KLASIFIKASI *TWEET CYBERBULLYING* DI APLIKASI X DENGAN ALGORITMA NAÏVE BAYES

Esti Mulyani[1], Hurul Aini Putri Prasetya[2]

[1,2] Jurusan Teknik Informatika, Politeknik Negeri Indramayu
[1]estimulyani@polindra.ac.id, [2] hurulprasetya22@student.polindra.ac.id

.

Abstrak

Perkembangan media sosial memberikan pengaruh besar dalam membentuk pola komunikasi masyarakat modern. Namun, kemudahan interaksi tersebut juga menghadirkan tantangan baru, salah satunya berupa meningkatnya kasus cyberbullying. Penelitian ini bertujuan mengembangkan sistem klasifikasi dua tahap untuk mendeteksi dan mengidentifikasi bentuk-bentuk cyberbullying pada tweet berbahasa Indonesia. Pengumpulan data dilakukan melalui proses crawling menggunakan autentikasi sesi pengguna (auth token) dengan kata kunci tertentu yang berkaitan dengan ujaran negatif. Tahap pertama sistem merupakan klasifikasi biner, yaitu memisahkan tweet menjadi dua kategori utama: *cyberbullying* dan *non-cyberbullying*. Model Multinomial Naïve Bayes digunakan pada tahap ini dengan penerapan pra-pemrosesan teks serta ekstraksi fitur menggunakan *Term Frequency-Inverse Document Frequency* (TF-IDF), menghasilkan akurasi sebesar 95%. Tahap kedua merupakan klasifikasi multikelas, di mana tweet yang sebelumnya terdeteksi sebagai *cyberbullying* dikategorikan lebih lanjut ke dalam tujuh label, yaitu: (1) SARA, (2) Pelecehan, (3) Penghinaan, (4) Body Shaming, (5) Ancaman, (6) Kata Kasar, dan (7) Ambiguitas, yakni kategori tambahan bagi teks dengan makna ganda atau sarkasme yang sulit ditentukan secara pasti. Pada tahap ini, algoritma Multinomial Naïve Bayes kembali digunakan dan menghasilkan akurasi sebesar 82%. Seluruh hasil klasifikasi disajikan secara interaktif melalui aplikasi web berbasis Python Flask. Hasil penelitian menunjukkan bahwa pendekatan dua tahap dengan algoritma Naïve Bayes mampu memberikan performa yang baik dalam identifikasi konten cyberbullying dan berpotensi digunakan sebagai alat pendukung pencegahan perundungan daring di media sosial.

Kata kunci: *Cyberbullying,* Naïve Bayes, Klasifikasi Teks, Media Sosial, NLP

1. Pendahuluan

Media sosial telah menjadi bagian tak terpisahkan dari kehidupan masyarakat modern dan membentuk cara individu berinteraksi dalam ruang digital. Platform seperti X (sebelumnya Twitter) membuka ruang komunikasi yang luas dan bebas, memungkinkan pengguna untuk menyampaikan pendapat serta berekspresi tanpa batasan fisik. Namun, kebebasan ini juga membawa konsekuensi serius berupa penyebaran konten negatif, salah satunya adalah *cyberbullying*—yakni tindakan perundungan yang dilakukan secara daring melalui kata-kata yang merendahkan, mengancam, atau menghina (Rosa et al., 2019a). Fenomena ini semakin mengkhawatirkan di tengah pesatnya pertumbuhan pengguna internet. Berdasarkan laporan Asosiasi Penyelenggara Jasa Internet Indonesia (APJII), pada tahun 2024 tercatat lebih dari 221 juta pengguna internet di Indonesia, atau sekitar 79,5% dari total populasi nasional. Fakta ini menjadikan Indonesia sebagai salah satu negara dengan tingkat penetrasi media sosial tertinggi di dunia (APJII, 2024), yang secara tidak langsung memperbesar potensi terjadinya cyberbullying di ruang daring.

Seiring meningkatnya ancaman tersebut, deteksi otomatis terhadap konten *cyberbullying* menjadi salah satu topik utama dalam penelitian bidang *Natural Language Processing* (NLP) dan *machine learning*. Berbagai metode telah dikembangkan untuk mengklasifikasikan ujaran negatif, mulai dari pendekatan berbasis statistik klasik seperti algoritma Naïve Bayes hingga teknik modern berbasis pembelajaran mendalam. Di Indonesia, studi mengenai klasifikasi *cyberbullying* telah dilakukan dengan menggabungkan teknik NLP dan *scraping* data media sosial secara kontekstual untuk mengidentifikasi *tweet* bermuatan negatif secara efektif (Azzahra & Majid, 2025). Penelitian internasional pun menunjukkan hasil yang menjanjikan. Chia et al. (2021) membuktikan bahwa kombinasi antara *feature engineering* dan *machine learning* mampu mendeteksi sarkasme dan ironi yang sering muncul dalam ujaran perundungan. Sementara itu, Ogunleye & Dharmaraj (2023) mengevaluasi kinerja BERT dan RoBERTa dalam mendeteksi *cyberbullying*, dengan hasil yang kompetitif. Namun, sebagian besar studi yang ada masih terbatas pada bahasa Inggris dan hanya berfokus pada klasifikasi biner, tanpa mengidentifikasi jenis-jenis *cyberbullying* secara spesifik.

Beberapa penelitian lokal telah menunjukkan bahwa algoritma Naïve Bayes, khususnya varian Multinomial Naïve Bayes, cukup efektif dalam mengklasifikasikan komentar bernada negatif pada media sosial Indonesia (Rifai et al., 2023; Wibisono et al., 2025). Di sisi lain, berbagai studi terkini telah menerapkan pendekatan *deep learning* seperti ensemble model, arsitektur transformer, mekanisme perhatian, hingga model hybrid seperti XLNet dengan BiLSTM untuk meningkatkan akurasi deteksi cyberbullying (Chen et al., 2024; Cuzzocrea et al., 2025; Fati et al., 2023; Muneer et al., 2023). Walaupun metode-metode tersebut memberikan peningkatan performa yang signifikan, teknik tersebut memerlukan sumber daya komputasi yang jauh lebih besar, ukuran dataset yang lebih luas, serta proses pelatihan yang kompleks. Kebutuhan tersebut tidak selalu sesuai untuk implementasi yang mengutamakan efisiensi, interpretabilitas, dan kemampuan berjalan secara ringan pada sistem real-time.

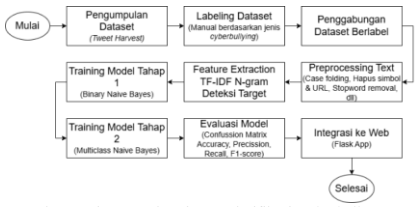
Berdasarkan pertimbangan tersebut, penelitian ini memilih Multinomial Naïve Bayes sebagai model klasifikasi karena sifatnya yang efisien, stabil untuk bahasa Indonesia yang cenderung memiliki struktur teks pendek seperti tweet, serta menghasilkan performa yang kompetitif pada dataset berskala menengah. Untuk ekstraksi fitur, TF-IDF digunakan karena mampu menghasilkan representasi teks yang terukur, mudah diinterpretasikan, dan secara empiris selaras dengan model probabilistik seperti Naïve Bayes.

Selanjutnya, gap penelitian yang ingin diisi oleh studi ini muncul dari ketidaksesuaian antara kompleksitas pendekatan canggih yang ada dan kebutuhan akan sistem deteksi yang ringan, cepat, namun tetap akurat. Selain itu, sebagian besar penelitian sebelumnya masih menggunakan pendekatan klasifikasi satu tahap atau berfokus pada kategori ujaran kebencian tertentu, sehingga belum menyediakan kerangka kategorisasi yang komprehensif untuk konteks cyberbullying di Indonesia. Oleh karena itu, penelitian ini mengusulkan pendekatan berbasis Naïve Bayes dengan dua tahap klasifikasi, yakni: (1) mendeteksi apakah sebuah tweet termasuk *cyberbullying* atau bukan, dan (2) mengidentifikasi kategori spesifik dari tujuh label *cyberbullying* yang relevan dengan konteks Indonesia—SARA, pelecehan, penghinaan, body shaming, ancaman, kata kasar, dan ambiguitas. Dengan demikian, penelitian ini berupaya menjembatani kebutuhan akan solusi yang lebih sederhana namun tetap memberikan hasil yang andal dalam memetakan berbagai bentuk cyberbullying pada platform X.

Dengan mengintegrasikan algoritma Naïve Bayes dan metode ekstraksi fitur TF-IDF, serta pendekatan pelabelan data berbasis analisis semantik secara manual, sistem ini diharapkan mampu memberikan kontribusi signifikan dalam pengembangan alat bantu pendeteksi konten negatif dalam konteks sosial media Indonesia (Farasalsabila et al., 2024; Philipo et al., 2024). Penelitian ini tidak hanya bersifat teknis, namun juga menjadi bagian dari upaya preventif dalam menangani ancaman perundungan siber yang kian masif di ruang digital.

2. Metode

Penelitian ini dilaksanakan melalui beberapa tahapan utama sebagaimana ditunjukkan pada Gambar 1.



Gambar 1. Diagram Alur Sistem Klasifikasi *Cyberbullying* Dua Tahap

Alur penelitian terdiri atas delapan tahap, yaitu: pengumpulan dataset, pelabelan dataset, pra-pemrosesan teks, ekstraksi fitur, pelatihan model tahap 1 (klasifikasi biner), pelatihan model tahap 2 (klasifikasi multikelas), evaluasi model, dan integrasi sistem ke web. Penjelasan setiap tahap dijabarkan pada subbab berikut.

2.1 Pengumpulan Data

Data berupa tweet berbahasa Indonesia dikumpulkan dari platform media sosial X (sebelumnya Twitter). Proses dilakukan secara otomatis dengan metode *crawling* menggunakan autentikasi sesi pengguna (auth_token) melalui skrip Python Tweet Harvest di Google Colab. Alat ini dipilih karena bersifat open-source, fleksibel, dan tidak dibatasi kuota harian seperti API resmi, sehingga dapat mengumpulkan data dalam jumlah besar secara efisien.

Pengambilan data difokuskan pada kata kunci yang mencerminkan tujuh kategori utama *cyberbullying*, yaitu SARA, pelecehan, penghinaan, *body shaming*, ancaman, kata kasar dan *ambiguitas*. Dataset mentah yang diperoleh memiliki atribut *id_str, created_at, full_text, username,* dan jenis_*cyberbullying*, tetapi hanya tiga atribut (*created_at, full_text*, jenis_*cyberbullying*) yang digunakan dalam penelitian ini.

2.2 Labeling Data

Proses pelabelan dilakukan dalam dua tahap untuk menyesuaikan pendekatan klasifikasi yang digunakan dalam penelitian ini, yaitu klasifikasi biner dan multikelas.

Tahap 1 — Label Biner
Seluruh tweet terlebih dahulu diberi label:

- *Cyberbullying*
- Bukan *cyberbullying*

Tahap ini penting sebagai filter utama, sebagaimana praktik pada penelitian dua tahap oleh (Fati et al., 2023; López-Vizcaíno et al., 2021).

Tahap 2 — Label Multikelas Cyberbullying
Untuk tweet yang masuk kategori *Cyberbullying*, annotator melakukan pelabelan lanjutan ke dalam tujuh jenis:

1. SARA
2. Pelecehan
3. Penghinaan
4. Body shaming
5. Ancaman
6. Kata kasar
7. Ambiguitas

Pendekatan ini sama seperti yang digunakan oleh (Azzahra & Majid, 2025), yang menekankan pentingnya *hierarchical labeling* agar model dapat menangani variasi intensitas dan bentuk serangan. Contoh hasil labeling ditunjukkan pada Tabel 1.

Tabel 1. Contoh Tweet yang Sudah di Labeling

Created_at	Full_text	Jenis_*cyberbullying*
Mon May 12 19:22:19 +0000 2025	ELU SIH BABI NGEYEL BGT DASAR V*N**T	Kata kasar
Fri May 09 02:49:34 +0000 2025	(((((Nasrani sekte sesat)))))) Awokwokwokwokwokwokwokwok Bilang aj lo bodoh tentang sejarah agama lo dan lo benci am yg berbau bau Islam	SARA
Mon Dec 16 10:43:52 +0000 2024	Dasar ba bi aerrrr gendut obesitas penyakitan badan lo kek kudanil	*Body shaming*
Sun Sep 24 11:14:32 +0000 2023	Buka baju gih	Pelecehan
Tue Jun 24 05:10:07 +0000 2025	xavier mukanya bloon banget sumpah (affectionate)	penghinaan
Wed Jun 04 09:24:49 +0000 2025	GUE BUNUH LO ANJ. N***T*T TU MULUT	Ancaman
	ngapain make celana ketat gitu? mau mancing wkwk	Ambiguitas
	a***ng cuacanya panas bener parah	Bukan *cyberbullying*

Catatan: baris terakhir **bukan *cyberbullying*** karena tidak menyerang individu/kelompok tertentu.

2.3 Penggabungan Dataset

Dataset hasil pelabelan awal berasal dari lebih dari **250 file terpisah** (format CSV/Excel). Setiap file berisi kumpulan tweet berdasarkan kata kunci tertentu yang mewakili kategori *cyberbullying*. Untuk keperluan analisis dan pelatihan model, seluruh file ini kemudian digabungkan secara bertahap ke dalam satu *dataframe* utama menggunakan pustaka Pandas pada Python. Hasil penggabungan dataset menghasilkan total 8.627 tweet, dengan distribusi kelas biner ditunjukkan pada Tabel 2.

Tabel 2.Tabel Distribusi Kelas Setelah Penggabungan

Kelas	Jumlah
Bukan Cyberbullying	4.377
Cyberbullying	4.250
Total	8.627

Sebanyak 4.250 tweet yang dikategorikan sebagai *cyberbullying* dipetakan kembali ke dalam tujuh jenis serangan sebagaimana ditunjukkan pada Tabel 3.

Tabel 3. Tabel Distribusi Kelas Multikelas Cyberbullying

Jenis Cyberbullying	Jumlah
SARA	669
Pelecehan	500
Penghinaan	571
Body shaming	495
Ancaman	566
Kata kasar	508
Ambiguitas	167

Dataset akhir kemudian dibagi menjadi data latih dan data uji menggunakan rasio 80:20. Dengan jumlah total 8.627 tweet, diperoleh data latih sebanyak 6.901 tweet dan data uji sejumlah 1.726 tweet.

2.4 *Preprocessing Text*

Data mentah hasil proses *crawling* tidak dapat langsung digunakan dalam pelatihan model klasifikasi karena masih mengandung banyak elemen yang tidak relevan atau mengganggu. Oleh karena itu, dilakukan tahap pra-pemrosesan (*preprocessing*) untuk meningkatkan kualitas data. Tujuan tahap ini adalah:

1. Menyederhanakan representasi teks,
2. Mengurangi *noise*, dan
3. M'enyamakan format teks sehingga dapat meningkatkan akurasi algoritma klasifikasi.

Teknik pra-pemrosesan yang digunakan antara lain:

1. *Case folding*: mengubah seluruh huruf menjadi kecil.
2. *Cleansing*: menghapus simbol, angka, URL, dan karakter non-alfabet.
3. *Tokenization*: memecah kalimat menjadi kata.
4. *Stopword removal*: menghapus kata umum yang tidak memiliki makna penting, seperti "yang", "dan", "itu".

5. *Stemming*: mengembalikan kata ke bentuk dasarnya menggunakan pustaka Sastrawi.

Penggunaan teknik pra-pemrosesan seperti *tokenization*, *stopword removal*, dan *stemming* juga telah diterapkan dalam penelitian lokal (Ridwan & Muzakir, 2022; Rifai et al., 2023; Wibisono et al., 2025) untuk memastikan kualitas fitur klasifikasi lebih optimal.

2.5 *Feature Extraction*

Setelah teks dipra-proses, tahap berikutnya adalah ekstraksi fitur menggunakan metode *Term Frequency-Inverse Document Frequency (TF-IDF)* dengan pendekatan *n-gram* (1-gram dan 2-gram). TF-IDF dipilih karena mampu menurunkan bobot kata umum sekaligus menaikkan bobot kata yang lebih spesifik terhadap konteks *cyberbullying* (Abdulloh & Hidayatullah, 2021; Prabowo & Azizah, 2020). Metode ini menggabungkan dua komponen utama:

1. *Term Frequency* (TF): frekuensi kemunculan suatu *term* dalam dokumen.
2. *Inverse Document Frequency* (IDF): ukuran seberapa jarang suatu *term* muncul di seluruh kumpulan dokumen.

Persamaan 1 menunjukkan rumus yang digunakan untuk perhitungan TF-IDF.

$$TF - IDF\ (t, d, D) = TF(t, d) \times IDF(t, D) \quad (1)$$

Dengan:

- t adalah term (kata),
- d adalah dokumen (*tweet*),
- D adalah kumpulan semua dokumen (*tweet*),

$$TF(t, d) = \frac{ft, d}{\sum k\ fk, d} \quad (2)$$

Persamaan 2 yaitu frekuensi kemunculan term t dalam dokumen d dibagi total semua term dalam dokumen tersebut

$$IDF(t, D) = \log(\frac{N}{nt}) \quad (3)$$

Pada Persamaan 3 di mana N adalah jumlah total dokumen dan n_t adalah jumlah dokumen yang mengandung term t.

Hasil TF-IDF berupa vektor numerik berdimensi sesuai jumlah kata unik dalam korpus. Implementasi teknis dilakukan dengan pustaka Scikit-learn (fungsi TfidfVectorizer), sebagaimana juga diterapkan oleh (Alfarizi et al., 2022).

2.6 Training Model Tahap 1 (Klasifikasi Biner)

Tahap pertama adalah klasifikasi biner dengan algoritma Multinomial Naïve Bayes. Tujuannya adalah menyaring tweet menjadi dua kelas utama, yaitu:

- 0 = Bukan *cyberbullying*
- 1 = *Cyberbullying*

Dengan pendekatan ini, sistem dapat berfungsi sebagai filter awal untuk memastikan hanya tweet bermuatan negatif yang akan diproses lebih lanjut.

Secara matematis, klasifikasi mengikuti Teorema Bayes dengan formula seperti ditunjukkan pada Persamaan 4.

$$P(H|X) = \frac{P(X|H) \cdot P(H)}{P(X)} \quad (4)$$

di mana:

- $P(H|X)$: probabilitas hipotesis H (kelas) terhadap kata X
- $P(X|H)$: probabilitas data X muncul dalam kelas H
- $P(H)$: probabilitas awal dari kelas H
- $P(X)$: probabilitas data X secara keseluruhan

2.7 Training Model Tahap 2 (Klasifikasi Multikelas)

Tweet yang telah dikategorikan sebagai *cyberbullying* pada tahap biner kemudian diproses pada tahap kedua, yaitu klasifikasi multikelas. Pada tahap ini, tweet dikelompokkan ke dalam **7** kategori spesifik, yaitu:

1. SARA
2. Pelecehan
3. Penghinaan
4. Body shaming
5. Ancaman
6. Kata kasar
7. Ambiguitas (label tambahan untuk teks yang sulit ditentukan ke salah satu kelas dengan jelas, misalnya mengandung unsur sarkasme atau makna ganda)

Algoritma yang digunakan tetap Multinomial Naïve Bayes (Azumah et al., 2024; Mubeen et al., 2025). Kriteria klasifikasi menggunakan pendekatan *Maximum A Posteriori* (MAP):

$$V_{MAP} = \arg\max_{v_j \epsilon V} P(v_j) \prod_{i=1}^{n} P(x_i|v_j) \quad (5)$$

Keterangan:

- V: himpunan semua kelas yang mungkin
- x_i: fitur ke-i dari dokumen
- $P(x_i|v_j)$: probabilitas kemunculan fitur pada kelas

Penambahan label ambiguitas dimaksudkan untuk mengurangi kesalahan klasifikasi pada data abu-abu (*borderline case*), sehingga sistem dapat memberikan hasil yang lebih realistis dibanding memaksakan klasifikasi ke enam kategori utama.

2.8 Evaluasi Model

Evaluasi memakai *confusion matrix* seperti ditunjukkan pada Tabel 4 serta metrik akurasi, presisi, *recall*, F1 (Bilgin & Bekar, 2025; Rosa et al., 2019b).

Tabel 4. Tabel Dasar Confusion Matrix

	Prediksi Positif	**Prediksi Negatif**
Kelas Positif	True Positive (TP)	False Negative (FN)
Kelas Negatif	False Positive (FP)	True Negative (TN)

Berdasarkan nilai-nilai ini, dapat dihitung metrik performa sebagai berikut:

Akurasi *(Accuracy)* menunjukkan proporsi prediksi yang benar terhadap seluruh jumlah data. Rumusnya ditunjukkan pada Persamaan 6.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

Presisi *(Precision)* mengukur seberapa banyak dari prediksi positif yang benar-benar positif. Rumusnya ditunjukkan pada Persamaan 7.

$$Precision = \frac{TP}{TP + FP} \quad (7)$$

Recall mengukur seberapa banyak dari seluruh data positif yang berhasil diklasifikasikan dengan benar. Rumusnya ditunjukkan pada Persamaan 8.

$$Recall = \frac{TP}{TP + FN} \quad (8)$$

F1-score merupakan rata-rata harmonik dari presisi dan recall, digunakan ketika distribusi kelas tidak seimbang. Rumusnya ditunjukkan pada Persamaan 9.

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (9)$$

Model klasifikasi diuji menggunakan data uji yang telah dipisahkan dari data latih dengan metode split. Pada tahap klasifikasi biner, model menunjukkan akurasi mencapai 95%, sedangkan pada tahap klasifikasi multikelas, diperoleh nilai akurasi, presisi, dan *recall* yang bervariasi untuk masing-masing kategori *cyberbullying*.

Namun, akurasi 95% ini perlu diuji lebih lanjut terhadap data nyata di luar sampel pelatihan, untuk memastikan model tidak mengalami *overfitting* terhadap data uji saat ini. Evaluasi menyeluruh ini penting sebagai dasar untuk mengukur keandalan sistem saat diterapkan dalam konteks dunia nyata yang lebih kompleks dan dinamis.

2.9 Integrasi ke *Web*

Sistem klasifikasi *cyberbullying* diimplementasikan dalam bentuk aplikasi web interaktif menggunakan *framework* Flask berbasis Python. Flask dipilih karena ringan, fleksibel, serta mendukung integrasi dengan pustaka *Natural Language Processing* (NLP) dan *machine learning* seperti Scikit-learn. Aplikasi dirancang untuk menerima input berupa teks *tweet*, memprosesnya otomatis, dan menampilkan hasil klasifikasi secara real-time.

Alur kerja sistem dibangun berdasarkan *pipeline* klasifikasi dua tahap dengan algoritma Multinomial Naïve Bayes (Bilgin & Bekar, 2025; Rosa et al., 2019b). Teks yang telah dipra-proses dikonversi menjadi representasi numerik menggunakan metode TF-IDF. Pra-pemrosesan mencakup *case folding*, *cleansing*, *tokenization*, *stopword removal*, dan *stemming*. Klasifikasi dilakukan dalam dua tahap:

1. Tahap pertama membedakan antara *tweet* yang mengandung *cyberbullying* dan yang tidak.
2. Tahap kedua mengklasifikasikan *tweet* bermuatan *cyberbullying* ke dalam enam kategori: SARA, pelecehan, penghinaan, *body shaming*, ancaman, dan kata kasar (Azumah et al., 2024; Mubeen et al., 2025).

Sistem dilengkapi fungsi detect_target() untuk membedakan ujaran yang ditujukan ke orang lain atau reflektif terhadap diri sendiri. *Tweet* yang bersifat *self-harm* tidak dikategorikan sebagai *cyberbullying*.

Antarmuka pengguna dirancang sederhana dan responsif, memungkinkan input manual, visualisasi hasil klasifikasi biner dan multikelas, serta tampilan *confidence score*. Sistem ini diharapkan berkontribusi dalam deteksi otomatis konten negatif berbahasa Indonesia untuk edukasi, penelitian, dan mitigasi risiko sosial digital.

3. Hasil dan Pembahasan

3.1 Hasil Evaluasi Klasifikasi Biner

Evaluasi pada tahap pertama bertujuan untuk membedakan antara tweet yang mengandung unsur *cyberbullying* dan yang bukan *cyberbullying*. Model klasifikasi yang digunakan adalah Multinomial Naïve Bayes, dengan fitur teks hasil ekstraksi TF-IDF dan N-gram.

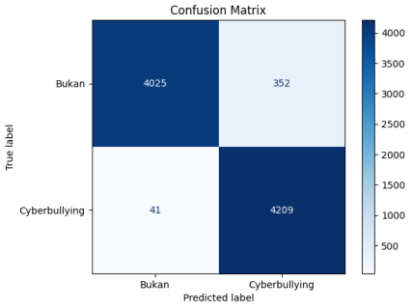
Hasil evaluasi menunjukkan performa yang sangat baik, dengan akurasi keseluruhan sebesar 95% pada data uji. Rincian metrik evaluasi ditampilkan pada Tabel 5, sementara Gambar 2 menampilkan confusion matrix sebagai representasi visual dari distribusi prediksi model terhadap data uji.

Nilai precision dan recall pada kedua kelas relatif seimbang. Kelas *Bukan Cyberbullying* memiliki precision sebesar 0.99, namun recall lebih rendah (0.92), artinya masih terdapat beberapa tweet normal yang salah terklasifikasi sebagai *cyberbullying* (false positive). Sebaliknya, kelas *Cyberbullying* memiliki recall yang sangat tinggi (0.99), menunjukkan bahwa hampir semua tweet yang mengandung unsur *cyberbullying* berhasil terdeteksi dengan baik.

Hal ini penting karena sistem lebih diutamakan untuk sensitif terhadap deteksi konten negatif, sehingga meminimalkan risiko terlewatnya tweet berbahaya. Meski demikian, adanya kesalahan klasifikasi pada tweet normal tetap menjadi catatan, karena dapat menimbulkan *false alarm*.

Tabel 5. Evaluasi Klasifikasi Biner

Label	Precision	Recall	F1-Score	Support
Bukan *Cyberbullying*	0.99	0.92	0.95	4377
Cyberbullying	0.92	0.99	0.96	4250
Accuracy			0.95	8627



Gambar 2. Confusion Matrix pada Klasifikasi Biner (*Cyberbullying* vs Bukan)

Gambar 2 menunjukkan bahwa mayoritas prediksi model berada pada diagonal utama confusion matrix, menandakan performa yang konsisten tinggi. Namun, masih terdapat 352 tweet normal yang salah diklasifikasikan sebagai *cyberbullying* serta 41 tweet *cyberbullying* yang salah diklasifikasikan sebagai normal. Jika ditinjau lebih dalam, kesalahan ini kemungkinan disebabkan oleh beberapa faktor:

- *False Positive* (352 tweet normal → *cyberbullying*): terjadi karena adanya penggunaan kata-kata bernuansa negatif dalam konteks bercanda, sarkasme, atau sindiran, sehingga model menganggapnya sebagai *cyberbullying*. Contoh kasus seperti penggunaan kata kasar dalam interaksi pertemanan, yang secara literal bersifat ofensif tetapi dalam konteks percakapan sebenarnya tidak bermaksud menyerang.
- *False Negative* (41 *tweet cyberbullying* → normal): meskipun jumlahnya relatif kecil, hal ini terjadi ketika tweet *cyberbullying* menggunakan bahasa halus, ironi, atau bentuk ujaran tidak langsung. Misalnya, tweet yang bersifat merendahkan tetapi dengan pilihan kata sopan, sehingga pola kata yang digunakan tidak cukup kuat untuk dikenali sebagai konten berbahaya oleh model.

Temuan ini konsisten dengan studi (Abdulloh & Hidayatullah, 2021) yang menunjukkan bahwa pendekatan Naïve Bayes efektif dalam mendeteksi konten negatif, namun masih memiliki keterbatasan dalam menangani variasi bahasa yang bersifat kontekstual.

Secara keseluruhan, hasil klasifikasi biner ini cukup andal untuk dijadikan tahap awal sebelum klasifikasi lebih lanjut ke kategori jenis *cyberbullying*. Tingginya recall pada kelas *Cyberbullying* menjadikan sistem ini sensitif dalam mendeteksi konten berbahaya, meski tetap diperlukan strategi tambahan untuk menekan jumlah *false positive*, seperti pemanfaatan analisis konteks kalimat yang lebih dalam.

3.2 Hasil Evaluasi Klasifikasi Multikelas

Commented [P1]: Untuk justifikasi pada strategi "dua tahap" ini berhasil, perlu dibandingkan dengan tanpa strategi "dua tahap"

Tahap kedua dari sistem klasifikasi dilakukan untuk mengidentifikasi jenis spesifik dari *cyberbullying* pada tweet yang sebelumnya telah difilter oleh klasifikasi biner. Model yang digunakan adalah Multinomial Naïve Bayes Multikelas, yang mengklasifikasikan tweet ke dalam tujuh kategori: *ambiguitas, ancaman, body shaming, kata kasar, pelecehan, penghinaan,* dan *SARA*. Evaluasi dilakukan menggunakan metrik *precision, recall,* dan *F1-score* terhadap data uji yang telah dilabeli secara manual.

Pada eksperimen pembanding, model satu tahap menghasilkan akurasi sebesar 74%. Nilai ini lebih rendah dibandingkan strategi dua tahap yang mencapai akurasi 95% pada klasifikasi biner dan 82% pada klasifikasi multikelas, seperti terlihat dalam Tabel 6.

Tabel 6. Perbandingan Performa Strategi

Pendekatan	Tahap Klasifikasi	Akurasi	Precision Rata-rata	Recall Rata-rata	F1-Score
Satu Tahap	Multikelas	74%	0.76	0.73	0.74
Dua Tahap	Biner	95%	0.96	0.95	0.95
	Multikelas	82%	0.84	0.85	0.84

CB = *Cyberbullying*

Hasil evaluasi ditampilkan pada Tabel 7, sedangkan distribusi visual dari hasil prediksi model divisualisasikan dalam bentuk *confusion matrix* pada Gambar 3.

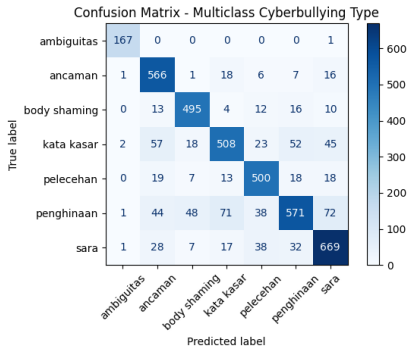
Secara keseluruhan, sistem ini mencapai akurasi global sebesar 82%, dengan nilai rata-rata precision 0.84, recall 0.85, dan F1-score 0.84. Nilai ini menunjukkan bahwa model bekerja cukup andal dalam mengidentifikasi jenis-jenis *cyberbullying* yang berbeda di media sosial. Beberapa temuan penting dari evaluasi:

1. **Ambiguitas**
   Kategori ini memperoleh hasil sangat baik dengan precision 0.97 dan recall 0.99. Hal ini menunjukkan bahwa model hampir tidak pernah salah mengenali tweet ambiguitas, meskipun jumlah datanya relatif lebih sedikit.

2. **Ancaman**
   Recall kategori ancaman sangat tinggi (0.92), menandakan hampir semua tweet ancaman berhasil terdeteksi. Namun, precision lebih rendah (0.78), yang berarti masih ada cukup banyak tweet non-ancaman yang salah diklasifikasikan sebagai ancaman.
3. **Body Shaming, Pelecehan, dan SARA**
   Ketiga kategori ini menunjukkan performa relatif stabil dengan F1-score antara 0.82–0.88. Model cukup mampu mengenali pola linguistik yang khas dari masing-masing kategori.
4. **Kata Kasar dan Penghinaan**
   Kedua kategori ini cenderung memiliki performa lebih rendah (F1-score 0.74–0.76). Hal ini disebabkan karena adanya tumpang tindih makna semantik antara keduanya, di mana kata-kata kasar sering digunakan dalam konteks penghinaan, sehingga model kesulitan membedakannya secara tegas.

Tabel 7. Evaluasi Klasifikasi Multikelas *Cyberbullying*

Label	Precision	Recall	F1-Score	Support
Ambiguitas	0.97	0.99	0.98	168
Ancaman	0.78	0.92	0.84	615
Body shaming	0.86	0.90	0.88	550
Kata Kasar	0.81	0.72	0.76	705
Pelecehan	0.81	0.87	0.84	575
Penghinaan	0.82	0.68	0.74	845
SARA	0.81	0.84	0.82	792
Accuracy			0.82	4250



Gambar 3. Confusion Matrix pada Klasifikasi Multikelas *Cyberbullying*

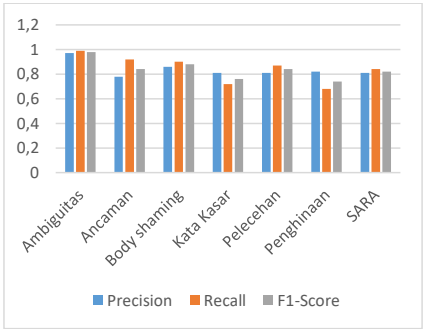
Gambar 3 memperlihatkan bahwa sebagian besar prediksi model berada pada diagonal utama, menandakan performa klasifikasi yang cukup konsisten. Namun, masih terlihat adanya kesalahan klasifikasi, khususnya:

- **Kata Kasar vs Penghinaan**: banyak terjadi salah prediksi karena kedua kategori ini memiliki kemiripan makna. Tweet dengan kata-kata ofensif bisa dianggap sekadar kasar, atau bisa juga dianggap sebagai hinaan langsung tergantung konteks.
- **Ancaman vs Kategori Lain**: beberapa tweet pelecehan atau hinaan diklasifikasikan sebagai ancaman karena sama-sama menggunakan kata-kata keras, meskipun maksudnya berbeda.
- **SARA vs Penghinaan/Kata Kasar**: tumpang tindih muncul ketika ujaran bernuansa SARA disampaikan dengan kata-kata kasar, sehingga model kadang salah mengklasifikasikannya sebagai penghinaan.

Analisis ini menunjukkan bahwa kesalahan klasifikasi umumnya muncul karena ambiguitas linguistik dan kontekstual. Bahasa di media sosial cenderung informal, penuh singkatan, dan mengandung sarkasme, sehingga sulit dipetakan ke satu kategori yang jelas. Temuan ini sejalan dengan penelitian Chamidah (2021) yang menjelaskan bahwa Naïve Bayes memiliki keterbatasan dalam menangani ambiguitas semantik pada teks, khususnya dalam bahasa informal seperti media sosial.

3.3 Visualisasi dan Interpretasi

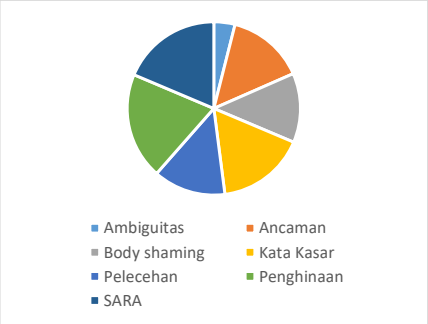
Visualisasi hasil klasifikasi dilakukan untuk memberikan gambaran menyeluruh terhadap performa model Multinomial Naïve Bayes dalam tahap klasifikasi multikelas. Dua jenis visualisasi digunakan dalam subbab ini, yaitu diagram batang (bar chart) untuk membandingkan nilai precision, recall, dan F1-score pada setiap kategori, serta diagram lingkaran (pie chart) untuk menampilkan distribusi jumlah data (*support*) pada masing-masing kategori.



Gambar 4. Visualisasi Precision, Recall, dan F1-Score pada Klasifikasi Multikelas

Gambar 4 menyajikan bar chart yang membandingkan metrik evaluasi pada masing-masing kategori *cyberbullying*. Berdasarkan grafik tersebut, kategori Body shaming, Pelecehan, dan SARA menunjukkan performa yang paling stabil dan tinggi, dengan nilai precision dan *recall* relatif seimbang dan berada di atas 0.82. Hal ini mengindikasikan bahwa model cukup andal dalam mengenali pola linguistik yang khas pada kategori-kategori tersebut.

Sebaliknya, kategori Ancaman memiliki nilai precision yang relatif lebih rendah dibandingkan recall. Kondisi ini menunjukkan bahwa model cenderung memberikan prediksi ancaman lebih sering, meskipun tidak seluruhnya benar, sehingga menyebabkan *false positive* lebih tinggi. Selain itu, kategori Kata Kasar dan Penghinaan juga memiliki performa lebih rendah dibanding kategori lainnya. Hal ini wajar karena secara semantik kedua kategori tersebut sering saling tumpang tindih, sehingga model kesulitan membedakannya secara tegas.



Gambar 5. Diagram Lingkaran Distribusi Data (Support) pada Setiap Kategori *Cyberbullying*

Gambar 5 menampilkan *pie chart* yang memperlihatkan distribusi jumlah tweet pada masing-masing kategori. Terlihat bahwa kategori Penghinaan dan SARA mendominasi dengan jumlah data yang lebih besar, sedangkan kategori Ambiguitas dan Ancaman memiliki proporsi paling sedikit. Ketidakseimbangan jumlah data antar kategori ini menjadi salah satu faktor penyebab turunnya performa model, terutama pada kategori minoritas seperti Ancaman dan Penghinaan.

Secara keseluruhan, visualisasi ini menunjukkan bahwa model bekerja cukup baik pada sebagian besar kategori. Namun, peningkatan akurasi masih diperlukan pada kategori dengan nilai precision atau recall yang rendah, serta pada kategori dengan distribusi data yang kecil. Untuk mengatasi hal ini, langkah selanjutnya dapat mencakup strategi *data balancing* atau *feature enrichment* agar model dapat melakukan generalisasi yang lebih baik pada kategori-kategori tersebut.

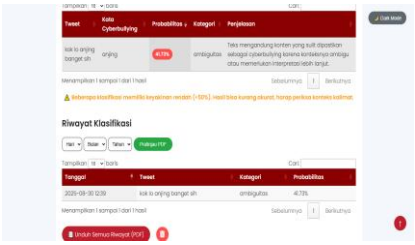
3.4 Implementasi Sistem Web

Aplikasi web dikembangkan menggunakan Flask dengan pendekatan klasifikasi dua tahap. Input berupa teks tweet terlebih dahulu melalui tahapan *pre-processing* (pembersihan teks, tokenisasi, *stopword removal*, dan stemming) lalu direpresentasikan menggunakan TF-IDF sebelum diproses oleh model Multinomial Naïve Bayes.

Jika tweet terdeteksi sebagai *cyberbullying*, sistem melanjutkan ke tahap kedua untuk mengklasifikasikan teks tersebut ke dalam tujuh kategori spesifik (Ambiguitas, Ancaman, Body shaming, Kata Kasar, Pelecehan, Penghinaan, dan SARA). Hasil klasifikasi kemudian ditampilkan dalam bentuk kategori, probabilitas (*confidence score*), serta penjelasan singkat mengenai alasan prediksi.

Selain itu, sistem dilengkapi dengan fitur riwayat klasifikasi yang mencatat setiap prediksi, termasuk tanggal, teks tweet, kategori hasil klasifikasi, serta nilai probabilitasnya. Dengan adanya fitur ini, pengguna dapat meninjau kembali hasil klasifikasi sebelumnya, termasuk mengunduh seluruh riwayat dalam format PDF.

Gambar 6 menampilkan antarmuka aplikasi web, di mana pengguna dapat memasukkan teks tweet, memperoleh hasil klasifikasi secara langsung, melihat tingkat probabilitas, serta memantau riwayat klasifikasi yang tersimpan.



Gambar 6. Antarmuka Aplikasi Web Klasifikasi *Cyberbullying* Berbasis Flask

Dengan keberadaan sistem ini, pendekatan klasifikasi dua tahap yang diusulkan tidak hanya terbukti secara teoritis, tetapi juga telah diimplementasikan secara praktis untuk kebutuhan riset, monitoring konten daring, maupun edukasi digital. Pengembangan sistem ke dalam bentuk aplikasi web juga membuka peluang integrasi dengan API media sosial untuk deteksi *cyberbullying* secara otomatis dan berskala besar.

3.5 Perbandingan dengan Penelitian Sebelumnya

Hasil klasifikasi dua tahap yang dikembangkan dalam penelitian ini menunjukkan performa yang kompetitif dibandingkan studi-studi terdahulu di bidang serupa. Pada tahap klasifikasi biner, sistem mampu mencapai akurasi mendekati sempurna, dengan *precision* dan *recall* yang tinggi pada kedua kelas, yaitu tweet yang mengandung *cyberbullying* dan yang tidak. Hal ini menunjukkan bahwa model efektif dalam mendeteksi seluruh konten negatif sekaligus meminimalisasi kesalahan klasifikasi terhadap tweet non-*cyberbullying*.

Capaian ini lebih baik dibandingkan penelitian (Abdulloh & Hidayatullah, 2021), yang melaporkan akurasi di atas 96% dalam klasifikasi konten negatif menggunakan pendekatan Naïve Bayes. Walaupun sama-sama menggunakan data dari media sosial, penelitian tersebut hanya menerapkan pendekatan satu tahap tanpa eksplorasi lebih lanjut ke dalam jenis-jenis ujaran negatif secara spesifik.

Pada tahap klasifikasi multikelas, sistem yang dikembangkan berhasil mengelompokkan tweet ke dalam tujuh kategori dengan akurasi keseluruhan sebesar 82%. Hasil ini menunjukkan bahwa pendekatan dua tahap memberikan keuntungan dalam memisahkan proses identifikasi awal *cyberbullying* dari pengelompokan kategori spesifik, sehingga meningkatkan akurasi sekaligus fleksibilitas dalam interpretasi hasil.

Keunggulan lain dari penelitian ini terletak pada pemanfaatan representasi fitur berbasis TF-IDF dan n-gram, serta integrasi fungsi *detect_target()* untuk menyaring ujaran reflektif atau bersifat *self-harm* agar tidak diklasifikasikan secara keliru. Pendekatan ini masih jarang diterapkan dalam penelitian terdahulu, sehingga menjadi kontribusi tambahan dalam mengembangkan sistem klasifikasi yang lebih kontekstual dan adaptif terhadap dinamika bahasa informal di media sosial.

Dengan membandingkan hasil evaluasi dan strategi teknis yang digunakan, dapat disimpulkan bahwa pendekatan klasifikasi dua tahap berbasis Multinomial Naïve Bayes dalam penelitian ini memberikan peningkatan signifikan baik dalam akurasi maupun granularitas hasil klasifikasi, dibandingkan studi-studi sebelumnya yang masih mengandalkan pendekatan konvensional.

3.6 Keterbatasan Sistem

Sistem klasifikasi dua tahap yang dikembangkan dalam penelitian ini sudah menunjukkan performa yang cukup baik, namun tetap memiliki sejumlah keterbatasan. Salah satu kelemahan utama terletak pada metode pengumpulan data berbasis kata kunci. Pendekatan ini cenderung hanya menangkap tweet dengan kata kasar atau ujaran negatif yang eksplisit, sementara bentuk *cyberbullying* yang lebih halus, implisit, atau sarkastik berpotensi tidak terdeteksi dengan baik.

Fungsi *detect_target()* yang dirancang untuk membedakan apakah ujaran ditujukan pada diri sendiri atau orang lain juga masih terbatas dalam mengenali konteks kalimat yang kompleks. Pada beberapa kasus, sistem masih rentan salah dalam mengklasifikasikan teks yang bersifat ambigu atau memiliki makna ganda, sehingga *confidence score* sering kali rendah (<50%).

Selain itu, sistem masih menghadapi kendala ketidakseimbangan distribusi data antar kategori. Misalnya, kategori ancaman dan penghinaan memiliki jumlah data yang relatif sedikit, sehingga performa model pada kelas minoritas ini lebih rendah dibandingkan kategori dominan seperti kata kasar atau SARA. Kondisi ini berdampak pada nilai precision dan recall yang tidak merata antar kelas.

Dari sisi teknis, model Naïve Bayes yang digunakan juga memiliki keterbatasan mendasar karena mengasumsikan independensi antar fitur. Dalam pemrosesan teks, makna sering kali bergantung pada urutan kata dan hubungan antarfrasa, yang tidak dapat sepenuhnya ditangkap oleh pendekatan ini. Akibatnya, sistem masih kalah dalam menangani konteks dibandingkan pendekatan berbasis *deep learning* (misalnya LSTM atau *transformer-based models*) yang lebih mampu memahami struktur semantik dan konteks linguistik. Secara keseluruhan, keterbatasan ini membuka ruang untuk pengembangan lebih lanjut, baik dari sisi pengayaan data, penanganan kelas minoritas, maupun eksplorasi algoritme yang lebih kompleks agar sistem dapat memberikan hasil klasifikasi yang lebih akurat dan kontekstual.

3.7 Implikasi Penelitian

Penelitian ini memperlihatkan bahwa algoritma klasik seperti Multinomial Naïve Bayes masih relevan dan efektif dalam mendeteksi *cyberbullying* berbahasa Indonesia. Dengan menerapkan pendekatan klasifikasi dua tahap, sistem mampu mencapai performa tinggi pada deteksi awal (biner) sekaligus memberikan pemetaan lebih detail ke dalam kategori perundungan spesifik.

Secara praktis, sistem yang dikembangkan berpotensi dimanfaatkan oleh berbagai pihak, seperti lembaga pendidikan, organisasi perlindungan anak, maupun pengelola platform media sosial, sebagai alat pemantau dan pencegah konten negatif. Tampilan aplikasi *web* yang interaktif juga mempermudah pengguna umum untuk menguji teks dan memahami jenis perundungan yang terdeteksi.

Dari sisi akademik, penelitian ini membuka arah baru bagi studi *multi-level classification* dalam konteks bahasa lokal, khususnya bahasa Indonesia. Integrasi antara *Natural Language Processing* (NLP) dan aplikasi web menjadikan sistem tidak hanya terbatas pada ranah penelitian, tetapi juga lebih aplikatif dan mudah diakses oleh publik.

Keberhasilan model ini sekaligus memberikan landasan bagi eksplorasi metode yang lebih maju, seperti IndoBERT atau model berbasis *deep learning* lainnya, untuk meningkatkan pemahaman konteks linguistik serta akurasi prediksi. Dengan demikian, penelitian ini diharapkan menjadi dasar pengembangan sistem pemantauan konten daring yang lebih adaptif, kontekstual, dan peka terhadap dinamika bahasa serta budaya digital di Indonesia.

4. Kesimpulan

Penelitian ini berhasil mengembangkan sistem klasifikasi dua tahap berbasis *Multinomial Naïve*

Bayes untuk mendeteksi konten *cyberbullying* dalam tweet berbahasa Indonesia. Pada tahap klasifikasi biner, sistem mencapai akurasi sebesar 95%, yang menunjukkan kemampuan yang andal dalam membedakan tweet yang mengandung *cyberbullying* dan yang tidak. Sementara itu, pada tahap klasifikasi multikelas, sistem memperoleh akurasi keseluruhan sebesar 82%, dengan performa terbaik ditunjukkan pada kategori Body shaming, Pelecehan, dan SARA. Implementasi dalam bentuk aplikasi web berbasis Flask memungkinkan proses klasifikasi berlangsung secara *real-time*, dilengkapi dengan tampilan kategori hasil, *confidence score*, serta riwayat prediksi. Dengan capaian tersebut, sistem ini dapat dikatakan efektif dalam mendukung deteksi dini konten bermuatan negatif, sekaligus membuka peluang pengembangan ke arah sistem yang lebih adaptif dan kontekstual sesuai dinamika bahasa media sosial.

Berdasarkan keterbatasan penelitian ini, beberapa arah pengembangan dapat dipertimbangkan untuk penelitian selanjutnya. Pertama, penerapan teknik *data balancing* seperti *oversampling* atau SMOTE disarankan untuk memperbaiki representasi kelas minoritas, khususnya kategori *Ancaman* dan *Ambiguitas*, yang cenderung kurang terwakili dalam dataset. Kedua, penggunaan representasi teks yang lebih kontekstual melalui word embeddings (misalnya *FastText* atau *Word2Vec*) maupun model berbasis *transformer* seperti IndoBERT berpotensi meningkatkan performa klasifikasi, terutama dalam menangani kasus ambiguitas dan konteks semantik yang kompleks.

Selain itu, pengembangan selanjutnya dapat mencakup perluasan fungsi *detect_target()* agar sistem tidak hanya mengidentifikasi jenis *cyberbullying*, tetapi juga mampu mendeteksi target spesifik dari ujaran. Uji coba sistem pada data real-time melalui integrasi dengan API media sosial juga penting dilakukan untuk memastikan robustness dan reliabilitas kinerja sistem dalam skenario aplikasi nyata.

Daftar Pustaka:

Abdulloh, N., & Hidayatullah, A. F. (2021). Deteksi Cyberbullying pada Cuitan Media Sosial Twitter. *Automata*, *Vol 1*(1), 1–5.

Alfarizi, M. I., Syafaah, L., & Lestandy, M. (2022). Emotional Text Classification Using TF-IDF (Term Frequency-Inverse Document Frequency) And LSTM (Long Short-Term Memory). *JUITA : Jurnal Informatika*, *10*(2), 225. https://doi.org/10.30595/juita.v10i2.13262

Azumah, S. W., Elsayed, N., Elsayed, Z., Ozer, M., & Guardia, A. La. (2024). Deep Learning Approaches for Detecting Adversarial Cyberbullying and Hate Speech in Social Networks. *2024 2nd International Conference on Artificial Intelligence, Blockchain, and Internet of Things, AIBThings 2024 - Proceedings*. https://doi.org/10.1109/AIBThings63359.2024.10863625

Azzahra, S. A., & Majid, N. W. A. (2025). Klasifikasi dan Analisis Semantik Cyberbullying Sosial Media X: Integrasi Web Scraping dan Natural Language Processing (NLP). *Jurnal Educatio FKIP UNMA*, *11*(2), 353–360. https://doi.org/10.31949/educatio.v11i2.12725

Bilgin, M., & Bekar, B. N. (2025). Turkish Cyberbullying Detection with Fine-Tuned Pre-Trained Language Models. *Bilişim Teknolojileri Dergisi*, *18*(2), 115–127. https://doi.org/10.17671/gazibtd.1528238

Chen, S., Wang, J., & He, K. (2024). Chinese Cyberbullying Detection Using XLNet and Deep Bi-LSTM Hybrid Model. *Information (Switzerland)*, *15*(2). https://doi.org/10.3390/info15020093

Chia, Z. L., Ptaszynski, M., Masui, F., Leliwa, G., & Wroczynski, M. (2021). Machine Learning and feature engineering-based study into sarcasm and irony classification with application to cyberbullying detection. *Information Processing and Management*, *58*(4), 1–33. https://doi.org/10.1016/j.ipm.2021.102600

Cuzzocrea, A., Akter, M. S., Shahriar, H., & Garcia Bringas, P. (2025). Cyberbullying Detection, Prevention, and Analysis on Social Media via Trustable LSTM-Autoencoder Networks over Synthetic Data: The TLA-NET Approach †. *Future Internet*, *17*(2). https://doi.org/10.3390/fi17020084

Farasalsabila, F., Utami, E., & Hanafi, H. (2024). Deteksi Cyberbullying Menggunakan BERT dan Bi-LSTM. *Jurnal Teknologi*, *17*(1), 1–6. https://doi.org/10.34151/jurtek.v17i1.4636

Fati, S. M., Muneer, A., Alwadain, A., & Balogun, A. O. (2023). Correction to: Cyberbullying Detection on Twitter Using Deep Learning-Based Attention Mechanisms and Continuous Bag of Words Feature Extraction (Mathematics, (2023), 11, 16, (3567), 10.3390/math11163567). *Mathematics*, *11*(21). https://doi.org/10.3390/math11214494

López-Vizcaíno, M. F., Nóvoa, F. J., Carneiro, V., & Cacheda, F. (2021). Early detection of cyberbullying on social media networks. *Future Generation Computer Systems*, *118*, 219–229. https://doi.org/10.1016/j.future.2021.01.006

Mubeen, M., Muskan, A., Akram, A., Rashid, J., Alshalali, T. A. N., & Sarwar, N. (2025). Cyberbullying-Related Automated Hate Speech Detection on Social Media Platforms Using Stack Ensemble Classification Method. *International Journal of Computational Intelligence Systems*, *18*(1). https://doi.org/10.1007/s44196-025-00919-z

Muneer, A., Alwadain, A., Ragab, M. G., & Alqushaibi, A. (2023). Cyberbullying Detection on Social Media Using Stacking Ensemble Learning and Enhanced BERT. *Information (Switzerland)*, *14*(8). https://doi.org/10.3390/info14080467

Ogunleye, B., & Dharmaraj, B. (2023). The Use of a Large Language Model for Cyberbullying Detection. *Analytics*, *2*(3), 694–707. https://doi.org/10.3390/analytics2030038

Philipo, A. G., Sarwatt, D. S., Ding, J., Daneshmand, M., & Ning, H. (2024). *Assessing Text Classification Methods for Cyberbullying Detection on Social Media Platforms*. 1–15. https://arxiv.org/pdf/2412.19928

Prabowo, W. A., & Azizah, F. (2020). Sentiment Analysis for Detecting Cyberbullying Using TF-IDF and SVM. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *4*(6). https://doi.org/10.29207/resti.v4i6.2753

Ridwan, M., & Muzakir, A. (2022). Model Klasifikasi Ujaran Kebencian pada Data Twitter dengan Menggunakan CNN-LSTM. *Teknomatika: Jurnal Teknologi & Informatika*, *12*(02), 209–218. https://ojs.palcomtech.ac.id/index.php/teknomatika/article/view/604

Rifai, H. S., Febrianti, S., & Santoso, I. (2023). Analisis Sentimen Tanggapan Masyarakat Terhadap Cyberbullying Di Media Sosial Menggunakan Algoritma Naïve Bayes (Nb). *Jurnal Ikraith-Informatika*, *7*(2), 183–196.

Rosa, H., Pereira, N., Ribeiro, R., Ferreira, P. C., Carvalho, J. P., Oliveira, S., Coheur, L., Paulino, P., Veiga Simão, A. M., & Trancoso, I. (2019a). Automatic cyberbullying detection: A systematic review. *Computers in Human Behavior*, *93*(October 2018), 333–345. https://doi.org/10.1016/j.chb.2018.12.021

Rosa, H., Pereira, N., Ribeiro, R., Ferreira, P. C., Carvalho, J. P., Oliveira, S., Coheur, L., Paulino, P., Veiga Simão, A. M., & Trancoso, I. (2019b). Automatic cyberbullying detection: A systematic review. *Computers in Human Behavior*, *93*, 333–345. https://doi.org/10.1016/J.CHB.2018.12.021

Wibisono, B., Machmud, A., Suryani, N., & Yunita, Y. (2025). Analisis Sentimen Cyberbullying Pada Komentar X Menggunakan Metode Naïve Bayes. *Computer Science (CO-SCIENCE)*, *5*(1), 8–16. https://doi.org/10.31294/coscience.v5i1.5152

Halaman ini sengaja dikosongkan