

Multimodalitas Ujaran Kebencian dalam Toxic Comments di Media Sosial

Agustiansyah Amanda¹⁾

Suhartono²⁾

Agusniar Dian Savitri³⁾

Universitas Negeri Surabaya^{1,2,3}

*) Penulis Korespondensi: Jl. Hortatori No. XX, Kab/Kota, Kode Pos, Negara

Posel: xxxx@xxxxx.xxx

Abstrak: Penelitian ini bertujuan untuk (1) menganalisis peran bahasa verbal dan berbagai unsur multimodal dalam membentuk dan mengonstruksi makna ujaran kebencian (*hate speech*) dalam *toxic comments* di media sosial; dan (2) menganalisis ujaran kebencian dalam *toxic comments* merefleksikan dinamika sosial, relasi antarkelompok, serta pola interaksi yang berkembang dalam ruang komunikasi digital. Penelitian ini menggunakan metode *Systematic Literature Review* (SLR) dengan pedoman PRISMA 2020. Data diperoleh melalui penelusuran artikel jurnal internasional periode 2021–2026 yang terindeks di Google Scholar. Sebanyak 10 artikel yang memenuhi kriteria inklusi dianalisis secara tematik. Hasil penelitian menunjukkan bahwa makna ujaran kebencian dibangun melalui interaksi antara teks, gambar, meme, emoji, simbol visual, dan elemen tipografis yang bekerja secara simultan. Integrasi berbagai moda tersebut memperkuat daya ilokusi, ekspresi emosi, dan pesan ideologis yang terkandung dalam ujaran kebencian. Selain itu, ujaran kebencian multimodal ditemukan merefleksikan polarisasi identitas, konflik sosial-politik, relasi antarkelompok, serta perubahan norma komunikasi dalam masyarakat digital. Temuan penelitian menegaskan bahwa *hate speech* merupakan fenomena komunikasi multimodal yang tidak dapat dipahami hanya melalui analisis linguistik. Kajian ini berkontribusi pada pengembangan perspektif multimodalitas dalam studi komunikasi digital serta memberikan dasar bagi penguatan literasi digital dan pengembangan sistem moderasi konten yang lebih kontekstual.

Kata Kunci: Multimodalitas; *hate speech*; *toxic comments*; media sosial.

Multimodal Hate Speech in Toxic Comments on Social Media

Abstract: This study aimed to (1) analyze the roles of verbal language and various multimodal elements in shaping and constructing the meaning of hate speech in toxic comments on social media, and (2) examine how hate speech in toxic comments reflected social dynamics, intergroup relations, and interaction patterns within digital communication spaces. This study employed a Systematic Literature Review (SLR) method following the PRISMA 2020 guidelines. Data were collected from international journal articles published between 2021 and 2026 and indexed in Google Scholar. A total of ten articles that met the inclusion criteria were analyzed thematically. The findings revealed that the meaning of hate speech was constructed through the interaction of text, images, memes, emojis, visual symbols, and typographic elements operating simultaneously. The integration of these modes strengthened the illocutionary force, emotional expression, and ideological messages embedded in hate speech. Furthermore, multimodal hate speech was found to reflect identity polarization, socio-political conflicts, intergroup relations, and shifts in communication norms within digital society. The findings demonstrated that hate speech is a multimodal communication phenomenon that cannot be fully understood through linguistic analysis alone. This study contributes to the advancement of multimodality perspectives in digital communication research and provides a foundation for strengthening digital literacy and developing more context-sensitive content moderation systems.

Keywords: multimodality, hate speech, toxic comments, social media.

Proses artikel: Dikirim: 21-01-2026; Direvisi: 10-07-2026; Diterima: 11-07-2026; Diterbitkan: 13-07-2026

Gaya sitasi (MLA edisi ke-7): Amanda, Agustiansyah, Suhartono, and Agusniar Dian Savitri. "Multimodalitas Ujaran Kebencian dalam Toxic Comments di Media Sosial." *Hortatori: Jurnal Pendidikan Bahasa dan Sastra Indonesia* 10.01 (2026): 12–21. Print/Online. **Pemegang Hak Cipta:** Penulis. **Publikasi Utama:** Hortatori: Jurnal Pendidikan Bahasa dan Sastra Indonesia (2026).



Proses ini berada di bawah lisensi *Creative Commons Attribution-ShareAlike 4.0 International License*.

Pendahuluan

Perkembangan teknologi digital membawa perubahan mendasar dalam pola komunikasi masyarakat modern. Kehadiran media sosial telah menciptakan ruang interaksi baru yang memungkinkan individu untuk berbagi informasi, menyampaikan opini, serta berpartisipasi dalam berbagai diskusi publik secara cepat dan tanpa batas geografis. Platform media sosial, seperti Facebook, X (Twitter), dan Instagram, tidak hanya berfungsi sebagai sarana pertukaran informasi, namun sebagai ruang pembentukan makna, identitas, dan relasi sosial dalam kehidupan sehari-hari (Boulianne et al.). Peningkatan intensitas penggunaan media sosial menjadikan bentuk komunikasi yang terjadi di dalamnya juga semakin kompleks. Pengguna tidak lagi mengandalkan teks sebagai satu-satunya sarana penyampaian pesan, melainkan memanfaatkan berbagai unsur lain seperti emoji, meme, GIF, gambar, maupun simbol visual yang dapat memperkuat, memperjelas, bahkan mengubah makna suatu pesan (Reena Nagpal et al.). Dapat disimpulkan, komunikasi digital pada era media sosial berlangsung secara multimodal, yaitu melalui perpaduan berbagai sumber semiotik yang bekerja secara bersamaan dalam proses penyampaian dan pemaknaan pesan (Castaldi). Pemahaman terhadap interaksi di media sosial tidak lagi cukup dilakukan melalui analisis teks semata, melainkan perlu mempertimbangkan keterlibatan berbagai elemen multimodal yang membentuk pesan secara utuh. Kompleksitas komunikasi multimodal tidak hanya memperkaya cara pengguna berinteraksi, namun membuka peluang munculnya berbagai bentuk komunikasi yang bersifat negatif di ruang digital.

Kompleksitas komunikasi multimodal di media sosial memberikan peluang bagi pengguna untuk mengekspresikan gagasan dan emosi secara lebih beragam. Dampaknya, media sosial menjadi ruang yang rentan terhadap munculnya berbagai bentuk komunikasi negatif. Fenomena yang semakin sering ditemukan adalah toxic comments, yaitu komentar yang mengandung unsur penghinaan, serangan verbal, pelecehan, maupun bentuk komunikasi agresif lainnya. Ujaran kebencian (*hate speech*) dapat dipahami sebagai salah satu bentuk *toxic comments* yang secara khusus menargetkan individu atau kelompok berdasarkan identitas tertentu, seperti etnis, agama, ras, gender, maupun karakteristik sosial lainnya (Unlu et al.). Kemudahan akses dan tingginya tingkat konektivitas antarpengguna menyebabkan penyebaran ujaran kebencian berlangsung dengan cepat serta menjangkau audiens yang lebih luas (Kentmen-Cin, 2025). Akibatnya, media sosial tidak hanya menjadi sarana komunikasi dan pertukaran informasi, tetapi juga berpotensi menjadi ruang reproduksi berbagai bentuk intoleransi, diskriminasi, dan konflik sosial yang berkembang dalam masyarakat. Kondisi tersebut menunjukkan bahwa perkembangan teknologi komunikasi digital tidak selalu menghasilkan interaksi yang konstruktif, tetapi juga dapat memperluas penyebaran perilaku komunikasi yang bersifat agresif dan diskriminatif.

Hate speech dalam *toxic comments* menjadi perhatian penting karena dampaknya tidak hanya dirasakan oleh individu yang menjadi sasaran, tetapi juga memengaruhi kehidupan sosial secara lebih luas. Hasil penelitian menunjukkan paparan ujaran kebencian dapat menimbulkan tekanan psikologis, perasaan tidak aman, penurunan kepercayaan diri, hingga gangguan kesehatan mental pada korban (Shmulewitz et al.). Hasil penelitian juga menunjukkan ujaran kebencian meningkatnya polarisasi sosial, menguatnya stereotip negatif antarkelompok, serta menurunnya kualitas dialog publik yang sehat dan konstruktif (Liu et al.). Komunikasi agresif jika terus muncul dan diterima sebagai bagian yang lazim dari interaksi digital, kondisi tersebut berpotensi membentuk normalisasi perilaku tidak beradab (*uncivil behavior*) dalam ruang publik digital (Bührer et al.). Dapat disimpulkan, *hate speech* tidak dapat dipandang semata-mata sebagai persoalan bahasa atau perilaku individu, melainkan sebagai fenomena sosial yang berkaitan erat dengan relasi kekuasaan, identitas sosial, serta dinamika komunikasi dalam masyarakat digital. Mempertimbangkan besarnya dampak tersebut, kajian mengenai *hate speech* perlu dilakukan secara lebih komprehensif dengan mempertimbangkan konteks sosial dan bentuk komunikasi yang berkembang di media digital.

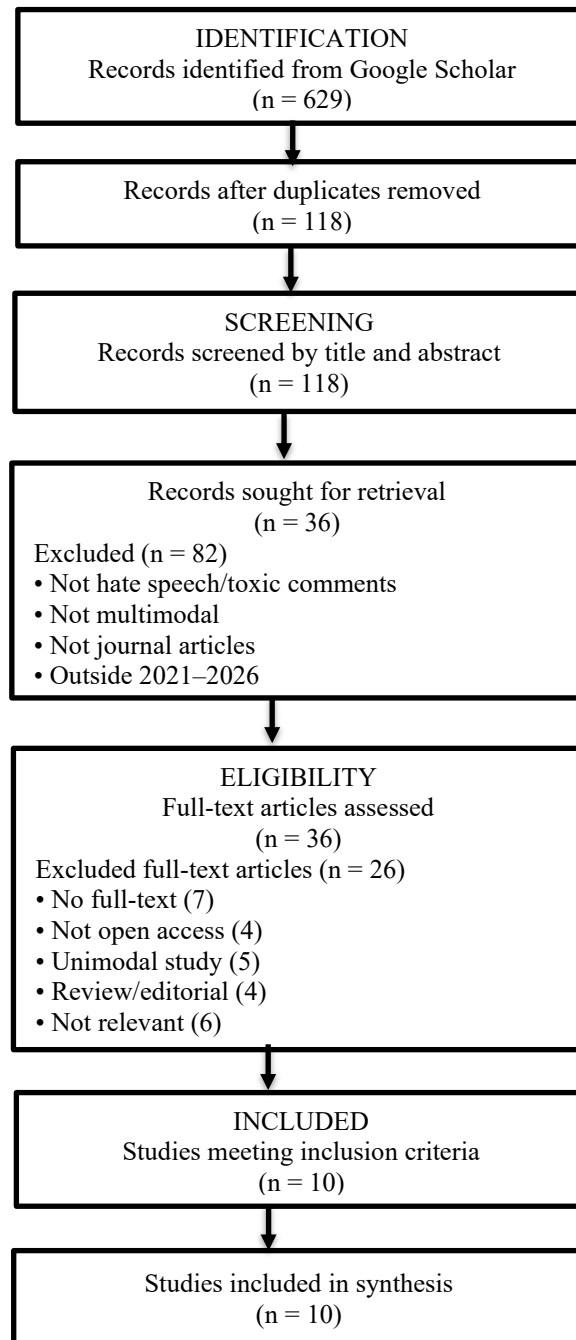
Penelitian *hate speech* di media sosial selama lima tahun terakhir masih didominasi oleh pendekatan linguistik dan komputasional yang berfokus pada deteksi otomatis serta analisis unsur-unsur tekstual (Unlu et al.; Fillies et al.). Hasil penelitian tersebut efektif dalam mengidentifikasi pola ujaran kebencian, namun pendekatan tersebut belum mampu menjelaskan bagaimana makna kebencian dikonstruksi dalam komunikasi digital yang bersifat multimodal. Pengguna media sosial sering mengombinasikan teks dengan emoji, meme, GIF, dan simbol visual untuk memperkuat makna pesan (Reena Nagpal et al.). Perspektif multimodalitas, makna dibangun melalui interaksi berbagai elemen komunikasi, baik verbal maupun nonverbal, yang saling melengkapi dalam proses pemaknaan pesan (Wang). Integrasi antara pendekatan multimodal dan analisis dinamika sosial dalam penelitian *hate speech* masih belum banyak dilakukan, sehingga menyisakan kesenjangan pengetahuan terkait bagaimana makna kebencian dikonstruksi dan direpresentasikan dalam lingkungan digital. Penelitian ini mengkaji *hate speech* yang muncul dalam *toxic comments* melalui perspektif multimodalitas dan komunikasi digital.

Penelitian ini penting dikarenakan peningkatan penggunaan media sosial sebagai ruang komunikasi utama masyarakat dan semakin kompleksnya bentuk penyebaran ujaran kebencian di lingkungan digital (Boulianne et al.). Pemahaman mengenai konstruksi multimodal *hate speech* diperlukan tidak hanya untuk memperkaya kajian akademik di bidang komunikasi digital, tetapi juga untuk mendukung pengembangan strategi literasi digital, kebijakan moderasi konten, dan upaya pencegahan penyebaran ujaran kebencian yang lebih efektif. Penelitian ini juga berkontribusi teoretis melalui pengintegrasian perspektif multimodalitas dan komunikasi digital, penelitian ini diharapkan dapat memberikan kontribusi praktis bagi peneliti, pengelola platform media sosial, maupun pembuat kebijakan dalam memahami dan merespons berbagai bentuk komunikasi agresif yang berkembang di ruang digital kontemporer. Penelitian ini menawarkan perspektif yang lebih komprehensif dalam menjelaskan *hate speech* sebagai praktik komunikasi multimodal yang tidak hanya merepresentasikan agresi linguistik, tetapi juga mencerminkan berbagai dinamika sosial yang berkembang dalam masyarakat digital saat ini.

Tujuan penelitian ini untuk mengidentifikasi dan menganalisis peran bahasa verbal serta berbagai unsur multimodal dalam membentuk dan mengonstruksi makna ujaran kebencian (*hate speech*) yang muncul dalam *toxic comments* di media sosial. Penelitian ini juga bertujuan untuk mengkaji bagaimana ujaran kebencian merefleksikan dinamika sosial, relasi antarkelompok, serta pola interaksi yang berkembang dalam ruang komunikasi digital. Melalui pendekatan *Systematic Literature Review* (SLR), penelitian ini berupaya memberikan pemahaman yang lebih komprehensif mengenai konstruksi makna ujaran kebencian sebagai praktik komunikasi multimodal sekaligus sebagai manifestasi berbagai fenomena sosial yang terjadi di lingkungan digital. Mempertimbangkan latarbelakang dan tujuan penelitian di tersebut, rumusan masalah penelitian sebagai berikut: (1) bagaimana peran bahasa verbal dan berbagai unsur multimodal dalam membentuk dan mengonstruksi makna ujaran kebencian (*hate speech*) dalam *toxic comments* di media sosial; dan (2) bagaimana ujaran kebencian dalam *toxic comments* merefleksikan dinamika sosial, relasi antarkelompok, serta pola interaksi yang berkembang dalam ruang komunikasi digital.

Metode

Penelitian ini menggunakan metode *Systematic Literature Review* (SLR) dengan mengacu pada pedoman PRISMA 2020 (*Preferred Reporting Items for Systematic Reviews and Meta-Analyses*) untuk mengidentifikasi, mengevaluasi, dan menyintesis penelitian mengenai deteksi *hate speech* dalam *toxic comments* berbasis multimodal di media sosial. Kajian ini dilatarbelakangi oleh meningkatnya penyebaran ujaran kebencian di platform digital serta berkembangnya pemanfaatan berbagai modalitas data, seperti teks, gambar, audio, dan video, dalam meningkatkan akurasi deteksi konten berbahaya. Penelusuran literatur dilakukan melalui Google Scholar menggunakan kata kunci yang relevan, dengan fokus pada artikel jurnal internasional yang dipublikasikan pada periode 2021–2026. Proses seleksi artikel dilakukan melalui tahapan *identification*, *screening*, *eligibility*, dan *included* sesuai alur PRISMA 2020:

Gambar 1 Diagram Alur Seleksi Artikel untuk *Systematic Literature Review* (SLR)

Hasil dan Diskusi

Penelitian ini bertujuan untuk menganalisis bagaimana bahasa verbal dan unsur multimodal dalam membentuk makna ujaran kebencian dan menganalisis bagaimana ujaran kebencian sebagai cerminan dinamika sosial dan pola interaksi digital. Dari hasil studi literatur, ditemukan 10 artikel yang sejalan dengan tema tersebut.

Tabel 1 Hasil Review Artikel

Judul Jurnal	Tahun	Author	Metode	Kontribusi terhadap Tema SLR
An inter-modal attention-based deep learning framework	2024	Eniafe Festus Ayetiran,	Inter-modal attention mechanism,	Ujaran kebencian lebih efektif diidentifikasi melalui integrasi teks dan gambar dibandingkan analisis

Judul Jurnal	Tahun	Author	Metode	Kontribusi terhadap Tema SLR
using unified modality for multimodal fake news, hate speech and offensive language detection		Özlem Özgöbek	BiLSTM-CNN, unified modality (text, image, image-text fusion)	unimodal. Temuan ini mengindikasikan bahwa unsur visual berperan dalam memperkuat interpretasi pesan ofensif dan membentuk makna hate speech secara multimodal.
Decoding Digital Discourse Through Multimodal Text and Image Machine Learning Models to Classify Sentiment and Detect Hate Speech in Race- and LGBTQIA+ Community-Related Posts	2025	Thu T Nguyen et al.	Multimodal models (CLIP, VisualBERT, intermediate fusion), BERT for text, VGG16 for images, SMOTE, LDA topic modeling	Kombinasi representasi teks dan visual mampu menangkap konteks ujaran kebencian yang berkaitan dengan identitas sosial, seperti ras dan LGBTQIA+. Temuan ini memperlihatkan bahwa <i>hate speech</i> tidak hanya dibangun melalui bahasa verbal, tetapi juga melalui representasi visual yang merefleksikan dinamika sosial dan relasi antarkelompok.
CrisisHateMM: Multimodal Analysis of Directed and Undirected Hate Speech in Text-Embedded Images From Russia-Ukraine Conflict	2023	Aashish Bhandari et al.	Dataset multimodal dengan 4,700 text-embedded images, benchmark menggunakan unimodal dan multimodal algorithms	Konstruksi makna hate speech terbentuk melalui interaksi antara teks dan gambar dalam media digital. Studi ini menegaskan pentingnya pendekatan multimodal untuk memahami ujaran kebencian serta menunjukkan bahwa bentuk <i>directed</i> dan <i>undirected hate speech</i> merefleksikan relasi sosial dan pola interaksi yang berbeda di ruang digital.
Recent Advances in Online Hate Speech Moderation: Multimodality and the Role of Large Models	2024	Ming Shan Hee et al.	Survey literatur sistematis, analisis LLMs dan LMMs untuk moderasi hate speech	Perkembangan deteksi dan moderasi hate speech semakin mengarah pada pendekatan multimodal yang mengintegrasikan teks, gambar, dan konteks komunikasi digital. Hasil penelitian ini menegaskan bahwa pemanfaatan Large Language Models (LLMs) dan Large Multimodal Models (LMMs) memungkinkan pemahaman makna ujaran kebencian yang lebih komprehensif dibandingkan pendekatan berbasis teks saja. Selain itu, penelitian ini mengidentifikasi pentingnya konteks sosial, budaya, dan bahasa dalam interpretasi hate speech, sehingga mendukung pandangan bahwa ujaran kebencian merupakan fenomena komunikasi multimodal yang merefleksikan dinamika sosial di ruang digital.
A comprehensive framework for multi-modal hate speech detection in social	2025	R. Prabhu, V. Seethalakshmi	Hybrid CNN-RNN architecture (LSTM), attention-based fusion	Artikel ini mengidentifikasi bahwa pendekatan multimodal berbasis fusi teks dan gambar menghasilkan performa deteksi yang lebih baik dibandingkan pendekatan unimodal. Temuan tersebut menunjukkan

Judul Jurnal	Tahun	Author	Metode	Kontribusi terhadap Tema SLR
media using deep learning			mechanism, ensemble voting	bahwa ujaran kebencian di media sosial sering dikonstruksi melalui kombinasi elemen visual dan verbal yang membentuk makna secara bersamaan.
Multimodal Sentiment Analysis in Indonesian: A Comparative Study of Deep Learning Models for Hate Speech Detection on social media	2025	Mas'ud Muhammadih et al.	Perbandingan CNN, LSTM, dan Transformer-based models pada data teks dan gambar dari media sosial Indonesia	Model multimodal berbasis Transformer lebih efektif dibandingkan model berbasis teks saja dalam mendeteksi ujaran kebencian. Temuan tersebut menegaskan bahwa makna ujaran kebencian di media sosial dibentuk melalui interaksi antara elemen verbal dan visual, sehingga pendekatan multimodal diperlukan untuk menangkap makna secara lebih komprehensif.
Deciphering Implicit Hate: Evaluating Automated Detection Algorithms for Multimodal Hate (Findings of ACL)	2021	Austin Botelho, Bertie Vidgen, Scott A. Hale	Eksperimen dengan context-aware transformer dan model multimodal; dataset 5.000 tweet multimodal	Studi ini menunjukkan bahwa ujaran kebencian implisit dikonstruksi melalui kombinasi teks, gambar, dan konteks. Temuan tersebut menegaskan bahwa makna kebencian sering muncul dari interaksi antar-modalitas, sehingga pendekatan multimodal lebih mampu mengidentifikasi ujaran kebencian yang tidak terdeteksi melalui analisis teks saja.
A Hybrid Deep Learning Approach for Multi-Class Cyberbullying Classification Using Multi-Modal Social Media Data (Applied Sciences)	2024	Israt Tabassum & Vimala Nunavath	Hybrid DL (BERT/DistilBERT /RoBERTa + CNN/ResNet/ViT) dengan late fusion	Penelitian ini mengidentifikasi bahwa makna konten bermuatan kebencian atau perundungan tidak hanya tercermin dalam teks, tetapi juga dibentuk melalui representasi visual yang menyertainya. Temuan ini mendukung pola bahwa interaksi antar-modalitas menjadi faktor penting dalam pembentukan dan interpretasi makna ujaran kebencian di media sosial.
A Multi-Modal Dataset for Hate Speech Detection on social media: Case-study of Russia-Ukraine Conflict (CASE Workshop)	2022	Surendrabikram Thapa et al.	Konstruksi dataset 5.680 pasangan teks-gambar (Twitter); baseline CNN, LSTM, dan multimodal fusion	Studi ini menunjukkan bahwa ujaran kebencian dalam konteks konflik politik tidak hanya diekspresikan melalui teks, tetapi juga melalui kombinasi teks dan gambar yang saling melengkapi. Temuan tersebut mengindikasikan bahwa elemen visual berperan dalam memberikan konteks, memperkuat pesan, dan membentuk interpretasi makna kebencian yang tidak selalu dapat dipahami melalui teks saja.
Deciphering Implicit Hate: Evaluating	2021	Austin Botelho,	Eksperimen dengan context-aware	Penelitian ini menunjukkan bahwa makna ujaran kebencian implisit

Judul Jurnal	Tahun	Author	Metode	Kontribusi terhadap Tema SLR
Automated Detection Algorithms for Multimodal Hate (Findings of ACL)		Bertie Vidgen, Scott A. Hale	transformer dan model multimodal; dataset 5.000 tweet multimodal	dikonstruksi melalui interaksi antara teks, gambar, dan konteks komunikasi. Temuan tersebut menegaskan bahwa multimodalitas membantu mengungkap makna kebencian yang tersembunyi dan tidak dapat dipahami secara utuh melalui analisis teks saja.

Tabel 1 menunjukkan bahwa ujaran kebencian di media sosial merupakan fenomena komunikasi yang bersifat multimodal dan berkaitan erat dengan dinamika sosial di ruang digital. Makna hate speech tidak hanya dibentuk melalui bahasa verbal, tetapi juga melalui interaksi dengan unsur visual seperti gambar, meme, emoji, simbol, dan elemen tipografis. Hasil penelitian diatas juga menunjukkan bahwa ujaran kebencian merefleksikan polarisasi identitas, konflik sosial-politik, serta perubahan pola interaksi dalam masyarakat digital. Berdasarkan kecenderungan temuan tersebut, hasil kajian ini disajikan dalam dua tema utama, yaitu peran bahasa verbal dan unsur multimodal dalam konstruksi makna ujaran kebencian serta representasi dinamika sosial dan pola interaksi digital yang tercermin dalam toxic comments di media sosial.

Peran Bahasa Verbal dan Unsur Multimodal dalam Konstruksi Makna Ujaran Kebencian pada Toxic Comments

Hasil sintesis terhadap sepuluh artikel menunjukkan bahwa ujaran kebencian dalam media sosial tidak lagi dibangun hanya melalui bahasa verbal, melainkan melalui integrasi berbagai moda semiotik yang bekerja secara simultan. Teks, gambar, meme, emoji, simbol visual, warna, dan elemen tipografis berfungsi sebagai sumber makna yang saling melengkapi dalam menyampaikan pesan kebencian. Makna *hate speech* dalam *toxic comments* bersifat multimodal karena interpretasinya bergantung pada hubungan antarmoda, bukan pada unsur linguistik secara terpisah. Perspektif multimodalitas menjelaskan bahwa makna komunikasi digital dibentuk melalui interaksi berbagai sumber semiotik yang memiliki kontribusi berbeda tetapi saling terhubung dalam proses pemaknaan (Castaldi).

Hasil penelitian diatas memperlihatkan bahwa kombinasi teks dan visual menghasilkan pemahaman yang lebih akurat terhadap ujaran kebencian dibandingkan pendekatan berbasis teks saja. Model deteksi multimodal yang menggabungkan teks dan gambar memiliki kemampuan lebih tinggi dalam mengidentifikasi konten ofensif dibandingkan model unimodal (Ayetiran and Özgöbek). Temuan tersebut didukung hasil penelitian Prabhu and Seethalakshmi [NO_PRINTED_FORM] menjelaskan mekanisme fusi multimodal memungkinkan identifikasi pesan kebencian yang tersembunyi di balik hubungan antara narasi verbal dan representasi visual. Hasil tersebut mengindikasikan bahwa unsur visual tidak hanya berfungsi sebagai pelengkap informasi, tetapi turut membangun makna ideologis dan emosional dari suatu ujaran kebencian.

Representasi visual memainkan peran penting dalam ujaran kebencian yang berkaitan dengan identitas sosial, khususnya isu ras dan komunitas LGBTQIA+ (Nguyen et al.). Gambar sering digunakan untuk memperkuat stereotip, mempertegas kategorisasi sosial, serta menampilkan simbol-simbol yang secara implisit mengandung pesan diskriminatif. Bahasa verbal yang secara literal tampak netral dapat berubah menjadi ujaran kebencian ketika dikombinasikan dengan visual tertentu. Fenomena ini menunjukkan bahwa makna ujaran kebencian sering kali muncul melalui relasi antarmodalitas yang hanya dapat dipahami apabila konteks visual dan verbal dianalisis secara bersamaan.

Temuan lain yang konsisten ditemukan dalam artikel-artikel yang direview adalah keberadaan ujaran kebencian implisit. Hasil penelitian menunjukkan bahwa banyak bentuk hate speech tidak disampaikan secara eksplisit melalui kata-kata kasar atau penghinaan langsung, melainkan melalui kombinasi teks, gambar, dan konteks sosial yang membentuk makna diskriminatif secara tidak langsung (Boulianne et al.). Bentuk komunikasi semacam ini sering memanfaatkan humor, ironi, sarkasme, maupun meme sebagai strategi penyampaian pesan. Akibatnya, analisis berbasis kata kunci sering gagal mengidentifikasi kebencian yang tersembunyi karena makna sebenarnya muncul dari hubungan antara berbagai unsur multimodal.

Perspektif pragmatik, integrasi antara teks dan visual berkontribusi terhadap peningkatan daya ilokusi ujaran. Bahasa verbal menyampaikan proposisi utama, sementara unsur visual memperkuat efek emosional dan intensi komunikatif yang ingin dicapai. Gambar, emoji, atau simbol visual berfungsi sebagai penguat tindak tutur yang meningkatkan kekuatan ekspresif suatu pesan. Temuan tersebut didukung dari hasil penelitian Wang (2024), komunikasi digital tidak hanya mentransmisikan informasi, tetapi juga membangun sikap, emosi, dan posisi ideologis tertentu melalui kombinasi berbagai sumber semiotik. Selain gambar dan meme, emoji juga ditemukan memiliki kontribusi penting dalam pembentukan makna kebencian. Emoji dapat berfungsi sebagai penanda ironi, sindiran, ejekan, maupun ekspresi agresi yang memperkuat interpretasi pesan verbal. Emoji digunakan untuk menyamarkan ujaran kebencian sehingga tampak sebagai candaan atau bentuk komunikasi informal. Emoji ketika dipahami dalam konteks sosial tertentu, penggunaan emoji juga mempertegas pesan diskriminatif yang disampaikan.

Dapat disimpulkan bahwa makna *hate speech* dalam media sosial bersifat kontekstual dan tidak dapat direduksi hanya pada struktur bahasa verbal. Bahasa verbal dan unsur multimodal membentuk sistem komunikasi yang terintegrasi dalam konstruksi ujaran kebencian. Makna *hate speech* muncul dari hubungan dinamis antara teks, visual, simbol, dan konteks sosial yang melingkupinya. Pemahaman terhadap *toxic comments* memerlukan pendekatan multimodal yang mampu mengungkap hubungan antar elemen komunikasi secara menyeluruh, bukan sekadar menganalisis unsur leksikal atau sintaksis secara terpisah.

Ujaran Kebencian sebagai Refleksi Dinamika Sosial dan Pola Interaksi dalam Ruang Komunikasi Digital

Hasil sintesis literatur menunjukkan bahwa ujaran kebencian tidak hanya merupakan fenomena linguistik, tetapi juga merepresentasikan berbagai dinamika sosial yang berkembang dalam masyarakat digital. *Toxic comments* muncul sebagai refleksi konflik identitas, polarisasi kelompok, ketimpangan kekuasaan, dan ketegangan sosial yang telah ada sebelumnya dalam kehidupan masyarakat. Media sosial menyediakan ruang bagi individu untuk mengekspresikan sikap, emosi, dan ideologi secara terbuka sehingga berbagai bentuk kebencian dapat diproduksi, disebarluaskan, dan direproduksi secara berkelanjutan (Kentmen Cin). Identitas kelompok merupakan faktor utama yang mendorong munculnya *hate speech* di media sosial. Ujaran kebencian sering diarahkan kepada kelompok yang dianggap berbeda secara rasial, etnis, agama, gender, maupun orientasi seksual (Nguyen et al.). *Hate speech* berfungsi sebagai mekanisme simbolik untuk membangun batas sosial antara kelompok “kami” dan “mereka”. Bahasa verbal dan representasi visual digunakan untuk memperkuat identitas kelompok sendiri sekaligus mendeligitimasi kelompok lain. Proses tersebut memperlihatkan bahwa ujaran kebencian tidak hanya menyampaikan penghinaan, tetapi juga mereproduksi relasi sosial yang bersifat eksklusif dan diskriminatif.

Hasil penelitian mengenai studi konflik Rusia–Ukraina juga menunjukkan bahwa *hate speech* berkaitan erat dengan konteks politik dan konflik sosial. Teks dan gambar digunakan secara bersamaan untuk membangun narasi yang mendukung kelompok tertentu sekaligus menyerang kelompok lawan (Bhandari et al.). Ujaran kebencian dalam konteks konflik tidak hanya berfungsi sebagai ekspresi emosi individu, tetapi juga sebagai instrumen mobilisasi sosial yang memperkuat solidaritas internal kelompok. *Hate speech* dapat dipahami sebagai bagian dari praktik komunikasi politik yang mencerminkan relasi kekuasaan dan pertarungan identitas di ruang digital.

Hasil penelitian menunjukkan media sosial memfasilitasi terbentuknya pola interaksi yang berbeda dari komunikasi tatap muka. Karakteristik komunikasi digital yang asinkron, anonim, dan multilateral memungkinkan individu menyampaikan ujaran agresif dengan risiko sosial yang lebih rendah. Kondisi tersebut mendorong fenomena disinhibisi daring, yaitu kecenderungan individu untuk mengekspresikan perilaku yang lebih ekstrem dibandingkan ketika berinteraksi secara langsung. Dampaknya, komentar bernada kebencian lebih mudah diproduksi dan disebarluaskan tanpa mempertimbangkan dampaknya terhadap pihak lain. Selain faktor sosial media, interpretasi *hate speech* dipengaruhi oleh konteks sosial, budaya, dan bahasa yang melatarbelakanginya. Simbol atau gambar yang tampak netral pada satu kelompok dapat memiliki makna ofensif bagi kelompok lain. Dapat disimpulkan, ujaran kebencian tidak dapat dipahami secara universal tanpa mempertimbangkan latar sosial pengguna dan komunitas digital tempat pesan tersebut beredar. Temuan ini menunjukkan bahwa konstruksi makna *hate speech* selalu berkaitan dengan praktik sosial yang berkembang dalam masyarakat.

Hasil review menunjukkan bahwa media sosial berperan sebagai arena resonansi yang memungkinkan penyebaran simbol dan narasi kebencian secara masif. Algoritma platform cenderung

memperkuat distribusi konten yang memicu keterlibatan pengguna, termasuk konten yang bersifat kontroversial atau provokatif. Situasi ini menyebabkan ujaran kebencian lebih mudah memperoleh perhatian, dukungan, dan reproduksi oleh pengguna lain. Seiring waktu, praktik tersebut berpotensi menciptakan normalisasi terhadap komunikasi agresif sehingga perilaku yang sebelumnya dianggap tidak pantas menjadi diterima sebagai bagian dari budaya interaksi digital (Bührer et al.). Perspektif komunikasi digital, *hate speech* dalam *toxic comments* dapat dipahami sebagai manifestasi hubungan timbal balik antara struktur teknologi dan dinamika sosial. Teknologi menyediakan medium penyebaran pesan, sedangkan faktor sosial seperti identitas kelompok, konflik politik, stereotip, dan emosi kolektif menentukan isi serta arah komunikasi yang berkembang. Ujaran kebencian bukan sekadar masalah individu atau bahasa, melainkan fenomena sosial yang terbentuk melalui interaksi kompleks antara pengguna, platform digital, dan konteks masyarakat yang lebih luas.

Hasil review diatas dapat disimpulkan *hate speech* dalam *toxic comments* merefleksikan berbagai dinamika sosial kontemporer, termasuk polarisasi identitas, konflik politik, fragmentasi sosial, dan perubahan norma komunikasi publik. Media sosial tidak hanya menjadi saluran penyebaran ujaran kebencian, tetapi juga menjadi ruang tempat makna kebencian dinegosiasikan, diperkuat, dan dinormalisasi melalui partisipasi kolektif pengguna. Temuan ini mempertegas bahwa analisis ujaran kebencian perlu dilakukan secara multidisipliner dengan mengintegrasikan perspektif multimodalitas, pragmatik, komunikasi digital, dan kajian sosial agar mampu menjelaskan fenomena tersebut secara lebih komprehensif.

Simpulan

Hasil penelitian ini menunjukkan bahwa ujaran kebencian dalam *toxic comments* di media sosial merupakan praktik komunikasi multimodal yang maknanya dibentuk melalui interaksi bahasa verbal dengan unsur visual, seperti gambar, meme, emoji, dan simbol. *Hate speech* tidak hanya berfungsi sebagai ekspresi kebencian, tetapi juga merefleksikan polarisasi identitas, relasi antarkelompok, serta dinamika sosial yang berkembang di ruang digital. Kebaruan penelitian yang diperoleh yaitu sintesis yang menghubungkan konstruksi multimodal ujaran kebencian dengan konteks sosial dan pola interaksi digital secara terpadu. Hasil penelitian ini berkontribusi pada pengembangan teori multimodalitas dalam studi komunikasi digital serta dapat menjadi dasar bagi penguatan literasi digital dan pengembangan sistem moderasi konten yang lebih kontekstual. Penelitian selanjutnya diharapkan dapat mengkaji data empiris lintas platform dan lintas budaya untuk memahami fenomena ujaran kebencian multimodal secara lebih mendalam.

Ucapan Terima Kasih

Penulis mengucapkan terima kasih kepada Program Studi Magister Pendidikan Bahasa Indonesia, Universitas Negeri Surabaya, atas dukungan akademik dan fasilitas yang diberikan selama proses penyusunan penelitian ini. Penulis juga menyampaikan apresiasi kepada seluruh pihak yang telah memberikan masukan dan dukungan dalam proses penelusuran, seleksi, dan analisis literatur sehingga penelitian ini dapat diselesaikan dengan baik.

Daftar Rujukan

- Ayetiran, E. F., and Ö. Özgöbek. "An Inter-Modal Attention-Based Deep Learning Framework Using Unified Modality for Multimodal Fake News, Hate Speech and Offensive Language Detection." *Information Systems*, vol. 123, 2024, doi:10.1016/j.is.2024.102378.
- Bhandari, A., et al. "CrisisHateMM: Multimodal Analysis of Directed and Undirected Hate Speech in Text-Embedded Images from Russia-Ukraine Conflict." *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE, 2023, pp. 1994–2003. doi:10.1109/CVPRW59228.2023.00193.
- Boulianne, S., C. P. Hoffmann, and M. Bossetta. "Social Media Platforms for Politics: A Comparison of Facebook, Instagram, Twitter, YouTube, Reddit, Snapchat, and WhatsApp." *New Media & Society*, vol. 27, no. 11, 2025, pp. 6006–6037. doi:10.1177/14614448241262415.
- Bührer, S., K. Koban, and J. Matthes. "Misreading Uncivil and Intolerant Anti-Immigration Content in Digital Spheres: The Role of Recent Online Perpetration." *New Media & Society*, 2025. doi:10.1177/14614448251401865.

- Castaldi, J. "Refining Concepts for Empirical Multimodal Research: Defining Semiotic Modes and Semiotic Resources." *Frontiers in Communication*, vol. 9, 2024. doi:10.3389/fcomm.2024.1336325.
- Fillies, J., et al. "A Novel German TikTok Hate Speech Dataset: Far-Right Comments against Politicians, Women, and Others." *Discover Data*, vol. 3, no. 1, 2025, p. 4. doi:10.1007/s44248-025-00020-y.
- Kentmen Cin, C. "Hate Speech on Social Media: A Systemic Narrative Review of Political Science Contributions." *Social Sciences*, vol. 14, no. 10, 2025, p. 610. doi:10.3390/socsci14100610.
- Liu, Z., C. Luo, and J. Lu. "Hate Speech in the Internet Context: Unpacking the Roles of Internet Penetration, Online Legal Regulation, and Online Opinion Polarization from a Transnational Perspective." *Information Development*, vol. 40, no. 4, 2024, pp. 533–549. doi:10.1177/026666669221148487.
- Nguyen, T. T., et al. "Decoding Digital Discourse Through Multimodal Text and Image Machine Learning Models to Classify Sentiment and Detect Hate Speech in Race- and Lesbian, Gay, Bisexual, Transgender, Queer, Intersex, and Asexual Community–Related Posts on Social Media: Quantitative Study." *Journal of Medical Internet Research*, vol. 27, 2025, e72822. doi:10.2196/72822.
- Prabhu, R., and V. Seethalakshmi. "A Comprehensive Framework for Multi-Modal Hate Speech Detection in Social Media Using Deep Learning." *Scientific Reports*, vol. 15, no. 1, 2025. doi:10.1038/s41598-025-94069-z.
- Nagpal, Reena, Ishfaq Ahmad Trambo, and Sinoj Antony. "The Postmodern Transformation of Language through Social Media, Memes, Emojis, and GIFs: A Study Among Selected North Indian University Students." *Forum for Linguistic Studies*, vol. 6, no. 5, 2024, pp. 574–587. doi:10.30564/fls.v6i5.7057.
- Shmulewitz, D., et al. "Exposure to Online Hate Speech Is Positively Associated with Post-Traumatic Stress Disorder Symptom Severity." *Scientific Reports*, vol. 15, no. 1, 2025, p. 29869. doi:10.1038/s41598-025-16168-1.
- Unlu, A., et al. "From Prejudice to Marginalization: Tracing the Forms of Online Hate Speech Targeting LGBTQ+ and Muslim Communities." *New Media & Society*, vol. 28, no. 3, 2026, pp. 1338–1369. doi:10.1177/14614448241312900.
- Wang, S. "The Power of Images: How Multimodal Hate Speech Shapes Prejudice and Prosocial Behavioral Intentions." *Social Media + Society*, vol. 10, no. 4, 2024. doi:10.1177/20563051241292990.