasifikasi hingga mencapai titik konvergensi terendah pada fasa komputasi.

Evaluasi dan Validasi Kinerja Model

Berpijak pada landasan teori Temu Kembali Informasi (*Information Retrieval*) dan Sains Sosial Komputasional, validasi matematis mutlak diperlukan untuk membuktikan bahwa temuan fenomena sosial tidak dipengaruhi oleh bias algoritmik. Kinerja algoritma IndoBERT pada 77 sampel data uji dievaluasi secara komprehensif menggunakan instrumen *Confusion Matrix* (matriks kebingungan). Matriks ini memetakan secara transparan titik temu dan penyimpangan antara label aktual (kognisi manusia) dengan kelas prediksi (penalaran probabilistik mesin).

Keandalan klasifikasi model kemudian dihitung melalui metrik evaluasi matematis berikut:

1. **Precision (Presisi):** Parameter yang mengukur akurasi prediksi positif untuk meminimalkan insiden positif palsu (*False Positives*), menjamin tingkat kepercayaan sistem dalam melabeli teks.
2. **Recall (Sensitivitas):** Parameter yang mengukur proporsi sampel positif aktual yang berhasil diidentifikasi, berfungsi untuk meminimalkan negatif palsu (*False Negatives*). Hal ini krusial agar tidak ada suara kritis atau peringatan publik yang terlewat dari radar deteksi algoritma.
3. **F1-Score:** Nilai rata-rata harmonis antara *precision* dan *recall*. Pengukuran metrik ini bertindak sebagai penyeimbang utama untuk menghindari ilusi akurasi semu (*misleading accuracy*) yang sering terjadi pada set data dengan distribusi kelas yang tidak proporsional (*imbalanced data*). Evaluasi melalui *F1-Score* memastikan bahwa ketangguhan model dinilai secara objektif di seluruh kelas sentimen, terlepas dari polaritas mana yang bertindak sebagai mayoritas maupun minoritas.

HASIL

Hasil Ekstraksi Data dan Distribusi Sentimen Publik

Berdasarkan proses penambangan data (data mining) pada platform TikTok, penelitian ini berhasil mengekstraksi 385 sampel data tekstual yang merepresentasikan jejak digital dan opini publik terhadap fenomena deepfake. Melalui tahapan pra-pemrosesan dan pelabelan data, distribusi polaritas sentimen masyarakat terpetakan dengan sangat kontras. Hasil klasifikasi menunjukkan dominasi mutlak pada kelas sentimen netral yang mencapai 225 data, disusul oleh sentimen positif sebanyak 95 data. Sebaliknya, sentimen negatif berada pada posisi minoritas ekstrem dengan hanya 65 data.

Dominasi sentimen netral dan positif ini mengindikasikan sebuah pola perilaku digital di mana pengguna media sosial merespons kehadiran manipulasi kecerdasan buatan murni sebagai tren kreativitas visual dan hiburan. Alih-alih melihatnya sebagai potensi ancaman siber atau penipuan identitas, masyarakat cenderung terpukau oleh kecanggihan teknologi sintesis tanpa tendensi kritis untuk mempertanyakan implikasi etis maupun keamanan di baliknya.

Analisis Tingkat Interaksi dan Viralitas Konten

Tingkat penerimaan publik yang permisif terhadap manipulasi realitas berbanding lurus dengan tingginya metrik interaksi pengguna. Hasil analisis kuantitatif memotret tingkat viralitas yang sangat masif, di mana secara keseluruhan, konten *deepfake* mencatatkan rata-rata penayangan yang menembus angka 980.000 pemutaran per video dengan jumlah *likes* mencapai lebih dari 68.000 interaksi.

Lebih jauh, analisis membedah bahwa konten bersentimen netral yang menyajikan hiper-realitas tanpa sisipan peringatan bahaya justru meraih rekor penyebaran tertinggi, dengan rata-rata 1,15 juta tayangan dan dibagikan ulang (*share*) hingga lebih dari 8.300 kali per video. Pola distribusi eksponensial ini membuktikan agresivitas algoritma rekomendasi TikTok dalam mendistribusikan manipulasi visual ke

ruang publik. Dalam ekosistem ini, pengguna tidak hanya bertindak sebagai penonton pasif, tetapi secara sadar maupun tidak sadar bertransformasi menjadi agen distributor yang menyebarkan ulang konten rekayasa tersebut tanpa melalui proses verifikasi kredibilitas.

Di sisi lain, meskipun jumlah konten bersentimen negatif sangat minim (65 data), konten yang menyuarakan kewaspadaan terhadap penipuan identitas tetap mampu mendulang atensi yang signifikan dengan rata-rata 83.000 *likes*. Fenomena ini menandakan keberadaan kelompok masyarakat yang mulai menyadari ancaman krisis kepercayaan digital, meskipun suara mereka masih tertutup oleh arus utama.

Analisis Tingkat Interaksi dan Viralitas Konten

Guna memastikan bahwa pemetaan sentimen tersebut valid secara komputasional, kinerja algoritma IndoBERT dievaluasi menggunakan 20% alokasi data uji (77 sampel teks) yang belum pernah dikenali oleh mesin sebelumnya. Kinerja model dievaluasi menggunakan instrumen Confusion Matrix untuk mengukur tingkat presisi prediksi kecerdasan buatan dibandingkan dengan label aktual.

Tabel 1.1 Confusion Matrix Pengujian Model IndoBERT

Kelas Aktual (True Label)	Prediksi Negatif	Prediksi Netral	Prediksi Positif	Total Aktual
Negatif	10 (True Positive)	3	0	13
Netral	2	40 (True Positive)	3	45
Positif	1	2	16 (True Positive)	19
Total Prediksi	13	45	19	77 (Total Data Uji)

Sumber: Hasil Olah Data

Berdasarkan Tabel 1.1, ketangguhan arsitektur deep learning ini terlihat jelas pada nilai di sepanjang diagonal utama. Algoritma IndoBERT berhasil memprediksi secara akurat 10 dari 13 komentar negatif, 40 dari 45 komentar netral, dan 16 dari 19 komentar positif.

Tabel 1.2 Kinerja Klasifikasi Model IndoBERT berdasarkan Kelas Sentimen

Kelas Sentimen	Precision	Recall	F1-Score
Negatif	76.92%	76.92%	76.92%
Netral	88.89%	88.89%	88.89%
Positif	84.21%	84.21%	84.21%
Rata-rata Makro (Macro Avg)	83.34%	83.34%	83.34%
Akurasi Keseluruhan (Accuracy)	{85.71%}		

Sumber: Hasil Olah Data

Berdasarkan abstraksi perhitungan dari *confusion matrix* tersebut, Tabel 1.2 menyajikan rincian metrik evaluasi model secara komprehensif. Arsitektur IndoBERT menunjukkan performa klasifikasi yang sangat andal dengan mencatatkan tingkat akurasi keseluruhan (*overall accuracy*) sebesar 85,71%. Secara lebih spesifik, kelas sentimen 'Netral'—yang merupakan kelas mayoritas—mencatat metrik tertinggi (*Precision*, *Recall*, dan *F1-Score* masing-masing 88,89%). Tingginya metrik pada kelas ini menunjukkan bahwa model sangat akurat dalam mengenali teks tanpa tendensi emosi. Di sisi lain, meskipun kelas 'Negatif' memiliki jumlah data uji yang paling minoritas, model tetap mampu menghasilkan *F1-Score* yang solid di angka 76,92%. Secara agregat, nilai rata-rata makro (*macro average*) yang stabil di angka 83,34% untuk seluruh metrik evaluasi membuktikan bahwa model tidak mengalami *overfitting* terhadap kelas dominan. Tingginya angka presisi, sensitivitas (*recall*), dan rata-rata harmonis (*F1-Score*) ini memvalidasi secara matematis bahwa arsitektur IndoBERT yang diimplementasikan bebas dari ilusi akurasi semu (*misleading accuracy*), sehingga sangat reliabel untuk

digunakan sebagai instrumen pemetaan opini publik dalam penelitian ini.

PEMBAHASAN

Bila ditarik ke dalam perspektif sosiologis dan teori media baru, temuan klasifikasi teks dan metrik viralitas ini menyingkap sebuah paradoks digital yang tajam di Indonesia. Tingginya angka adopsi teknologi dan interaksi terhadap konten deepfake, yang justru berbenturan dengan minimnya sentimen kewaspadaan, mengonfirmasi terjadinya krisis literasi media baru, khususnya pada aspek kesadaran algoritmik. Temuan ini tidak sekadar sejalan, melainkan memperdalam dan mengontekstualisasikan postulat dari penelitian-penelitian terdahulu. Dalam kajian sebelumnya, Nugroho dan Wulandari (2023) memperingatkan secara teoretis bahwa disrupsi kecerdasan buatan berkorelasi langsung dengan kerentanan penyebaran hoaks di media sosial. Melalui riset ini, peringatan tersebut mendapatkan pembuktian empiris yang spesifik: kerentanan publik ternyata berakar dari "kelumpuhan kritis" (*critical paralysis*) yang secara nyata dibuktikan oleh dominasi sentimen netral dan positif terhadap konten manipulatif. Lebih lanjut, Setiawan et al. (2024) sebelumnya memetakan bahwa opini publik secara umum sangat terpengaruh oleh eksistensi teknologi deepfake. Namun, dialog dengan temuan riset ini membongkar dimensi baru dari pengaruh tersebut. Publik di platform algoritmik seperti TikTok tidak hanya sekadar "terpengaruh" secara pasif, melainkan secara sukarela menormalisasi manipulasi realitas tersebut sebagai bentuk konsumsi hiburan (*entertainment*). Ketidadaan resistensi kognitif yang tercermin dari minimnya sentimen negatif (hanya 65 data) menunjukkan bahwa ancaman deepfake kini telah bertransformasi dari sekadar isu keamanan informasi menjadi budaya pop-digital yang diterima masyarakat tanpa filter. Paradoks ini semakin diperburuk oleh mekanisme rekomendasi algoritma platform seperti TikTok yang dirancang dengan pendekatan *user-centered*, di mana konten disajikan secara spesifik untuk mengeksploitasi preferensi dan rekam jejak interaksi penggunaanya (Wicaksono et al., 2024).

Pengguna media sosial saat ini beroperasi di dalam "ilusi validitas visual". Mereka secara reaktif menyerap informasi yang tersaji meyakinkan secara visual dan auditori tanpa mendekonstruksi proses sintesis di balik pembuatannya. Manipulasi realitas sering kali dianggap lumrah karena masyarakat gagal menarik demarkasi yang tegas antara realitas faktual dan fabrikasi kecerdasan buatan, menjadikannya ancaman konten ilegal yang memerlukan upaya penanggulangan serius di Indonesia (Hidayat & Wijaya, 2024). Kegagalan kognitif ini dapat dijelaskan melalui dua faktor. Pertama, kecanggihan teknologi *deepfake* yang memanfaatkan jaringan saraf tiruan dari *deep learning* mampu menganalisis kumpulan data besar untuk mempelajari dan meniru ekspresi wajah, perilaku, hingga suara seseorang secara sangat presisi (Fernandes & Fatma, 2025; Patrick et al., 2024). Kedua, secara neuropsikologis, paparan konten video pendek yang dipersonalisasi secara masif terbukti mengaktifkan sistem penghargaan di otak (*reward system*), seperti *ventral tegmental area* (VTA) dan *default mode network* (DMN) (Wicaksono et al., 2024). Stimulasi ini terkait erat dengan pelepasan dopamin yang memicu perilaku adiktif seperti *doom scrolling*, yang secara langsung mendegradasi fungsi eksekutif pengguna dan membuat mereka mengambil keputusan secara impulsif tanpa pemikiran reflektif (Wicaksono et al., 2024).

Kendati demikian, tingginya angka *likes* pada minoritas konten yang bersentimen negatif membuktikan bahwa embrio pemikiran kritis sesungguhnya telah lahir di kalangan sekelompok kecil masyarakat. Kelompok ini telah memiliki pijakan literasi untuk memahami ancaman misinformasi dari disrupsi kecerdasan buatan. Sayangnya, resonansi suara kritis tersebut masih sering tenggelam oleh amplifikasi algoritma yang mengeksploitasi kepuasan hiburan. Sistem distribusi algoritma beroperasi dengan memberikan peringkat pada konten berdasarkan metrik interaksi semu seperti retensi waktu tonton, jumlah suka, dan aktivitas membagikan video sehingga mengutamakan viralitas di atas kredibilitas (Wicaksono et al., 2024). Realitas interaktif ini sangat selaras dengan fenomena psikologis *Fear of Missing Out* (FOMO), di mana dorongan impulsif untuk mengikuti tren digital yang viral dan masif sering kali mengesampingkan dorongan kognitif untuk melakukan verifikasi fakta secara mandiri (Maulana et al., 2025).

Kondisi ini menyimpulkan bahwa ketahanan informasi masyarakat masih sangat rapuh. Oleh karena itu, intervensi strategis mutlak diperlukan; pendidikan literasi digital nasional tidak boleh lagi mandek pada kecakapan operasional gawai (*digital skills*). Hal ini sejalan dengan peringatan dini dari Hobbs (2010) dan Buckingham (2007) bahwa literasi digital sejati harus berbasis pada evaluasi kritis terhadap institusi dan teknologi media itu sendiri, bukan sekadar kemampuan menekan tombol. Konsep

ini harus segera direorientasi menuju literasi kecerdasan buatan (*AI literacy*) yang membekali publik dengan kemampuan dekonstruksi konten sintesis. Mengingat metode forensik digital untuk deteksi *deepfake* akan terus berkejaran dengan kemampuan arsitektur *Generative Adversarial Networks* (GAN) yang semakin sempurna dalam menghasilkan citra palsu (Fernandes & Fatma, 2025), sistem deteksi cerdas berbasis *Natural Language Processing* (NLP) menggunakan model pra-latih seperti BERT maupun IndoBERT dapat menjadi alat bantu krusial dalam menyaring misinformasi di level tekstual (Koto et al., 2020; Pratama & Yulianti, 2023). Namun, benteng pertahanan utama tidak bisa lagi hanya disandarkan pada mesin pendeteksi semata. Ketahanan sejati harus dibangun pada kognisi publik untuk merestorasi durasi atensi berkelanjutan dan kemampuan komprehensi informasi (Wicaksono et al., 2024), demi menghadapi era manipulasi realitas yang persisten di masa depan.

PENUTUP

Penelitian ini menyimpulkan bahwa penyebaran konten *deepfake* di TikTok secara nyata mengekspos kerentanan literasi media baru masyarakat Indonesia, di mana manipulasi realitas lebih banyak diterima sebagai inovasi visual dan hiburan semata. Hal ini dibuktikan melalui pengujian model IndoBERT yang secara presisi memetakan dominasi sentimen netral (225 data) dan positif (95 data) dibandingkan sentimen negatif (65 data) dari 385 sampel, menunjukkan minimnya kesadaran kritis publik terhadap potensi penyalahgunaan dan pelanggaran etika. Secara praktis, temuan ini mendesak penerapan implikasi strategis berupa reorientasi kurikulum literasi digital nasional menuju literasi kecerdasan buatan (*AI literacy*) yang komprehensif, serta keharusan bagi penyedia platform untuk segera mengimplementasikan mitigasi sistemik seperti pelabelan algoritmik (*algorithmic labeling*).

Secara teoretis, penelitian ini memberikan implikasi penting dengan memperluas kerangka konseptual literasi media baru dalam menghadapi disrupsi kecerdasan buatan. Mengintegrasikan algoritma komputasional (IndoBERT) dengan tinjauan sosiologis membuktikan bahwa model pemrosesan bahasa alami (NLP) tidak hanya berfungsi sebagai alat evaluasi sentimen komersial, melainkan juga sebagai instrumen diagnostik yang andal untuk mengukur "paradoks digital" di tengah masyarakat. Temuan ini memperkaya literatur komunikasi komputasional dengan menegaskan bahwa kelumpuhan kritis (*critical paralysis*) publik saat ini dipicu oleh normalisasi hiper-realitas yang difasilitasi oleh algoritma platform.

Meskipun arsitektur komputasi ini terbukti tangguh menjembatani data tidak terstruktur, penelitian ini memiliki sejumlah keterbatasan penelitian (*limitations*) yang perlu dipertimbangkan secara objektif. Pertama, penelitian menggunakan data *cross-sectional* tunggal dari platform TikTok, sehingga membatasi kemampuan generalisasi opini publik di ekosistem digital lainnya. Kedua, teknik penarikan sampel yang hanya bergantung pada tiga tagar teknis spesifik (*#deepfake*, *#AI*, dan *#faceswap*) memunculkan potensi bias seleksi, sehingga riset ini mungkin belum sepenuhnya menjangkau percakapan dari masyarakat awam yang mengonsumsi manipulasi visual tanpa kesadaran label teknis. Ketiga, mesin komputasi yang digunakan masih menghadapi kendala teknis dalam mendeteksi sarkasme linguistik yang dinamis di media sosial. Oleh karena itu, penelitian selanjutnya disarankan untuk memperluas cakupan dataset lintas platform, menggunakan leksikon kata kunci pencarian yang lebih variatif dan organik, mengintegrasikan analisis multimodal yang menggabungkan fitur teks, visual, maupun auditori untuk meningkatkan akurasi, serta mengembangkan kerangka riset longitudinal guna memantau efektivitas program peningkatan ketahanan informasi masyarakat secara berkelanjutan.

DAFTAR PUSTAKA

- Buckingham, D. (2007). Digital media literacies: Rethinking media education in the age of the Internet. *Research in Comparative and International Education*, 2(1), 43-55.
- Hobbs, R. (2010). *Digital and media literacy: A plan of action*. The Aspen Institute.
- Cahya, R. R., & Ichwani, A. (2026). Komparasi kinerja model Support Vector Machine dan Indonesian Bidirectional Encoder Representations from Transformers dalam klasifikasi sentimen ulasan aplikasi Wondr by BNI. *Journal of Science and Social Research*, 9(2), 1965-1972.
- Fernandes, Y. A., & Fatma, Y. (2025). Metode deep learning dalam teknologi *deepfake*: Systematic literature review. *JATI (Jurnal Mahasiswa Teknik Informatika)*, 9(2), 3403-3410.

- Hidayat, T., & Wijaya, K. (2024). Konten ilegal kecerdasan buatan berbasis deepfake dan upaya penanggulangannya di Indonesia. *Jurnal Ilmiah Sistem Informasi*, 8(1), 22-35.
- Ihalauw, S. A., Lapatta, N. T., Nugraha, D. W., Wirdayanti, & Lamasitudju, C. A. (2025). Analisis sentimen terhadap kinerja awal pemerintahan menggunakan IndoBERT dan SMOTE pada media sosial X. *Jurnal Algoritma*, 22(2), 803-815.
- Koto, F., Rahimi, A., Lau, J. H., & Baldwin, T. (2020). IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP. *Proceedings of the 28th International Conference on Computational Linguistics*, 757-770.
- Maharani, M., & Ihsan, M. C. (2026). Kategorisasi komentar live streaming TikTok menggunakan Support Vector Machine. *Journal of Artificial Intelligence and Digital Business (RIGGS)*, 5(1), 5782-5788.
- Maulana, M. T. F., Nugroho, B. I., & Utami, E. U. S. (2025). Analisis sentimen ulasan aplikasi TikTok di Google Playstore menggunakan algoritma Naive Bayes. *Journal of Artificial Intelligence and Digital Business (RIGGS)*, 4(3), 6627-6635.
- Nugroho, A., & Wulandari, S. (2023). Analisis pengaruh teknologi kecerdasan buatan terhadap penyebaran berita hoax di media sosial. *Jurnal NAAFI (STKIP Majenang)*, 4(2), 112-125.
- Patrick, M., Lubis, C., & Dharmawan, A. B. (2024). Pendeteksian citra deepfake wajah di smartphone menggunakan MobileNetV3-Small dan LBP. *Jurnal Ilmu Komputer dan Sistem Informasi*, 12(1). <https://doi.org/10.24912/jiksi.v12i1.28214>
- Pratama, D. A., & Yulianti, E. (2023). Sistem deteksi berita hoaks berbasis algoritma Natural Language Processing (NLP) menggunakan BERT. *Open Jurnal STMIK Amika*, 6(1), 34-45.
- Setiawan, A. A., Santoso, T. A., Prawira, G. J., & Rakhmawati, N. A. (2024). Analisis sentimen publik terhadap penggunaan teknologi deepfake pada era digital. *Etika Teknologi Informasi*, 1(1), 1-15.
- Wicaksono, T. A., Anggelina, E., Reformasianto, V. L., Wimbarti, S., Mufidah, D., & Kusrohmaniah, S. (2024). The impact of TikTok social media on users: A neuropsychological perspective. *Psikostudia: Jurnal Psikologi*, 13(3), 455-463.[cite: 2]