INTI NUSA MANDIRI
DOI: https://doi.org/10.33480/inti.v21i1.8699

VOL. 21. NO. 1 AGUSTUS 2026
P-ISSN: 0216-6933 | E-ISSN: 2685-807X
Diterbitkan Oleh: LPPM Universitas Nusa Mandiri

INVESTIGASI HATE SPEECH PADA PLATFORM "X" MENGGUNAKAN METODE HYBRID INDOBERT–GRAPH ATTENTION NETWORK

Weslie Austin[1] ' Puguh Hiskiawan[2]*

Program Studi Data Sains[1,2]
Universitas Bunda Mulia, Kota Tangerang, Indonesia[1,2]
ubm.ac.id[1,2]
s36230013@student.ubm.ac.id[1], phiskiawan@bundamulia.ac.id[2*]

(*) Corresponding Author

Abstract—*The rapid growth of social media, particularly the X (formerly Twitter) platform, has made it a primary space for public discourse in Indonesia, yet its openness has also become a systematic loophole for the spread of hate speech that threatens social cohesion. Conventional text-based detection models fail to capture hidden linguistic nuances such as slang and regional euphemisms, and they disregard the social dimension of hate propagation reflected in user interaction patterns. This study proposes a hybrid IndoBERT-GAT architecture that integrates textual semantic representation with social graph structure into a single unified framework. The research method employed 16,646 government-related tweets collected via Apify crawling, labeled through a hybrid approach (manual annotation and pseudo-labeling), then represented through IndoBERT embeddings for textual features and a Graph Attention Network to model a heterogeneous tweet-user graph, before being combined via a feature fusion mechanism and evaluated through an ablation study across four model scenarios. The results show that the full Hybrid IndoBERT-GAT model achieved the highest test F1-score (65%), outperforming the same architecture without metadata (61.8%), the GAT+metadata-only baseline (42.2%), and the MLP+metadata baseline (39.9%), demonstrating that IndoBERT's semantic representation is the dominant contributor while graph structure and metadata serve as complementary signals, although the model still tends to overpredict hate speech in politically sarcastic content that uses sharp language without genuinely hateful intent.*

Keywords*: Ablation Study, Graph Attention Network, Hate Speech, Hybrid IndoBERT-GAT*

Abstrak—Ekosistem media sosial, khususnya platform X (dahulu Twitter), telah menjadi ruang diskursus publik utama di Indonesia namun rentan terhadap penyebaran ujaran kebencian yang mengancam kohesi sosial. Model deteksi berbasis teks konvensional gagal menangkap nuansa linguistik tersembunyi seperti bahasa gaul dan eufemisme daerah, serta mengabaikan dimensi sosial penyebaran kebencian berupa pola interaksi antar pengguna. Penelitian ini mengusulkan arsitektur hybrid IndoBERT-GAT yang mengintegrasikan representasi semantik teks dan struktur graf sosial ke dalam satu kerangka terpadu. Metode penelitian menggunakan 16.646 tweet terkait isu pemerintahan yang dikumpulkan melalui crawling Apify, dilabeli secara hybrid (anotasi manual dan pseudo-labeling), kemudian direpresentasikan melalui embedding IndoBERT untuk fitur teks dan Graph Attention Network untuk memodelkan graf heterogen tweet-pengguna, sebelum digabungkan melalui mekanisme fusi fitur dan dievaluasi melalui ablation study terhadap empat skenario model. Hasil penelitian menunjukkan model Hybrid IndoBERT-GAT penuh mencapai f1-score uji tertinggi (65%), mengungguli arsitektur yang sama tanpa metadata (61,8%), baseline GAT+metadata (42,2%), dan baseline MLP+metadata (39,9%), membuktikan bahwa representasi semantik IndoBERT merupakan kontributor dominan sementara struktur graf dan metadata berperan sebagai sinyal pelengkap, meskipun model masih cenderung overpredict pada konten kritik politik yang menggunakan bahasa tajam namun tidak bermuatan kebencian.

Kata kunci: *Ablation Study, Graph Attention Network, Ujaran Kebencian, Hybrid IndoBERT-GAT*

PENDAHULUAN

Media sosial, khususnya platform X telah berkembang menjadi ruang utama bagi wacana publik di Indonesia. Indonesia menempati peringkat keempat di dunia dengan lebih dari 24,4 juta pengguna Twitter aktif, dengan penetrasi tertinggi di kalangan kelompok usia 18–34 tahun. Namun, keterbukaan platform ini sekaligus menjadi celah sistematis bagi penyebaran ujaran kebencian yang mengancam kohesi sosial dan keselamatan individu, sejalan dengan tinjauan oleh (Rawat dkk., 2024), yang menyoroti tren dan tantangan penelitian deteksi ujaran kebencian di media sosial secara umum.

Urgensi masalah ini diperkuat oleh beberapa fakta empiris. Selama pemilihan daerah tahun 2024, AJI Indonesia dan Monash University menganalisis 479.350 teks yang dikumpulkan dari X dan TikTok, di mana 49.587 teks (10,34%) diidentifikasi sebagai ujaran kebencian, dengan X menyumbang 20.335 kasus (41,01%) (Aliansi Jurnalis Independen (AJI) Indonesia; Monash University Indonesia, 2024). Pada penelitian sebelumnya, pengembangan dataset dan metode deteksi ujaran kebencian berbahasa Indonesia menunjukkan bahwa karakteristik anotasi dan keragaman kategori konten menjadi tantangan penting dalam pengembangan sistem deteksi otomatis. Temuan ini sejalan dengan studi oleh (Ibrohim & Budi, 2023) mengenai pengembangan dan tantangan dalam mendeteksi ujaran kebencian serta bahasa kasar di media sosial Indonesia.

Kegagalan pendekatan deteksi konvensional berakar pada dua kelemahan mendasar. Pertama, model berbasis teks seperti TF-IDF, Naive Bayes, dan bahkan Long Short-Term Memory (LSTM) sama sekali mengabaikan konteks semantik dari bahasa gaul, eufemisme daerah, dan kode komunitas yang umum digunakan dalam ujaran kebencian berbahasa Indonesia yang terselubung—tantangan terkait bahasa gaul dan pencampuran kode ini juga ditekankan oleh (Pamungkas & Chiril, 2025), sementara kinerja terbatas model Bi-LSTM konvensional pada teks bahasa Indonesia semakin dikonfirmasi oleh (Perwira Joan Dwitama dkk., 2023). Deteksi ujaran kebencian telah berkembang dari pendekatan berbasis fitur dan machine learning konvensional menuju metode deep learning dan pre-trained language models yang mampu menangkap representasi bahasa secara lebih kontekstual (Jahan & Oussalah, 2023). Kedua, pendekatan yang hanya memanfaatkan representasi tekstual belum sepenuhnya mempertimbangkan konteks sosial dan pola interaksi pengguna yang dapat memberikan informasi tambahan dalam memahami penyebaran ujaran kebencian di media sosial (Ibrohim & Budi, 2023). Perkembangan Pemrosesan Bahasa Alami (NLP) berbasis Transformer telah mendorong penggunaan pre-trained language models seperti BERT sebagai pendekatan utama untuk mempelajari representasi bahasa secara kontekstual dan kemudian mengadaptasikannya ke berbagai tugas *downstream* (Min dkk., 2024). Model bahasa berbasis Transformer yang telah dipra-latih pada korpus bahasa Indonesia telah menunjukkan kinerja yang kuat dalam berbagai tugas klasifikasi teks, termasuk deteksi ujaran kebencian pada Twitter berbahasa Indonesia (Hakim dkk., 2024; Kusuma & Chowanda, 2023). Menggabungkan IndoBER*tweet* dengan BiLSTM-CNN dan mencapai akurasi 85,45% serta F1-score sebesar 85,06% pada dataset ujaran kebencian di Twitter berbahasa Indonesia, sejalan dengan temuan (Kusuma & Chowanda, 2023), yang juga melaporkan kinerja tinggi dari IndoBER*tweet* yang dipadukan dengan BiLSTM untuk deteksi ujaran kebencian berbahasa Indonesia di Twitter. Namun, semua pendekatan ini masih bergantung sepenuhnya pada fitur tekstual.

Di sisi lain, Graph Attention Network (GAT) merupakan salah satu pendekatan Graph Neural Network yang menggunakan mekanisme attention untuk memberikan bobot berbeda terhadap informasi dari node tetangga, sehingga memungkinkan representasi graf mempertimbangkan kontribusi relatif dari setiap hubungan (Vrahatis dkk., 2024), sebuah pendekatan berbasis graf yang juga diterapkan oleh (Nagar dkk., 2022) menggunakan Graph Convolutional Network untuk deteksi ujaran kebencian di media sosial. Penelitian mengenai Perilaku Tidak Otentik yang Terkoordinasi (CIB) menunjukkan jaringan bot dan buzzer terkoordinasi dapat memperkuat konten kebencian hingga 1.200% lebih cepat daripada konten organik (Cinelli dkk., 2022). Bukti empiris lebih lanjut menunjukkan bahwa 78% ujaran kebencian yang lolos dari deteksi berbasis teks dapat diidentifikasi jika riwayat perilaku akun dan pola jejaring sosial diperhitungkan.

Penelitian sebelumnya memperlakukan pemodelan semantik dan sosial secara terpisah: studi berbasis IndoBERT (Hakim dkk., 2024; Kusuma & Chowanda, 2023) mengabaikan struktur jaringan, sedangkan detektor berbasis graf (Nagar dkk., 2022) menggunakan agregasi seragam GCN alih-alih perhatian, dan tidak dipasangkan dengan model bahasa Indonesia. Penelitian ini mengisi kesenjangan tersebut melalui tiga kontribusi: (1) graf *tweet*-pengguna heterogen dengan mention, balasan, dan *retweet* sebagai jenis tepi yang

VOL. 21. NO. 1 AGUSTUS 2026
P-ISSN: 0216-6933 | E-ISSN: 2685-807X
Diterbitkan Oleh: LPPM Universitas Nusa Mandiri

berbeda; (2) mekanisme perhatian GAT untuk menilai hubungan sosial mana yang paling berpengaruh terhadap penyebaran ujaran kebencian; dan (3) studi ablasi yang mengisolasi kontribusi marjinal komponen graf saat dipasangkan dengan IndoBERT, yang belum pernah dilaporkan untuk data bahasa Indonesia (Hiskiawan dkk., 2025).

BAHAN DAN METODE

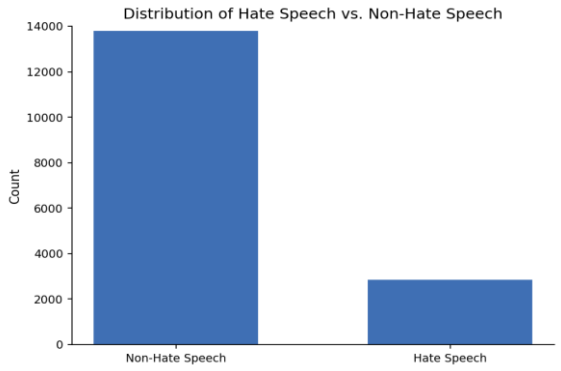
Pengumpulan dan Pra-pemrosesan Data

Data dikumpulkan menggunakan metode *crawling* melalui Apify dengan aktor Kaitoeasyapi, dengan fokus pada wacana terkait pemerintah menggunakan kata kunci "Prabowo", "Pemerintah", "Program Pemerintah", dan "Makan Bergizi Gratis". *Retweet* dipertahankan untuk melestarikan pola interaksi guna pemodelan graf, sementara *tweet* duplikat dihapus; tidak ada penyaringan akun bot yang diterapkan. Nama pengguna dianonimkan menjadi ID pengguna sebelum analisis untuk menjaga privasi data. Kumpulan data akhir terdiri dari 16.646 *tweet*, sebagaimana disajikan pada Tabel 1.

Tabel 1. Komposisi Kumpulan Data Penelitian

Kategori	Jumlah	Deskripsi
tweet	16.646	Data dikumpulkan pada Januari 2026, bahasa: Bahasa Indonesia.
Label manual	1.070	Anotasi peneliti
Label semu	15.576	Pelatihan mandiri IndoBERT

Sumber: (Hasil Penelitian, 2026)



Sumber: (Hasil Penelitian, 2026)

Gambar 1. Distribusi Kelas Ucapan Kebencian

Proses pelabelan menggunakan pendekatan campuran dengan mengacu pada kategori ujaran kebencian dan bahasa kasar yang digunakan dalam literatur deteksi ujaran kebencian berbahasa Indonesia (Ibrohim & Budi, 2023). Anotasi manual terhadap 1.070 *tweet* oleh peneliti, dan pelabelan semu otomatis terhadap 15.576 *tweet* sisanya menggunakan mekanisme pelatihan mandiri IndoBERT yang telah disempurnakan berdasarkan data yang dilabeli secara manual. Prosedur penandaan semu tersebut mengikuti skema pelatihan mandiri berulang. Kumpulan data yang dihasilkan menunjukkan ketidakseimbangan kelas yang signifikan. Dari 16.646 *tweet* yang telah diberi label, 13.645 *tweet* (81,97%) dikategorikan sebagai bukan ujaran kebencian, sedangkan 3.001 *tweet* (18,03%) dikategorikan sebagai ujaran kebencian, sebagaimana digambarkan pada Gambar 1. Ketidakseimbangan ini sejalan dengan distribusi alami konten ujaran kebencian di media sosial, di mana postingan yang secara eksplisit mengandung ujaran kebencian biasanya merupakan minoritas dibandingkan dengan wacana umum, dan kemudian diatasi selama pelatihan model melalui kerugian entropi silang tertimbang.

Tweet ditokenisasi menggunakan tokenizer IndoBERT (*WordPiece*, panjang maksimum 128 token). Model disempurnakan pada subset berlabel, lalu memprediksi label untuk *tweet* yang belum berlabel; hanya prediksi dengan kepercayaan >90% diterima sebagai *pseudo-label*. Siklus "latih–prediksi–saring" ini diulang tiga iterasi hingga seluruh 15.576 *tweet* mendapat *pseudo-label.* Untuk mencegah kebocoran data, model IndoBERT baru diinisialisasi dan dilatih ulang dari awal pada dataset lengkap sebelum digunakan untuk ekstraksi embedding [CLS] pada arsitektur hibrida.

Untuk menilai keandalan proses anotasi manual, kesesuaian antar-anotator diukur di antara tiga anotator yang secara independen memberi label pada subset 1.070 tweet menggunakan kategori ujaran kebencian dan bahasa kasar yang digunakan dalam literatur deteksi ujaran kebencian berbahasa Indonesia (Ibrohim & Budi, 2023). Tanpa adanya diskusi di antara para anotator, Cohen's Kappa dihitung untuk setiap pasangan anotator guna mengukur tingkat kesepakatan antar-anotator pada data kategorikal dengan memperhitungkan kemungkinan kesepakatan yang terjadi secara kebetulan (Benomar dkk., 2023). Berdasarkan 1.058 *tweet* dengan triplet label lengkap, sebagaimana disajikan dalam Tabel 2. Semua koefisien pasangan berkisar antara $\kappa = 0{,}808$ hingga $\kappa = 0{,}840$ (rata-rata $\kappa = 0{,}827$), Nilai Cohen's Kappa yang diperoleh berada pada kisaran 0,808–0,840 dengan rata-rata 0,827, yang menunjukkan tingkat kesepakatan antar-anotator yang tinggi berdasarkan kategori interpretasi Kappa yang umum digunakan dalam literatur inter-rater reliability (Benomar dkk., 2023), sehingga menegaskan keandalan proses pelabelan. Ketidaksepakatan yang terjadi bersifat minor dan lebih bersifat interpretatif daripada sistematis,

VOL. 21. NO. 1 AGUSTUS 2026
P-ISSN: 0216-6933 | E-ISSN: 2685-807X
Diterbitkan Oleh: LPPM Universitas Nusa Mandiri
INTI NUSA MANDIRI
DOI: https://doi.org/10.33480/inti.v21i1.8699

sejalan dengan tingginya kesesuaian antar-pasangan anotator yang telah dilaporkan sebelumnya. Mengingat rendahnya tingkat ketidaksepakatan tersebut, *ground-truth* akhir ditentukan oleh Anotator 1, yang bertindak sebagai peneliti utama pada penelitian ini dan karenanya memegang tanggung jawab utama dalam interpretasi taksonomi, sehingga memastikan standar pelabelan yang tunggal dan konsisten. Kumpulan data yang telah difinalisasi kemudian dibagi menjadi set pelatihan, validasi, dan pengujian dengan perbandingan 70:15:15.

Tabel 2. Kesesuaian Anotator

Pasangan Anotator	Koefisien Kappa Cohen	Tingkat Kesesuaian
Anotator1–Anotator2	0,832	Kesepakatan Hampir Sempurna
Anotator1–Anotator3	0,808	Kesepakatan yang Signifikan
Anotator2–Anotator3	0,840	Kesepakatan Hampir Sempurna
Rata-rata	0,827	Kesepakatan Hampir Sempurna

Sumber: (Hasil Penelitian, 2026)

Rekayasa Fitur

Penelitian ini menggunakan dua jenis representasi fitur secara bersamaan. Fitur tekstual diekstraksi menggunakan IndoBERT (indobenchmark/indobert-base-p1) dengan tokenizer sub-kata dan panjang maksimum 128 token. Representasi [CLS] berdimensi 768 digunakan sebagai representasi semantik setiap tweet (Kusuma & Chowanda, 2023). Hal ini mencerminkan pola yang lebih luas di mana teknik NLP terapan seperti terjemahan mesin dan pelatihan keterampilan ekstraksi kata kunci bagi siswa sekolah kejuruan Indonesia (Hiskiawan, Indarwan, dkk., 2026) terus menunjukkan relevansi praktis pemrosesan teks berbasis transformer di berbagai konteks pendidikan dan terapan. Hal ini memainkan peran penting dalam memungkinkan jaringan saraf untuk menangkap sinyal yang tertanam secara budaya dan linguistik. Fitur metadata pengguna terdiri dari delapan fitur: *followers_count*, *following_count*, *statuses_count*, *media_count*, *account_age*, *is_blue_verified*, *favorites_count*, dan *can_dm*. Fitur numerik dinormalisasi menggunakan standarisasi *z-score*.

Konstruksi Grafik Heterogen

Graf heterogen G = (V, E) dibangun untuk merepresentasikan interaksi *tweet*-pengguna dalam satu struktur terintegrasi. Spesifikasi simpul dan tepi ditunjukkan pada Tabel 3.

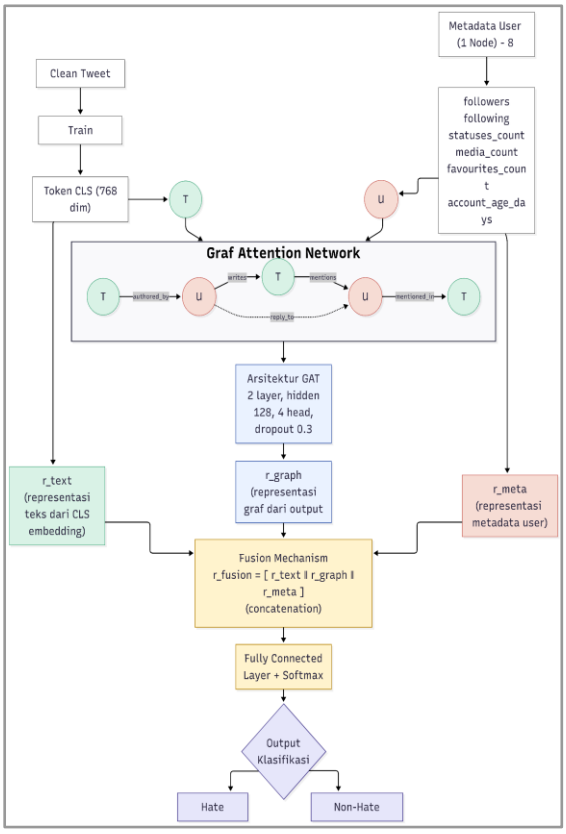
Tabel 3. Spesifikasi Grafik Heterogen

Komponen	Jenis	Deskripsi
Node tweet	*Node*	Embedding [CLS] berdimensi 768
Node Pengguna	*Node*	Metadata profil (8 fitur)
Authored_by	*Edge*	*tweet* → Pengguna (*author_Relation*)
mentions	*Edge*	*tweet* → Pengguna (disebutkan dalam teks)
mentioned_in	*Edge*	Pengguna → *tweet* (pengguna yang disebutkan)
Menulis	*Edge*	Pengguna → *tweet* (*author_relation* terbalik)

Sumber: (Hasil Penelitian, 2026)

Arsitektur Hibrida IndoBERT-GAT

Gambar 2. mengilustrasikan arsitektur keseluruhan dari model Hybrid IndoBERT-GAT yang diusulkan, yang menunjukkan bagaimana cabang teks, cabang graf, dan cabang metadata diproses secara paralel sebelum digabungkan melalui mekanisme fusi fitur.



Sumber: (Hasil Penelitian, 2026)

Gambar 2. Arsitektur Hybrid IndoBERT & GAT

Arsitektur yang diusulkan mengintegrasikan tiga komponen melalui mekanisme fusi fitur, dengan definisi formal yang dirangkum dalam Tabel 4. Pertama, cabang teks memproses *tweet* melalui

INTI NUSA MANDIRI
DOI: https://doi.org/10.33480/inti.v21i1.8699
VOL. 21. NO. 1 AGUSTUS 2026
P-ISSN: 0216-6933 | E-ISSN: 2685-807X
Diterbitkan Oleh: LPPM Universitas Nusa Mandiri

IndoBERT untuk menghasilkan $rtextr_{text}rtext$, yang disesuaikan secara khusus menggunakan laju pembelajaran sebesar 2×10^{-5} dan ukuran batch sebesar 32 berdasarkan konfigurasi eksperimen yang digunakan. Kedua, cabang graf memproses graf heterogen $G = (V, E)G = (V, E)G = (V, E)$ menggunakan GAT, menghasilkan $rgraphr_{graph}rgraph$ melalui agregasi informasi dari node tetangga dengan bobot attention yang berbeda sesuai kontribusi masing-masing hubungan (Vrahatis dkk., 2024), sejalan dengan penelitian terapan terkini mengenai klasifikasi teks media sosial Indonesia berbasis transformer (Winson & Hiskiawan, 2026). Ketiga, mekanisme fusi menggabungkan ketiga representasi tersebut menjadi $rfusionr_{fusion}rfusion$, yang diklasifikasikan oleh lapisan terhubung penuh dengan aktivasi softmax, mengikuti prinsip fusi fitur yang diterapkan oleh (Miao dkk., 2024) dan (Hiskiawan, Sari, dkk., 2026). Pendekatan transfer learning berbasis BERT menunjukkan efektivitas yang tinggi dalam deteksi ujaran kebencian pada data media sosial, termasuk pada kondisi sumber daya berlabel yang terbatas (Ali dkk., 2022). Untuk mengatasi ketidakseimbangan kelas, model dilatih menggunakan kerugian entropi silang tertimbang (Tabel 4).

Tabel 4. Komponen Arsitektur Hybrid Dan Definisi Formal

Komponen	Rumus.	Deskripsi
GAT attention	$\alpha ij \propto exp(aT[Whi\|Whj])$	*Attention weight node* i–j
Graph rep.	$rgraph = \sigma(\Sigma\alpha ijWhj)$	2 *layers*, dim 128, 4 *heads*, *dropout* 0.3
Text rep.	$rtext = [CLS]$	768-dim, lr=2×10^{-5}, batch=32
Feature fusion	$rfusion = [rtext\|rgraph\|rm$	*Concatenation*, 3 *reps*.
Classifier	$\hat{y} = softmax(Wcrfusion +$	*FC layer* + *softmax*
Loss fn.	$L = -\Sigma wcyclog(\hat{y}c)$	*Weighted cross-entropy*
Normalization	$z = (x - \mu)/\sigma$	*Z-score*, 8 *features*

Source: (Hasil Penelitian, 2026)

Skenario Pengujian

Evaluasi dilakukan melalui studi ablasi menggunakan empat skenario eksperimental, seperti yang ditunjukkan pada Tabel 5. Metrik utama adalah F1-score, yang memberikan bobot seimbang kepada setiap kelas dan sangat penting untuk menangani distribusi kelas yang tidak seimbang dalam dataset ujaran kebencian. Metrik pendukung meliputi Presisi, Recall, dan AUC-ROC.

Tabel 5. Skenario Studi Ablasi

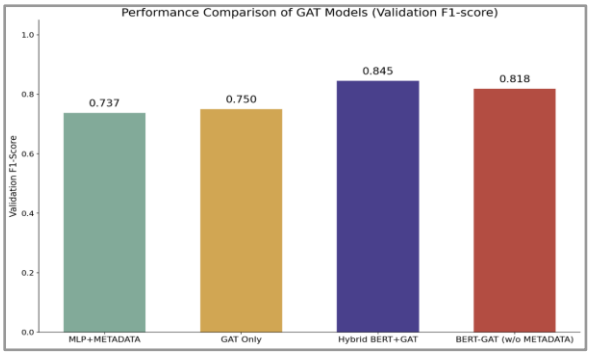
Model	Deskripsi
IndoBERT-GAT Hibrida	Model yang diusulkan: IndoBERT + grafik heterogen + metadata pengguna
GAT Murni (dengan METADATA)	Eksperimen 1: metadata dan struktur graf hanya dengan METADATA, tanpa IndoBERT
MLP + METADATA	Eksperimen 2: Multi Layer Perceptron sebagai pengganti Hybrid IndoBERT+GAT
Hybrid IndoBERT-GAT (tanpa METADATA)	Eksperimen 3: IndoBERT + graf heterogen (tanpa METADATA)

Sumber: (Hasil Penelitian, 2026)

HASIL DAN PEMBAHASAN

Kinerja Model pada Data Pelatihan dan Validasi

Sebagaimana diilustrasikan pada Gambar 3, perbandingan F1-score validasi dilakukan pada empat skenario ablasi sebelum pemilihan model akhir: MLP+METADATA mencapai F1-score validasi sebesar 0,737, GAT murni mencapai 0,750, Hybrid IndoBERT-GAT mencapai skor tertinggi kedua sebesar 0,818, dan model Hybrid IndoBERT-GAT mencapai skor tertinggi sebesar 0,845.



Sumber: (Hasil Penelitian, 2026)

Gambar 3. Perbandingan Model GAT

Perbandingan ini menegaskan bahwa penggabungan representasi semantik IndoBERT ke dalam arsitektur berbasis graf menghasilkan peningkatan yang jelas dibandingkan dengan dua konfigurasi dasar, bahkan pada tahap validasi sebelum proses pelatihan akhir selesai. Berdasarkan perbandingan awal ini, model Hybrid IndoBERT-GAT kemudian dilatih selama 50 epoch menggunakan optimizer AdamW, yang dipilih karena mekanisme laju pembelajaran adaptif dan regularisasi *weight decay* yang terpisah, yang sangat cocok untuk penyempurnaan arsitektur berbasis *transformer* yang dikombinasikan dengan

VOL. 21. NO. 1 AGUSTUS 2026
P-ISSN: 0216-6933 | E-ISSN: 2685-807X
Diterbitkan Oleh: LPPM Universitas Nusa Mandiri
INTI NUSA MANDIRI
DOI: https://doi.org/10.33480/inti.v21i1.8699

komponen jaringan saraf graf. Sepanjang proses pelatihan, F1-score pada data validasi menunjukkan tren kenaikan yang secara umum stabil, yang mengindikasikan bahwa model secara bertahap mampu mempelajari pola-pola diskriminatif dari representasi gabungan teks dan struktur graf.

Terlepas dari tren positif ini, terjadi fluktuasi pada epoch ke-10 hingga ke-13, di mana F1-score validasi sempat menurun sebelum pulih kembali. Pola ini lazim terjadi pada arsitektur hibrida, karena mekanisme perhatian pada IndoBERT dan lapisan graf memerlukan beberapa epoch untuk menstabilkan bobot sebelum mencapai keseimbangan yang saling melengkapi. Setelah fase ini, kinerja model meningkat konsisten dan mendekati konvergensi, menunjukkan mekanisme fusi telah mencapai keadaan optimasi yang stabil. Untuk memastikan bahwa model akhir mencerminkan kemampuan generalisasi optimalnya, bukan keadaan *overfitting* yang berpotensi terjadi pada epoch pelatihan terakhir, snapshot bobot model yang sesuai dengan kinerja validasi terbaik disimpan dan kemudian digunakan sebagai model akhir untuk evaluasi pada set uji. Snapshot ini mencapai F1-Score validasi terbaik sebesar 0,8697, yang lebih baik lagi dibandingkan 0,845 yang tercatat pada perbandingan awal yang ditunjukkan pada Gambar 3, dan berfungsi sebagai titik pemeriksaan yang dibawa ke dalam perbandingan kinerja model dan analisis studi ablasi selanjutnya.

Perbandingan Kinerja Model

Tabel 6. Perbandingan Kinerja Model (Set Uji)

Model	F1-Score	Presisi	Recall	Akurasi	AUC-ROC
Hibrida IndoBERT-GAT	65%	51,6%	89,5 %	82,9 %	91,5%
Hanya GAT+METADATA	42,2 %	28,5%	81,3 %	59,8 %	69,4%
MLP + METADATA	39,9 %	34,4%	46,6 %	74,4 %	77,9%
Hibrida indoBERT-GAT (tanpa METADATA)	61,8 %	46,8%	90,9 %	79,9 %	91,2%

Sumber: (Hasil Penelitian, 2026)

Analisis Studi Ablasi

1) Kontribusi representasi semantik IndoBERT: perbandingan kinerja antara Hybrid IndoBERT-GAT (F1 65%) dan MLP+METADATA (F1 39,9%) menunjukkan selisih lebih dari 25 poin, meskipun kedua model tersebut menggunakan struktur graf dan metadata pengguna yang identik. Selisih ini menunjukkan bahwa representasi semantik IndoBERT berkontribusi secara substansial dalam menangkap nuansa linguistik tersembunyi dari ujaran kebencian seperti eufemisme regional dan bahasa gaul yang tidak dapat ditangkap oleh MLP, sejalan dengan perbandingan kinerja berbagai pendekatan pembelajaran mendalam berbasis embedding IndoBER*tweet* yang dilaporkan oleh (Mufva dkk., 2025) dalam deteksi ujaran kebencian di Twitter berbahasa Indonesia.

2) Perilaku struktur graf dan metadata tanpa representasi teks yang bermakna: perbandingan antara GAT murni (F1 42,2%; recall 81,3%; presisi 28,5%; akurasi 59,8%) dan MLP+METADATA (F1 39,9%; recall 46,6%; presisi 34,4%; akurasi 74,4%) menunjukkan kinerja keseluruhan yang relatif serupa dan jauh di bawah model hibrida, namun dengan pola kesalahan yang berbeda. GAT murni cenderung memprediksi berlebihan kelas ujaran kebencian (recall tinggi, presisi dan akurasi rendah), kemungkinan karena sinyal jaringan dan metadata akun (misalnya, akun baru, jumlah pengikut rendah) memiliki korelasi lemah dengan indikasi bot/buzzer tanpa dapat memastikan adanya konten ujaran kebencian yang sesungguhnya. Sebaliknya, baseline MLP lebih konservatif, menghasilkan akurasi yang lebih tinggi tetapi recall yang jauh lebih rendah pada kelas minoritas.

3) Kontribusi terpisah dari metadata pengguna tidak dapat diukur dalam penelitian ini, karena keempat skenario yang dievaluasi selalu menyertakan metadata pengguna sebagai bagian dari struktur graf. Penelitian selanjutnya disarankan untuk menambahkan skenario ablasi lebih lanjut (misalnya, IndoBERT+GAT tanpa metadata pengguna) guna mengisolasi kontribusi fitur ini secara terpisah dari struktur jaringan.

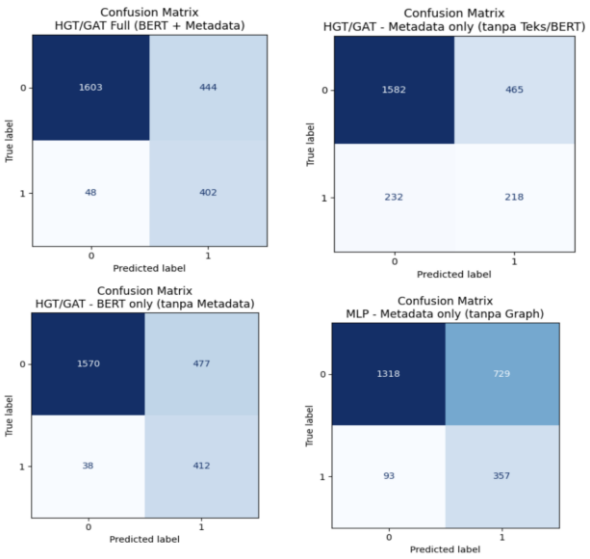
Analisis Kesalahan Klasifikasi

Berdasarkan nilai presisi (51,6%) dan recall (89,5%) dari model Hybrid IndoBERT-GAT pada Tabel 6, model tersebut menunjukkan kecenderungan untuk terlalu sering mengklasifikasikan sebagai ucapan kebencian: proporsi false positive relatif lebih tinggi daripada false negative. Pola ini konsisten dengan nilai recall yang tinggi namun presisi yang lebih moderat, dan kemungkinan besar terjadi pada kritik tajam atau sarkasme politik terhadap isu-isu pemerintah yang menggunakan bahasa yang tajam tanpa konten kebencian yang sesungguhnya, di mana sinyal media sosial semakin mendorong klasifikasi ke arah

kelas ujaran kebencian. Asimetri presisi-recall ini terutama dapat dikaitkan dengan fungsi kerugian entropi silang tertimbang yang diterapkan selama pelatihan untuk mengatasi ketidakseimbangan kelas antara *tweet* ujaran kebencian dan non-ujaran kebencian.

Dengan memberi bobot lebih tinggi pada kelas ujaran kebencian minoritas, tujuan pelatihan memberikan penalti lebih berat pada false negatives, mendorong batas keputusan model ke arah sensitivitas lebih tinggi dengan mengorbankan spesifisitas. Perilaku ini diperparah dua faktor sekunder: mekanisme graph attention dapat menyebarkan sinyal condong ke ujaran kebencian melalui tepi *authored_by* dan *mentions* saat *tweet* berinteraksi dengan akun mirip bot/buzzer, dan representasi IndoBERT kurang diskriminatif terhadap permusuhan implisit seperti sarkasme politik, di mana bahasa tajam digunakan tanpa niat kebencian sesungguhnya. Strategi pembobotan kelas tetap menjadi pendorong utama, dengan faktor graf dan semantik memperkuat bukan menyebabkan secara independen bias tersebut, sebuah trade-off yang bisa dibilang lebih disukai dalam moderasi konten, di mana kegagalan mendeteksi ujaran kebencian sesungguhnya memiliki biaya sosial lebih tinggi daripada menandai berlebihan konten ambang batas untuk ditinjau manusia.

Gambar 4 menyajikan *Confusion Matrix* dari keempat skenario yang dievaluasi, yang menawarkan perbandingan visual dari distribusi positif palsu dan negatif palsu yang mendasari *trade-off Precision-recall* yang dilaporkan dalam Tabel 6.



Sumber: (Hasil Penelitian, 2026)

Gambar 4. Matriks Kebingungan dari Keempat Eksperimen

KESIMPULAN

Penelitian ini mengusulkan arsitektur Hybrid IndoBERT-GAT yang menggabungkan representasi semantik IndoBERT, struktur interaksi sosial melalui Jaringan Perhatian Graf (*Graph Attention Network*), dan metadata pengguna untuk mengatasi kelemahan model berbasis teks dalam mendeteksi ujaran kebencian terselubung serta dimensi sosial dari penyebaran ujaran kebencian. Model ini mencapai F1-score validasi sebesar 0,8697 dan F1-score pengujian sebesar 65%, melampaui *baseline* GAT murni (42,2%) dan baseline MLP+METADATA (39,9%), yang menunjukkan bahwa representasi semantik IndoBERT merupakan kontributor utama sementara struktur graf berfungsi sebagai sinyal pelengkap, meskipun model ini masih cenderung melebih-lebihkan prediksi ujaran kebencian dalam konten politik. Penelitian lebih lanjut disarankan untuk mengeksplorasi Heterogeneous Graph *transformer*, *co-conversation edges* untuk deteksi tingkat utas, serta studi ablasi tambahan guna mengisolasi kontribusi metadata.

DAFTAR PUSTAKA

Ali, R., Farooq, U., Arshad, U., Shahzad, W., & Beg, M. O. (2022). Hate speech detection on Twitter using transfer learning. *Computer Speech & Language*, *74*, 101365. https://doi.org/10.1016/j.csl.2022.101365

Aliansi Jurnalis Independen (AJI) Indonesia; Monash University Indonesia. (2024). *Report on Hate Speech Monitoring in the 2024 Indonesia Regional Elections*. https://www.monash.edu/indonesia/news/report-on-hate-speech-monitoring-in-the-2024-indonesia-regional-elections

Benomar, A., Zarour, E., Létourneau-Guillon, L., & Raymond, J. (2023). Measuring Interrater Reliability. *Radiology*, *309*(3). https://doi.org/10.1148/radiol.230492

Cinelli, M., Cresci, S., Quattrociocchi, W., Tesconi, M., & Zola, P. (2022). Coordinated inauthentic behavior and information spreading on Twitter. *Decision Support Systems*, *160*, 113819. https://doi.org/10.1016/j.dss.2022.113819

Hakim, A. N., Sibaroni, Y., & Prasetyowati, S. S. (2024). Detection of Hate-Speech Text on Indonesian Twitter Social Media Using IndoBERTweet-BiLSTM-CNN. *2024 12th International Conference on Information and Communication Technology (ICoICT)*, 374–381. https://doi.org/10.1109/ICoICT61617.2024.10698615

Hiskiawan, P., Geasela, Y. M., Heryanto, H., Stephanie, E., Ardianti, M., & Sukarno, F. A. (2025). Trustworthy Data Science Framework for Non-Invasive Nutritional Screening Using Computer Vision. *2025 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS)*, 1743–1748. https://doi.org/10.1109/ICIMCIS68501.2025.11327268

Hiskiawan, P., Indarwan, M. S., Ho, C. A., & Sutojo, K. S. B. (2026). Pemberdayaan Siswa SMK DKV melalui Praktikum Interaktif Natural Language Processing. *PengabdianMu: Jurnal Ilmiah Pengabdian kepada Masyarakat*, *11*(3), 949–956. https://doi.org/10.33084/pengabdianmu.v11i3.11700

Hiskiawan, P., Sari, M. K., Siregar, R. E., Valencia, N. A., & Wijayanti, T. P. (2026). A Deep Learning Data Fusion Approach for Prostate MRI Zonal Segmentation Using Dual-Channel U-Net with T2 and ADC Images. *2026 International Conference on Current Research in Artificial Intelligence and Data Science (ICCRAIDS)*, 1–7. https://doi.org/10.1109/ICCRAIDS67816.2026.11519671

Ibrohim, M. O., & Budi, I. (2023). Hate speech and abusive language detection in Indonesian social media: Progress and challenges. *Heliyon*, *9*(8), e18647. https://doi.org/10.1016/j.heliyon.2023.e18647

Jahan, M. S., & Oussalah, M. (2023). A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, *546*, 126232. https://doi.org/10.1016/j.neucom.2023.126232

Kusuma, J. F., & Chowanda, A. (2023). Indonesian Hate Speech Detection Using IndoBERTweet and BiLSTM on Twitter. *JOIV : International Journal on Informatics Visualization*, *7*(3), 773–780. https://doi.org/10.30630/joiv.7.3.1035

Miao, Z., Chen, X., Wang, H., Tang, R., Yang, Z., Huang, T., & Tang, W. (2024). Detecting Offensive Language Based on Graph Attention Networks and Fusion Features. *IEEE Transactions on Computational Social Systems*, *11*(1), 1493–1505. https://doi.org/10.1109/TCSS.2023.3250502

Min, B., Ross, H., Sulem, E., Veyseh, A. P. Ben, Nguyen, T. H., Sainz, O., Agirre, E., Heintz, I., & Roth, D. (2024). Recent Advances in Natural Language Processing via Large Pre-trained Language Models: A Survey. *ACM Computing Surveys*, *56*(2), 1–40. https://doi.org/10.1145/3605943

Mufva, P. A., Chandra, K. H., Aji, K. F., Iswanto, I. A., & Joddy, S. (2025). Performance comparison of deep learning approaches for Indonesian twitter hate speech detection using IndoBERTweet embedding. *Procedia Computer Science*, *269*, 1663–1671. https://doi.org/10.1016/j.procs.2025.09.109

Nagar, S., Gupta, S., Bahushruth, C. S., Barbhuiya, F. A., & Dey, K. (2022). *Hate Speech Detection on Social Media Using Graph Convolutional Networks* (hlm. 3–14). https://doi.org/10.1007/978-3-030-93413-2_1

Pamungkas, E. W., & Chiril, P. (2025). Ngalawan Ujaran Sengit: hate speech detection in indonesian code-mixed social media data. *Language Resources and Evaluation*, *59*(3), 2387–2414. https://doi.org/10.1007/s10579-025-09810-x

Perwira Joan Dwitama, A., Hatta Fudholi, D., & Hidayat, S. (2023). Indonesian Hate Speech Detection Using Bidirectional Long Short-Term Memory (Bi-LSTM). *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, *7*(2), 302–309. https://doi.org/10.29207/resti.v7i2.4642

Rawat, A., Kumar, S., & Samant, S. S. (2024). Hate speech detection in social media: Techniques, recent trends, and future challenges. *WIREs Computational Statistics*, *16*(2). https://doi.org/10.1002/wics.1648

Vrahatis, A. G., Lazaros, K., & Kotsiantis, S. (2024). Graph Attention Networks: A Comprehensive Review of Methods and Applications. *Future Internet*, *16*(9), 318. https://doi.org/10.3390/fi16090318

Winson, W., & Hiskiawan, P. (2026). An Applied Data Science Approach for Detecting Depression Symptoms in Indonesian Social Media Text Using Transformer Models. *Journal of Applied Informatics and Computing*, *10*(3), 2115–2127. https://doi.org/10.30871/jaic.v10i3.12644

