

Ujaran Kebencian di Media Sosial dalam Perspektif Linguistik Forensik: Kajian Literatur

Putri Octavia Simbolon¹, Ida Binneka², Irma Yanti Sitorus³, Lili Tansliova⁴

^{1,2,3,4}Universitas Negeri Medan

*Email Korespondensi: putrioctaviaas@gmail.com¹, binnekaida@gmail.com²,
irmayantisitorus4@gmail.com³

ARTICLE INFORMATION

Article History:	ABSTRAK
Published: 24 October 2026	Ujaran kebencian di media sosial tidak selalu tampil sebagai kata kasar, tetapi dapat terbentuk melalui rujukan identitas, tindak tutur, konteks percakapan, dan unsur visual. Penelitian ini bertujuan menjelaskan kontribusi linguistik forensik dalam menganalisis ujaran kebencian digital serta merumuskan unsur pemeriksaan yang relevan bagi pembuktian. Penelitian menggunakan kajian literatur kualitatif terarah terhadap artikel ilmiah, buku, dokumen kebijakan, dan peraturan yang relevan, terutama terbitan 2016–2025. Data dianalisis melalui pembacaan, pengodean tematik, dan sintesis konseptual secara sistematis untuk menemukan hubungan antartema. Hasil kajian menunjukkan bahwa ujaran kebencian perlu dibedakan dari kritik dan ujaran kasar dengan mempertimbangkan sasaran identitas, bentuk leksikal dan gramatikal, fungsi pragmatik, serta risiko diskriminasi. Penafsiran juga memerlukan konteks percakapan, unsur multimodal, dan dokumentasi bukti digital yang dapat ditelusuri. Sistem deteksi otomatis membantu penyaringan awal, tetapi belum menggantikan pemeriksaan manusia. Penelitian ini menyimpulkan bahwa analisis linguistik forensik harus kontekstual, transparan, dan membatasi kesimpulan pada bukti kebahasaan tanpa mengambil alih penilaian hukum.
Kata Kunci: ujaran kebencian; media sosial; linguistik forensik; pragmatik; bukti digital.	
Keyword: hate speech; social media; forensic linguistics; pragmatics; digital evidence.	ABSTRACT Hate speech on social media does not always appear as offensive language; it can also emerge through identity references, speech acts, conversational context, and visual elements. This study examines the contribution of forensic linguistics to interpreting digital hate speech and identifies analytical components relevant to evidentiary assessment. It employs a focused qualitative literature review of scholarly articles, books, policy documents, and regulations, primarily published between 2016 and 2025. Sources were examined through close reading, thematic coding, and conceptual synthesis. Findings indicate that hate speech must be distinguished from criticism and offensive expressions by considering identity targets, lexical and grammatical features, pragmatic functions, and risks of discrimination. Interpretation also requires conversational context, multimodal elements, and traceable documentation of digital evidence. Automated detection supports preliminary screening but cannot replace human analysis. The study concludes that forensic linguistic assessment should remain contextual, transparent, and limited to linguistic evidence without substituting for legal judgment.
Page: 165 - 179	
DOI: https://doi.org/10.36709/pesastra.v3i4.184	

© 2026 The Author(s). PESAstra: Pendidikan Bahasa dan Sastra by PT Casa Cendekia Media

PESAstra: Pendidikan Bahasa dan Sastra an open access article under terms of Creative Commons Attribution-Share Alike International License (CC BY 4.0)

Pendahuluan

Media sosial memperluas kesempatan warga untuk menyampaikan pendapat, membangun jejaring, dan menanggapi peristiwa publik secara cepat. Pada saat yang sama, ruang ini memungkinkan penghinaan, pelabelan kelompok, ajakan permusuhan, dan narasi yang mengubah perbedaan identitas menjadi alasan untuk mengecualikan pihak lain. Persoalan tersebut tidak cukup dipahami sebagai masalah etika berkomunikasi. Ujaran yang beredar melalui komentar, unggahan, video singkat, atau percakapan berantai dapat memengaruhi persepsi sosial, memperkuat prasangka, dan dalam keadaan tertentu membangun risiko nyata terhadap diskriminasi, permusuhan, atau kekerasan. Karena itu, pembacaan atas ujaran kebencian harus menjaga dua kepentingan sekaligus, yaitu perlindungan atas martabat kelompok rentan dan kehati-hatian agar kritik, ketidaksetujuan, atau ekspresi yang tidak menyenangkan tidak langsung diposisikan sebagai tindak pidana (Council of Europe, 2022; Undang-Undang Republik Indonesia Nomor 1 Tahun 2024).

Kehati-hatian itu semakin penting dalam konteks Indonesia yang ditandai oleh keragaman bahasa, etnis, agama, dan praktik komunikasi digital. Revisi kedua Undang-Undang Informasi dan Transaksi Elektronik memberi batas yang lebih tegas terhadap muatan kebencian berdasarkan identitas tertentu. Dalam penjelasan yang tersedia pada basis data peraturan, norma tersebut dihubungkan dengan tindakan yang dilakukan secara sengaja dan di depan umum serta memunculkan risiko nyata terhadap diskriminasi, permusuhan, atau kekerasan (Undang-Undang Republik Indonesia Nomor 1 Tahun 2024). Rumusan ini menunjukkan bahwa pemeriksaan tidak boleh berhenti pada keberadaan kata yang bernada negatif. Penilai perlu menelaah siapa atau kelompok apa yang menjadi sasaran, bagaimana ujaran itu bekerja dalam konteksnya, dan apakah bentuk bahasa tersebut memuat dorongan, penghasutan, atau konstruksi permusuhan yang relevan secara hukum.

Dalam kajian internasional, istilah **hate speech** sendiri tidak selalu dipakai dengan batas yang sama. Sejumlah penelitian membedakan ujaran kebencian dari bahasa kasar, pelecehan, toksisitas, dan serangan personal. Perbedaan tersebut muncul karena setiap kategori mempunyai sasaran, derajat ancaman, dan fungsi pragmatik yang berbeda. Fortuna dan Nunes (2018) memperlihatkan bahwa perbedaan definisi, skema pelabelan, dan karakter korpus membentuk variasi besar dalam penelitian deteksi ujaran kebencian. Poletto et al. (2021) juga menunjukkan bahwa sumber data dan kerangka anotasi menentukan apa yang akhirnya diakui sebagai ujaran kebencian. Dengan demikian, satu kata atau frasa tidak dapat diberi makna hukum tanpa melihat hubungan bentuk kebahasaan dengan konteks penggunaannya.

Penelitian tentang ujaran kebencian di media sosial berkembang melalui dua arus utama. Arus pertama berfokus pada pengembangan sumber data dan teknik deteksi otomatis. Penelitian ini menghasilkan korpus, model klasifikasi, serta pengujian terhadap kemampuan sistem untuk mengenali teks eksplisit maupun implisit. Arus kedua menempatkan ujaran sebagai tindakan sosial dan bukti bahasa yang memerlukan pembacaan makna, relasi penutur, sasaran, serta konteks peredaran. Kedua arus tersebut saling melengkapi, tetapi tidak identik. Sistem komputasional dapat menyaring volume data yang besar, sedangkan linguistik forensik menjelaskan mengapa suatu ungkapan bermakna merendahkan, menghasut, atau tidak memenuhi unsur ujaran kebencian dalam konteks tertentu (Coulthard, Johnson, & Wright, 2016; Rawat, Kumar, & Samant, 2024).

Dalam studi Indonesia, perhatian terhadap bentuk kebahasaan, tindak tutur, dan hubungan dengan norma hukum telah muncul pada data Instagram, Facebook, dan komentar politik. Suryani, Istianingrum, dan Hanik (2021) menunjukkan penggunaan linguistik forensik untuk membaca ujaran terhadap figur publik di Instagram. Nuraeni, Harisanti, dan Maksum (2022) menganalisis tuturan kebencian dalam komentar politik melalui perspektif pragmatik. Syahid dan Datang (2024) memperlihatkan bahwa teks dan gambar dapat bekerja

bersama membangun sentimen etnis dalam unggahan Facebook. Kajian-kajian tersebut penting karena menggeser perhatian dari kata tunggal menuju fungsi ujaran dan lingkungan digitalnya. Namun, fokus pada satu kasus atau satu platform masih menyisakan kebutuhan untuk menyatukan unsur kebahasaan, konteks digital, dokumentasi bukti, dan batas hukum dalam satu kerangka pembacaan.

Artikel ini mengambil posisi pada kebutuhan tersebut. Kajian ini tidak mengklaim membangun definisi hukum baru atau menggantikan penilaian aparat penegak hukum. Tujuannya adalah menyintesis literatur untuk menjawab dua pertanyaan. Pertama, unsur apa saja yang diperlukan dalam analisis ujaran kebencian media sosial dari perspektif linguistik forensik. Kedua, bagaimana unsur tersebut dapat digunakan secara bertahap agar pembacaan bahasa tidak menyederhanakan masalah menjadi sekadar pencarian kata kasar. Kontribusi artikel terletak pada penyusunan kerangka analitis yang menghubungkan bentuk ujaran, tindakan komunikatif, target, konteks percakapan, moda visual, dan pengelolaan bukti digital. Kerangka ini relevan bagi peneliti bahasa, pemeriksa konten, dan pihak yang memerlukan alasan analitis yang dapat ditelusuri ketika menilai suatu unggahan.

Metode

Penelitian ini menggunakan kajian literatur kualitatif terarah. Desain tersebut dipilih karena tujuan penelitian bukan mengukur frekuensi ujaran pada satu populasi ataupun menguji kinerja sebuah model, melainkan menafsirkan dan menyintesis konsep, temuan, serta prosedur analitis yang tersebar dalam literatur. Sumber utama mencakup artikel jurnal, prosiding ilmiah, buku pengantar linguistik forensik, dokumen kebijakan, dan peraturan yang berhubungan langsung dengan ujaran kebencian, bahasa digital, pragmatik, multimodalitas, serta pembuktian bahasa. Publikasi yang diprioritaskan terbit pada 2016–2025 agar pembahasan merekam perkembangan media sosial dan sistem deteksi terkini. Dokumen hukum Indonesia digunakan sebagai rujukan normatif untuk menjelaskan konteks, bukan sebagai objek analisis putusan.

Penelusuran sumber dilakukan melalui laman penerbit, repositori ilmiah, dan basis publikasi yang dapat diverifikasi. Kata kunci yang digunakan memadukan istilah **hate speech**, **online hate**, **forensic linguistics**, **social media**, **implicit hate speech**, **multimodal evidence**, serta padanan Indonesia seperti ujaran kebencian, linguistik forensik, dan media sosial. Sumber dipilih apabila membahas sedikitnya satu dari empat fokus: pembeda ujaran kebencian dan ujaran ofensif, analisis kebahasaan atau pragmatik, konteks serta multimodalitas, atau keterbatasan deteksi otomatis dan bukti digital. Artikel yang hanya membahas arsitektur model tanpa implikasi bahasa atau konteks tetap dipakai secara terbatas ketika diperlukan untuk menjelaskan batas penggunaan teknologi.

Analisis berlangsung melalui empat tahap. Pertama, penulis membaca judul, abstrak, dan bagian pembahasan sumber untuk menetapkan relevansi. Kedua, informasi penting dikodekan ke dalam kategori awal, yaitu definisi dan batas kategori, bentuk kebahasaan, tindak tutur dan target, konteks interaksi, multimodalitas, dokumentasi bukti, serta otomatisasi. Ketiga, kategori dibandingkan untuk menemukan hubungan, ketegangan, dan bagian yang berulang di antara sumber. Keempat, hasil perbandingan dirumuskan menjadi tema pembahasan. Keterlacakan analisis dijaga melalui pencatatan sumber pada setiap klaim yang berasal dari literatur. Kajian ini tidak menggunakan wawancara, observasi, atau data unggahan baru sehingga tidak menyajikan kutipan pengguna media sosial sebagai data primer.

Batas kajian perlu ditegaskan. Artikel ini tidak menilai legalitas kasus individual dan tidak menyediakan prosedur pengambilan data untuk perkara tertentu. Kesimpulan yang disajikan berupa sintesis konseptual dari sumber yang dipilih, sehingga tidak dapat digeneralisasi sebagai gambaran prevalensi ujaran kebencian di seluruh platform. Meski

demikian, pendekatan ini memadai untuk memperjelas keputusan analitis yang harus dibuat sebelum sebuah ujaran dipetakan ke dalam kategori hukum atau dimasukkan ke dalam sistem moderasi konten.

Hasil

Sintesis literatur menghasilkan empat tema yang saling berkaitan. Tema pertama menempatkan ujaran kebencian sebagai kategori yang lebih sempit daripada bahasa kasar. Tema kedua menjelaskan kebutuhan membaca bentuk bahasa bersama tindakan komunikatif dan target. Tema ketiga menunjukkan bahwa konteks percakapan dan moda visual membentuk makna. Tema keempat menguraikan hubungan antara analisis manusia, teknologi deteksi, dan tata kelola bukti digital. Keempat tema tersebut tidak bekerja secara berurutan secara kaku. Pemeriksa dapat memulai dari teks yang tersedia, tetapi kesimpulan baru dapat dibangun setelah seluruh unsur terkait dibaca bersama.

Tabel 1. Matriks Penelitian Terdahulu yang Digunakan dalam Kajian

No .	Penulis Tahun	Sumber Data	Platform	Bentuk Ujaran	Pendekatan Linguistik	Temuan	Relevansi Forensik
1	Suryani et al. (2021)	Teks unggahan akun @mantanaurelhermansyah dan respons terkait Aurel Hermansyah	Instagram	Penghinaan, pencemaran nama baik, kata tabu	Linguistik forensik; tindak tutur ilokusi dan perlokusi	Ujaran berbentuk teks digunakan untuk menghina dan mencemarkan nama baik serta menandai akun lain.	Menunjukkan cara menghubungkan bentuk teks, fungsi tutur, dan dugaan pelanggaran.
2	Nuraeni et al. (2022)	Komentar warganet pada akun @obrolanpolitik	Instagram	Tuturan kebencian dalam komentar politik	Pragmatik; bentuk dan fungsi tuturan	Penelitian memetakan bentuk serta fungsi tuturan kebencian yang ditulis warganet.	Memperkuat penggunaan analisis tindak tutur untuk membaca fungsi ujaran.
3	Syahid & Datang (2024)	Satu postingan mengenai gesekan Tionghoa dan pribumi	Facebook	Sentimen etnis dalam teks dan gambar	Tindak tutur, semiotika sosial, dan gramatika visual	Relasi elaboratif antara teks dan gambar membangun sentimen etnis.	Menegaskan bahwa bukti ujaran kebencian digital dapat bersifat multimodal.
4	Ibrohim & Budi (2023)	Literatur dan sumber data penelitian media sosial berbahasa Indonesia	Multi-platform	Ujaran kebencian dan bahasa abusif	Tinjauan deteksi berbasis pengolahan bahasa alami	Tantangan utama mencakup dataset yang valid, ketahanan model, target, kategori, dan tingkat kebencian.	Memberi konteks Indonesia bagi penggunaan sistem deteksi sebagai penyaringan awal.

No .	Penulis Tahun	Sumber Data	Platform	Bentuk Ujaran	Pendekatan Linguistik	Temuan	Relevansi Forensik
5	ElSherief et al. (2021)	Korpus unggahan Twitter berlabel	Twitter	Kebencian implisit, kode, dan tidak langsung	Taksonomi ujaran implisit dan analisis korpus	Bentuk implisit memerlukan pembacaan implikasi, bukan hanya kata kasar eksplisit.	Mendukung analisis pragmatik terhadap ujaran yang terselubung.
6	Sap et al. (2019)	Data Twitter dan anotasi ujaran ofensif	Twitter	Bahasa yang dipersepsi ofensif dan variasi dialek	Analisis bias anotasi dan evaluasi model	Ketidakpekaan pada variasi dialek dapat meningkatkan pelabelan ofensif terhadap kelompok tertentu.	Mengingatkan perlunya kehati-hatian terhadap dialek dan variasi bahasa Indonesia.
7	Röttger et al. (2021)	HateCheck, set uji fungsional berbahasa Inggris	Tidak spesifik	Ancaman, penghinaan, rujukan kelompok, dan bentuk nonkebencian	Pengujian fungsional model deteksi	Akurasi umum tidak selalu mengungkap kelemahan model dalam fungsi bahasa tertentu.	Mendukung kebutuhan alasan analitis yang dapat diuji, bukan hanya label otomatis.
8	Yu et al. (2022)	Komentar Reddit dengan konteks percakapan sebelumnya	Reddit	Kebencian, kontra-ujaran, dan komentar netral	Analisis konteks percakapan dan klasifikasi tiga kategori	Penilaian manusia dapat berubah ketika konteks komentar ditampilkan.	Menegaskan bahwa tangkapan layar satu kalimat sering belum cukup sebagai bahan analisis.
9	Ghosh et al. (2023)	Enam dataset ujaran kebencian dalam percakapan daring	Multi-platform	Ujaran kebencian implisit	Pemodelan konteks pengguna dan percakapan	Konteks penggunaan dan alur percakapan membantu pengenalan ujaran implisit.	Memperkuat ko-teks dan riwayat interaksi dalam pembacaan forensik.
10	Risquallah et al. (2026)	Artikel tentang komentar media sosial sebagai korpus forensik	Beragam, Instagram dominan	Ujaran kebencian, pencemaran nama baik, dan kejahatan siber	Tinjauan sistematis; semantik leksikal dan pragmatik tindak tutur	Masih terdapat kekurangan kajian multimodal dan standardisasi ekstraksi bukti digital.	Mendasari kebutuhan dokumentasi konteks dan pengelolaan bukti digital yang tertib.

Tabel 1 memperlihatkan bahwa kajian terdahulu mencakup data Instagram, Facebook, Twitter, Reddit, dan korpus lintas platform. Bentuk ujaran yang dibahas tidak hanya berupa

penghinaan eksplisit, tetapi juga kebencian implisit, sentimen etnis, pencemaran nama baik, dan kontra-ujaran. Pendekatan yang digunakan meliputi linguistik forensik, pragmatik, semiotika sosial, gramatika visual, serta evaluasi sistem deteksi. Pola ini mengarah pada empat hasil utama: ujaran kebencian perlu dibedakan dari ujaran kasar, makna perlu dibaca bersama target dan tindak tutur, konteks percakapan serta unsur visual berpengaruh terhadap penafsiran, dan teknologi deteksi memerlukan pemeriksaan manusia serta tata kelola bukti digital.

Pembahasan

Temuan pertama menunjukkan bahwa ujaran kebencian tidak identik dengan setiap bentuk penghinaan atau kata yang dianggap tidak sopan. Bahasa kasar dapat diarahkan kepada individu tanpa merujuk pada keanggotaan kelompok, dapat muncul sebagai luapan emosi, atau bahkan dipakai secara akrab dalam komunitas tertentu. Sebaliknya, ujaran kebencian lazim berkaitan dengan pemosisian negatif terhadap seseorang karena identitasnya atau dengan penggambaran sebuah kelompok sebagai ancaman, rendah, tidak layak, atau patut disingkirkan. Pembedaan ini bukan usaha mengecilkan dampak penghinaan personal, tetapi cara menjaga ketepatan kategori. Fortuna dan Nunes (2018) mengingatkan bahwa definisi dan label yang tidak konsisten membuat hasil penelitian serta sistem deteksi mudah saling bertentangan. Ibrohim dan Budi (2023) juga mencatat bahwa penelitian bahasa Indonesia masih menghadapi tantangan untuk membedakan sasaran, kategori, dan tingkat kebencian dalam data media sosial.

Dari sudut linguistik forensik, pembedaan tersebut dapat dimulai dengan menanyakan tiga hal: siapa sasaran ujaran, prediksi negatif apa yang dilekatkan kepadanya, dan tindakan apa yang didorong oleh penutur. Sasaran dapat berupa individu, tetapi alasan penyerangannya mungkin bertumpu pada agama, suku, ras, asal, jenis kelamin, orientasi seksual, atau ciri kelompok lain. Prediksi terlihat ketika kelompok digambarkan dengan generalisasi yang menghilangkan keragaman anggotanya. Tindakan dapat muncul melalui penghinaan, dehumanisasi, pengusiran, penolakan, ancaman, atau ajakan kepada pihak lain untuk memperlakukan kelompok secara negatif. Council of Europe (2022) memosisikan tingkat keparahan dengan mempertimbangkan konteks, niat, isi, cakupan, dan kemungkinan dampak. Kerangka ini membantu mencegah dua kesalahan yang berlawanan, yaitu meremehkan ekspresi diskriminatif yang terselubung atau mempidanakan secara berlebihan ekspresi yang tidak memenuhi ambang kebencian.

Pada ruang digital, batas kategori semakin kabur karena ekspresi kebencian tidak selalu memakai leksikon yang secara terbuka menghina. ElSherief et al. (2021) menunjukkan bahwa ujaran implisit dapat menggunakan bahasa berkode, generalisasi, atau penyampaian tidak langsung. Ahn et al. (2024) menegaskan bahwa sifat halus dan ketergantungan pada konteks membuat klasifikasi biner menjadi tidak memadai. Karena itu, pemeriksa tidak seharusnya menyimpulkan bahwa tidak ada ujaran kebencian hanya karena teks tidak memuat kata tabu. Sebaliknya, kata tabu juga tidak otomatis membuktikan kebencian apabila dipakai dalam kutipan, pelaporan, bantahan, humor yang tidak menyerang kelompok, atau percakapan yang konteksnya mengubah fungsi literalnya.

Bagi kebutuhan hukum Indonesia, kehati-hatian kategorisasi mendapat dasar tambahan. Pasal mengenai penyebaran informasi yang menimbulkan rasa kebencian atau permusuhan dipahami secara bersyarat sebagai muatan kebencian berdasarkan identitas tertentu yang dilakukan dengan sengaja dan di depan umum serta menimbulkan risiko nyata terhadap diskriminasi, permusuhan, atau kekerasan (Undang-Undang Republik Indonesia Nomor 1 Tahun 2024). Formulasi tersebut memerlukan bukti mengenai relasi antara ujaran dan identitas, bukan hanya reaksi emosional pembaca. Dengan demikian, kategori linguistik

forensik yang baik harus mampu menjelaskan hubungan tersebut melalui bukti teks dan konteks, bukan hanya melalui label umum seperti “negatif” atau “toxic”.

Tema kedua memperlihatkan bahwa analisis perlu memeriksa lapis leksikal, gramatikal, pragmatik, dan diskursif secara bersamaan. Pada tingkat leksikal, pemeriksa dapat mengidentifikasi kata evaluatif, julukan, metafora hewan, istilah penghinaan, atau penanda kelompok. Pada tingkat gramatikal, perhatian dapat diarahkan pada bentuk generalisasi, penghilangan pelaku, modalitas kewajiban, kalimat imperatif, dan pola sebab-akibat yang memosisikan kelompok tertentu sebagai sumber masalah. Namun, bentuk-bentuk itu belum memiliki nilai tunggal. Metafora, pronomina, atau nominalisasi memperoleh makna melalui urutan percakapan dan pengetahuan bersama yang diaktifkan oleh penutur. Coulthard et al. (2016) menempatkan analisis bahasa sebagai kegiatan berbasis bukti yang menguji bagaimana makna dibangun, bukan sekadar mencocokkan teks dengan daftar istilah.

Analisis tindak tutur melengkapi pemeriksaan bentuk bahasa. Ujaran dapat berfungsi untuk mengejek, merendahkan, mengancam, memerintahkan, menghasut, menormalisasi penolakan, atau mengajak audiens membayangkan tindakan terhadap kelompok sasaran. Satu ungghahan yang secara literal berbentuk pertanyaan, misalnya, dapat menjadi sindiran atau ajakan terselubung apabila disertai konteks peristiwa, tagar, dan respons percakapan tertentu. Penelitian Indonesia yang memakai pendekatan pragmatik memperlihatkan bahwa makna ujaran kebencian dapat hadir melalui fungsi ilokusi, bukan hanya melalui kosakata negatif (Nuraeni et al., 2022). Karena itu, ahli bahasa perlu menjelaskan fungsi yang didukung oleh unsur tekstual, bukan menyimpulkan niat pribadi penulis tanpa dasar.

Target juga perlu dipetakan secara teliti. Ada perbedaan antara serangan terhadap orang tertentu sebagai individu dan serangan yang menjadikan identitas kelompok sebagai dasar penilaian. Dalam kasus yang kedua, nama orang dapat berfungsi sebagai pintu masuk untuk menyerang komunitas yang lebih luas. Sebaliknya, ujaran terhadap tokoh publik tidak selalu berarti ujaran kebencian terhadap kelompok identitasnya. Pemetaan target sebaiknya memuat rujukan eksplisit, rujukan implisit, pronomina, tagar, serta ko-teks yang membantu pembaca mengenali kelompok yang dimaksud. Pendekatan ini sesuai dengan kebutuhan untuk membedakan penghinaan, pencemaran nama baik, kritik keras, dan ujaran kebencian tanpa menyamakan semuanya.

Keterkaitan antara bentuk, fungsi, dan target menjadi dasar argumentasi forensik. Argumentasi yang kuat tidak berbunyi “ungkapan ini buruk, maka ia kebencian”. Argumentasi yang lebih dapat diuji menunjukkan, misalnya, bahwa pilihan nomina tertentu menggeneralisasi kelompok, prediksi tertentu mengonstruksi kelompok sebagai bahaya, dan bentuk ajakan atau pengandaian mendorong penolakan terhadap mereka. Penjelasan ini dapat kemudian dibandingkan dengan konteks percakapan dan norma hukum. Hasilnya bukan kepastian otomatis tentang kesalahan pidana, melainkan uraian kebahasaan yang memungkinkan penilai hukum melihat dasar makna secara transparan.

Tema ketiga menegaskan bahwa ungghahan media sosial tidak pernah berdiri sendiri. Makna komentar dapat dipengaruhi oleh ungghahan induk, berita yang dibagikan, gambar, video, waktu peredaran, balasan sebelum dan sesudahnya, serta pengetahuan bersama dalam komunitas. Yu, Blanco, dan Hong (2022) menemukan bahwa penilaian manusia terhadap komentar dapat berubah ketika konteks percakapan disediakan. Ghosh et al. (2023) juga menunjukkan bahwa konteks pengguna dan konteks percakapan relevan untuk mendeteksi ujaran implisit. Dua temuan ini memperkuat prinsip forensik yang sederhana tetapi penting: tangkapan layar satu kalimat sering belum cukup untuk menjelaskan fungsi sebuah ujaran.

Konteks bukan alasan untuk mengabaikan teks. Ia berfungsi untuk menguji apakah pembacaan terhadap teks konsisten dengan lingkungan komunikasinya. Dalam praktik pemeriksaan, konteks minimal dapat mencakup tautan atau akun sumber, tanggal dan waktu, ungghahan induk, urutan komentar, tagar, akun yang ditandai, serta deskripsi tampilan visual.

Apabila ujaran menggunakan kata ganti seperti “mereka”, konteks dibutuhkan untuk menentukan rujukannya. Jika seseorang menulis “singkirkan saja”, konteks menentukan siapa yang dimaksud dan apakah kalimat itu merupakan deskripsi, candaan, kutipan, atau dorongan tindakan. Tanpa dokumentasi tersebut, penafsiran berisiko mengisi kekosongan dengan asumsi pemeriksa sendiri.

Multimodalitas memperluas kebutuhan tersebut. Banyak unggahan menggabungkan teks, foto, meme, musik, emoji, tipografi, atau rekaman visual. Syahid dan Datang (2024) menunjukkan bahwa sentimen etnis dalam sebuah unggahan Facebook dapat dibaca dari hubungan elaboratif antara teks dan gambar, bukan dari salah satunya secara terpisah. Dalam konteks yang sama, gambar dapat menyediakan sasaran yang tidak disebutkan dalam teks, sedangkan teks dapat memberi arah evaluatif terhadap gambar. Oleh sebab itu, pemeriksa perlu mendeskripsikan unsur visual secara konkret, lalu menjelaskan relasinya dengan unsur verbal. Analisis tidak boleh mengubah interpretasi visual menjadi fakta yang tidak terlihat dalam materi asli.

Prinsip konteks juga mengubah cara memahami jangkauan publik. Ujaran di ruang yang tampak tertutup dapat tetap menyebar melalui fitur bagikan, tangkapan layar, atau unggahan ulang. Namun, kemungkinan penyebaran tidak sama dengan bukti bahwa penyebaran tertentu sudah terjadi. Dalam penilaian forensik, perlu dibedakan antara fakta teknis yang terdokumentasi, misalnya jumlah komentar yang tampak atau fitur platform yang tersedia, dan dugaan mengenai dampak sosial. Pembedaan ini penting agar keterangan ahli tidak berubah menjadi spekulasi tentang akibat. Penilaian risiko sebaiknya didasarkan pada isi, sasaran, konteks peristiwa, khalayak, dan pola sirkulasi yang benar-benar terlihat dalam bukti.

Tema keempat memperlihatkan bahwa linguistik forensik memerlukan tata kelola bukti digital, bukan hanya kemampuan menafsirkan kata. Data media sosial mudah dihapus, disunting, diunggah ulang, atau dipisahkan dari konteksnya. Karena itu, dokumen kerja sebaiknya membedakan salinan untuk pemeriksaan dari materi asal, mencatat sumber dan waktu pengambilan, serta menyimpan tampilan yang cukup untuk memperlihatkan hubungan teks dengan konteksnya. Praktik tersebut tidak dengan sendirinya menjamin keaslian secara hukum, tetapi membantu menjaga keterlacakan proses analisis. Kajian tentang komentar media sosial sebagai korpus forensik juga menunjukkan masih terbatasnya standarisasi ekstraksi bukti yang dapat diterima secara hukum, terutama ketika data bersifat multimodal (Maulana et al., 2025).

Perkembangan sistem deteksi otomatis menawarkan manfaat nyata pada tahap penyaringan. Model dapat membantu memilah volume besar komentar, menandai bentuk eksplisit, atau mengidentifikasi pola yang perlu diperiksa kembali. Ramos et al. (2024) dan Rawat et al. (2024) mencatat perkembangan metode dari pembelajaran mesin konvensional menuju model *transformer*. Akan tetapi, kemajuan teknis tidak menghapus masalah konseptual. Kualitas model bergantung pada definisi, korpus, label, bahasa, dan konteks yang dipakai saat pelatihan. Model yang akurat pada satu kumpulan data tidak otomatis akurat pada dialek, komunitas, atau bentuk ujaran lain.

Risiko tersebut terlihat jelas pada masalah bias. Sap et al. (2019) menunjukkan bahwa ketidakpekaan anotator terhadap variasi dialek dapat menghasilkan model yang lebih sering melabeli bahasa kelompok tertentu sebagai ofensif. Temuan itu bermakna penting bagi Indonesia yang memiliki keragaman dialek, campur kode, dan ekspresi lokal. Jika sistem memperlakukan penanda kebahasaan kelompok sebagai indikator kebencian tanpa memahami fungsi sosialnya, moderasi dapat justru memperbesar ketidakadilan. Oleh sebab itu, keluaran otomatis perlu diperlakukan sebagai sinyal awal yang harus diperiksa oleh manusia, bukan sebagai putusan semantik atau hukum.

Kebutuhan evaluasi berbasis fungsi mendukung posisi tersebut. Röttger et al. (2021) memperlihatkan bahwa ukuran akurasi umum tidak selalu mengungkap kelemahan model pada fungsi kebahasaan tertentu. Mereka mengembangkan pengujian yang menilai berbagai kemampuan, termasuk pengenalan ancaman, penghinaan, referensi kelompok, dan bentuk non-kebencian yang menyerupai ujaran kebencian. Bagi linguistik forensik, pelajaran utamanya bukan kewajiban memakai perangkat tertentu, melainkan pentingnya menguji alasan di balik keputusan. Sistem atau pemeriksa harus dapat menjelaskan bukti apa yang dipakai, konteks mana yang diperhitungkan, dan unsur apa yang masih tidak pasti.

Dengan demikian, kerja sama antara teknologi dan analisis ahli sebaiknya bersifat berlapis. Tahap pertama dapat menggunakan sistem untuk pencarian atau pengelompokan data. Tahap kedua melakukan pemeriksaan manusia terhadap bentuk bahasa, target, konteks, dan bahan visual. Tahap ketiga menyusun laporan yang membedakan fakta dokumenter, deskripsi linguistik, dan inferensi. Tahap terakhir menghubungkan temuan linguistik dengan norma yang relevan tanpa mengambil alih kewenangan penilai hukum. Susunan ini menjaga kegunaan teknologi sekaligus membatasi kecenderungan untuk memperlakukan skor klasifikasi sebagai bukti final.

Berdasarkan empat tema tersebut, artikel ini merumuskan kerangka kerja enam langkah. Langkah pertama adalah mengamankan unit data secara proporsional: teks, identitas akun yang tampak, tautan, waktu, unggahan induk, dan unsur visual yang berkaitan. Langkah kedua adalah membuat transkripsi atau deskripsi yang memisahkan apa yang benar-benar tampak dari penafsiran awal. Langkah ketiga adalah mengidentifikasi bentuk bahasa, termasuk rujukan kelompok, prediksi, generalisasi, metafora, modalitas, dan struktur ajakan. Langkah keempat adalah menafsirkan tindak tutur dengan membaca ko-teks serta urutan percakapan. Langkah kelima adalah memetakan target, jangkauan, dan risiko yang dapat dijelaskan melalui bukti. Langkah keenam adalah menulis simpulan yang memisahkan temuan linguistik dari penilaian hukum akhir.

Kerangka ini tidak dimaksudkan sebagai daftar centang mekanis. Sebuah komentar singkat mungkin hanya memungkinkan pembacaan terbatas karena konteksnya hilang. Dalam keadaan demikian, laporan yang bertanggung jawab harus menyebutkan keterbatasan dan menghindari kepastian yang tidak ditopang bukti. Sebaliknya, unggahan dengan teks, gambar, tagar, dan rangkaian balasan dapat memperlihatkan bukti yang lebih kaya tetapi juga menuntut analisis lebih hati-hati. Ketelitian bukan berarti memperpanjang laporan dengan istilah teknis. Ketelitian berarti menunjukkan jalur argumentasi dari data yang dapat diperiksa menuju kesimpulan yang proporsional.

Kajian terdahulu mendukung pendekatan integratif tersebut. Penelitian berbasis kasus di Indonesia telah memperlihatkan nilai analisis tindak tutur dan multimodalitas, sedangkan riset internasional memperlihatkan masalah definisi, konteks, dan bias dalam sistem deteksi. Jika dua kelompok literatur ini dibaca bersama, terlihat bahwa kebutuhan utama bukan memilih antara teknologi atau analisis kualitatif. Kebutuhan utamanya ialah membangun prosedur yang dapat menerjemahkan sinyal digital menjadi penjelasan kebahasaan yang relevan dan dapat diuji. Posisi ini memperluas penggunaan linguistik forensik dari sekadar mengidentifikasi kata bermasalah menjadi praktik penalaran berbasis bukti.

Penggunaan istilah forensik kerap menimbulkan salah paham bahwa ahli bahasa bertugas menyatakan seseorang bersalah atau tidak bersalah. Padahal, keterangan linguistik terutama menerangkan ciri bentuk dan makna ujaran beserta alasan analitisnya. Ahli dapat menjelaskan bahwa suatu pronomina merujuk kepada kelompok tertentu, bahwa kalimat imperatif berfungsi mengajak, atau bahwa relasi teks dan gambar membangun dehumanisasi. Akan tetapi, pemenuhan seluruh unsur tindak pidana, pembuktian kesengajaan sebagai keadaan batin, penilaian alat bukti, dan putusan akhir berada pada ranah hukum. Perbedaan

peran ini justru meningkatkan nilai keterangan ahli karena simpulannya tidak melampaui kompetensi dan materi yang diperiksa (Coulthard et al., 2016).

Batas tersebut penting ketika laporan menggunakan istilah seperti “niat”, “dampak”, atau “risiko”. Niat komunikatif dapat dibahas sebagai inferensi pragmatik apabila teks dan konteks memberikan dasar yang memadai, tetapi tidak boleh diperlakukan sebagai pengetahuan langsung mengenai batin penulis. Dampak aktual perlu dibedakan dari kemungkinan dampak. Sebuah unggahan dapat dianalisis sebagai memiliki kapasitas untuk memperkuat permusuhan berdasarkan isi dan jangkauannya, tetapi klaim bahwa unggahan itu telah menyebabkan tindakan tertentu membutuhkan bukti tambahan di luar analisis bahasa. Kerangka Council of Europe (2022) membantu dengan menempatkan konteks, pembicara, isi, cakupan, kemungkinan membahayakan, dan karakter publik sebagai unsur yang perlu ditimbang secara bersama.

Dalam laporan yang baik, setiap simpulan dapat ditandai tingkat kepastiannya. Deskripsi seperti “unggahannya memuat kata X”, “komentar muncul sebagai balasan terhadap unggahan Y”, atau “gambar menampilkan Z” merupakan fakta dokumenter sepanjang materi dapat diverifikasi. Pernyataan bahwa pilihan kata tertentu “mengonstruksi kelompok sebagai ancaman” merupakan interpretasi linguistik yang harus dijelaskan melalui bentuk bahasa dan konteks. Sementara itu, pernyataan bahwa ujaran “memenuhi semua unsur pidana” merupakan penilaian normatif yang sebaiknya tidak dinyatakan sebagai hasil tunggal ahli bahasa. Pemisahan ini membuat pembaca dapat menguji tiap tahap argumentasi dan mengurangi risiko kesimpulan yang terlalu jauh.

Literatur tentang deteksi otomatis memperlihatkan masalah yang relevan bagi pembacaan bahasa Indonesia, yakni kecenderungan sistem mengenali pola permukaan tanpa memahami variasi sosial. Bahasa Indonesia di media sosial hadir bersama bahasa daerah, dialek urban, singkatan, ejaan kreatif, campur kode, dan penanda afektif seperti emoji. Unsur tersebut dapat mengubah intensitas, sasaran, atau bahkan orientasi sebuah ujaran. Kata yang kasar dalam satu komunitas mungkin dipakai sebagai penanda keakraban pada komunitas lain, sementara ekspresi yang tampak netral dapat menjadi sangat merendahkan karena merujuk pada stereotip lokal. Ibrohim dan Budi (2023) menempatkan ketersediaan data yang valid serta kemampuan mengenali target dan tingkat kebencian sebagai tantangan penting dalam penelitian Indonesia.

Karena itu, pemeriksa perlu menghindari dua cara baca yang sama-sama lemah. Cara pertama memaksakan standar bahasa baku dan menilai semua deviasi ejaan sebagai agresivitas. Cara kedua terlalu mengandalkan pengetahuan lokal yang tidak terdokumentasi sehingga penafsiran sulit diuji oleh pembaca lain. Jalan tengahnya ialah menjelaskan bentuk lokal yang relevan, menyatakan dasar pemahaman konteksnya, dan membedakan informasi yang tersedia dari asumsi. Jika sebuah akronim atau istilah komunitas dianggap bermakna tertentu, sumber konteksnya perlu ditunjukkan melalui rangkaian percakapan, definisi komunitas yang dapat diperiksa, atau pola penggunaan yang konsisten.

Masalah ini berkaitan dengan temuan Sap et al. (2019) mengenai bias ketika anotator tidak peka terhadap dialek. Walaupun penelitian tersebut menggunakan konteks bahasa Inggris Amerika, pelajarannya bersifat metodologis: keputusan klasifikasi dapat mencerminkan prasangka penilai terhadap varietas bahasa, bukan semata sifat ujaran. Penelitian dan moderasi konten di Indonesia perlu menguji apakah korpus merepresentasikan variasi geografis dan sosial secara memadai. Tanpa upaya tersebut, kategori kebencian berisiko menjadi alat untuk menandai suara kelompok tertentu sebagai masalah hanya karena tidak sesuai dengan ragam dominan.

Moderasi konten membutuhkan keputusan yang cepat, sedangkan penafsiran linguistik sering membutuhkan waktu dan konteks. Ketegangan ini tidak dapat diselesaikan hanya dengan menaikkan ambang sensitivitas sistem. Penghapusan yang terlalu luas dapat

membatasi kritik, pelaporan pengalaman diskriminasi, kutipan edukatif, atau kontra-ujaran yang sebenarnya menolak kebencian. Sebaliknya, standar yang terlalu sempit dapat membiarkan bahasa berkode dan serangan terhadap kelompok rentan tetap beredar. Literatur tentang data pelatihan bahasa abusif menekankan bahwa definisi, instruksi anotasi, dan siapa yang memberi label sangat menentukan arah moderasi yang dihasilkan (Vidgen & Derczynski, 2020; Vidgen et al., 2019).

Pada titik ini, pendekatan linguistik forensik menawarkan prinsip proporsionalitas. Konten yang jelas berupa ancaman atau ajakan kekerasan terhadap kelompok dapat ditandai sebagai prioritas tinggi. Konten yang ambigu atau sangat bergantung pada konteks sebaiknya masuk ke pemeriksaan lanjutan, bukan langsung diputus hanya dari indikator leksikal. Konten yang merupakan pelaporan, dokumentasi, kritik, atau kontra-ujaran perlu dibaca melalui fungsi komunikatifnya. Yu et al. (2022) menunjukkan bahwa konteks percakapan bahkan dapat mengubah penilaian manusia terhadap komentar. Artinya, desain moderasi yang mengabaikan konteks berpotensi keliru baik dalam arah penghapusan maupun pembiaran.

Kebutuhan akan prosedur berlapis juga dapat dilihat sebagai bentuk akuntabilitas. Platform atau institusi yang menggunakan perangkat deteksi perlu mampu menjelaskan kategori yang dipakai, data apa yang dianalisis, jalur banding yang tersedia, dan kapan penilaian manusia dilakukan. Walaupun artikel ini tidak mengevaluasi kebijakan platform tertentu, literatur yang disintesis mendukung pandangan bahwa transparansi analitis adalah syarat penting ketika bahasa menjadi dasar pembatasan konten. Dengan cara ini, moderasi tidak hanya mengejar kecepatan, tetapi juga memberi alasan yang dapat diuji dan diperbaiki.

Arah pengembangan pertama ialah pembangunan korpus yang menyimpan konteks secara etis dan proporsional. Banyak data ujaran kebencian dikumpulkan sebagai kalimat tunggal sehingga rujukan, ironi, respons terhadap unggahan sebelumnya, dan fungsi kontra-ujaran hilang. Korpus masa depan perlu mempertimbangkan unit percakapan, metadata yang tidak melanggar privasi, serta penanda modalitas seperti gambar atau emoji ketika unsur tersebut memengaruhi makna. Poletto et al. (2021) telah menunjukkan bahwa perbedaan sumber dan skema anotasi merupakan faktor utama dalam keragaman sumber daya penelitian. Korpus berbahasa Indonesia yang transparan tentang keputusan pelabelannya akan lebih berguna bagi penelitian maupun pelatihan pemeriksa.

Arah kedua ialah pengembangan pedoman laporan ahli yang mudah diikuti. Pedoman tersebut tidak harus menyeragamkan seluruh analisis, tetapi dapat menetapkan unsur minimum: identifikasi bahan yang diperiksa, batas data, transkripsi, deskripsi konteks, analisis bentuk bahasa, analisis tindak tutur, peta sasaran, unsur multimodal, dan tingkat kepastian simpulan. Standar semacam ini membantu membedakan laporan yang berangkat dari bukti dengan komentar umum tentang moralitas bahasa. Ia juga memungkinkan pihak lain menilai apakah inferensi yang dibuat cukup didukung oleh bahan yang tersedia.

Arah ketiga ialah kolaborasi antardisiplin. Peneliti bahasa dapat bekerja bersama ahli hukum, ilmu komputer, psikologi sosial, dan pengelola platform tanpa mencampurkan peran. Ahli bahasa menjelaskan makna dan variasi; ahli hukum menilai relevansi normatif; ahli komputasi membangun alat dan menguji keterbatasannya; sedangkan peneliti sosial membantu memahami pola dampak dan relasi kuasa. Keterhubungan tersebut penting karena ujaran kebencian adalah fenomena bahasa, tetapi tidak selesai dijelaskan pada tingkat bahasa saja. Sintesis ini menunjukkan bahwa justru dengan menjaga batas keahlian, kolaborasi dapat menghasilkan pemeriksaan yang lebih cermat dan bertanggung jawab.

Pemeriksaan ujaran kebencian juga memunculkan persoalan etika. Materi yang dianalisis dapat mengulang penghinaan, menampilkan identitas pengguna, atau memperluas kembali jangkauan konten yang semula terbatas. Oleh sebab itu, kebutuhan dokumentasi harus diseimbangkan dengan prinsip minimisasi. Laporan cukup memuat bagian yang relevan

untuk menjelaskan analisis dan dapat menggunakan penyamaran identitas apabila publikasi ilmiah tidak memerlukan penyebutan akun. Ketelitian ini tidak mengurangi kekuatan bukti. Sebaliknya, ia membantu mencegah penelitian berubah menjadi sarana reproduksi ujaran yang sedang dikritisi. Dalam kajian berbasis publikasi, keputusan etis perlu dijelaskan secara singkat apabila penulis mereproduksi contoh yang berpotensi memicu atau merugikan kelompok tertentu.

Etika juga menyangkut posisi penafsir. Peneliti atau ahli tidak datang sebagai pembaca yang netral tanpa latar sosial, sehingga keputusan tentang apa yang dianggap mengancam atau merendahkan perlu dibuka kepada pemeriksaan. Instruksi anotasi yang jelas, pembacaan oleh lebih dari satu penilai ketika memungkinkan, dan pencatatan alasan analitis dapat mengurangi ketergantungan pada intuisi tunggal. Vidgen dan Derczynski (2020) mengingatkan bahwa masalah data pelatihan bukan semata persoalan jumlah, melainkan juga keputusan manusia yang melekat dalam proses pelabelan. Prinsip serupa berguna dalam analisis kualitatif: alasan harus terlihat, bukan hanya hasil kategorinya.

Bagi kelompok yang menjadi sasaran, pendekatan kontekstual memiliki arti praktis. Pembacaan yang hanya mengejar kata-kata eksplisit dapat gagal mengenali stereotip, pengucilan, dan ancaman terselubung yang dialami berulang oleh kelompok tertentu. Namun, perlindungan yang efektif tidak mensyaratkan pengabaian terhadap ketelitian. Justru analisis yang memperlihatkan cara kerja ujaran memberi dasar yang lebih kuat untuk memahami bentuk kerugian simbolik dan sosial yang mungkin ditimbulkannya. Dengan demikian, etika pemeriksaan bersatu dengan kualitas metode: keduanya menuntut perhatian terhadap konteks, bukti, dan dampak dari cara materi disajikan.

Secara keseluruhan, literatur memperlihatkan pergeseran dari pencarian kata ofensif menuju pembacaan ujaran sebagai fenomena yang berlapis. Studi komputasional memperjelas perlunya sumber data yang beragam, pengujian fungsi, dan kehati-hatian terhadap bias. Studi linguistik forensik dan pragmatik memperjelas arti tindak tutur, sasaran, rujukan, serta konteks peredaran. Studi multimodal memperingatkan bahwa unsur visual dapat mengubah atau melengkapi makna teks. Peraturan Indonesia menambahkan kebutuhan untuk membedakan muatan kebencian berdasarkan identitas dari ekspresi negatif yang tidak memiliki unsur tersebut. Setiap kelompok sumber menyumbang bagian yang berbeda, dan tidak satu pun cukup jika digunakan sendirian.

Sintesis ini memperlihatkan alasan mengapa penilaian ujaran kebencian perlu ditulis sebagai argumentasi, bukan hanya sebagai hasil label. Argumentasi harus dimulai dari apa yang terdokumentasi, menjelaskan alat kebahasaan yang digunakan, menghubungkan alat itu dengan target dan konteks, lalu menyatakan batas inferensinya. Apabila salah satu mata rantai tidak tersedia, simpulan perlu disesuaikan. Pola tersebut dapat diterapkan dalam penelitian, moderasi konten, maupun penyusunan keterangan ahli. Ia tidak menjanjikan jawaban yang selalu cepat, tetapi menawarkan jalur keputusan yang lebih dapat dipertanggungjawabkan ketika bahasa digital dipersoalkan di ruang publik maupun hukum.

Kajian ini memiliki keterbatasan yang melekat pada desainnya. Sumber yang disintesis mencakup penelitian dengan definisi, bahasa, platform, dan tujuan yang berbeda, sehingga tidak tepat untuk diperlakukan sebagai satu kumpulan data yang seragam. Artikel ini juga tidak mengukur prevalensi ujaran kebencian, tidak menguji sebuah model, dan tidak memeriksa berkas perkara. Karena itu, kerangka yang dirumuskan harus dipahami sebagai panduan analitis konseptual yang perlu disesuaikan dengan bahan, yurisdiksi, dan prosedur pembuktian yang berlaku pada setiap penggunaan. Keterbatasan ini sekaligus menunjukkan perlunya studi empiris yang mendokumentasikan praktik pemeriksaan ujaran kebencian berbahasa Indonesia secara lebih rinci.

Meski demikian, keberagaman sumber menjadi kekuatan untuk melihat persoalan dari beberapa sisi. Literatur teknik membantu menjelaskan alasan sistem otomatis perlu diawasi,

sedangkan literatur forensik dan pragmatik membantu menerangkan apa yang luput jika teks dipisahkan dari konteks. Penyatuan kedua perspektif tersebut tidak menghapus perbedaan metodologis, tetapi memungkinkan pembaca memahami bahwa keputusan tentang ujaran kebencian selalu melibatkan pilihan kategorisasi, interpretasi bahasa, dan tanggung jawab terhadap dampak sosialnya.

Kesimpulan

Kajian ini menyimpulkan bahwa ujaran kebencian di media sosial perlu dianalisis sebagai tindakan bahasa yang terkait dengan target identitas, konteks komunikasi, dan risiko sosial, bukan sebagai daftar kata terlarang. Sintesis literatur menunjukkan bahwa perbedaan dari ujaran kasar merupakan langkah awal yang penting. Setelah itu, pemeriksa perlu membaca pilihan leksikal dan gramatikal bersama tindak tutur, rujukan sasaran, urutan percakapan, serta unsur visual. Bukti digital harus didokumentasikan secara cukup agar penafsiran dapat ditelusuri dan tidak menggantikan fakta dengan dugaan.

Linguistik forensik memberi kontribusi dengan menyediakan cara untuk menjelaskan bagaimana makna kebencian dibangun dan mengapa suatu ujaran dapat relevan bagi penilaian hukum. Kontribusi ini bersifat melengkapi, bukan menggantikan, kewenangan penilai hukum. Sistem otomatis dapat membantu menyaring data, tetapi sifat implisit, kontekstual, dan beragamnya bahasa digital membuat pemeriksaan manusia tetap diperlukan. Bagi riset berikutnya, pengembangan korpus berbahasa Indonesia perlu memasukkan variasi dialek, campur kode, percakapan berantai, dan data multimodal dengan prosedur dokumentasi yang jelas. Langkah tersebut dapat memperkuat ketepatan analisis sekaligus mengurangi risiko pelabelan berlebihan terhadap ekspresi yang tidak memenuhi unsur ujaran kebencian.

PERNYATAAN TAMBAHAN PENULIS

Konflik Kepentingan

Penulis menyatakan bahwa tidak terdapat konflik kepentingan yang berkaitan dengan penelitian, penulisan, maupun publikasi artikel ini.

Pernyataan Ketersediaan Data

Keseluruhan data dapat diakses dengan menghubungi korespondensi.

Pernyataan Penggunaan Artificial Intelligence

Penulis menggunakan alat bantu berbasis Artificial Intelligence (AI) secara terbatas hanya untuk membantu perbaikan bahasa, tata bahasa, struktur kalimat, dan keterbacaan naskah. AI tidak digunakan untuk menghasilkan, merekayasa, atau memodifikasi data dan temuan penelitian. Seluruh data penelitian, seperti kutipan atau penggalan teks, unsur kebahasaan, transkrip, hasil observasi, wawancara, angket, hasil tes, karya atau respons peserta didik, serta bentuk data lainnya sesuai dengan objek dan metode penelitian, diperoleh dari sumber penelitian yang sebenarnya dan telah diverifikasi oleh penulis. Analisis, interpretasi, pembahasan, dan kesimpulan merupakan tanggung jawab sepenuhnya dari penulis.

Daftar Pustaka

- Ahn, H., Kim, Y., Kim, J., & Han, Y.-S. (2024). SharedCon: Implicit hate speech detection using shared semantics. In *Findings of the Association for Computational Linguistics: ACL 2024* (pp. 10444–10455). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.findings-acl.622>
- Council of Europe. (2022). *Recommendation CM/Rec(2022)16 of the Committee of Ministers to member States on combating hate speech*. <https://search.coe.int/cm?i=0900001680a67955>
- Coulthard, M., Johnson, A., & Wright, D. (2016). *An introduction to forensic linguistics: Language in evidence* (2nd ed.). Routledge. <https://doi.org/10.4324/9781315630311>
- Davidson, T., Warmley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 11, pp. 512–515).
- ElSherief, M., Ziems, C., Muchlinski, D., Anupindi, V., Seybolt, J., De Choudhury, M., & Yang, D. (2021). Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 345–363). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.emnlp-main.29>
- Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys*, 51*(4), Article 85, 1–30. <https://doi.org/10.1145/3232676>
- Ghosh, S., Suri, M., Chiniya, P., Tyagi, U., Kumar, S., & Manocha, D. (2023). CoSyn: Detecting implicit hate speech in online conversations using a context synergized hyperbolic network. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 6159–6173). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.emnlp-main.377>
- 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>
- Kamalludin, I., & Arief, B. N. (2019). Kebijakan formulasi hukum pidana tentang penanggulangan tindak pidana penyebaran ujaran kebencian (*hate speech*) di dunia maya. *Law Reform*, 15*(1), 113–129. <https://doi.org/10.14710/lr.v15i1.23358>
- Kiela, D., Firooz, H., Mohan, A., Goswami, V., Singh, A., Fitzpatrick, C., ... Lipstein, G. (2020). The hateful memes challenge: Detecting hate speech in multimodal memes. In *Advances in Neural Information Processing Systems*, 33*, 2611–2624.
- Lim, M. (2017). Freedom to hate: Social media, algorithmic enclaves, and the rise of tribal nationalism in Indonesia. *Critical Asian Studies*, 49*(3), 411–427. <https://doi.org/10.1080/14672715.2017.1341188>
- Risquallah, M. A., Rosmaya, M., & Assiroj, P. (2026). Social media comments as forensic linguistic corpus in cybercrime handling: A systematic literature review. *MULTITEK Indonesia: Jurnal Ilmiah*, 20*(1).
- Mullah, N. S., & Zainon, W. M. N. W. (2021). Advances in machine learning algorithms for hate speech detection in social media: A review. *IEEE Access*, 9*, 88364–88376. <https://doi.org/10.1109/ACCESS.2021.3089515>
- Nuraeni, I., Harisanti, N. L. R., & Maksum, H. (2022). Tuturan kebencian dalam komentar warganet pada akun Instagram @obrolanpolitik: Kajian pragmatik. *Ranah: Jurnal Kajian Bahasa*, 11*(1), 189–197. <https://doi.org/10.26499/rnh.v11i1.4198>
- Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., & Patti, V. (2021). Resources and benchmark corpora for hate speech detection: A systematic review. *Language

- Resources and Evaluation, 55*(2), 477–523. <https://doi.org/10.1007/s10579-020-09502-8>
- Ramos, G., Batista, F., Ribeiro, R., Fialho, P., Moro, S., Fonseca, A., Guerra, R., Carvalho, P., Marques, C., & Silva, C. (2024). A comprehensive review on automatic hate speech detection in the age of the transformer. **Social Network Analysis and Mining*, 14*, Article 204. <https://doi.org/10.1007/s13278-024-01361-3>
- 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), e1648. <https://doi.org/10.1002/wics.1648>
- Röttger, P., Vidgen, B., Nguyen, D., Waseem, Z., Margetts, H., & Pierrehumbert, J. (2021). HateCheck: Functional tests for hate speech detection models. In **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing** (pp. 41–58). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.acl-long.4>
- Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2019). The risk of racial bias in hate speech detection. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics** (pp. 1668–1678). Association for Computational Linguistics. <https://doi.org/10.18653/v1/P19-1163>
- Suryani, Y., Istianingrum, R., & Hanik, S. U. (2021). Linguistik forensik ujaran kebencian terhadap artis Aurel Hermansyah di media sosial Instagram. **Belajar Bahasa: Jurnal Ilmiah Program Studi Pendidikan Bahasa dan Sastra Indonesia*, 6*(1), 107–118. <https://doi.org/10.32528/bb.v6i1.4167>
- Syahid, A., & Datang, F. A. (2024). Multimodalitas dalam linguistik forensik: Studi kasus ujaran kebencian terhadap etnis Tionghoa di Facebook. **Kandai*, 20*(2). <https://doi.org/10.26499/jk.v20i2.7033>
- Undang-Undang Republik Indonesia Nomor 1 Tahun 2024 tentang Perubahan Kedua atas Undang-Undang Nomor 11 Tahun 2008 tentang Informasi dan Transaksi Elektronik. <https://peraturan.bpk.go.id/Details/274494/uu-no-1-tahun-2024>
- Vidgen, B., & Derczynski, L. (2020). Directions in abusive language training data, a systematic review: Garbage in, garbage out. **PLoS ONE*, 15*(12), e0243300. <https://doi.org/10.1371/journal.pone.0243300>
- Vidgen, B., Harris, A., Nguyen, D., Tromble, R., Hale, S., & Margetts, H. (2019). Challenges and frontiers in abusive content detection. In **Proceedings of the Third Workshop on Abusive Language Online** (pp. 80–93). Association for Computational Linguistics. <https://doi.org/10.18653/v1/W19-3509>
- Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In **Proceedings of the NAACL Student Research Workshop** (pp. 88–93). Association for Computational Linguistics. <https://doi.org/10.18653/v1/N16-2013>
- Yu, X., Blanco, E., & Hong, L. (2022). Hate speech and counter speech detection: Conversational context does matter. In **Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies** (pp. 5918–5930). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.naacl-main.433>