



Klasifikasi Ujaran Kebencian Teks Bahasa Gaul Terkait Isu Israel Menggunakan Support Vector Machine

Kamila Insani P^{1*}, Mahardika Abdi Prawira Tanjung²

^{1*,2}Program Studi Sistem Informasi, Fakultas Ilmu Komputer dan Teknologi Informasi, Universitas Muhammadiyah Sumatera Utara, Medan, Indonesia

Email: ^{1*}kamilainsani07@gmail.com, ²mahardikaabdiprawira@umsu.ac.id

Abstrak

Penelitian ini bertujuan mengembangkan sistem klasifikasi ujaran kebencian berbahasa Indonesia pada teks bahasa gaul yang berkaitan dengan isu Israel di platform YouTube. Tingginya potensi polarisasi opini pada kolom komentar terkait isu geopolitik tersebut, serta terbatasnya penelitian yang menyoroti variasi bahasa informal, menjadi latar belakang utama penelitian ini. Pendekatan kuantitatif dengan desain penelitian eksperimen digunakan, berbasis metode Support Vector Machine (SVM). Data dikumpulkan melalui teknik web scraping menggunakan YouTube Data API v3 terhadap 79 video, menghasilkan 15.096 komentar mentah yang kemudian melalui tahap praproses berupa case folding, normalisasi bahasa gaul, tokenizing, stopword removal, dan stemming sehingga menghasilkan 13.314 data bersih. Data dilabeli secara otomatis berbasis leksikon, kemudian diekstraksi fitur menggunakan TF-IDF sebanyak 5.000 fitur sebelum dilatih dengan model SVM berkernel linear. Hasil pengujian terhadap 2.663 data uji menunjukkan performa model yang sangat baik, dengan akurasi 99,10%, presisi 97,28%, recall 93,63%, dan F1-Score 95,42% pada kelas ujaran kebencian. Ujaran kebencian yang ditemukan didominasi kata kasar dan kata bermuatan hasutan, dengan proporsi 10,02% dari keseluruhan data. Model selanjutnya diimplementasikan pada sistem web SIKUK berbasis PHP native untuk mendukung moderasi konten secara cepat dan objektif.

Kata Kunci: Ujaran Kebencian, Bahasa Gaul, Support Vector Machine, TF-IDF, YouTube.

PENDAHULUAN

YouTube sebagai salah satu platform media sosial terbesar telah menjadi ruang interaksi publik yang aktif melalui fitur kolom komentar. Dalam konteks isu global seperti konflik Israel–Palestina, kolom komentar pada video YouTube sering dipenuhi berbagai opini yang bersifat pro dan kontra, yang tidak jarang mengandung ujaran kebencian. Ujaran tersebut dapat muncul dalam bentuk ekspresi negatif, provokasi, maupun serangan terhadap kelompok tertentu, sehingga berpotensi memperburuk polarisasi sosial di ruang digital (Hofmann et al., 2025).

Fenomena ujaran kebencian pada platform YouTube didukung oleh data kuantitatif yang menunjukkan tingginya persentase kemunculan konten negatif dalam kolom komentar. Dari 4.983 komentar YouTube yang dianalisis oleh Hofmann et al. (2025), sekitar 40,4% komentar pada sumber publik dan 31,6% pada sumber privat terindikasi mengandung ujaran kebencian terkait konflik Israel–Palestina. Tingginya angka tersebut menunjukkan bahwa YouTube dapat dikategorikan sebagai platform dengan tingkat interaksi yang tinggi sekaligus rentan terhadap penyebaran ujaran kebencian, khususnya pada isu-isu sensitif.

Volume data yang sangat besar pada media sosial menyebabkan proses identifikasi ujaran kebencian secara manual menjadi kurang efektif. Metode manual membutuhkan waktu yang lama serta rentan terhadap subjektivitas penilai. Oleh karena itu, diperlukan pendekatan otomatis yang mampu melakukan klasifikasi teks secara cepat dan akurat menggunakan teknik komputasi. Dalam bidang text mining dan Natural Language Processing (NLP), berbagai algoritma machine learning telah digunakan untuk mendeteksi ujaran kebencian secara otomatis pada data teks dari media sosial (Khatib Sulaiman et al., 2024).

Salah satu metode yang sering digunakan dalam klasifikasi teks adalah Support Vector Machine (SVM). Metode ini dikenal memiliki kemampuan yang baik dalam menangani data berdimensi tinggi seperti teks serta mampu menghasilkan performa klasifikasi yang stabil. Penelitian yang membandingkan beberapa algoritma machine learning menunjukkan bahwa SVM memiliki performa yang kompetitif dalam klasifikasi hate speech pada data berbahasa Indonesia (Suwarno & Kusnadi, 2021). Selain itu, penggunaan

SVM dengan pembobotan fitur seperti TF-IDF mampu meningkatkan performa klasifikasi teks secara signifikan (Kamalia, 2021).

Meskipun berbagai penelitian mengenai deteksi ujaran kebencian telah dilakukan, sebagian besar masih berfokus pada data teks umum di media sosial tanpa mempertimbangkan variasi bahasa, seperti bahasa gaul atau slang yang sering muncul dalam komunikasi digital. Penelitian sebelumnya juga umumnya menggunakan dataset yang berkaitan dengan isu domestik. Penelitian yang secara khusus membahas ujaran kebencian dalam konteks isu geopolitik internasional seperti konflik Israel masih sangat terbatas (Khatib Sulaiman et al., 2024). Padahal, diskusi mengenai konflik internasional sering memicu reaksi emosional yang kuat sehingga berpotensi menghasilkan ujaran kebencian dalam berbagai bentuk bahasa yang kompleks.

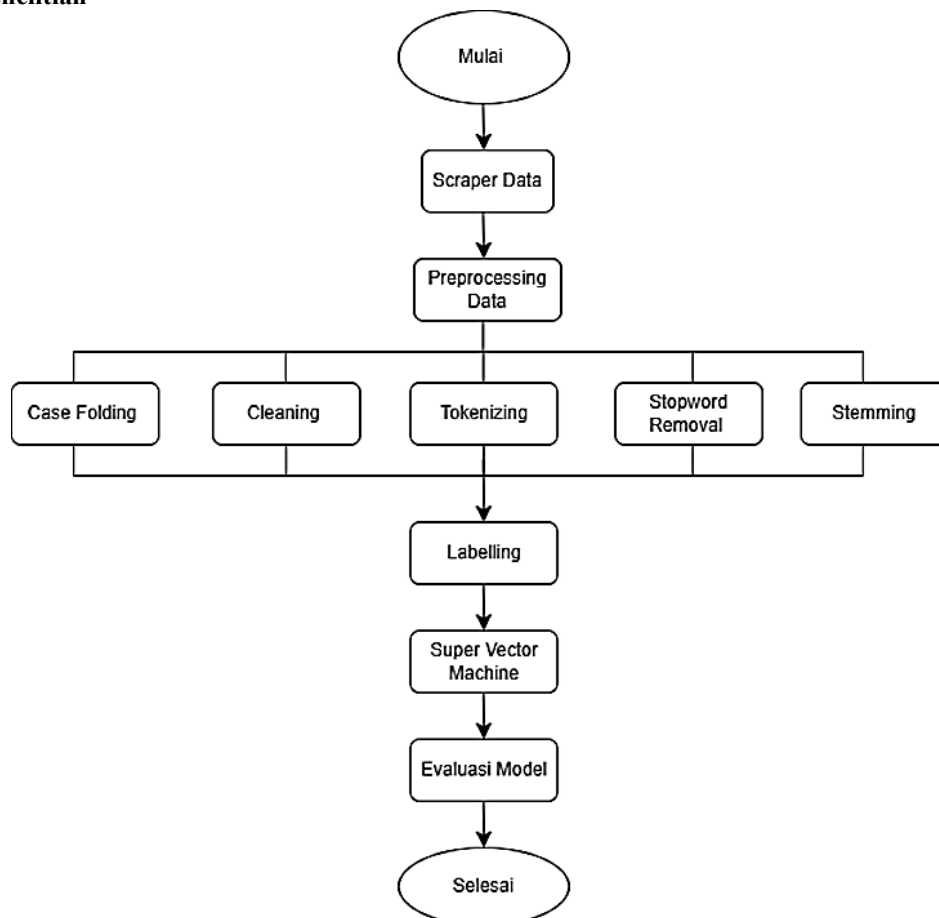
Berdasarkan gap penelitian tersebut, penelitian ini bertujuan: (1) menerapkan metode SVM dalam mengklasifikasikan ujaran kebencian pada teks bahasa gaul berbahasa Indonesia terkait isu Israel; (2) mengukur tingkat performa model SVM berdasarkan nilai accuracy, precision, recall, dan F1-score; serta (3) mengidentifikasi karakteristik ujaran kebencian yang muncul pada teks bahasa gaul dalam diskusi isu Israel di media sosial. Penelitian ini diharapkan dapat memberikan kontribusi dalam pengembangan sistem deteksi ujaran kebencian yang lebih efektif untuk mendukung moderasi konten digital secara lebih cepat dan objektif (Susanto et al., 2025).

METODE

Jenis dan Pendekatan Penelitian

Penelitian ini menggunakan pendekatan kuantitatif dengan jenis penelitian eksperimen. Eksperimen dilakukan dengan membangun model klasifikasi menggunakan algoritma Support Vector Machine (SVM) untuk mengklasifikasikan komentar YouTube menjadi kategori ujaran kebencian (hate speech) dan bukan ujaran kebencian (non-hate speech). Proses eksperimen mencakup pengumpulan data, praproses teks, ekstraksi fitur, penerapan algoritma klasifikasi, dan evaluasi performa model.

Alur Penelitian



Gambar 1. Alur Penelitian

Secara keseluruhan, alur penelitian terdiri dari lima tahap utama: (1) pengumpulan data melalui web scraping YouTube; (2) praproses teks; (3) pelabelan data berbasis leksikon; (4) ekstraksi fitur TF-IDF dan pelatihan model SVM; serta (5) evaluasi model dan implementasi sistem web SIKUK.

Pengumpulan Data

Data dikumpulkan dari platform YouTube menggunakan YouTube Data API v3. Sebanyak 15 kata kunci pencarian digunakan untuk menemukan video yang relevan dengan isu Israel-Palestina, antara lain: "Israel Palestina", "Israel Gaza", "zionis", "boikot produk Israel", "dukung Palestina", "genosida Israel", "Israel vs Hamas", "propaganda Israel", "tentara Israel", "perang Gaza", "serangan Israel", "gencatan senjata Gaza", "anak Palestina", "Netanyahu", dan "berita Israel Palestina". Rentang waktu video dibatasi mulai pertengahan tahun 2024 hingga saat pengumpulan data. Proses pencarian berhasil menemukan 344 video unik, yang selanjutnya dipilih 79 video untuk diambil komentarnya. Jumlah komentar utama (top-level comment) beserta balasan (reply) yang berhasil dikumpulkan sebanyak 15.096 komentar mentah.

Praproses Teks

Data mentah yang diperoleh melalui proses dan praproses teks terdiri dari enam tahapan berurutan sebagai berikut:

1. *leaning dasar* — membuang baris kosong, komentar duplikat berdasarkan comment_id maupun isi teks yang identik, serta komentar yang terlalu pendek (kurang dari tiga kata). Tahap ini mengurangi data dari 15.096 menjadi 13.402 baris.
2. *Case folding dan pembersihan simbol* — mengubah seluruh huruf menjadi huruf kecil, membuang URL, mention (@), hashtag (#), emoji/karakter non-ASCII, mengoreksi elongasi huruf (misalnya "anjirrrr" menjadi "anjirr"), serta membuang simbol dan tanda baca. Tahap ini menghasilkan 13.315 baris.
3. *Normalisasi bahasa gaul* — memetakan kata-kata tidak baku ke bentuk bakunya menggunakan Kamus Alay — Colloquial Indonesian Lexicon (Salsabila et al., 2018) yang memuat 4.331 entri pasangan kata slang dan padanan bakunya.
4. *Tokenizing* — memecah teks menjadi unit-unit kata (token) yang dapat diproses secara individual.
5. *Stopword removal* — menghapus kata-kata umum yang tidak memiliki kontribusi makna signifikan terhadap klasifikasi (seperti "yang", "dan", "di", "ke") menggunakan daftar stopwords Bahasa Indonesia dari pustaka Sastrawi (Rianto et al., 2021).
6. *Stemming* — mengubah kata berimbuhan menjadi bentuk kata dasarnya menggunakan algoritma stemming Bahasa Indonesia dari pustaka Sastrawi, dengan strategi cache per kata unik untuk efisiensi pemrosesan.

Setelah seluruh tahap praproses, diperoleh 13.314 data bersih yang siap dilabeli.

Pelabelan Data Otomatis Berbasis Leksikon

Mengingat jumlah data yang besar (13.314 baris), pelabelan dilakukan secara otomatis menggunakan pendekatan *weak supervision* berbasis leksikon. Setiap komentar dicocokkan terhadap dua kelompok kamus kata: (a) kata kasar/makian dengan bobot 1, mencakup 27 kata seperti "anjing", "goblok", "bangsat", "bego", dan "tolol"; serta (b) kata hasutan/kekerasan dengan bobot 2, mencakup 23 kata seperti "bunuh", "bantai", "benci", "jajah", dan "musuh". Komentar diberi label *hate speech* apabila skor totalnya ≥ 1 , dan *non hate speech* apabila tidak.

Untuk menjaga validitas, diambil sampel acak sebanyak 120 komentar (60 dari tiap kelas) untuk diperiksa secara manual sebagai validasi hasil pelabelan otomatis.

Ekstraksi Fitur TF-IDF

Teks hasil praproses diubah menjadi representasi numerik menggunakan metode Term Frequency-Inverse Document Frequency (TF-IDF). Ukuran kosakata (vocabulary) dibatasi maksimal 5.000 fitur (max_features=5.000, min_df=2, ngram_range=(1,1)) agar model tetap ringkas dan dapat diekspor ke format JSON untuk implementasi di sistem web. Nilai TF-IDF diperoleh dari perkalian nilai TF dan IDF suatu term (Septiani & Isabela, 2022), yang menghasilkan matriks data latih berukuran (10.651, 5.000).

Pelatihan Model SVM

Dataset berlabel dibagi dengan proporsi 80:20 menggunakan *stratified split*, menghasilkan 10.651 data latih dan 2.663 data uji. Model dilatih menggunakan LinearSVC dari scikit-learn dengan parameter C=1,0, class_weight="balanced" (untuk mengatasi ketidakseimbangan kelas 90%:10%), random_state=42, dan max_iter=10.000. Parameter class_weight="balanced" dipilih untuk memberi bobot lebih besar pada kelas minoritas (*hate speech*) selama proses pelatihan (Kowsari et al., 2019).

Implementasi Sistem Web SIKUK

Model yang telah dilatih diekspor ke format JSON (tfidf_vocab.json dan svm_model.json) untuk diintegrasikan ke dalam sistem web SIKUK (Sistem Klasifikasi Ujaran Kebencian) yang dibangun menggunakan PHP native dan MySQLi. Sistem mencakup halaman publik untuk klasifikasi teks secara real-time dan halaman admin untuk pengelolaan dataset, pemantauan statistik, dan riwayat proses klasifikasi.

Evaluasi Model

Evaluasi performa model menggunakan confusion matrix untuk menghitung nilai accuracy, precision, recall, dan F1-score. Confusion matrix terdiri dari empat komponen: True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN) (Powers & Ailab, 2020). Formula masing-masing metrik adalah:

$$\begin{aligned}\text{Accuracy} &= (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN}) \\ \text{Precision} &= \text{TP} / (\text{TP} + \text{FP}) \\ \text{Recall} &= \text{TP} / (\text{TP} + \text{FN}) \\ \text{F1-Score} &= 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})\end{aligned}$$

HASIL DAN PEMBAHASAN

Hasil Pengumpulan dan Praproses Data

Proses pengumpulan data menggunakan YouTube Data API v3 dengan 15 kata kunci berhasil menemukan 344 video unik. Dari 79 video yang dipilih, diperoleh 15.096 komentar mentah. Setelah melewati seluruh tahapan praproses teks, jumlah data bersih yang dihasilkan adalah 13.314 komentar. Rincian perubahan jumlah data pada setiap tahap praproses disajikan pada Tabel 1.

Tabel 1. Ringkasan Tahapan Praproses Data

Tahap Praproses	Jumlah Data
Data mentah (raw)	15.096
Setelah cleaning dasar (dedup + filter)	13.402
Setelah case folding & cleaning simbol	13.315
Setelah stopword removal	13.314
Data bersih final (setelah stemming)	13.314

Sebelum: 15096 -> Sesudah cleaning dasar: 13402 (buang 1694 baris)

Gambar 2. Contoh output perbandingan teks sebelum dan sesudah normalisasi slang

Hasil Pelabelan Data

Pelabelan otomatis berbasis leksikon menghasilkan distribusi label sebagaimana disajikan pada Tabel 2. Sebanyak 1.334 komentar (10,02%) berlabel *hate speech* dan 11.980 komentar (89,98%) berlabel *non-hate speech*. Proporsi ini mencerminkan kondisi imbalanced class yang umum terjadi pada data ujaran kebencian di media sosial, di mana komentar yang terindikasi kebencian merupakan minoritas dari keseluruhan diskusi.

Tabel 2. Distribusi Label Hasil Pelabelan Otomatis

Label	Jumlah Data	Persentase
non_hate_speech	11.980	89,98%
hate_speech	1.334	10,02%
Total	13.314	100%

Validasi sampel acak terhadap 120 komentar (60 per kelas) menunjukkan bahwa hasil pelabelan otomatis berbasis leksikon cukup representatif untuk digunakan sebagai data latih, meski diakui belum sepenuhnya mampu menangkap konteks, sarkasme, atau negasi yang kompleks.

Hasil Ekstraksi Fitur TF-IDF

Proses ekstraksi fitur TF-IDF menghasilkan vocabulary sebanyak 5.000 term dari 10.651 data latih. Matriks representasi fitur data latih berukuran (10.651, 5.000) dan data uji berukuran (2.663, 5.000).

Tabel 3. Hasil Pembagian Data Latih dan Data Uji

Keterangan	Jumlah
Data latih (train)	10.651 (80%)
Data uji (test)	2.663 (20%)
Proporsi non_hate_speech (data latih)	90%
Proporsi hate_speech (data latih)	10%

Hasil Evaluasi Model SVM

Model SVM dengan kernel linear dievaluasi terhadap 2.663 data uji. Hasil evaluasi ditampilkan pada Tabel 4.

Tabel 4. Hasil Evaluasi Model SVM

Metrik	Nilai (Kelas hate speech)
Accuracy (keseluruhan)	99,10%
Precision	97,28%
Recall	93,63%
F1-Score	95,42%

Rincian classification report untuk masing-masing kelas ditampilkan pada Tabel 5.

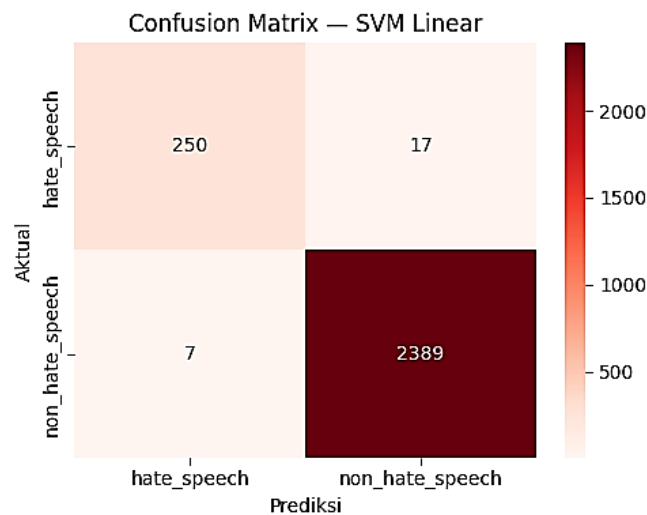
Tabel 5. Classification Report Lengkap Model SVM

Kelas	Precision	Recall	F1-Score	Support
hate_speech	0,97	0,94	0,95	267
non_hate_speech	0,99	1,00	1,00	2.396
Accuracy			0,99	2.663
Macro avg	0,98	0,97	0,97	2.663
Weighted avg	0,99	0,99	0,99	2.663

Confusion matrix hasil pengujian model disajikan pada Tabel 6 dan divisualisasikan sebagai heatmap pada Gambar 3.

Tabel 6. Confusion Matrix Model SVM pada Data Uji

Aktual \ Prediksi	hate_speech	non_hate_speech
hate_speech	250 (TP)	17 (FN)
non_hate_speech	7 (FP)	2.389 (TN)



Gambar 3. Visualisasi Heatmap Confusion Matrix

Pembahasan Performa Model

Nilai precision sebesar 97,28% menunjukkan bahwa dari seluruh komentar yang diprediksi sebagai ujaran kebencian, hanya 7 dari 2.396 komentar non-hate speech yang keliru ditandai sebagai hate speech (false positive). Nilai recall sebesar 93,63% menunjukkan bahwa model mampu mengenali 250 dari 267 komentar yang benar-benar mengandung ujaran kebencian pada data uji, dengan 17 komentar hate speech yang tidak terdeteksi (false negative).

Perlu dicatat bahwa tingginya nilai accuracy (99,10%) juga dipengaruhi oleh ketidakseimbangan proporsi kelas pada dataset (90% non_hate_speech berbanding 10% hate_speech). Oleh karena itu, metrik precision, recall, dan F1-score pada kelas minoritas (hate_speech) menjadi indikator yang lebih representatif untuk menilai kemampuan model (Chicco & Jurman, 2020). Nilai F1-score sebesar 95,42% menunjukkan keseimbangan yang sangat baik antara precision dan recall pada kelas hate speech meskipun data tidak seimbang.

Hasil penelitian ini sejalan dengan penelitian Suwarno & Kusnadi (2021) yang menunjukkan SVM memiliki performa kompetitif dalam klasifikasi hate speech berbahasa Indonesia, serta penelitian Rahman et al. (2021) yang memperoleh akurasi tinggi menggunakan SVM dengan TF-IDF pada data Twitter berbahasa Indonesia. Keunggulan model pada penelitian ini dibandingkan penelitian sebelumnya antara lain terletak pada penambahan tahap normalisasi bahasa gaul yang membuat model lebih adaptif terhadap variasi bahasa informal, serta penggunaan data dari isu geopolitik spesifik yang belum banyak diteliti.

Hasil Sanity Check Model

Sebelum model diekspor ke sistem web, dilakukan pengujian cepat (sanity check) terhadap dua contoh teks untuk memastikan konsistensi prediksi. Hasil sanity check disajikan pada Tabel 7.

Tabel 7. Hasil Sanity Check Model

Teks Uji	Hasil Prediksi	Decision Score
"semoga cepat damai ya kasian banget anak-anak di sana"	non hate speech	+0,9942
"anjing banget nih orang goblok bacot doang"	hate speech	-2,7171

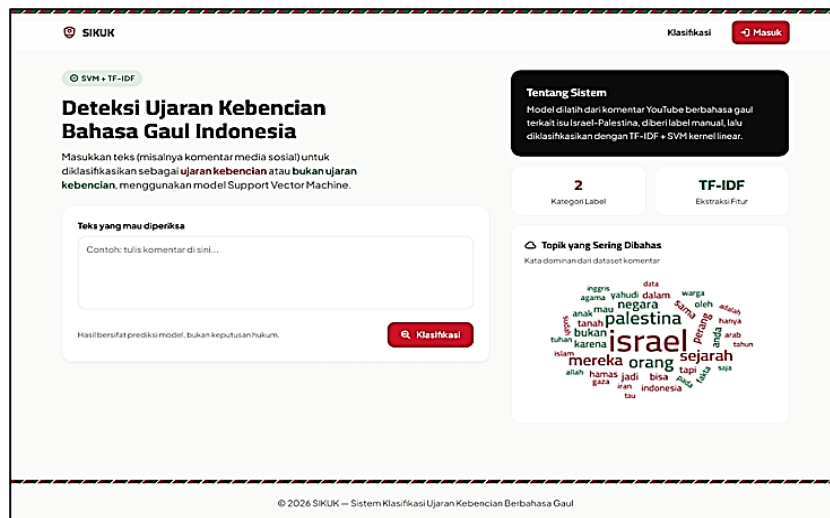
Karakteristik Ujaran Kebencian pada Teks Bahasa Gaul

Berdasarkan hasil pencocokan leksikon pada tahap pelabelan, ujaran kebencian pada komentar YouTube terkait isu Israel-Palestina didominasi oleh dua kelompok kata: (1) kata kasar/makian seperti "anjing", "goblok", "tolol", dan "bangsat"; serta (2) kata yang mengandung unsur hasutan atau kekerasan seperti "bunuh", "bantai", "musnahkan", "benci", dan "penjajah". Proporsi komentar yang terindikasi sebagai ujaran kebencian adalah 10,02% dari keseluruhan dataset, menunjukkan bahwa meskipun bukan mayoritas, keberadaannya tetap signifikan untuk dipantau dan dimoderasi (Jefriyanto Saud & Ratna Dewanti, 2025).

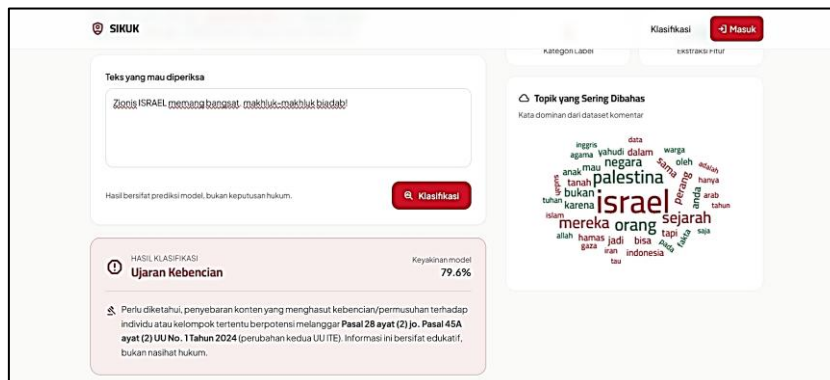
Komentar ujaran kebencian juga cenderung menggunakan bahasa gaul yang telah dinormalisasi pada tahap praproses, seperti penyingkatan kata dan variasi ejaan tidak baku, yang jika tidak dinormalisasi terlebih dahulu akan menyulitkan proses pencocokan leksikon maupun ekstraksi fitur TF-IDF. Hal ini mengkonfirmasi temuan Valencia Suci Handayani (2026) bahwa penggunaan bahasa gaul menyebabkan ujaran kebencian sulit dikenali karena maknanya sering bersifat kontekstual.

Hasil Implementasi Sistem Web SIKUK

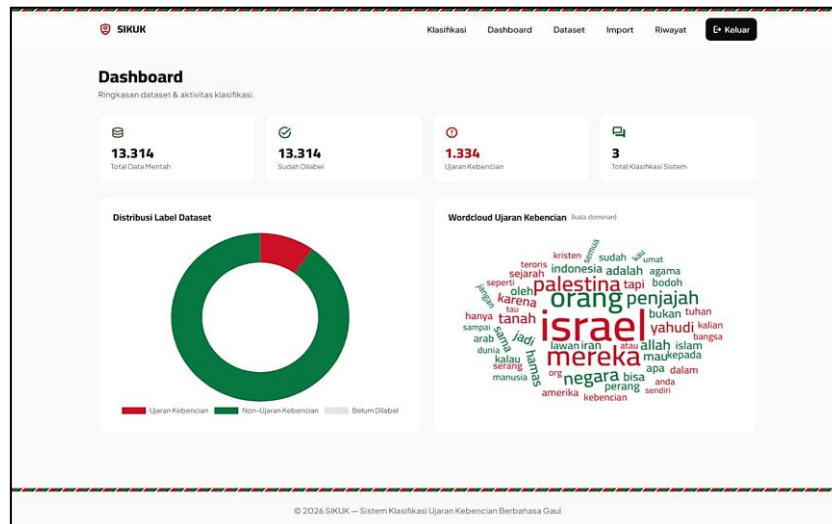
Model yang telah diekspor (tfidf_vocab.json berisi 5.000 term dan svm_model.json berisi 5.000 bobot fitur) berhasil diintegrasikan ke dalam sistem web SIKUK berbasis PHP native. Sistem ini mengimplementasikan pipeline praproses yang identik dengan tahap pelatihan di Python (case folding → cleaning → normalisasi slang → tokenizing → stopword removal → stemming menggunakan pustaka Sastrawi versi PHP) untuk memastikan konsistensi prediksi antara lingkungan pengembangan dan produksi.



Gambar 4. Tampilan Halaman Klasifikasi Publik SIKUK



Gambar 5. Tampilan Hasil Klasifikasi Teks Ujaran Kebencian (dengan rujukan UU ITE)



Gambar 6. Tampilan Dashboard Admin SIKUK

Hasil pengujian black-box testing terhadap 10 skenario uji menunjukkan seluruhnya berstatus "Berhasil", mencakup pengujian klasifikasi teks hate speech dan non-hate speech, validasi input kosong, autentikasi admin, pengelolaan dataset, dan riwayat klasifikasi.

Sistem memiliki beberapa keunggulan utama: (a) klasifikasi real-time tanpa reload halaman menggunakan AJAX; (b) inferensi model SVM diimplementasikan murni di PHP tanpa memerlukan runtime Python pada server produksi; (c) praproses PHP dibuat identik langkah-demi-langkah dengan training Python; (d) menampilkan rujukan dasar hukum (Pasal 28 ayat (2) jo. Pasal 45A ayat (2) UU No. 1 Tahun 2024) ketika teks terdeteksi sebagai ujaran kebencian; serta (e) antarmuka responsif yang dapat diakses dari berbagai perangkat.

KESIMPULAN

Berdasarkan hasil penelitian dan pembahasan yang telah diuraikan, dapat ditarik tiga kesimpulan utama. Pertama, proses klasifikasi ujaran kebencian berbahasa Indonesia pada teks bahasa gaul terkait isu Israel menggunakan metode Support Vector Machine (SVM) berhasil diimplementasikan secara menyeluruh, mulai dari pengumpulan 15.096 komentar mentah dari 79 video YouTube melalui YouTube Data API v3, praproses teks (case folding, cleaning, normalisasi bahasa gaul, tokenizing, stopword removal, dan stemming) yang menghasilkan 13.314 data bersih, pelabelan otomatis berbasis leksikon, ekstraksi fitur TF-IDF sebanyak 5.000 fitur, pelatihan model SVM kernel linear, hingga integrasi model ke dalam sistem web SIKUK berbasis PHP native.

Kedua, model SVM yang dibangun menunjukkan performa yang sangat baik dalam mengklasifikasikan ujaran kebencian pada teks bahasa gaul, dengan nilai accuracy sebesar 99,10%, precision sebesar 97,28%, recall sebesar 93,63%, dan F1-score sebesar 95,42% pada kelas ujaran kebencian, berdasarkan pengujian terhadap 2.663 data uji.

Ketiga, karakteristik ujaran kebencian pada teks bahasa gaul dalam diskusi isu Israel di YouTube didominasi oleh penggunaan kata kasar/makian dan kata bermuatan hasutan atau kekerasan, dengan proporsi komentar yang terindikasi ujaran kebencian sebesar 10,02% dari keseluruhan dataset.

Penelitian selanjutnya disarankan untuk: (1) melakukan anotasi manual penuh dengan pengukuran inter-annotator agreement; (2) memperluas kamus leksikon seiring perkembangan bahasa gaul di media sosial; (3) membandingkan performa SVM dengan algoritma lain seperti IndoBERT; (4) menerapkan teknik penyeimbangan data seperti SMOTE secara eksplisit; serta (5) memperluas cakupan kata kunci dan sumber data dari platform media sosial lain.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Universitas Muhammadiyah Sumatera Utara, khususnya Fakultas Ilmu Komputer dan Teknologi Informasi Program Studi Sistem Informasi, atas dukungan akademis yang diberikan selama proses penelitian ini berlangsung.

DAFTAR PUSTAKA

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). A brief survey of text mining: Classification, clustering and extraction techniques. *arXiv preprint*. <https://doi.org/10.48550/arXiv.1707.02919>

- Chakravarthi, B. R. (2024). Detection of homophobia and transphobia in YouTube comments. *International Journal of Data Science and Analytics*, 18(1), 49–68. <https://doi.org/10.1007/s41060-023-00400-0>
- Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21(1), 6. <https://doi.org/10.1186/s12864-019-6413-7>
- Chimkhilai, P., & Panyametheekul, S. (2023). Hate speech in YouTube comments on Rohingya refugees in Thailand and Syrian refugees in Europe. *LEARN Journal: Language Education and Acquisition Research Network*, 17(1), 133–161.
- Hofmann, S., Sommermann, C., Kraus, M., Zschech, P., & Rosenberger, J. (2025). *Hate speech and sentiment of YouTube video comments from public and private sources covering the Israel-Palestine conflict*. arXiv. <https://doi.org/10.48550/arXiv.2503.10648>
- Jefriyanto Saud, & Ratna Dewanti. (2025). Hate speech communication patterns in social media: A case study on Instagram during the 2024 political year. *International Journal of Integrative Sciences*, 4(7), 1533–1548. <https://doi.org/10.55927/ijis.v4i7.408>
- Kamalia, H. (2021). Sentiment analysis of hate speech using SVM method. *Khatulistiwa Smart: Journal of Artificial Intelligence*, 35(1).
- Kartika, S., & Nurhayati, N. (2023). Ujaran kebencian (hate speech) di media sosial dalam konteks hukum dan perubahan sosial (Studi kasus pada masyarakat kota Medan). *JURNAL MERCATORIA*, 16(1), 99–106. <https://doi.org/10.31289/mercatoria.v16i1.7668>
- Khatib Sulaiman, J., Ariska, A., Kamayani, M., & Muhammadiyah DrHamka, U. (2024). Deteksi hate speech pada kolom komentar TikTok dengan menggunakan SVM. *Indonesian Journal of Computer Science*, 13(1). <https://doi.org/10.33022/ijcs.v13i1.3791>
- Kowsari, K., Meimandi, K. J., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. *Information*, 10(4), 150. <https://doi.org/10.3390/info10040150>
- Kurniawan, S., Gata, W., Puspitawati, D. A., Parthama, I. K. S., Setiawan, H., & Hartini, S. (2020). Text mining pre-processing using Gata framework and RapidMiner for Indonesian sentiment analysis. *IOP Conference Series: Materials Science and Engineering*, 835(1). <https://doi.org/10.1088/1757-899X/835/1/012057>
- Mustofa, B. A., Laksito, W., & Saptomo, Y. (2025). Use of natural language processing in social media text analysis. *Journal of Artificial Intelligence and Engineering Applications*, 4(2). <https://doi.org/10.59934/jaica.v4i2.755>
- 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>
- Powers, D. M. W. (2020). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2010.16061>
- Purnama Sari, I., Krianto Sulaiman, O., Apdilah, D., & Simanjuntak, P. (2025). Analisis komparatif Support Vector Machine dan Random Forest untuk deteksi email phishing. *AICOMS*, 4(2), 18–27. <https://doi.org/10.58466/aicoms.v4i1.1806>
- Rahman, O. H., Abdillah, G., & Komarudin, A. (2021). Klasifikasi ujaran kebencian pada media sosial Twitter menggunakan Support Vector Machine. *Jurnal RESTI*, 5(1), 17–23. <https://doi.org/10.29207/resti.v5i1.2700>
- Rianto, Mutiara, A. B., Wibowo, E. P., & Santosa, P. I. (2021). Improving the accuracy of text classification using stemming method, a case of non-formal Indonesian conversation. *Journal of Big Data*, 8(1), 48. <https://doi.org/10.1186/s40537-021-00413-1>
- Saifullah, S., Dreżewski, R., Dwiyanto, F. A., Aribowo, A. S., Fauziah, Y., & Cahyana, N. H. (2024). Automated text annotation using a semi-supervised approach with meta vectorizer and machine learning algorithms for hate speech detection. *Applied Sciences*, 14(3), 1078. <https://doi.org/10.3390/app14031078>
- Salsabila, U. N., Winatmoko, Y. A., Septiandri, A. A., & Jamal, A. (2018). Colloquial Indonesian lexicon. 2018 *International Conference on Asian Language Processing (IALP)*. <https://doi.org/10.1109/IALP.2018.8629151>
- Santos-Ríos, R., Vilares, J., & Alonso, M. A. (2023). Some experiments on the use of natural language processing for sexism detection and classification in social media. *Kalpa Publications in Computing*, 14. <https://easychair.org/publications/paper/Hqwq>
- Sari, H., Leonarde Ginting, G., & Zebua, T. (2021). Penerapan algoritma text mining dan TF-IDF untuk pengelompokan topik skripsi pada aplikasi repository STMIK Budi Darma. *Jurnal Informatika*, 2(7), 414–432.

- Septiani, D., & Isabela, I. (2022). Analisis term frequency inverse document frequency (TF-IDF) dalam temu kembali informasi pada dokumen teks. *SINTESIA: Jurnal Sistem dan Teknologi Informasi Indonesia*, 1(2), 14–19.
- 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. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2406.19349>
- Suwarno, & Kusnadi, R. (2021). Comparative analysis of SVM, XGBoost and neural network on hate speech classification. *Jurnal RESTI*, 5(5), 896–903. <https://doi.org/10.29207/resti.v5i5.3506>
- Tahir, I., & Ramadhan, M. G. F. (2024). Hate speech on social media: Indonesian netizens' hate comments of presidential talk shows on YouTube. *LLT Journal: Journal on Language and Language Teaching*, 27(1), 230–251. <https://doi.org/10.24071/llt.v27i1.8180>
- Valencia Suci Handayani, M. (2026). Arsitektur hibrida IndoBERTweet-CNN untuk klasifikasi ujaran kebencian berbahasa gaul di media sosial. *Infotekmesin*, 17(01). <https://doi.org/10.35970/infotekmesin.v17i1.3026>
- Venkata Mahesh Babu Batta. (2024). Text classification in natural language processing. *International Journal of Advanced Research in Science, Communication and Technology*, 282–290. <https://doi.org/10.48175/ijarset-17645>
- Vidgen, B., & Derczynski, L. (2021). 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>