Perspektif: Jurnal Pendidikan dan Ilmu Bahasa
Volume 4, Nomor 2, Juni 2026
e-ISSN: 2988-5914; p-ISSN: 3025-0641, Hal. 42-59
DOI: https://doi.org/10.59059/perspektif.v4i2.3063
Tersedia: https://journal.staiypiqbaubau.ac.id/index.php/Perspektif

Bias Algoritma dan Kegagalan Pragmatik AI dalam Mengidentifikasi Ujaran Kebencian Berbasis Budaya Lokal di Indonesia

Qinthara Khairun Azida[1*], Zakiyatul Marwa[2], Nazarena Putri Narahita[3], Elsa Rahma Sari[4], Ahmad Arzani Ibnul Hikam[5], Bohdan Filipov[6]
[1-3]Program Studi Farmasi, Universitas Gadjah Mada, Indonesia
[4-5]Program Studi Filsafat, Universitas Gadjah Mada, Indonesia
[6]Program Studi ND Sastra dan Bahasa Indonesia, Universitas Gadjah Mada, Indonesia
Email: qintharakhairunazida@mail.ugm.ac.id [1], zakiyatulmarwa@mail.ugm.ac.id[2], nazarenaputrinarahita@mail.ugm.ac.id[3], elsarahmasari@mail.ugm.ac.id[4], ahmadarzaniibnulhikam@mail.ugm.ac.id[5], bohdanfilipov@mail.ugm.ac.id[6]*
**Penulis Korespondensi: qintharakhairunazida@mail.ugm.ac.id*

Abstract*. This study aims to identify the pragmatic failures of Large Language Models (LLMs) and the biases of Anglophone-based AI moderation algorithms in detecting Indonesian hate speech expressed through sarcasm, satire, euphemism, and local cultural metaphors. It also examines the extent to which AI systems understand and interpret the pragmatic meanings within the corpus. This study employs a qualitative descriptive approach with a comparative design. Data were collected through the documentation of hate speech expressions on social media containing elements of local cultural hatred. The data were analyzed using qualitative descriptive methods with pragmatic and thematic approaches. The findings show that all corpus data contain political satire and indirect hate expressed through irony, sarcasm, absurd metaphors, and popular culture wordplay. Testing with Claude AI showed that the system was capable of identifying the data as implicit criticism and recognizing the pragmatic functions of emoticons and contextual meanings in the utterances. However, the analysis also demonstrated limitations in understanding local sociocultural contexts, particularly the metaphors "daun nangka" and "daun sawit," which were interpreted merely as absurd humor. These findings indicate that AI detection accuracy does not necessarily reflect a deep pragmatic and cultural understanding within the Indonesian context.*

Keywords*: Algorithmic Bias; Indonesian Hate Speech; Large Language Models (LLMs); Local Cultural Sarcasm; Pragmatics.*

Abstrak. Penelitian ini bertujuan untuk mengidentifikasi kegagalan pragmatik LLM dan bias algoritma moderasi AI berbasis perspektif Anglofon dalam mendeteksi kebencian berbahasa Indonesia yang menggunakan sarkasme, sindiran, eufemisme, dan metafora budaya lokal, serta mengukur pemahaman interpretasi dan AI terhadap korpus tersebut. Penelitian ini menggunakan pendekatan deskriptif kualitatif dengan desain komparatif. Teknik pengumpulan data dilakukan melalui penelusuran dan pencatatan kebencian pada media sosial yang mengandung unsur kebencian berbasis budaya lokal. Teknik analisis data dilakukan secara deskriptif kualitatif dengan pendekatan pragmatik dan analisis tematik. Hasil penelitian menunjukkan bahwa seluruh data korpus mengandung sindiran politik dan kebencian tidak langsung yang diwujudkan melalui ironi, sarkasme, metafora absurd, dan plesetan budaya populer. Pengujian terhadap Claude AI menampilkan bahwa sistem mampu mengidentifikasi seluruh data sebagai kritik terselubung serta mengenali fungsi pragmatik emotikon dan keberadaan makna dalam tuturan. Namun, analisis menunjukkan adanya keterbatasan pada pemahaman konteks sosiokultural lokal, khususnya pada metafora "daun nangka" dan "daun sawit" yang hanya dianggap sebagai humor yang absurd. Temuan ini menunjukkan bahwa akurasi deteksi AI belum sepenuhnya mencerminkan pemahaman pragmatik dan budaya yang mendalam dalam konteks Indonesia.

Kata kunci: Bias Algoritma; Kebencian Berbahasa Indonesia; *Large Language Model* (LLM); Pragmatik; Sarkasme Budaya Lokal.

1. LATAR BELAKANG

Perkembangan dunia digital di Indonesia ditandai denga peningkatan jumlah pengguna sosial media, dilansir dari data Statista (Yonatan, 2026) presentase penduduk Indonesia pengguna sosial media mencapai 81%. Banyaknya pengguna sosial media menjadi tantangan tersendiri di tengah arus informasi yang cepat menyebar. Penyebaran komentar negatif seperti

Naskah Masuk: 06 April 2026; Revisi: 04 Mei 2026; Diterima: 20 Juni 2026; Terbit: 26 Juni 2026

ujaran kebencian (*hate speech*) menjadi salah satu contoh tantangan yang berpotensi mengancam integrasi sosial. Ujaran kebencian merujuk pada bentuk komunikasi yang mengintimidasi, mengancam, dan menghina suatu individu maupun kelompok (Farwati et al., 2023). Dalam rangka memilah dan menyaring teks yang mengandung ujaran kebencian digunakan teknologi *Artificial Intelligence* (AI) berbasis *Large Languange Models* (LLM) sebagai alat yang bekerja secara otomatis.

Beberapa studi terdahulu telah meneliti efektivitas AI maupun LLM dalam mitigasi konten di ruang media sosial. Penelitian sebelumnya menyebutkan bahwa seluruh LLM yang diuji oleh peneliti menciptakan disinformasi dan dapat dieksploitasi dalam rangka untuk menciptakan disinformasi (Vinay et al., 2025). ChatGPT sebagai salah satu model AI berbasis LLM, ketika diberikan data terkait Amerika akan menunjukkan keselarasan yang kuat dengan budaya Amerika, tetapi kurang efektif beradaptasi dengan konteks budaya lain, salah satu penyebab hal ini adalah penggunaan bahasa Inggris yang digunakian untuk menguji LLM (Cao et al., 2023). Penggunaan *code-mixing* (percampuran bahasa) dalam satu ucapan membuat LLM gagal memahami makna yang akurat (khususnya dalam ucapan bernuansa sarkasme), penelitian ini berfokus pada percampuran bahasan antara Tamil-Inggris dan Malayalam-Inggris (Deroy & Maity, 2025).

LLM dianggap sebagai model AI dengan kekuatan pemahaman semantik, penalaran kontekstual, dan kemampuan generatif yang memperkenalkan paradigma transformatif untuk moderasi konten (Chen et al., 2025). Padahal model AI masih gagal mendeteksi ujaran kebencian terselubung yang dibungkus dengan sarkasme halus, metafora, maupun dialek lokal khas Indonesia. Celah ini memicu bias algoritma yang berisiko meloloskan konten provokatif dan salah menyensor konten dengan ekspresi budaya atau humor lokal yang lahir dalam tradisi masyarakat Indonesia. Celah sosiokultural inilah yang menjadi fokus utama dalam penelitian ini sekaligus menjadi kebaruan dalam penelitian ini.

Penelitian ini bertujuan untuk mengidentifikasi bentuk-bentuk kegagalan pragmatik LLM dalam memproses ujaran kebencian yang menggunakan strategi kebahasaan lokal seperti sarkasme, sindiran, dan metafora budaya. Selanjutnya, penelitian ini bertujuan untuk menganalisis bias algoritma pada sistem moderasi otomatis berbasis AI yang didominasi perspektif Anglofon (penutur) berdampak pada kegagalan pendeteksian ujaran kebencian yang disampaikan melalui sarkasme, eufemisme, dan metafora budaya khas Indonesia. Serta untuk mengukur tingkat kesesuaian interpretasi manusia dan AI terhadap korpus ujaran kebencian berbasis konteks budaya Indonesia.

2. KAJIAN TEORITIS

Sistem moderasi konten berbasis kecerdasan buatan terus berkembang, namun ada masalah mendasar yang sering diabaikan. Model-model ini tidak benar-benar memahami sistem moderasi otomatis dibangun hampir sepenuhnya dari korpus data berbahasa Inggris, dengan perspektif Anglofon yang sangat kental. Ketika model semacam itu dipaksakan bekerja pada bahasa Indonesia dengan segala dialek, kearifan lokal, dan lapisan pragmatiknya banyak hal yang luput dari deteksi. Menurut Sonni (2025), menunjukkan bahwa ekosistem media digital Indonesia menunjukkan pola amplifikasi ujaran kebencian yang tidak terdeteksi optimal oleh sistem moderasi berbasis kecerdasan buatan, salah satunya karena ketidaksesuaian antara model dengan karakteristik bahasa lokal. Hal ini bukan sekadar soal teknis yang bisa ditambal ini soal kerangka pikir yang sudah keliru sejak awal.

Tantangan ini tidak bisa dianggap enteng. Ujaran kebencian di Indonesia jarang hadir dalam bentuk serangan verbal yang langsung dan eksplisit kebencian justru lebih sering bersembunyi di balik sarkasme, eufemisme, atau metafora yang hanya bisa ditangkap oleh mereka yang tumbuh dalam konteks budayanya. Sementara itu, Wijanarko et al., (2024) mencatat bahwa kompleksitas tuturan yang menargetkan kelompok minoritas di Indoensia begitu berlapis. Satu teks bisa sekaligus membawa label kebencian etnis, agama, dan politik tanpa satupun kata kasar di dalamnya. Model NLP yang dilatih untuk mengenali pola permukaan tidak akan pernah cukup untuk ini. Lebih mendasar lagi, penelitian pragmatik terhadap LLM menunjukkan bahwa kemampuan model-model ini dalam menyelesaikan implikatur percakapan masih jauh dari memadai pada kondisi *zero-shot*, akurasi model pra-latih mendekati angka acak, sementara manusia mencapai 86% (Ruis et al., 2023). Dengan kata lain, inti dari ujaran kebencian pragmatis makna yang tidak pernah diucapkan secara harfiah adalah justru titik buta yang paling sulit diatasi oleh sistem deteksi otomatis saat ini. Lee et al., (2024) menunjukkan bahwa model kecerdasan buatan mempunyai keterbatasan serius dalam mendeteksi sarkasme. Hanya ketika mereka menguji pendekatan *pragmatic metacognitive prompting* akurasi deteksi naik secara bermakna. Tanpa panduan konstekstual yang ekplisit, model standar memang praktis buta terhadap dimensi pragmatik dari tuturan ambigu.

Bias dalam sistem deteksi ujaran kebencian bukan spekulasi karena ini sudah terdokumentasi empiris. Davani et al., (2023) menunjukkan bahwa klasifikasi ujaran kebencian yang dilatih dari anotasi manusia secara sistematis menyerap dan mereproduksi stereotip sosial terhadap kelompok-kelompok marginal. Giorgi et al., (2025) melangkah lebih jauh tentang bias anotator manusia dan model kecerdasan buatan ternyata saling berinteraksi, menciptakan distorsi sistematis dalam pelabelan data. Di Indonesia, dengan lebih dari 700 bahasa daerah

dan puluhan kelompok etnis, akibatnya jauh lebih serius. Sistem moderasi yang bias tidak sekadar tidak akurat. Ia secara aktif merugikan komunitas yang sudah rentan, sambil membiarkan ujaran kebencian dari kelompok yang lebih dominan lolos begitu saja. Ini bukan kegagalan teknis tetapi kegagalan keadilan sosial.

Adapun lapisan bias lain yang juga perlu diperhatikan, yaitu bias geografis. Personal. Piot et al., (2025) menunjukkan bahwa wilayah yang kurang terwakili dalam data pelatihan secara konsisten mendapat akurasi deteksi yang lebih rendah. Di Indonesia, artinya wilayah luar Jawa, kawasan pedesaan, dan komunitas yang belum sepenuhnya masuk eksosistem digital nasional menanggung risiko lebih tinggi akibat kesalahan sistem. Teknik debias tuning pada model kecerdasan buatan memang tersedia, tetapi penerapannya membutuhkan data berlabel berkualitas tinggi dalam bahasa-bahasa daerah Indonesia yang hingga kini masih sangat langka. Keterbatasan data ini justru memperparah siklus bias yang sudah ada, dan komunitas paling rentan yang terus menanggung akibatnya.

Beberapa peneliti mulai membangun fondasi yang lebih solid di tengah semua keterbatasan. Susanto et al., (2025) mengembangkan dataset ujaran kebencian berbahasa Indonesia yang diperkaya dengan informasi demografis anotator sebuah pengakuan bahwa persepsi terhadap ujaran kebencian tidak universal, melainkan arah berbeda dengan mengembangkan kerangka pengujian fungsional yang dirancang khusus untuk bahasa-bahasa *low resource* di Asia Tenggara, termasuk bahasa Indonesia. Kedua inisiatif ini mempunyai satu benang merah yang penting. Solusi untuk tantangan moderasi konten di kawasan ini harus tumbuh dari dalam konteksnya sendiri, bukan dari adaptasi model yang dikembangkan untuk realitas budaya yang berbeda.

Pendekatan sinergis antara manusia dan AI bukan sekadar kompromi, ini adalah pengakuan jujur bahwa sistem otomatis memiliki batas yang tidak bisa dilompati begitu saja. Park et al., (2025) bahkan sudah mengujinya secara langsung sistem kolaboratif manusia-LLM untuk moderasi ujaran kebencian lintas budaya terbukti lebih akurat dibanding sistem otomatis murni. Dari temuan-temuan ini, ada satu kesimpulan yang sulit dibantah terkait paradigma "AI sebagai pengganti manusia" sudah saatnya diganti dengan "AI sebagai mitra kerja", dimana sistem otomatis hanya berperan sebagai penyaring awal, lalu dilengkapi dengan tinjauan manusia yang benar-benar paham konteks budaya dan linguistik setempat. Untuk Indonesia, ini bukan pilihan tamabahan tetapi ini keharusan. Moderator yang ememhami dialek Jawa, Sunda, Batak, atau Bugis bukan pelengkap sistem. Mereka adalah inti dari sistem moderasi yang bisa dianggap adil.

Bahkan kolaborasi manusia-AI pun tidak otomatis berjalan mulus. Salah satu tantangan metodologis yang paling sulit dipecahkan adalah ketidakselarasan antara cara manusia dan AI membaca ujaran yang secara harfiah terdengar biasa, tetapi secara pragmatik bersifat menyerang. Singh et al., (2025) memperjelas mengapa ini bisa terjadi, berhubungan pada pendekatan *pragmatic metacognitive prompting* yang dikembangkan untuk konteks bahasa Inggris varian Australia dan India ternyata tidak bisa begitu saja dipindahkan ke bahasa lian yang punya logika pragmatik berbeda. Ia butuh adaptasi yang lebih mendalam, bukan sekadar terjemahan. Beberapa pelajaran yang perlu dicatat baik-baik ialah seberapa selaras manusia dan AI dalam menilai ujaran kebencian pragmatis sangat bergantung pada sejauh mana sistem AI dirancang untuk benar-benar mempertimbangkan konteks sosio budaya yang spesifik bukan menerapkan standar "universal" yang pada kenyataannya tidak pernah benar-benar universal.

Pada akhirnya, bias algoritma dan kegagalan pragmatik AI dalam konteks Indonesia bukan persoalan pinggiran yang bisa ditunda. Ini menyadarkan sesuatu yang jauh lebih mendasar, yakni keadilan digital dan kedaulatan linguistik masyarakat Indonesia atas ruang digitalnya sendiri. Wijanarko et al., (2024) sudah menunjukkan bahwa klasifikasi teks media sosial berbahasa Indonesia memerlukan pendekatan berlapis yang mempertimbangkan dimensi linguistik, budaya, dan sosial secara bersamaan tidak bisa hanya salah satu. Proses kedepannya ada tiga hal yang tidak bisa ditawar jika Indonesia ingin punya sistem moderasi yang benar-benar bekerja. Pertama, investasi serius dalam pembangunan data berlabel berkualitas tinggi untuk berbagai bahasa daerah. Kedua, pengembangan model yang secara eksplisit mengintegrasikan kompetensi pragmatik lintas budaya, dan ketiga yaitu pembentukan kerangka tata kelola yang menempatkan komunitas pengguna bukan sekadar sebagai objek moderasi, melainkan sebagai pemangku kepentingan aktif yang suaranya didengar. Tanpa langkah-langkah konkret itu sistem AI yang mengelola ruang digital Indonesia tidak akan menjadi pelindung warganya ia justru akan terus memproduksi ulang ketidakadilan yang sudah ada, hanya dengan kecepatan yang lebih tinggi dan wajah yang lebih saintifik.

3. METODE PENELITIAN

Penelitian ini menggunakan pendekatan kualitatif deskriptif dengan desain komparatif. Pendekatan ini bertujuan untuk membandingkan interpretasi manusia dengan respons kecerdasan buatan *Artificial Intelligence* (AI) dalam mengidentifikasi ujaran bencian berbasis budaya lokal di Indonesia. Fokus penelitian diarahkan pada kemampuan AI dalam memahami makna pragmatik, konteks sosial budaya, serta unsur implisit dalam suatu ujaran.

Data penelitian berupa ujaran yang mengandung indikasi ujaran kebencian tidak langsung yang diperoleh dari platform media sosial X dan TikTok selama periode pengamatan April–Mei 2026. Ujaran tersebut dapat berupa komentar, respons digital, maupun bentuk komunikasi singkat lain yang mengandung ujaran kebencian tidak langsung, satire, sarkasme, ironi, serta metafora budaya dalam konteks komunikasi digital masyarakat Indonesia. Korpus penelitian dibatasi pada ujaran berbahasa Indonesia nonformal yang menunjukkan unsur delegitimasi, sinisme, penghinaan terselubung, atau ekspresi kebencian implisit terhadap individu, kelompok, maupun otoritas tertentu.

Penelitian ini memfokuskan analisis pada strategi pragmatik dalam praktik komunikasi digital masyarakat Indonesia dan tidak mencakup ujaran kebencian eksplisit, analisis multimodal (gambar/video), maupun analisis mendalam terhadap variasi bahasa daerah. Kriteria identifikasi data mengacu pada keberadaan unsur sinisme, delegitimasi, penghinaan terselubung, atau ekspresi kebencian implisit yang disampaikan melalui ironi, hiperbola, eufemisme, pertanyaan retoris, maupun bentuk pragmatik lainnya. Data yang tidak memiliki konteks komunikasi yang jelas, ujaran duplikat, spam, atau respons yang tidak menunjukkan muatan pragmatik tidak dimasukkan ke dalam korpus penelitian.

Analisis data dilakukan melalui tiga tahap. Pertama, peneliti melakukan interpretasi manual terhadap korpus untuk mengidentifikasi makna literal, implikatur, target ujaran, serta konteks sosial budaya yang melatarbelakangi ujaran. Kedua, seluruh korpus dianalisis menggunakan Claude AI dengan instruksi (guided prompt) yang seragam untuk menjaga konsistensi pengujian. Ketiga, hasil interpretasi manusia dibandingkan dengan output AI berdasarkan kemampuan sistem dalam mengenali satire, memahami implikatur pragmatik, mengidentifikasi konteks sosial budaya lokal, dan menafsirkan makna implisit dalam ujaran.

Teknik analisis data dilakukan secara deskriptif kualitatif dengan pendekatan pragmatik dan analisis tematik. Pendekatan pragmatik digunakan untuk menafsirkan makna ujaran berdasarkan konteks situasi, maksud penutur, serta relasi sosial budaya yang terkandung dalam komunikasi digital masyarakat Indonesia. Sementara itu, analisis tematik digunakan untuk mengidentifikasi pola bias algoritmik, keterbatasan pemahaman kontekstual AI, serta potensi kesalahan klasifikasi dalam proses identifikasi ujaran kebencian. Validitas data dilakukan melalui perbandingan berbagai teori pragmatik dan sejumlah data ujaran untuk memastikan konsistensi interpretasi hasil analisis. Hasil penelitian kemudian disajikan secara deskriptif analitis untuk menjelaskan hubungan antara bias algoritmik, konteks sosial budaya lokal, dan keterbatasan AI dalam moderasi ujaran kebencian di ruang digital Indonesia.

4. HASIL DAN PEMBAHASAN

Bagian ini menyajikan data empiris yang diperoleh dari proses pengumpulan korpus data di media sosial, serta hasil pengujian mekanis menggunakan model bahasa besar (*Large Language Model*), yaitu Claude AI. Fokus penyajian hasil ini dibagi menjadi dua tahap: (1) klasifikasi karakteristik kebahasaan korpus data ujaran kebencian tidak langsung (*indirect hate speech*), dan (2) matriks evaluasi kemampuan pragmatik AI dalam mengidentifikasi teks-teks tersebut.

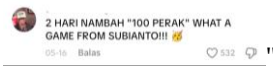
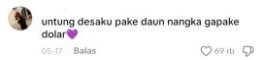
Klasifikasi Korpus Data Satire Politik dan Ujaran Kebencian Tidak Langsung

Berdasarkan hasil observasi di media sosial (X dan TikTok) terkait respons netizen terhadap isu fluktuasi nilai tukar rupiah dan pernyataan pejabat publik, peneliti mengumpulkan 8 korpus data utama. Seluruh data ini memiliki karakteristik unik, yaitu tidak menggunakan kata-kata kasar makian secara eksplisit (*direct hate speech*), melainkan menggunakan represi kebahasaan yang halus.

Karakteristik kebahasaan dan target sindiran dari masing-masing korpus data dijabarkan dalam Tabel 1 berikut:

Tabel 1. Klasifikasi Korpus Data Satire Politik dan Ujaran Kebencian Tidak Langsung.

No.	Kode Sampel	Teks Komentar	Karakteristik Kebahasaan / Gaya Bahasa	Target / Objek Sindiran
1	SK-01	 "...guys gak usah takut rupiah tembus ke 100 rb pun, kata ngab owo di desa gak hidup pake dollar😈 sungguh pikiran cerdas bapak presidenku ini😈 semoga jadi presiden seumur hidup😈"	Ironi & Sarkasme Ekstrem (Pujian Semu): Menggunakan jargon pujian dan emotikon kasih sayang (😈) untuk mendelegitimasi pernyataan figur otoritas.	Presiden / Otoritas Politik Tertinggi
2	SK-02	 "kita menang lagi nih? 😹"	Sinisme Retoris: Pertanyaan retoris singkat yang menggunakan emotikon kucing tertawa (😹) untuk meremehkan narasi keberhasilan pemerintah.	Narasi/Klaim Politik Pemerintah

3	SK-03	 "ga ngaruh bang, kita di desa masi pake sistem barter"	Satire Kontekstual (Reductio ad Absurdum): Menyederhanakan masalah ekonomi makro secara absurd dengan membawa narasi "kehidupan desa" yang primitif.	Kebijakan Ekonomi Pemerintah
4	SK-04	 "2 HARI NAMBAH "100 PERAK" WHAT A GAME FROM SUBIANTO!!! 🥳"	Sarkasme Berbasis Hiperbola: Menggunakan istilah dunia gaming ("What a game") dan emotikon pesta (🥳) untuk menyindir kegagalan stabilitas nilai tukar.	Presiden (Spesifik: Nama Belakang)
5	SK-05	 "gak usah pusing sama dollar kak, toh kita pake nya rupiah gak pake dollar wkwk 😂😭 manut aja apa kata pak presiden tercintaaaah"	Ironi Defensif Semu: Berpura-pura patuh ("manut aja") dan menggunakan tawa getir (😂😭) untuk mengkritik logika komunikasi publik otoritas.	Presiden / Kebijakan Ekonomi
6	SK-06	 "untung desaku pake daun nangka gapake dolar💜"	Metafora Kultural Lokal (Satire Absurd): Mengganti instrumen finansial dengan "daun nangka" sebagai simbol bahwa mata uang sudah tidak bernilai.	Krisis Finansial / Kebijakan Makro
7	SK-07	 "untung aku pakai daun sawit"	Pelesetan Kontekstual Geografis: Melanjutkan pola metafora daun dengan membawa komoditas lokal lain ("daun sawit") untuk menegaskan ketidakpedulian sinis.	Krisis Finansial / Kebijakan Makro
8	SK-08	 "sorry bro. GW orang desa. hal kayak gini gak ngaruh ke kehidupan GW. kita transaksi pake emerald"	Pelesetan Subkultur Digital / Pop Culture: Menggunakan istilah "emerald" (alat tukar fiktif dari game Minecraft) untuk memparodikan ketertinggalan logika ekonomi yang diucapkan otoritas.	Pernyataan Otoritas Politik

Hasil klasifikasi korpus temuan pada Tabel 1 menunjukkan bahwa ujaran kebencian dan kritik politik di media sosial Indonesia semakin banyak disampaikan melalui bentuk-bentuk tidak langsung (indirect hate speech). Seluruh bagian representatif yang dianalisis tidak menggunakan kata-kata kasar atau hinaan eksplisit, tetapi memanfaatkan metode pragmatik seperti ironi, sarkasme, metafora absurd, plesetan budaya populer, dan pertanyaan retoris. Fenomena ini memperlihatkan bahwa kebencian atau delegitimasi terhadap otoritas politik tidak selalu hadir dalam bentuk makian terbuka, melainkan disamarkan melalui humor, satire, dan permainan bahasa yang hanya dapat dipahami jika pembaca memiliki pemahaman terhadap konteks sosial dan budaya Indonesia.

Pada sampel SK-01, penggunaan frasa "sungguh pikiran cerdas bapak presidenku ini🥰" dan "semoga jadi presiden seumur hidup🥰" menunjukkan bentuk ironi dan sarkasme ekstrem. Secara literal, komentar tersebut tampak seperti pujian, tetapi secara pragmatik justru berfungsi untuk mengejek dan mendelegitimasi pernyataan pejabat publik terkait kondisi ekonomi. Penggunaan emotikon kasih sayang menjadi alat penguat sindiran karena bertolak belakang dengan maksud sebenarnya. Capaian ini menunjang pendapat Lee et al., (2024) bahwa model kecerdasan buatan masih memiliki keterbatasan serius dalam mendeteksi sarkasme, terutama ketika makna ujaran tidak berada pada level literal, melainkan tersembunyi dalam konteks pragmatik.

Hal serupa terlihat pada sampel SK-02 yang hanya berbentuk pertanyaan singkat "kita menang lagi nih? 😹". Meskipun sangat pendek dan tidak mengandung kata kasar, komentar tersebut memiliki muatan sinisme terhadap narasi keberhasilan pemerintah. Makna penghinaan muncul dari konteks sosial yang melatarbelakangi percakapan, bukan dari sintaks kalimat itu sendiri. Kondisi ini memperkuat argumentasi (Ruis et al., 2023) bahwa model NLP dan LLM modern masih lemah dalam memahami implikatur percakapan, terutama pada situasi zero-shot. Sistem moderasi otomatis memiliki pola perilaku untuk membaca teks seperti ini sebagai komentar biasa karena tidak terdapat indikator kebencian eksplisit.

Sampel SK-03, SK-06, SK-07, dan SK-08 memperlihatkan struktur satire absurd berbasis budaya lokal dan subkultur digital. Frasa seperti "kita di desa masih pake sistem barter", "pake daun nangka", "daun sawit", dan "transaksi pake emerald" merupakan metafora yang secara tidak langsung menyindir kondisi ekonomi dan logika komunikasi publik pemerintah. Penggunaan simbol-simbol lokal dan referensi budaya populer seperti Minecraft menunjukkan bahwa makna ujaran dibangun melalui pengetahuan kolektif komunitas digital. Bagi manusia yang memahami konteks budaya Indonesia dan budaya internet, makna sindiran

dapat segera dipahami. Namun, bagi sistem moderasi berbasis AI yang dilatih dominan menggunakan korpus bahasa Inggris, ujaran semacam ini sangat sulit dikategorikan sebagai satire politik atau ujaran kebencian tidak langsung.

Luaran ini sejalan dengan Sonni (2025) yang menyatakan bahwa sistem moderasi otomatis berbasis kecerdasan buatan masih gagal menangkap karakteristik pragmatik bahasa Indonesia karena mayoritas model dibangun menggunakan perspektif Anglofon. Bahasa Indonesia di media sosial tidak hanya menggunakan bahasa formal, tetapi juga bercampur dengan dialek daerah, bahasa gaul, plesetan, humor internet, dan representasi budaya lokal. Akibatnya, sistem AI sering kali tidak mampu membedakan antara candaan biasa dengan ujaran yang mengandung delegitimasi atau kebencian terselubung.

Pada sampel SK-04 dan SK-05 terlihat penggunaan hiperbola dan kepatuhan semu sebagai bentuk kritik politik. Kalimat "WHAT A GAME FROM SUBIANTO!!! 🥹" dan "manut aja apa kata pak presiden tercintaaaah" menunjukkan bentuk pujian palsu yang secara pragmatik bermakna negatif. Langkah-langkah semacam ini umum digunakan netizen Indonesia untuk menghindari sensor eksplisit sekaligus tetap menyampaikan kritik tajam. Secara linguistik, ujaran tersebut bersifat ambigu karena memiliki dua lapisan makna: makna literal yang tampak positif dan makna pragmatik yang bersifat menyerang. Ambiguitas inilah yang menjadi titik lemah utama sistem moderasi otomatis.

Selain persoalan pragmatik, keluaran penelitian ini juga memperlihatkan adanya potensi bias dalam sistem deteksi ujaran kebencian. Sap et al., (2019) memaparkan bahwa model deteksi cenderung menghasilkan pelabelan yang bias terhadap variasi bahasa tertentu. Dalam konteks Indonesia yang memiliki ratusan bahasa daerah dan variasi sosial, risiko kesalahan deteksi menjadi jauh lebih besar. Komentar seperti "daun nangka" atau "emerald" kemungkinan besar tidak akan dikenali sebagai satire politik karena sistem tidak memiliki pengetahuan budaya yang memadai. Pengaruhnya, ujaran bermuatan kebencian terselubung bisa lolos dari moderasi, sementara ujaran dari kelompok tertentu justru berpotensi diberi label negatif secara berlebihan.

Kondisi tersebut juga berkaitan dengan bias geografis sebagaimana dijelaskan oleh Piot et al., (2025). Bahasa media sosial Indonesia sangat dipengaruhi oleh konteks lokal, termasuk istilah daerah, komoditas lokal, dan budaya komunitas digital tertentu. Jika data pelatihan AI didominasi bahasa urban atau bahasa formal nasional, maka ujaran dari komunitas pedesaan atau luar Jawa akan lebih sulit dipahami oleh sistem. Dalam penelitian ini, penggunaan

representasi “desa”, “daun sawit”, dan “daun nangka” menunjukkan bagaimana identitas geografis dan budaya lokal menjadi bagian penting dalam konstruksi satire politik.

Temuan penelitian ini membuktikan perspektif bahwa moderasi konten tidak dapat sepenuhnya diserahkan pada sistem otomatis. Park et al., (2025) menegaskan bahwa pendekatan kolaboratif manusia dan AI menghasilkan akurasi moderasi yang lebih baik dibanding sistem otomatis murni. Dalam konteks Indonesia, moderator manusia yang memahami budaya lokal, humor digital, serta variasi pragmatik bahasa menjadi komponen penting untuk menafsirkan makna tersembunyi dalam ujaran media sosial. AI berpotensi digunakan sebagai penyaring awal, tetapi keputusan akhir tetap membutuhkan penafsiran manusia agar konteks budaya dan sosial tidak diabaikan.

Secara komperhensif, klasifikasi delapan korpus data menunjukkan bahwa ujaran kebencian dan satire politik di media sosial Indonesia semakin bergerak ke arah bentuk pragmatik yang implisit, ambigu, dan kontekstual. Strategi kebahasaan seperti ironi, sarkasme, metafora absurd, serta plesetan budaya populer menjadi alat utama untuk menyampaikan kritik dan delegitimasi politik tanpa menggunakan hinaan langsung. Capaian ini menegaskan bahwa tantangan moderasi konten di Indonesia bukan hanya persoalan teknis pendeteksian kata kasar, melainkan persoalan persepsi budaya, pragmatik, dan konteks sosial yang jauh lebih kompleks. Oleh karena itu, ekspansi sistem moderasi masa depan perlu mengintegrasikan strategi linguistik pragmatik, data lokal berbahasa Indonesia, serta kolaborasi aktif antara AI dan moderator manusia agar tercipta moderasi digital yang lebih adil dan kontekstual.

Matriks Evaluasi Kemampuan Pragmatik AI (Claude)

Setelah korpus data diklasifikasikan, peneliti melakukan pengujian mekanis dengan memasukkan 8 teks tersebut ke dalam sistem Claude AI menggunakan instruksi seragam (*guided prompt*) untuk mendeteksi adanya unsur ujaran kebencian tidak langsung, sinisme, atau satire politik. Hasil deteksi otomatis dan catatan kritis peneliti mengenai ketepatan analisis AI disajikan dalam Tabel 2 berikut:

Tabel 2. Matriks Evaluasi Kemampuan Pragmatik AI terhadap Satire Politik.

No.	Kode Sampel	Input Teks (Lokusi)	Deteksi AI	Ketepatan Analisis Pragmatik AI	Catatan Kritis Peneliti (Celah AI)
1	SK-01	"...ngab owo di desa gak hidup pake dollar🥰..."	YA (Satire)	Tepat	AI berhasil mengenali kontradiksi antara teks pujian dengan visualisasi krisis (rupiah 100rb).
2	SK-02	"kita menang lagi nih? 😸"	YA (Sinisme)	Tepat	AI mampu menerjemahkan fungsi emotikon 😸 sebagai instrumen meremehkan.
3	SK-03	"...kita di desa masi pake sistem barter"	YA (Satire)	Tepat	AI mengenali reduksi absurd masalah makroekonomi.
4	SK-04	"WHAT A GAME FROM SUBIANTO!!! 🥳"	YA (Sarkasme)	Tepat	AI mengenali sarkasme karena ada nama belakang target secara spesifik.
5	SK-05	"...manut aja apa kata pak presiden..."	YA (Satire)	Tepat	AI mengenali kontradiksi emosi pada gabungan emotikon 😂😭.
6	SK-06	"...untung desaku pake daun nangka..."	YA (Satire)	Cukup Tepat	AI mengenali devaluasi mata uang, namun gagal menjelaskan signifikansi budaya "daun nangka" di masyarakat lokal Indonesia.
7	SK-07	"untung aku pakai daun sawit"	YA (Satire)	Cukup Tepat	AI mendeteksi metafora berantai, tapi abai pada konteks geografis/komoditas lokal Indonesia.
8	SK-08	"...kita transaksi pake emerald"	YA (Satire)	Tepat	AI mengenali referensi subkultur pop (*Minecraft*).

Hasil asesmen pada Tabel 2 mengindikasikan bahwa kemampuan pragmatik AI, khususnya Claude, berada pada tingkat yang lebih maju dibandingkan asumsi umum mengenai kendala sistem moderasi otomatis. Seluruh korpus data berhasil dikenali sebagai bentuk satire politik, sinisme, maupun ujaran kebencian tidak langsung. Temuan ini menjadi menarik karena sebagian besar komentar tidak mengandung kata-kata kasar eksplisit, melainkan

memanfaatkan ironi, kontradiksi emosional, dan metafora implisit. Dengan kata lain, Claude AI tidak hanya bekerja pada level penelaahan leksikal, tetapi mulai mampu membaca relasi pragmatik antara konteks sosial, pilihan kata, dan ikon digital seperti emotikon.

Pada sampel SK-01 hingga SK-05, AI menunjukkan performa yang relatif baik dalam menangkap implikatur percakapan. Misalnya pada komentar "semoga jadi presiden seumur hidup😈", sistem berhasil memahami bahwa pujian tersebut sebenarnya merupakan bentuk sindiran politik. Hal ini mengindikasikan adanya kompetensi AI untuk mengenali kontradiksi antara makna literal dan maksud komunikatif penutur. Dalam kajian pragmatik, kejadian ini berkaitan dengan keahlian menginterpretasi illokusi, yaitu makna yang dimaksudkan penutur di balik ujaran literal. Hasil ini memperlihatkan perkembangan signifikan dibandingkan hasil kajian (Ruis et al., 2023) yang menunjukkan bahwa model bahasa pada kondisi zero-shot masih memiliki relevansi rendah dalam memahami implikatur percakapan.

Selain itu, keberhasilan AI dalam membaca kombinasi emotikon kontradiktif seperti 😂😭 atau penggunaan 😈 sebagai instrumen sindiran menunjukkan bahwa sistem mulai mampu mengenali fungsi pragmatik lambang visual digital. Dalam komunikasi media sosial, emotikon tidak lagi sekadar penanda emosi literal, tetapi menjadi alat retoris untuk membangun ironi dan sarkasme. Lee et al., (2024) memaparkan bahwa model AI umumnya mengalami kesulitan mendeteksi sarkasme jika hanya mengandalkan teks literal. Namun, penggunaan pendekatan contextual prompting dan pragmatic reasoning berpotensi meningkatkan sensitivitas model terhadap tuturan ambigu. Hasil riset ini mengindikasikan bahwa Claude AI tampaknya telah memiliki kapasitas reasoning pragmatik yang lebih baik dibanding model NLP konvensional, terutama dalam membaca susunan kontradiksi emosional dan satire berbasis konteks digital.

Meskipun demikian, keberhasilan tersebut belum sepenuhnya menunjukkan bahwa AI benar-benar memahami konteks budaya lokal secara mendalam. Pada sampel SK-06 dan SK-07, AI memang berhasil memberi label "satire", tetapi analisis pragmatik yang dihasilkan masih bersifat dangkal. Sistem hanya memahami bahwa "daun nangka" dan "daun sawit" merupakan metafora absurd yang menggantikan fungsi uang, tetapi gagal menangkap lapisan makna sosial yang melekat pada simbol tersebut. Dalam konteks masyarakat Indonesia, "daun nangka" dan "daun sawit" bukan sekadar objek acak, melainkan representasi kehidupan agraris, pekerja perkebunan, serta kelompok masyarakat pedesaan yang sering dijadikan representasi keterpinggiran ekonomi.

Kegagalan ini memvalidasi argumentasi Sonni (2025) bahwa sistem AI global masih dibangun dengan kerangka Anglofon yang kurang sensitif terhadap realitas linguistik lokal. AI dapat menginterpretasi sintaks umum satire karena memiliki data pelatihan yang luas mengenai humor internet dan sarkasme global, tetapi tidak memiliki kompetensi budaya yang cukup untuk menafsirkan simbol-simbol lokal Indonesia secara mendalam. Dengan demikian, AI hanya berhasil membaca struktur pragmatik permukaan, bukan makna sosial historis yang berada di balik pilihan kata tersebut.

Fenomena pada SK-06 dan SK-07 juga memperlihatkan bahwa pragmatik bahasa Indonesia di media sosial sangat dipengaruhi oleh konteks geografis dan ekonomi lokal. "Daun sawit" memiliki asosiasi masif dengan wilayah perkebunan dan masyarakat buruh sawit di Indonesia, sedangkan "daun nangka" merepresentasikan simbol keseharian masyarakat desa. Ketika simbol-simbol ini digunakan dalam satire politik, makna yang muncul bukan hanya humor absurd, tetapi juga kritik terhadap ketimpangan ekonomi dan ketidakpekaan elit terhadap realitas masyarakat bawah. AI gagal menangkap dimensi ini karena sistem tidak mempunyai pengetahuan sosiokultural yang cukup mengenai relasi antara simbol lokal dan kondisi sosial masyarakat Indonesia.

Temuan tersebut selaras dengan studi Piot et al., (2025) mengenai bias geografis dalam sistem AI. Wilayah dan budaya yang kurang terwakili dalam data pelatihan akan menghasilkan ketepatan pemaknaan yang lebih rendah. Dalam konteks Indonesia, mayoritas data digital yang digunakan untuk melatih model AI kemungkinan besar berasal dari bahasa Indonesia standar atau budaya urban populer, sementara ekspresi lokal berbasis komunitas agraris, daerah periferal, dan simbol budaya tradisional jauh lebih sedikit terwakili. Dampaknya, AI mampu memahami referensi global seperti "emerald" dari Minecraft pada SK-08, tetapi kesulitan memahami simbol lokal seperti "daun sawit" yang justru lebih dekat dengan realitas sosial Indonesia.

Perbedaan kapabilitas AI dalam memahami "emerald" dibanding "daun sawit" menjadi tolok ukur penting adanya ketimpangan representasi budaya dalam data pelatihan model. Referensi budaya populer global memiliki dokumentasi digital yang melimpah sehingga mudah dikenali AI, sedangkan simbol lokal Indonesia memiliki jejak data yang jauh lebih terbatas. Kondisi ini mengindikasikan bahwa kecerdasan buatan tidak benar benar netral, melainkan sangat dipengaruhi oleh dominasi budaya dalam ekosistem data global. Dengan kata lain, AI lebih "fasih" memahami budaya internet global dibanding memahami realitas budaya masyarakat lokal Indonesia.

Hasil penelitian ini juga memperlihatkan bahwa kecermatan klasifikasi tidak selalu identik dengan pengetahuan kontekstual yang mendalam. Secara teknis, Claude AI berhasil mengidentifikasi seluruh teks sebagai satire politik, tetapi pada beberapa kasus sistem gagal mendiskripsikan justifikasi sosiokultural di balik satire tersebut. Hal ini menunjukkan adanya perbedaan antara detection accuracy dan contextual comprehension. AI mampu mengenali pola linguistik bahwa sebuah teks mengandung sindiran, tetapi belum tentu memahami mengapa lambang ertentu dipilih dan apa makna sosial yang direpresentasikan oleh lambang tersebut.

Dalam perspektif moderasi konten digital, kondisi ini memiliki implikasi penting. Sistem AI mungkin cukup efektif digunakan sebagai alat penyaring awal untuk mendeteksi satire, sarkasme, dan ujaran kebencian tidak langsung. Namun, tanpa keterlibatan manusia yang memahami konteks budaya lokal, penguraian yang dihasilkan tetap berisiko dangkal dan bias. Luaran ini mendukung pandangan Park et al., (2025) bahwa pendekatan kolaboratif manusia dan AI merupakan model moderasi paling realistis untuk konteks multikultural. AI berfungsi untuk mempercepat penelusuran struktur linguistik, sedangkan manusia diperlukan untuk menafsirkan makna sosial, budaya, dan historis yang tidak dapat direduksi menjadi susunan statistik semata.

Secara komperhensif, hasil penilaian pragmatik terhadap Claude AI menunjukkan dua realitas yang berjalan bersamaan. Di satu sisi, AI telah mengalami perkembangan bermakna dalam mendeteksi satire politik dan ujaran kebencian implisit melalui pengertian kontradiksi emosional, ironi, dan emotikon digital. Namun di sisi lain, AI masih mengalami kekurangan mendasar dalam memahami simbol budaya lokal dan kritik sosial berbasis konteks geografis Indonesia. Capaian ini menegaskan bahwa tantangan utama moderasi konten masa depan bukan hanya meningkatkan ketepatan klasifikasi, tetapi membangun sistem AI yang memiliki sensitivitas pragmatik dan keahlian budaya yang benar-benar kontekstual terhadap masyarakat Indonesia.

5. KESIMPULAN DAN SARAN

Penelitian ini mengindikasikan bahwa ujaran kebencian dan satire politik di media sosial Indonesia semakin banyak disampaikan melalui bentuk pragmatik tidak langsung seperti ironi, sarkasme, metafora absurd, dan plesetan budaya populer. Seluruh korpus data yang dianalisis memperlihatkan bahwa kritik politik tidak lagi bergantung pada penggunaan kata kasar eksplisit, melainkan dibangun melalui konteks sosial, kontradiksi emosional, dan representasi budaya yang hanya dapat dipahami secara pragmatis. Hasil pengujian terhadap Claude AI menunjukkan bahwa sistem kecerdasan buatan telah mampu mendeteksi keberadaan

satire politik dan sinisme secara umum, termasuk mengenali penggunaan emotikon kontradiktif sebagai petunjuk makna implisit. Namun, kemampuan tersebut belum sepenuhnya disertai pemahaman kontekstual yang mendalam terhadap simbol lokal Indonesia. Kegagalan AI dalam menafsirkan makna sosiokultural pada metafora seperti "daun nangka" dan "daun sawit" mengindikasikan bahwa akurasi klasifikasi tidak selalu berarti keberhasilan menginterpretasi realitas budaya yang melatarbelakangi suatu ujaran.

Temuan ini memiliki sejumlah implikasi praktis bagi pengembangan sistem moderasi AI yang lebih adaptif terhadap konteks budaya Indonesia. Pertama, pengembang sistem moderasi perlu membangun basis data leksikal lokal (local cultural lexicon) yang memuat metafora, idiom, dan simbol kedaerahan yang umum digunakan dalam wacana politik digital, sehingga model dapat dilatih untuk mengenali makna implisit di balik referensi budaya tersebut, bukan hanya makna literalnya. Kedua, diperlukan mekanisme fine-tuning berbasis korpus lokal yang melibatkan anotator dari berbagai latar belakang etnis dan geografis di Indonesia, agar model tidak hanya mengenali pola bahasa Indonesia baku, tetapi juga variasi penggunaan bahasa di tingkat komunitas digital. Ketiga, sistem moderasi sebaiknya menerapkan pendekatan human-in-the-loop, di mana keputusan akhir terhadap konten yang terindikasi ambigu secara budaya tetap memerlukan validasi oleh moderator manusia yang memahami konteks lokal, sehingga AI berfungsi sebagai alat penyaring awal (first-pass filter) bukan pengambil keputusan tunggal. Keempat, perlu dikembangkan indikator kepercayaan kontekstual (contextual confidence score) yang memberi tanda ketika model mendeteksi adanya referensi budaya yang berada di luar batas pengetahuannya, sehingga sistem dapat secara otomatis mengeskalasi konten tersebut untuk ditinjau secara manual.

Implikasi-implikasi ini menegaskan bahwa pendekatan moderasi otomatis tidak dapat sepenuhnya menggantikan peran manusia, melainkan harus diposisikan sebagai sistem hibrida yang saling melengkapi. Kolaborasi antara AI dan moderator manusia yang memahami konteks linguistik serta budaya lokal menjadi cara pandang yang lebih realistis untuk menciptakan moderasi konten yang adil dan kontekstual di Indonesia.

Penelitian ini memiliki keterbatasan pada jumlah korpus data yang masih terbatas dan fokus penelitian yang hanya menguji satu model AI, sehingga temuan penelitian belum berpotensi digeneralisasikan untuk seluruh sistem kecerdasan buatan maupun seluruh bentuk ujaran kebencian di media sosial Indonesia. Selain itu, penelitian ini belum menguji variasi bahasa daerah, dialek regional, maupun respons AI dalam konteks multimodal seperti kombinasi teks, gambar, dan video. Penelitian selanjutnya disarankan untuk menggunakan dataset yang lebih luas dan beragam, melibatkan berbagai model AI, serta mengembangkan

pendekatan analisis pragmatik lintas budaya agar sistem moderasi digital di Indonesia berpotensi lebih sensitif terhadap konteks sosial dan kearifan lokal masyarakat.

DAFTAR REFERENSI

Cao, Y., Zhou, L., Lee, S., Cabello, L., Chen, M., & Hershcovich, D. (2023). Assessing Cross-Cultural Alignment between ChatGPT and Human Societies: An Empirical Study. *Cross-Cultural Considerations in NLP at EACL, C3NLP 2023 - Proceedings of the Workshop*, 53–67. https://doi.org/10.18653/v1/2023.c3nlp-1.7

Chen, C., Qu, W., Su, S., Feng, Y., & Li, T. (2025). A comprehensive review of LLM-based content moderation: advancements, challenges, and future directions. *Knowledge-Based Systems*, *330*, 114689. https://doi.org/https://doi.org/10.1016/j.knosys.2025.114689

Davani, A. M., Atari, M., Kennedy, B., & Dehghani, M. (2023). Hate Speech Classifiers Learn Normative Social Stereotypes. *Transactions of the Association for Computational Linguistics*, *11*(1), 300–319. https://doi.org/10.1162/tacl_a_00550

Deroy, A., & Maity, S. (2025). YouTube Comments Decoded: Leveraging LLMs for Low Resource Language Classification. *CEUR Workshop Proceedings*, *4054*, 244–254.

Farwati, R., Yuliyanti, W., & Ningsih, W. P. R. (2023). Ujaran Kebencian Dan Perundungan di Dunia Maya: Tantangan Etika dalam Ruang Digital Indonesia. *JISPENDIORA Jurnal Ilmu Sosial Pendidikan Dan Humaniora*, *2*(3), 213–225. https://doi.org/10.56910/jispendiora.v2i3.1001

Giorgi, T., Cima, L., Fagni, T., Avvenuti, M., & Cresci, S. (2025). Human and LLM Biases in Hate Speech Annotations: A Socio-Demographic Analysis of Annotators and Targets. *Proceedings of the International AAAI Conference on Web and Social Media*, *19*, 653–670. https://doi.org/10.1609/icwsm.v19i1.35837

Lee, J., Fong, W., Le, A., Shah, S., Han, K., & Zhu, K. (2024). Pragmatic Metacognitive Prompting Improves LLM Performance on Sarcasm Detection. *Proceedings of the 1st Workshop on Computational Humor (CHum)*, 1–8.

Park, J., Jeong, S., Song, S., Lee, Y., & Oh, A. (2025). *LLM-C3MOD: A Human-LLM Collaborative System for Cross-Cultural Hate Speech Moderation*. 71–88. https://doi.org/10.18653/v1/2025.c3nlp-1.7

Piot, P., Martín-Rodilla, P., & Parapar, J. (2025). *Personalisation or Prejudice? Addressing Geographic Bias in Hate Speech Detection using Debias Tuning in Large Language Models*.

Ruis, L., Khan, A., Biderman, S., Hooker, S., Rocktäschel, T., & Grefenstette, E. (2023). The Goldilocks of Pragmatic Understanding: Fine-Tuning Strategy Matters for Implicature Resolution by LLMs. *Advances in Neural Information Processing Systems*, *36*(NeurIPS).

Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2019). The risk of racial bias in hate speech detection. *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, 1668–1678. https://doi.org/10.18653/v1/p19-1163

Singh, I., Srirag, D., & Joshi, A. (2025). *Nek Minit: Harnessing Pragmatic Metacognitive Prompting for Explainable Sarcasm Detection of Australian and Indian English*. *May*.

Sonni, A. F. (2025). AI-based disinformation and hate speech amplification: analysis of Indonesia's digital media ecosystem. *Frontiers in Communication*, *Volume 10*. https://www.frontiersin.org/journals/communication/articles/10.3389/fcomm.2025.1603534

Susanto, L., Wijanarko, M. I., Pratama, P. A., Hong, T., Idris, I., Aji, A. F., & Wijaya, D. (2025). *IndoToxic2024: A Demographically-Enriched Dataset of Hate Speech and Toxicity Types for Indonesian Language*.

Vinay, R., Spitale, G., Biller-Andorno, N., & Germani, F. (2025). Emotional prompting amplifies disinformation generation in AI large language models. *Frontiers in Artificial Intelligence*, *Volume 8*-. https://doi.org/10.3389/frai.2025.1543603

Wijanarko, M. I., Susanto, L., Pratama, P. A., Idris, I., Hong, T., & Wijaya, D. (2024). Monitoring Hate Speech in Indonesia: An NLP-based Classification of Social Media Texts. *EMNLP 2024 - 2024 Conference on Empirical Methods in Natural Language Processing, Proceedings of System Demonstrations*, 142–152. https://doi.org/10.18653/v1/2024.emnlp-demo.15

Yonatan, A. Z. (2026). *Menilik Pengguna Media Sosial Indonesia 2017-2026*. https://data.goodstats.id/statistic/menilik-pengguna-media-sosial-indonesia-2017-2026-xUAlp