https://jurnal.pustakagalerimandiri.co.id/index.php/pustakadata
DOI: https://doi.org/10.55382/jurnalpustakadata.v6i4.2481

Jurnal Pustaka Data
Pusat Akses Kajian Database, Analisa Teknologi, dan Arsitektur Komputer
Pustaka Galeri Mandiri
Vol. 6 No. 4 (2026) 545 – 555
E ISSN: 2809-7483

Deteksi Cyberbullying pada Pesan Berbahasa Indonesia di Platform Discord Menggunakan Multinomial Naïve Bayes dan Pembobotan TF-IDF

Naufal Ikhsan Erman[1], Sarjon Defit[2], Abulwafa Muhammad[3]
[1,3]Teknik Informatika, Ilmu Komputer, Universitas Putra Indonesia YPTK Padang
[2]Sistem Informasi, Ilmu Komputer, Universitas Putra Indonesia YPTK Padang
[1]naufaliksan78@gmail.com. [2]sarjon_defit@upiyptk.ac.id [3]abulwafa@upiyptk.ac.id

Abstract

Cyberbullying on online communication platforms was shown to cause serious psychological harm, while moderation on Discord still relies on manual reporting that does not scale. This study aimed to build an automatic cyberbullying detection system for Indonesian-language messages on Discord using the Multinomial Naïve Bayes algorithm with TF-IDF weighting. A total of 71,591 raw messages were extracted from four Discord servers and cleaned into 49,712 messages. Candidate messages were screened through four layered stages, namely abusive lexicon matching, phrase pattern scanning, model-assisted screening, and identity marker scanning, yielding 3,027 labelled messages across six classes. The data were processed through seven preprocessing stages and classified under two schemes, six-class and binary. The results showed an accuracy of 0.802 with a macro-F1 of 0.531 for the six-class scheme, and an accuracy of 0.822 with a macro-F1 of 0.687 for the binary scheme. Five-fold cross validation produced 0.817 ± 0.010 and 0.823 ± 0.014, indicating stable performance. Compared with a majority-class predictor, the model improved macro-F1 by 55.9 percent. Error analysis revealed that 74.6 percent of misclassifications were missed bullying messages, whereas confusion between categories accounted for only 0.8 percent. An ablation test demonstrated that oversampling degraded performance by distorting class priors. It was concluded that the system is suitable as a moderator support tool rather than an autonomous enforcer.

Keywords: cyberbullying, Discord, Multinomial Naïve Bayes, TF-IDF, text mining

Abstrak

Perundungan siber pada platform komunikasi daring menimbulkan dampak psikologis yang serius bagi korban, sementara mekanisme moderasi Discord saat ini masih bertumpu pada pelaporan manual yang tidak skalabel untuk menangani volume pesan yang besar. Penelitian ini bertujuan membangun sistem deteksi perundungan siber otomatis untuk pesan berbahasa Indonesia pada platform Discord sebagai solusi atas keterbatasan tersebut, menggunakan algoritma Multinomial Naïve Bayes dengan pembobotan TF-IDF. Sebanyak 71.591 pesan mentah diekstraksi dari empat server Discord dan dibersihkan menjadi 49.712 pesan. Kandidat perundungan disaring melalui empat tahap berlapis, yaitu pencocokan leksikon kata kasar, pemindaian pola frasa, penyaringan dibantu model, serta pemindaian penanda identitas, sehingga diperoleh 3.027 pesan berlabel dalam enam kelas. Data diproses melalui tujuh tahap praproses dan diklasifikasikan pada dua skema, yaitu enam kelas dan biner. Hasil pengujian menunjukkan akurasi 0,802 dengan macro-F1 0,531 pada skema enam kelas, serta akurasi 0,822 dengan macro-F1 0,687 pada skema biner. Validasi silang lima lipat menghasilkan hasil yang stabil (0,817 ± 0,010 dan 0,823 ± 0,014). Dibandingkan penebak kelas mayoritas, model meningkatkan macro-F1 sebesar 55,9 persen. Analisis kesalahan memperlihatkan 74,6 persen kekeliruan berupa pesan perundungan yang terlewat,

Submitted : 13-08-2026 | Reviewed : 13-08-2026 | Accepted : 18-08-2026

sedangkan kekeliruan antar kategori hanya 0,8 persen. Uji ablasi membuktikan bahwa penyeimbangan kelas melalui penggandaan data justru menurunkan kinerja karena mendistorsi peluang awal. Dengan kinerja yang stabil dan tingkat kekeliruan antar kategori yang sangat rendah, sistem ini layak digunakan sebagai alat bantu moderator Discord untuk memprioritaskan pesan yang berpotensi mengandung perundungan, bukan sebagai penindak mandiri.

Kata kunci: *cyberbullying, Discord, Multinomial Naïve Bayes, TF-IDF, text mining*.

1. Pendahuluan

Platform komunikasi daring telah menjadi bagian yang tidak terpisahkan dari kehidupan sosial masyarakat modern. Discord, sebagai salah satu platform berbasis teks, suara, dan video yang paling banyak digunakan, memfasilitasi jutaan percakapan setiap hari pada komunitas permainan daring, pemrograman, pendidikan, hingga komunitas hobi. Kemudahan akses, anonimitas relatif, serta kekayaan fitur menjadikan Discord ruang interaksi yang dinamis sekaligus rentan terhadap penyalahgunaan. Salah satu bentuk penyalahgunaan yang paling merugikan adalah perundungan siber (cyberbullying), yaitu penggunaan teknologi digital untuk melakukan pelecehan, penghinaan, ancaman, atau pengucilan terhadap individu maupun kelompok secara sengaja dan berulang. Dampaknya terhadap korban bersifat nyata, mencakup depresi, kecemasan, isolasi sosial, hingga tindakan menyakiti diri sendiri pada kasus ekstrem [5].

Mekanisme moderasi yang tersedia pada Discord saat ini sebagian besar masih mengandalkan pelaporan pengguna secara manual serta peninjauan oleh tim moderator. Pendekatan tersebut memiliki tiga kelemahan mendasar, yaitu tidak skalabel untuk komunitas besar, bergantung pada kesediaan korban untuk melapor, dan kerap terlambat merespons insiden. Kajian menyeluruh mengenai sistem deteksi perundungan siber menunjukkan bahwa persoalan utama bidang ini terletak pada kualitas dataset, efisiensi waktu, dan biaya komputasi [3], sementara kajian lain menekankan sulitnya menangani ketidakseimbangan data serta memahami konteks sosial dan nuansa bahasa [4], [6]. Oleh sebab itu, sistem deteksi otomatis yang cepat, dapat ditelusuri, dan hemat sumber daya menjadi kebutuhan yang mendesak.

Text mining merupakan proses ekstraksi informasi bernilai tinggi dari data tekstual tidak terstruktur melalui identifikasi pola menggunakan teknik pembelajaran statistik [1], [2]. Pada tugas klasifikasi teks, algoritma Naïve Bayes telah lama terbukti efektif sekaligus efisien. Algoritma ini bekerja dengan menghitung peluang kelas berdasarkan fitur yang diekstraksi dari teks, memiliki kebutuhan komputasi yang rendah, mudah diimplementasikan, dan hasilnya dapat ditelusuri secara transparan. Meskipun asumsi independensi antarfiturnya bersifat menyederhanakan, kinerjanya tetap sebanding dengan metode yang lebih kompleks pada data berukuran kecil hingga menengah [9], [10]. Karakteristik tersebut sesuai dengan kondisi penelitian bahasa Indonesia yang umumnya menghadapi keterbatasan data berlabel dan sumber daya komputasi, sebagaimana juga ditemui pada penelitian bahasa dengan sumber daya terbatas lainnya [22].

Sejumlah penelitian telah menerapkan pembelajaran mesin untuk deteksi perundungan siber berbahasa Indonesia. Penelitian pada komentar Instagram menggunakan Naïve Bayes berbasis web melaporkan akurasi 98,5 persen [7], identifikasi perundungan pada media sosial melaporkan 93,10 persen [8], sedangkan optimasi Naïve Bayes dengan algoritma genetika pada komentar platform X memperoleh 77,34 persen [9]. Perbandingan Naïve Bayes dan Support Vector Machine pada data Twitter menunjukkan keduanya menghasilkan akurasi yang sebanding dengan waktu pelatihan Naïve Bayes yang jauh lebih singkat [10], sementara perbandingan varian Multinomial dan Bernoulli menunjukkan keunggulan varian Multinomial pada representasi berbasis frekuensi kata [11]. Pendekatan lain memakai Support Vector Machine [15], Random Forest [16], serta gabungan web scraping dan pemrosesan bahasa alami [14]. Pada ranah berbahasa Inggris, penelitian juga menggabungkan pembelajaran mesin dan pembelajaran mendalam untuk mengukur tingkat keparahan perundungan [12], [13].

Meskipun demikian, tinjauan terhadap penelitian tersebut memperlihatkan tiga kesenjangan. Pertama, hampir seluruhnya menggunakan data Twitter atau Instagram yang teksnya relatif panjang dan lebih tertata, sedangkan Discord — yang percakapannya sangat singkat, penuh singkatan, dan bercampur ragam daerah — sejauh penelusuran penulis belum banyak diteliti untuk bahasa Indonesia. Kedua,

mayoritas penelitian berhenti pada klasifikasi biner sehingga tidak dapat membedakan jenis perundungan yang terjadi, padahal jenis yang berbeda menuntut penanganan moderator yang berbeda pula. Ketiga, sebagian besar hanya melaporkan akurasi tanpa pembanding dasar (baseline) maupun uji kebocoran data, padahal pada data yang sangat timpang akurasi tinggi dapat diperoleh tanpa mendeteksi apa pun. Kesulitan anotasi pada data berbahasa Indonesia juga masih menjadi hambatan yang diakui secara luas [21].

Berdasarkan kesenjangan tersebut, penelitian ini menyusun tiga kontribusi. Pertama, dibangun dataset pesan Discord berbahasa Indonesia sebanyak 3.027 pesan yang dilabeli ke dalam enam kelas, yaitu penghinaan, ancaman, ujaran kebencian, pelecehan, pengucilan, dan bukan perundungan. Kedua, diperkenalkan prosedur penyaringan kandidat berlapis empat tahap untuk menjaring kelas minoritas yang tidak memuat kata kasar. Ketiga, kinerja model dilaporkan secara utuh disertai pembanding dasar, validasi silang, uji penempatan pembobotan terhadap kebocoran data, analisis kesalahan, serta uji ablasi terhadap penyeimbangan kelas. Tujuan penelitian ini adalah menerapkan tahapan text mining untuk mendeteksi perundungan siber pada pesan Discord berbahasa Indonesia, mengukur kinerja model Multinomial Naïve Bayes pada skema enam kelas dan biner, serta menilai keandalan sistem sebagai alat bantu moderasi.

2. Metode Penelitian

Penelitian ini bersifat kuantitatif eksperimental dengan tahapan berurutan mulai dari pengumpulan data, pembersihan, pelabelan, praproses, pembobotan TF-IDF, pelatihan model, hingga evaluasi. Alur pemrosesan sistem secara menyeluruh ditunjukkan pada Gambar 1.

Gambar 1. Alur sistem deteksi perundungan siber

2.1. *Pengumpulan dan Pembersihan Data*

Data berupa pesan teks berbahasa Indonesia diekstraksi dari empat server Discord yang dikelola langsung oleh peneliti. Proses ekspor menghasilkan delapan berkas CSV dengan total 71.591 pesan mentah. Pembersihan dijalankan secara terprogram melalui delapan langkah, yaitu penggabungan berkas, pembuangan pesan bot, normalisasi baris baru, penghapusan tautan, penghapusan emoji, anonimisasi identitas pengguna, pembuangan pesan tidak bermakna, serta pembuangan duplikat. Langkah terakhir bersifat penting karena pesan identik yang muncul pada data latih sekaligus data uji akan menimbulkan kebocoran data. Hasil pembersihan disajikan pada Tabel 1.

Tabel 1. Hasil pembersihan data

Tahap	Jumlah Pesan
Pesan mentah (gabungan delapan kanal)	71.591
Dibuang: pesan bot	351
Dibuang: pesan kosong	2.089
Dibuang: hanya tautan atau lampiran	714
Dibuang: hanya emoji	4.014
Dibuang: duplikat	14.711
Pesan bersih (siap diproses)	**49.712**

2.2. Pelabelan Data

Pelabelan menggunakan skema multikelas, yaitu setiap pesan diberi tepat satu label. Apabila sebuah pesan memuat lebih dari satu indikasi, label ditentukan menurut urutan prioritas ancaman > ujaran kebencian > pelecehan > penghinaan > pengucilan. Kelima kategori perundungan merupakan rumusan operasional yang menyesuaikan bentuk-bentuk yang lazim disebut pada literatur terhadap bentuk yang benar-benar teramati pada pesan teks Discord; bentuk yang menuntut pengamatan lintas waktu seperti penyamaran identitas tidak dijadikan kategori. Kelima kategori tersebut adalah penghinaan, yaitu pesan yang merendahkan, menghina, atau mencela individu maupun kelompok secara langsung; ancaman, yaitu pesan yang memuat ancaman fisik maupun psikologis, tersurat maupun tersirat, terhadap seseorang; ujaran kebencian, yaitu pesan provokatif atau diskriminatif berdasarkan suku, agama, ras, atau golongan tertentu sehingga sasarannya adalah identitas kelompok, bukan pribadi; pelecehan, yaitu pesan yang mengganggu, mempermalukan, atau merendahkan martabat seseorang, termasuk ejekan bermuatan seksual atau orientasi; dan pengucilan, yaitu pesan yang bertujuan mengucilkan, mengabaikan, atau mengeluarkan seseorang dari kelompok, dan kerap tidak memuat kata kasar sama sekali.

Karena pesan perundungan bersifat langka pada arsip percakapan, penjaringan kandidat dilakukan secara berlapis dalam empat tahap. Tahap pertama mencocokkan leksikon kata kasar dan kata kunci

ancaman setelah teks dinormalisasi. Leksikon disusun dari daftar kata kasar bahasa Indonesia terverifikasi [17], [18] dan daftar kata kasar dari pustaka perangkat lunak elang [19], yang disesuaikan dengan ragam komunikasi Discord, sedangkan normalisasi bentuk tidak baku memakai kamus bahasa sehari-hari berbahasa Indonesia [20]. Tahap kedua menyisir pola frasa untuk kelas minoritas yang umumnya tidak memuat kata kasar. Tahap ketiga memanfaatkan model biner terlatih untuk memberi skor pada seluruh pesan yang belum berlabel, lalu memunculkan pesan berpeluang tertinggi untuk ditinjau, sejalan dengan pendekatan anotasi semi-otomatis yang dilaporkan pada penelitian deteksi ujaran kebencian [21]. Tahap keempat memindai penanda identitas untuk menjaring kandidat ujaran kebencian yang paling langka.

Seluruh saran label yang dihasilkan sistem diverifikasi secara manual oleh peneliti mengacu pada pedoman anotasi tertulis, kemudian seluruh 3.027 pesan ditinjau ulang melalui diskusi dengan rekan sejawat. Pemeriksaan konsistensi berbasis kata kunci menunjukkan hanya 2 dari 166 pesan yang menyimpang dari pedoman, atau tingkat konsistensi 98,8 persen. Distribusi akhir dataset disajikan pada Tabel 2.

Tabel 2. Distribusi jumlah data per kelas

Kelas	Jumlah Data	Porsi (%)
Penghinaan (insult)	365	12,06
Ancaman (threat)	48	1,59
Ujaran kebencian (hate speech)	54	1,78
Pelecehan (harassment)	112	3,70
Pengucilan (exclusion)	56	1,85
Bukan perundungan (non-cyberbullying)	2.392	79,02
Total	**3.027**	**100,00**

2.3. Praproses Teks

Setiap pesan melewati tujuh tahap praproses berurutan, yaitu pembersihan berbasis ekspresi reguler, case folding, tokenisasi dengan pola [a-z]+, normalisasi melalui kamus 4.330 entri, penghapusan stopword, stemming dengan algoritma Nazief-Adriani, dan pembobotan TF-IDF. Penyusutan token akibat praproses tercatat mencapai 32 persen dari keseluruhan data. Contoh perubahan teks pada tiap tahap disajikan pada Tabel 3.

Tabel 3. Contoh hasil setiap tahap praproses

Tahap	Contoh A (penghinaan)	Contoh B (ancaman)
Teks asal	@user123 dasar BODOH!! gak ada gunanya lo di sini	Mau ku pukul????
Pembersihan	@user dasar BODOH gak ada gunanya lo di sini	Mau ku pukul
Case folding	@user dasar bodoh gak ada gunanya lo di sini	mau ku pukul
Tokenisasi	user, dasar, bodoh, gak, ada, gunanya, lo, di, sini	mau, ku, pukul
Normalisasi	user, dasar, bodoh, tidak, ada, gunanya, kamu, di, sini	mau, ku, pukul
Penghapusan stopword	dasar, bodoh, gunanya	ku
Stemming	dasar, bodoh, guna	ku

Contoh B memperlihatkan persoalan yang muncul pada pesan pendek. Pesan berlabel ancaman tersebut kehilangan kata kuncinya karena “pukul” termasuk dalam daftar stopword Sastrawi, sehingga hanya menyisakan satu token yang tidak bermakna. Kasus semacam ini menjadi salah satu penyebab 6,7 persen pesan tidak memiliki fitur aktif sama sekali, dan menjadi dasar uji ablasi stopword yang dibahas pada bagian pembahasan.

2.4. *Pembobotan TF-IDF dan Multinomial Naïve Bayes*

Teks hasil praproses diubah menjadi vektor numerik menggunakan pembobotan Term Frequency-Inverse Document Frequency dengan rentang n-gram (1, 2) dan ambang kemunculan minimum dua dokumen. Bobot sebuah kata dihitung sebagaimana Persamaan (1) dan (2).

$$\mathrm{idf}(t) = \log(N / \mathrm{df}(t)) + 1 \quad (1)$$

$$w(t, d) = \mathrm{tf}(t, d) \times \mathrm{idf}(t) \quad (2)$$

dengan N sebagai jumlah dokumen latih, df(t) jumlah dokumen yang memuat kata t, dan tf(t, d)

frekuensi kata t pada dokumen d. Vektor hasil pembobotan kemudian dinormalisasi panjangnya. Kosakata akhir yang terbentuk berjumlah 1.306 fitur dengan rata-rata 2,48 fitur aktif per pesan. Pemeriksaan sebaran fitur menunjukkan 199 pesan atau 6,7 persen sama sekali tidak memiliki fitur aktif, 1.655 pesan atau 55,5 persen hanya memiliki satu sampai dua fitur, dan hanya 19 pesan yang memiliki sepuluh fitur atau lebih. Kondisi tersebut merupakan konsekuensi langsung dari pesan Discord yang sangat singkat.

Matriks bobot menjadi masukan bagi Multinomial Naïve Bayes. Kelas akhir ditentukan melalui Persamaan (3), sedangkan peluang bersyarat tiap fitur dihitung memakai penghalusan Laplace pada Persamaan (4).

$$\hat{c} = \text{argmax } P(c) \cdot \Pi P(t \mid c) \quad (3)$$

$$P(t \mid c) = (\Sigma w(t, c) + \alpha) / (\Sigma w(c) + \alpha \cdot |V|) \quad (4)$$

dengan P(c) sebagai peluang awal kelas, w(t, c) jumlah bobot fitur t pada kelas c, |V| ukuran kosakata, dan α parameter penghalus. Agar nilai tidak hilang akibat keterbatasan presisi bilangan pecahan, perkalian dijalankan dalam bentuk penjumlahan logaritma.

2.5. Skenario dan Metrik Pengujian

Dataset berlabel dibagi menjadi data latih dan data uji dengan rasio 80:20 secara berstrata memakai random seed tetap agar hasil dapat direproduksi. Sebanyak 46 pesan gugur karena tidak menyisakan token setelah praproses, sehingga 2.981 pesan dipakai, terdiri atas 2.384 pesan latih dan 597 pesan uji. Pengujian dijalankan pada dua skema, yaitu klasifikasi enam kelas sesuai tujuan penelitian dan klasifikasi biner sebagai tolok ukur yang lebih andal mengingat distribusi data yang timpang. Pembobotan TF-IDF ditempatkan di dalam objek pipeline sehingga hanya dilatih pada lipatan data latih dan tidak pernah menyentuh data uji.

Metrik evaluasi dihitung dari empat komponen matriks kebingungan, yaitu true positive (TP), true negative (TN), false positive (FP), dan false negative (FN), sebagaimana Persamaan (5) sampai (8).

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \quad (5)$$

$$\text{Precision} = TP / (TP + FP) \quad (6)$$

$$\text{Recall} = TP / (TP + FN) \quad (7)$$

$$F1 = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (8)$$

Pada data yang timpang, akurasi dapat menyesatkan karena penebak kelas mayoritas pun memperoleh nilai tinggi tanpa mendeteksi apa pun. Oleh sebab itu rata-rata macro-F1, yaitu rata-rata F1 seluruh kelas tanpa pembobotan jumlah data, dijadikan acuan utama. Selain pembagian tunggal, pengujian diperkuat dengan validasi silang berstrata lima lipat.

3. Hasil dan Pembahasan

Rangkaian hasil penelitian berdasarkan urutan/susunan logis untuk membentuk sebuah cerita. Isinya menunjukan fakta/data dan jangan diskusikan hasilnya. Dapat menggunakan Tabel dan Angka tetapi tidak menguraikan secara berulang terhadap data yang sama dalam gambar, tabel dan teks. Untuk lebih memperjelas uraian, dapat mengunakan sub judul.

Pembahasan adalah penjelasan dasar, hubungan dan generalisasi yang ditunjukkan oleh hasil. Uraiannya menjawab pertanyaan penelitian. Jika ada hasil yang meragukan maka tampilkan secara objektif.

3.1. Penyetelan Parameter Penghalus

Nilai α dicari melalui validasi silang berstrata lima lipat pada data latih dengan metrik macro-F1. Nilai α = 0,01 memperoleh macro-F1 tertinggi sebesar 0,4929, diikuti 0,05 (0,4847), 0,1 (0,4673), dan 0,5 (0,2756), sedangkan nilai bawaan pustaka sebesar 1,0 hanya mencapai 0,1720 atau kurang dari sepertiganya. Selisih sebesar itu terjadi karena kosakata penelitian ini kecil dan mayoritas vektor pesan sangat jarang, sehingga penghalus yang terlalu besar menutupi bukti kebahasaan yang memang tipis. Temuan ini menegaskan bahwa penyetelan parameter berpengaruh nyata dan tidak dapat diabaikan.

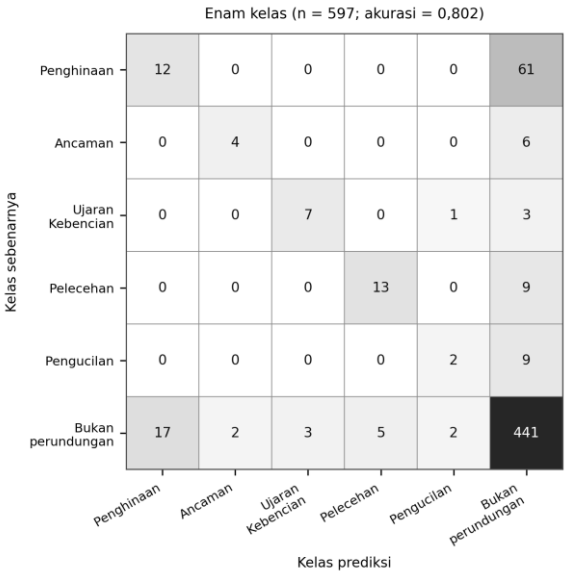
3.2. Klasifikasi Enam Kelas

Hasil pengujian pada skema enam kelas disajikan pada Tabel 4, sedangkan sebaran kesalahannya ditunjukkan pada Gambar 2.

Tabel 4. Hasil pengujian klasifikasi enam kelas

Kelas	Precision	Recall	F1-score	Data uji
Penghinaan	0,414	0,164	0,235	73
Ancaman	0,667	0,400	0,500	10
Ujaran kebencian	0,700	0,636	0,667	11
Pelecehan	0,722	0,591	0,650	22
Pengucilan	0,400	0,182	0,250	11
Bukan perundungan	0,834	0,938	0,883	470

Akurasi	—	—	0,802	597
Rata-rata macro	0,623	0,485	0,531	—
Rata-rata weighted	0,765	0,802	0,773	—



Gambar 2. Matriks kebingungan klasifikasi enam kelas

Model mengenali kelas mayoritas dengan precision 0,834 dan recall 0,938 sehingga akurasi keseluruhan tinggi. Yang lebih penting, seluruh kelas perundungan kini terdeteksi. Kelas ujaran kebencian memperoleh F1 0,667 dan kelas pelecehan 0,650, padahal jumlah data latihnya masing-masing hanya 43 dan 90 pesan. Sebaliknya kelas penghinaan justru paling lemah dengan F1 0,235 meskipun jumlah data latihnya paling banyak di antara kelas perundungan, yaitu 292 pesan.

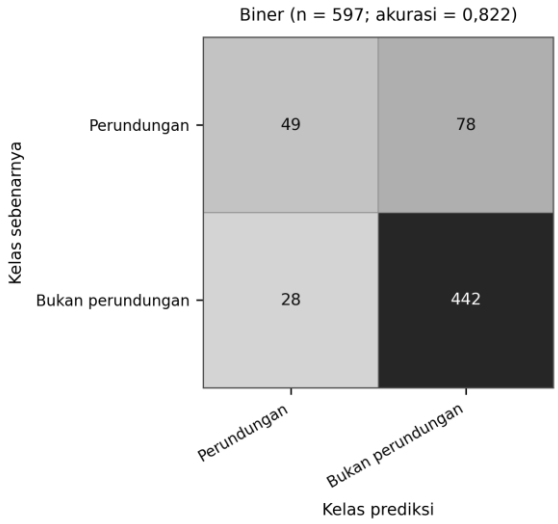
Pola tersebut menunjukkan bahwa penentu kinerja bukan semata-mata jumlah data, melainkan seberapa khas kosakata tiap kelas. Kelas ujaran kebencian memuat penanda identitas seperti nama suku atau agama yang hampir tidak pernah muncul pada percakapan biasa, sehingga peluang bersyaratnya sangat menonjol. Sebaliknya kelas penghinaan banyak memakai kata seperti “tolol” atau umpatan ringan yang justru juga sering muncul pada obrolan akrab yang dilabeli bukan perundungan, karena pada komunitas permainan daring kata tersebut kerap dipakai untuk bercanda tanpa sasaran. Keadaan itu diperberat oleh peluang awal: perbandingan peluang awal kelas bukan perundungan terhadap penghinaan adalah sekitar 6,6 berbanding 1, sehingga sebuah pesan baru hanya akan ditandai sebagai penghinaan apabila bukti kebahasaannya lebih dari enam kali lebih kuat. Akibatnya 61 dari 73 pesan penghinaan pada data uji berakhir diprediksi sebagai pesan biasa.

3.3. Klasifikasi Biner

Hasil pengujian pada skema biner disajikan pada Tabel 5 dan Gambar 3.

Tabel 5. Hasil pengujian klasifikasi biner

Kelas	Precision	Recall	F1-score	Data uji
Perundungan siber	0,636	0,386	0,480	127
Bukan perundungan	0,850	0,940	0,893	470
Akurasi	—	—	0,822	597
Rata-rata macro	0,743	0,663	0,687	—



Gambar 3. Matriks kebingungan klasifikasi biner

Pada tingkat biner model jauh lebih andal. Nilai precision 0,636 berarti sekitar enam dari sepuluh pesan yang ditandai memang benar perundungan, sedangkan recall 0,386 menunjukkan model menangkap sekitar 39 persen pesan perundungan. Perilaku konservatif ini dapat dijelaskan dari peluang awal data latih yang berbanding sekitar 3,8 berbanding 1 untuk kelas bukan perundungan, sehingga pesan bermuatan halus atau menyindir cenderung terlewat. Sifat tersebut sebenarnya sejalan dengan posisi sistem sebagai alat bantu moderator: pesan yang ditandai lebih dapat dipercaya sehingga moderator tidak dibanjiri peringatan palsu. Apabila recall yang lebih tinggi dibutuhkan, ambang keputusan dapat diturunkan dengan konsekuensi precision menurun.

3.4 Validasi Silang dan Perbandingan terhadap Pembanding Dasar

Karena data bersifat timpang, hasil dari satu kali pembagian dapat dipengaruhi keberuntungan pembagian. Validasi silang lima lipat dan perbandingan terhadap penebak kelas mayoritas disajikan pada Tabel 6.

Tabel 6. Hasil validasi silang lima lipat dan perbandingan terhadap pembanding dasar

Ukuran	Enam kelas	Biner	Pembanding dasar
Akurasi (holdout)	0,802	0,822	0,787
Macro-F1 (holdout)	0,531	0,687	0,441
Akurasi (validasi silang)	0,817 ± 0,010	0,823 ± 0,014	—
Macro-F1 (validasi silang)	0,523 ± 0,053	0,683 ± 0,029	—

Simpangan baku yang kecil pada akurasi, yaitu sekitar 0,01, menunjukkan kinerja model stabil di seluruh lipatan; akurasi antarlipatan bergerak pada rentang 0,807 sampai 0,829 untuk skema enam kelas dan 0,799 sampai 0,841 untuk skema biner. Meskipun demikian, kestabilan tersebut hanya berlaku pada akurasi. Simpangan baku macro-F1 skema enam kelas mencapai 0,053, jauh lebih besar dibanding skema biner yang hanya 0,029, sehingga kemampuan model mengenali kelas minoritas masih berubah-ubah bergantung pembagian data dan nilainya perlu dibaca sebagai perkiraan yang belum sepenuhnya mantap.

Perbandingan terhadap pembanding dasar memperlihatkan temuan yang penting. Selisih akurasi model terhadap penebak kelas mayoritas hanya 3,5 poin, sehingga apabila hanya akurasi yang dilihat sistem ini seolah tidak banyak memberi manfaat. Namun pada macro-F1 selisihnya menjadi 0,441 berbanding 0,687, atau peningkatan sekitar 55,9 persen secara relatif. Penebak kelas mayoritas memiliki akurasi tinggi tetapi tidak pernah mendeteksi satu pun pesan perundungan, sedangkan model yang dibangun benar-benar menemukan 49 dari 127 pesan perundungan pada data uji. Inilah alasan macro-F1 dan metrik per kelas dijadikan ukuran utama, bukan akurasi.

3.5 Uji Penempatan Pembobotan terhadap Kebocoran Data

Untuk memastikan tidak terjadi kebocoran data, dilakukan pengujian pembanding dengan meletakkan pembobotan TF-IDF di luar pipeline, yaitu dilatih pada seluruh data sebelum validasi silang dijalankan. Hasilnya disajikan pada Tabel 7.

Tabel 7. Perbandingan penempatan pembobotan TF-IDF

Penempatan TF-IDF	Macro-F1 (validasi silang)	Keterangan
Di dalam pipeline (dipakai penelitian ini)	0,6826	Sah secara metodologis
Di luar pipeline (dilatih pada seluruh data)	0,7166	Memuat keuntungan semu
Selisih	0,0340	Keuntungan semu

Selisih 0,0340 tersebut merupakan keuntungan semu yang timbul apabila informasi dari data uji ikut terpakai pada pembentukan fitur. Dengan demikian angka yang dilaporkan pada penelitian ini merupakan angka yang lebih rendah namun sah secara metodologis. Pengujian semacam ini jarang dilaporkan pada penelitian sejenis, padahal selisih sebesar tiga poin sudah cukup untuk mengubah kesimpulan perbandingan antaralgoritma.

3.6 Analisis Kesalahan Klasifikasi

Dari 597 pesan data uji pada skema enam kelas, 479 pesan diklasifikasikan dengan benar dan 118 pesan keliru. Kesalahan tersebut dikelompokkan menjadi tiga jenis sebagaimana Tabel 8.

Tabel 8. Pengelompokan jenis kesalahan klasifikasi

Jenis kesalahan	Jumlah	Persentase
A. Perundungan diprediksi sebagai pesan biasa (terlewat)	88	74,6%
B. Pesan biasa diprediksi sebagai perundungan	29	24,6%
C. Salah kategori antarkelas perundungan	1	0,8%
Jumlah kesalahan	**118**	**100,0%**

Temuan terpenting dari Tabel 8 adalah kesalahan jenis C hanya berjumlah satu kasus. Artinya, ketika model berhasil mengenali sebuah pesan sebagai perundungan, model hampir selalu menempatkannya pada kategori yang tepat. Persoalan utama model bukan membedakan jenis perundungan, melainkan membedakan perundungan dari percakapan biasa. Temuan ini mengarahkan perbaikan pada tempat yang benar, yaitu menambah kemampuan mengenali pesan menyerang secara umum akan jauh lebih berdampak daripada memperhalus pembedaan antarkategori.

Kesalahan jenis A yang mendominasi tersebar pada seluruh kelas perundungan, dengan rincian penghinaan 61 dari 73 pesan, pengucilan 9 dari 11, pelecehan 9 dari 22, ancaman 6 dari 10, dan ujaran kebencian 3 dari 11. Kelas pengucilan paling terdampak secara proporsional karena pesannya kerap tidak memuat kata kasar sama sekali. Adapun kesalahan jenis B berjumlah 29 pesan dari 470 pesan biasa, atau sekitar 6,2 persen, dan sebagian besar berupa pesan biasa yang diprediksi sebagai penghinaan. Tingkat kesalahan tersebut masih tergolong rendah sehingga moderator tidak akan dibanjiri peringatan palsu.

3.7 Uji Ablasi Penyeimbangan Kelas

Penyeimbangan kelas melalui penggandaan acak dokumen minoritas diuji melalui ablasi dengan konfigurasi yang identik. Penyeimbang ditempatkan di dalam pipeline sehingga penggandaan hanya terjadi pada lipatan latih. Hasil ablasi menunjukkan penggandaan data justru menurunkan kinerja secara tajam: akurasi turun dari 0,817 ± 0,010 menjadi 0,590 ± 0,025 dan macro-F1 dari 0,523 ± 0,053 menjadi 0,402 ± 0,041. Penyebabnya dapat ditelusuri pada dua komponen perhitungan Naïve Bayes. Ketika seluruh dokumen suatu kelas digandakan sebanyak k kali, pembilang dan penyebut pada perhitungan peluang bersyarat sama-sama terkali k sehingga nilainya praktis tidak berubah. Sebaliknya peluang awal berubah total karena jumlah dokumen tiap kelas menjadi sama, yaitu 0,1667 untuk keenam kelas. Dengan kata lain, penggandaan tidak menambah informasi kebahasaan apa pun, melainkan hanya menghapus pengetahuan model mengenai seberapa sering tiap kelas benar-benar muncul. Dampaknya nyata: jumlah pesan biasa yang keliru ditandai melonjak sekitar sembilan kali lipat. Dengan demikian dugaan bahwa penyeimbangan kelas akan memperbaiki kinerja tidak terdukung oleh data.

Sebagai gantinya, penanganan ketidakseimbangan dilakukan pada tahap pengumpulan data melalui penyaringan berlapis empat tahap yang menambah jumlah dokumen asli kelas minoritas. Pendekatan ini menambah keragaman kata sehingga peluang bersyarat ikut membaik, sedangkan penggandaan hanya mengubah peluang awal tanpa menambah keragaman kata sama sekali.

3.8 Pola yang Dipelajari Model dan Implementasi Sistem

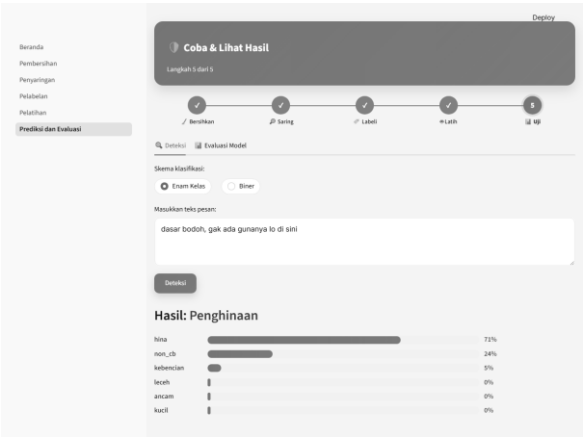
Untuk mengetahui pola yang benar-benar dipelajari model, dihitung istilah dengan bobot log-peluang tertinggi pada tiap kelas sebagaimana Tabel 9.

Tabel 9. Istilah paling menentukan pada tiap kelas

Kelas	Istilah paling menentukan
Ancaman	bunuh, hajar, gebuk, antem, awas, samperin
Ujaran kebencian	sesat, agama sesat, jawa tolol, jawa, sunda
Pelecehan	lesbi, bokep, homo, lonte, seks
Pengucilan	sok asik, usah ditemenin, ikut-ikut, cupu, dasar beban
Penghinaan	kayak tolol, tahi kau, bacot kau, najis, bangsat kau
Bukan perundungan	jam, yuk, valo, gila, makan, game, kasih

Tabel 9 memperlihatkan bahwa batas antarkelas yang ditetapkan pada pedoman anotasi ditemukan kembali oleh model secara mandiri dari data. Seluruh istilah teratas kelas pelecehan bermuatan seksual, sedangkan seluruh istilah teratas kelas ujaran kebencian merujuk pada suku dan agama. Hal ini menguatkan keputusan pemisahan kedua kelas tersebut, sebab pembedaannya tidak hanya berasal dari aturan pelabelan melainkan juga tercermin pada pola kata yang dipelajari model. Perlu dicatat bahwa sebagian istilah terikat pada kejadian tertentu di dalam korpus, sehingga kemampuan model pada data dari komunitas lain masih perlu diuji.

Model akhir diserialisasi dan dimuat oleh aplikasi web berbasis Streamlit yang memuat seluruh tahapan penelitian, mulai dari pembersihan, penyaringan, pelabelan, pelatihan, sampai pengujian. Antarmuka hasil deteksi ditunjukkan pada Gambar 4. Sistem menampilkan kategori keluaran beserta peluang tiap kelas dalam bentuk diagram batang sehingga keputusan model dapat ditelusuri oleh moderator, bukan disajikan sebagai vonis yang tertutup.



Gambar 4. Antarmuka hasil deteksi pada aplikasi web

3.9 Pembahasan

Akurasi pada skema biner, yaitu 82,2 persen pada pembagian tunggal dan 82,3 persen pada validasi silang, melampaui target minimal 75 persen maupun pembanding dasar 78,7 persen. Skema biner juga terbukti lebih seimbang daripada skema enam kelas, baik pada macro-F1 (0,687 berbanding 0,531) maupun akurasi. Terdapat pertukaran yang jelas antara keduanya: skema enam kelas memberi kategori terperinci tetapi kurang stabil pada kelas langka, sedangkan skema biner memberi deteksi yang andal tetapi kasar. Karena kesalahan antarkategori hanya satu kasus, model enam kelas dapat sekaligus melayani kebutuhan biner tanpa menurunkan akurasi, sehingga satu model saja memadai untuk kedua keperluan.

Apabila dibandingkan dengan penelitian terdahulu, akurasi 82,2 persen berada di bawah angka yang dilaporkan penelitian pada komentar Instagram [7] maupun identifikasi perundungan pada media sosial [8], namun di atas hasil optimasi Naïve Bayes dengan algoritma genetika pada platform X [9]. Perbandingan tersebut perlu dibaca dengan hati-hati karena tiga alasan. Pertama, penelitian terdahulu menggunakan data Twitter atau Instagram yang teksnya lebih panjang dan lebih tertata dibanding pesan Discord yang sangat singkat; pada penelitian ini rata-rata hanya 2,48 fitur aktif per pesan dan 6,7 persen pesan bahkan tidak memiliki fitur sama sekali. Kedua, sebagian penelitian tersebut memakai dataset dengan distribusi kelas yang lebih seimbang sehingga pembanding dasarnya jauh lebih rendah dan akurasi yang sama bermakna lebih tinggi. Ketiga, sebagian hanya melaporkan akurasi tanpa macro-F1, padahal pada data timpang kedua metrik tersebut dapat memberi kesimpulan yang berbeda. Perbandingan varian Naïve Bayes [11] maupun perbandingan terhadap Support Vector Machine [10] juga menunjukkan bahwa selisih kinerja antaralgoritma pada tugas ini umumnya kecil, sehingga kualitas dan representativitas data lebih menentukan daripada pemilihan algoritma.

3.10 Keterbatasan Penelitian

Sejumlah keterbatasan perlu dicatat secara terbuka agar hasil penelitian ini dibaca pada takaran yang tepat. Satu server menyumbang 92,98 persen data bersih, sehingga karakteristik bahasa dataset sangat dipengaruhi kebiasaan berkomunikasi satu komunitas dan hasilnya belum tentu berlaku umum. Sebagian data berlabel diperoleh melalui penyaringan berbantuan model, dan karena data tersebut ikut dipakai melatih dan menguji model, data uji berpotensi lebih banyak memuat pola yang memang mudah dikenali sehingga hasil pengujian dapat sedikit optimistis; cara ini tetap dipilih karena pesan perundungan sangat langka pada arsip percakapan. Kedua skema klasifikasi juga memakai pembagian data yang tidak sepenuhnya identik karena stratifikasi dilakukan terhadap label yang berbeda, sehingga selisih akurasi antarskema perlu dibaca sebagai gambaran umum, bukan perbandingan langsung pada data uji yang sama.

Dari sisi keandalan label, pelabelan dilakukan oleh peneliti dengan pedoman tertulis dan ditinjau ulang melalui diskusi rekan sejawat, namun belum diverifikasi ahli bahasa independen maupun diukur melalui uji kesepakatan antaranotator. Kelas minoritas juga hanya terwakili 10 sampai 11 pesan pada data uji, sehingga satu kesalahan klasifikasi dapat menggeser F1 kelas tersebut sekitar 0,1; penelitian ini pun belum melakukan uji signifikansi statistik, sehingga metrik per kelas sebaiknya dibaca sebagai indikasi awal. Dari sisi definisi, akademik perundungan siber memuat unsur pengulangan dan ketimpangan kuasa, sedangkan keduanya tidak dapat diukur dari satu pesan yang berdiri sendiri, sehingga yang dideteksi sistem ini adalah indikasi perundungan pada tingkat pesan, bukan perundungan sebagai pola perilaku berulang.

Dari sisi representasi fitur, bag-of-words tidak memodelkan urutan kata, dan penghapusan stopword menghilangkan penanda negasi; uji ablasi awal dengan mengecualikan delapan belas kata bermuatan menaikkan F1 pengucilan dari 0,182 menjadi 0,364 dan macro-F1 dari 0,531 menjadi 0,596, tetapi menurunkan pelecehan dari 0,591 menjadi 0,545 dan belum tegas secara statistik, sehingga model akhir tidak diubah. Emoji juga dibuang pada tahap pembersihan, padahal emoji dapat menjadi penanda penghinaan maupun pelecehan sehingga sebagian sinyal hilang sebelum pemodelan. Terakhir, kelas pelecehan tidak memiliki strategi penyaringan berbasis kamus maupun pola tersendiri sehingga kandidatnya sebagian besar berasal dari penyaringan berbantuan model; kelangkaan pesan perundungan eksplisit turut dipengaruhi komunitas yang relatif santai, penyaringan kata kunci bawaan Discord, serta perundungan yang berlangsung pada percakapan suara dan tidak terekam sebagai teks.

4. Kesimpulan

Penelitian ini berhasil membangun sistem deteksi perundungan siber pada pesan berbahasa Indonesia di platform Discord menggunakan Multinomial Naïve Bayes dengan pembobotan TF-IDF. Tahapan text mining yang terdiri atas tujuh langkah praproses terbukti mampu mengubah pesan yang sangat informal menjadi representasi fitur numerik yang dapat diklasifikasikan. Model memperoleh akurasi 0,822 dengan macro-F1 0,687 pada skema biner dan akurasi 0,802 dengan macro-F1 0,531 pada skema enam kelas, dengan hasil yang stabil pada validasi silang lima lipat. Dibandingkan penebak kelas mayoritas, model meningkatkan macro-F1 sebesar

55,9 persen sehingga terbukti benar-benar mempelajari pola pembeda antarkelas.

Tiga temuan metodologis diperoleh. Pertama, penyaringan kandidat berlapis empat tahap terbukti lebih efektif daripada penyeimbangan buatan untuk menangani kelas minoritas, karena menambah keragaman kata dan bukan sekadar mengubah peluang awal. Kedua, penggandaan dokumen minoritas justru menurunkan kinerja secara tajam, sehingga asumsi metodologis semacam ini perlu diuji secara empiris untuk setiap kombinasi algoritma dan data. Ketiga, penempatan pembobotan TF-IDF di luar pipeline menghasilkan keuntungan semu sebesar 0,0340 pada macro-F1, sehingga pelaporan kinerja tanpa uji kebocoran data berpotensi menyesatkan.

Kinerja yang diperoleh belum memadai untuk penerapan mandiri karena recall pada skema biner baru mencapai 0,386, sehingga sistem ini diposisikan sebagai alat bantu moderator, bukan sebagai penindak otomatis. Berdasarkan keterbatasan yang telah diuraikan, penelitian lanjutan disarankan menambah data berlabel pada kelas ancaman, ujaran kebencian, dan pengucilan dari komunitas yang lebih beragam; meninjau ulang pengaturan ekstraksi fitur, misalnya menurunkan ambang dokumen minimum, menambahkan n-gram karakter, serta mempertahankan kata negasi dan mempertimbangkan emoji sebagai fitur; menetapkan satu pembagian data uji yang sama untuk kedua skema agar perbandingan antarskema setara; membandingkan Naïve Bayes dengan algoritma lain pada data dan pembagian yang sama disertai uji signifikansi statistik; melibatkan anotator tambahan untuk mengukur tingkat kesepakatan antaranotator; menyeimbangkan penyaringan berbantuan model dengan pengambilan sampel acak serta menyusun strategi penyaringan khusus bagi kelas pelecehan; dan menerapkan sistem sebagai alat bantu moderasi dengan pengawasan manusia, misalnya menurunkan ambang keputusan agar recall meningkat, dengan pesan bertanda ditempatkan sebagai daftar prioritas tinjauan moderator dan bukan sebagai penindak otomatis.

Daftar Rujukan

[1] Chen, X., Xie, H., Qin, S. J., Wang, F. L., & Hou, Y. (2025). Artificial intelligence-supported student engagement research: Text mining and systematic analysis. European Journal of Education, 60(1). https://doi.org/10.1111/ejed.70008

[2] Torres-Cruz, F., Yucra-Mamani, Y. J., Mayta-Quispe, M. F., & Ibanez-Quispe, V. (2024). Integration of text mining and complex social network analysis for the quantitative characterisation of discursive evolution in artificial intelligence. Visual Review, 16(5), 271–278. https://doi.org/10.62161/revvisual.v16.5288

[3] Philipo, A. G., Sarwatt, D. S., Ding, J., Daneshmand, M., & Ning, H. (2025). Cyberbullying detection: Exploring datasets, technologies, and approaches on social media platforms. ACM Computing Surveys. https://doi.org/10.1145/3785654

[4] Kumar, R., & Bhat, A. (2022). A study of machine learning-based models for detection, control, and mitigation of cyberbullying in online social media. International Journal of Information Security, 21(6), 1409–1431. https://doi.org/10.1007/s10207-022-00600-y

[5] Harshitha, T. N., Prabu, M., Suganya, E., Sountharrajan, S., Bavirisetti, D. P., Gadde, N., & Uppu, L. S. (2024). ProTect: A hybrid deep learning model for proactive detection of cyberbullying on social media. Frontiers in Artificial Intelligence, 7. https://doi.org/10.3389/frai.2024.1269366

[6] Wang, S., Shibghatullah, A. S., Iqbal, T. J., & Keoy, K. H. (2024). A review of multimodal-based emotion recognition techniques for cyberbullying detection in online social media platforms. Neural Computing and Applications, 36(35), 21923–21956. https://doi.org/10.1007/s00521-024-10371-3

[7] Hutagalung, A. S., Negara, A. B. P., & Pratama, E. E. (2021). Aplikasi pendeteksi cyberbullying terhadap komentar postingan media sosial Instagram dengan metode Naive Bayes classifier berbasis website. Jurnal Sistem dan Teknologi Informasi (JUSTIN), 9(3), 364. https://doi.org/10.26418/justin.v9i3.44843

[8] Nita, T., & Alam, R. G. (2025). Identifikasi cyberbullying pada media sosial menggunakan algoritma Naive Bayes. Jurnal Teknologi Informasi, 6(2). https://doi.org/10.46576/djtechno

[9] Tahir, S. F., & Sugianto, C. A. (2024). Optimasi Naive Bayes menggunakan algoritma genetika pada klasifikasi komentar cyberbullying pada media sosial X. Jurnal Informatika dan Teknik Elektro Terapan, 12(3). https://doi.org/10.23960/jitet.v12i3.4834

[10] Lukman, K., & Novianto, S. (2025). Komparasi algoritma Naive Bayes dan SVM untuk identifikasi cyberbullying selebriti di media sosial Twitter. Jurnal Algoritma, 22(1), 970–981. https://doi.org/10.33364/algoritma/v.22-1.2196

[11] Dhuhita, W. M. P. (2023). Perbandingan kinerja algoritma Multinomial dan Bernoulli Naive Bayes dalam mengklasifikasikan komentar cyberbullying. Komputika: Jurnal Sistem Komputer, 12(2), 109–117. https://doi.org/10.34010/komputika.v12i2.9767

[12] Alabdulwahab, A., Haq, M. A., & Alshehri, M. (2023). Cyberbullying detection using machine learning and deep learning. International Journal of Advanced Computer Science and Applications, 14(10). https://doi.org/10.14569/IJACSA.2023.0141045

[13] Talpur, B. A., & O'Sullivan, D. (2020). Cyberbullying severity detection: A machine learning approach. PLOS ONE, 15(10), e0240924. https://doi.org/10.1371/journal.pone.0240924

[14] Azzahra, S. A., & Majid, N. W. A. (2025). Klasifikasi dan analisis semantik cyberbullying sosial media X: Integrasi web scraping dan natural language processing. Jurnal Educatio FKIP UNMA, 11(2). https://doi.org/10.31949/educatio.v11i2.12725

[15] Nugraha, D., & Astuti, P. (2023). Analisis sentimen cyberbullying pada sosial media Instagram menggunakan metode Support Vector Machine. Information System for Educators and Professionals, 8(2), 152–164.

[16] Nurnaryo, R. M. H., Suzanti, I. O., Fatah, D. A., Cahyani, A. D., & Mufarroha, F. A. (2022). Deteksi cyberbullying pada data tweet menggunakan metode Random Forest dan seleksi fitur information gain. Jurnal Simantec, 11(1).

[17] Ibrohim, M. O., & Budi, I. (2019). Multi-label hate speech and abusive language detection in Indonesian Twitter. In Proceedings of the 3rd Workshop on Abusive Language Online (pp. 46–57). Florence, Italy.

[18] Ibrohim, M. O., & Budi, I. (2018). A dataset and preliminaries study for abusive language detection in Indonesian social media. Procedia Computer Science, 135, 222–229. https://doi.org/10.1016/j.procs.2018.08.169

[19] Chan, S., & Tjandra, T. (2020). elang: Word embedding utilities for language models (English & Indonesian) (Version 0.1.1) [Software]. GitHub. https://github.com/onlyphantom/elang

[20] Salsabila, N. A., Winatmoko, Y. A., Septiandri, A. A., & Jamal, A. (2018). Colloquial Indonesian lexicon. In Proceedings of the International Conference on Asian

Language Processing (IALP) (pp. 226–229). Bandung, Indonesia. https://doi.org/10.1109/IALP.2018.8629151

[21] 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). https://doi.org/10.3390/app14031078

[22] Mukhim, B., Maji, A. K., & Das, S. (2025). Sentiment analysis with Khasi low-resource language through generation of sentiment words using machine learning. Computación y Sistemas, 29(3), 1225–1236. https://doi.org/10.13053/cys-29-3-5915
