Journal of Applied Informatics and Computing (JAIC)
Vol.10, No.3, June 2026, pp. 2775~2789
e-ISSN: 2548-6861

Comparative Analysis of the Performance of Machine Learning Methods and Text Embedding Techniques in Classifying Toxic Conversations in the Roblox Game

Syifa Alfariani [1]*, Octa Dama Yanti [2], Syifa Naura Milla Celesta [3], Ken Ditha Tania [4], Ahmad Rifai [5]
* Sistem Informasi, Universitas Sriwijaya
syifaalfariani87@gmail.com [1], octadamayanti23@gmail.com [2], syifanauramillacelestaaa@gmail.com [3]

Article Info

Article history:

Received 2026-03-07
Revised 2026-04-21
Accepted 2026-04-30

Keyword:

Bag-Of-Words,
Machine Learning,
Roblox,
Text Classification,
TF-IDF,
Toxic Chat.

ABSTRACT

Online games have evolved into digital social spaces where player interactions often include toxic communication, potentially affecting user experience and psychological well-being, especially among younger players. This research is intended to examine and compare the performance of various machine learning algorithms in classifying toxic chat on the Roblox platform and to identify underlying linguistic patterns. The dataset consists of 7,119 Indonesian-language chat data labeled into six categories: identity_hate, insult, obscene, severe_toxic, threat, and toxic. The methodology includes data preprocessing, text representation using Bag-of-Words (BoW) and TF-IDF, and classification using Naive Bayes, Support Vector Machine (SVM), and Random Forest. To assess how well the model performs, several metrics are used, including accuracy, precision, recall, F1-score, and 3-fold cross-validation. The results show that SVM with TF-IDF achieves the best performance with 84.48% accuracy, followed closely by SVM with BoW. The findings indicate that while classical machine learning models remain effective, challenges persist in distinguishing linguistically similar categories.

I. PENDAHULUAN

Permainan daring (*online games*) telah berkembang jauh melampaui sekadar hiburan interaktif, mereka menjadi ruang sosial digital di mana jutaan pemain berinteraksi setiap hari melalui komunikasi teks dan suara. Interaksi ini sejatinya memperkaya pengalaman bermain secara kolaboratif, namun ironisnya juga telah membuka pintu bagi perilaku komunikasi yang merusak, secara kolektif dikenal sebagai percakapan *toxic* [1]. Fenomena ini mencakup ujaran kebencian, penghinaan, serta ekspresi verbal agresif lainnya yang dapat merusak pengalaman bermain dan berdampak pada kondisi psikologis pemain, khususnya pada kelompok usia muda yang lebih rentan [2]. Tingginya prevalensi perilaku ini di berbagai komunitas game global menunjukkan bahwa *toxic behavior* telah menjadi isu signifikan dalam ekosistem permainan daring modern [3].

Platform permainan daring yang cukup populer di kalangan anak dan remaja, Roblox, sering dipersepsikan sebagai ruang bermain yang ramah anak karena konten kreatif dan interaktifnya yang luas. Roblox sendiri adalah platform *game multiplayer* berbasis konten buatan pengguna yang sangat menarik minat anak-anak usia sekolah, dan banyak penelitian menunjukkan bahwa anak usia sekolah dasar dan remaja merupakan bagian signifikan dari basis pengguna aktifnya. Penelitian fenomenologis tentang motif penggunaan Roblox menunjukkan tingginya tingkat partisipasi anak usia sekolah dalam platform ini, yang mengindikasikan penetrasi yang kuat di kalangan pemain muda [4]. Strategi penargetan audiens Roblox berkaitan erat dengan usia muda, memberikan bukti akademik bahwa pengguna muda merupakan bagian besar dari komunitas pemain platform ini [5]. Namun, popularitas ini membawa paradoks, meskipun platform ini dirancang dan diposisikan sebagai ruang bermain yang aman dan edukatif, komunikasi antarpemain seringkali mengandung bentuk percakapan yang kontradiktif dengan nilai-nilai itu. Interaksi dalam *game multiplayer* tidak terbatas pada instruksi strategi atau kolaborasi kreatif. Mereka juga menyertakan ejekan verbal, hinaan, dan komentar provokatif. Kondisi tersebut tidak

hanya mengurangi kualitas pengalaman bermain, tetapi juga berpotensi memberikan pengaruh yang kurang baik terhadap perkembangan sosial serta emosional pemain, khususnya pada kelompok usia muda, sebuah dilema yang memicu pertanyaan kritis tentang bagaimana komunitas, pengembang, dan sistem moderasi harus merespons [3].

Isu percakapan *toxic* pada permainan daring juga merupakan topik penelitian yang sedang *trending* dalam literatur *online gaming* dan *human-computer interaction* [6]. Perilaku *toxic* merupakan salah satu hambatan utama dalam menciptakan komunitas game yang inklusif dan menyenangkan, sekaligus menyiratkan urgensi pendekatan teknologi yang lebih efektif untuk deteksi dan mitigasi [7]. Dalam konteks tersebut, pemanfaatan *machine learning* menjadi penting karena berbagai algoritma klasifikasi mampu memproses kumpulan data teks dalam jumlah besar serta mengidentifikasi pola bahasa yang membedakan percakapan *toxic* dan *non-toxic* [8]. Kombinasi teknik representasi teks (*text embedding*) dan beberapa algoritma untuk klasifikasi seperti Naive Bayes, Support Vector Machine (SVM), dan Random Forest menawarkan potensi solusi yang terukur dan dapat diukur secara objektif terhadap masalah ini.

Berbagai penelitian telah mengkaji deteksi percakapan *toxic* dalam permainan daring dengan pendekatan yang mengacu pada *machine learning* dan *deep learning*. Salah satu penelitian mengusulkan integrasi *word embedding* dan *valence lexicon* untuk menangkap aspek emosional dalam percakapan pemain. Pendekatan ini terbukti mampu meningkatkan kemampuan model dalam mengidentifikasi ujaran *toxic*, termasuk yang bersifat implisit dan tidak disampaikan secara langsung [9]. Penelitian lain mengeksplorasi perbandingan dua model *deep learning*, yakni Long Short-Term Memory (LSTM) dan Convolutional Neural Network (CNN), dalam klasifikasi percakapan *toxic*. Hasilnya mengindikasikan bahwa CNN memiliki tingkat kestabilan yang lebih baik dalam memproses teks chat yang cenderung singkat dan dinamis, sedangkan LSTM mengalami penurunan performa ketika distribusi data tidak seimbang [10]. Selain pendekatan *deep learning*, metode klasik juga telah digunakan dalam analisis percakapan *toxic*. Salah satu studi menerapkan algoritma Naive Bayes yang dikombinasikan dengan pembobotan TF-IDF serta seleksi fitur menggunakan *Information Gain*. Hasilnya menunjukkan bahwa penggunaan TF-IDF mampu meningkatkan performa klasifikasi dengan akurasi mencapai 75%, sehingga menegaskan bahwa representasi fitur memiliki peran penting dalam proses analisis teks [11].

Meskipun berbagai pendekatan telah menunjukkan hasil yang baik, sebagian besar penelitian masih berfokus pada lingkungan permainan kompetitif dengan pola komunikasi yang relatif seragam. Selain itu, penelitian umumnya lebih menitikberatkan pada peningkatan akurasi tanpa mengeksplorasi karakteristik linguistik secara mendalam. Padahal, platform seperti Roblox memiliki variasi interaksi dan gaya bahasa yang lebih kompleks, namun masih jarang dikaji secara spesifik. Di sisi lain, algoritma klasik seperti Naive Bayes, SVM, dan Random Forest cenderung berperan sebagai pembanding, sehingga potensinya dalam analisis yang lebih interpretatif belum banyak dimanfaatkan.

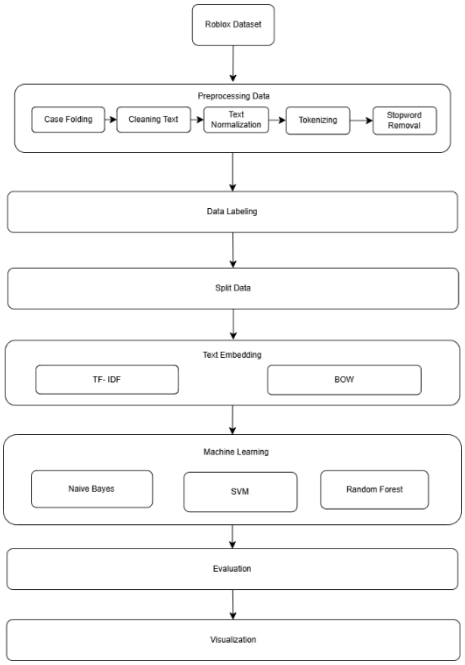
Berdasarkan hal tersebut, penelitian ini mengusulkan analisis komparatif algoritma *machine learning* klasik dalam klasifikasi percakapan *toxic* pada platform Roblox dengan pendekatan yang lebih komprehensif. Penelitian ini tidak hanya mengevaluasi performa model, tetapi juga mengintegrasikan eksplorasi data dalam kerangka *Knowledge Discovery in Databases* (KDD). Visualisasi *word cloud* digunakan untuk mengidentifikasi pola leksikal sebagai dasar pemahaman awal terhadap karakteristik bahasa dalam percakapan.

Penelitian ini bertujuan untuk mengevaluasi serta membandingkan performa algoritma Naive Bayes, SVM, dan Random Forest dalam proses klasifikasi percakapan *toxic* pada platform Roblox serta mengidentifikasi pola linguistik yang muncul dalam interaksi pengguna melalui pendekatan eksploratif. Kontribusi utama penelitian ini terletak pada integrasi antara analisis eksploratif dan proses klasifikasi dalam satu kerangka KDD, sehingga tidak hanya menghasilkan evaluasi performa model, tetapi juga pemahaman terhadap pola bahasa yang mendasari percakapan *toxic*. Selain itu, penelitian ini memberikan kontribusi kontekstual dengan mengkaji lingkungan Roblox yang memiliki karakteristik komunikasi lebih beragam dibandingkan platform game kompetitif, serta menunjukkan bahwa metode *machine learning* klasik tetap memiliki relevansi apabila dikombinasikan dengan analisis berbasis konteks dan eksplorasi data. Meskipun demikian, penelitian ini berfokus pada proses analisis dan evaluasi model, sehingga belum mencakup implementasi langsung pada sistem moderasi di lingkungan nyata.

II. METODE

Penelitian ini bertujuan untuk mengevaluasi serta membandingkan performa sejumlah algoritma *machine learning* dalam mendeteksi konten *toxic* pada percakapan dalam *game* Roblox. Dataset yang digunakan bersumber dari platform kaggle, mencakup percakapan berbahasa Indonesia yang terjadi selama permainan Roblox berlangsung. Secara keseluruhan, dataset tersebut mencakup 7.119 percakapan yang dikumpulkan selama aktivitas permainan.

Metodologi penelitian ini mencakup beberapa tahapan utama, yaitu *preprocessing data*, penerapan teknik *text embedding*, serta tahap klasifikasi yang memanfaatkan tiga pendekatan algoritma *machine learning*, yaitu Naive Bayes, Support Vector Machine, dan Random Forest.



Gambar 1. Alur penelitian

A. Preprocesssing Data

Tahap pra-pemrosesan data merupakan langkah awal dalam alur penelitian yang bertujuan mentransformasi data mentah menjadi data yang lebih terstruktur dan layak diguanakan pada tahap berikutnya. Pada proses ini, dilakukan seleksi terhadap data yang tidak relevan, data duplikat, maupun data yang berlebihan, sehingga hanya data yang memenuhi standar kualitas dan relevan dengan tujuan penelitian yang digunakan. Dengan demikian, tahap ini dapat meningkatkan kualitas dataset sebelum dilakukan proses analisis lebih lanjut [12]. Oleh karena itu, tahap ini berperan penting karena kinerja model sangat bergantung pada kualitas data yang digunakan.

1) *Case Folding* : Tahap ini seluruh teks diubah menjadi huruf kecil untuk menghindari perbedaan representasi kata akibat penggunaan huruf kapital [13]. Hal tersebut digunakan untuk mencegah terjadinya perbedaan makna atau pengenalan kata yang disebabkan oleh penggunaan huruf kapital.
2) *Cleaning Text* : Pada tahap ini, teks menjalani proses pembersihan dengan cara menghapus berbagai elemen yang tidak relevan, seperti tautan URL, angka, karakter-karakter khusus, serta spasi berlebih yang dapat memengaruhi proses pengolahan data selanjutnya [14].
3) *Text Normalization* : Tahap ini bertujuan untuk menyeragamkan bentuk kata, termasuk perbaikan ejaan, konversi kata tidak baku atau slang menjadi bentuk formal, serta penghapusan token yang tidak bermakna untuk meningkatkan konsistensi data [15].
4) *Tokenizing* : Tahap *tokenizing* teks dipecah menjadi unit-unit kecil (token) untuk mempermudah proses analisis dan pembentukan fitur [10]. Tahap ini berperan dalam mempermudah proses selanjutnya, seperti penghapusan *stopword* dan pembentukan representasi fitur melalui pendekatan seperti TF-IDF.
5) *Stopword Removal*: Tahap *stopword removal* memiliki tujuan untuk menyaring dan menghapus kata-kata umum dalam teks yang dinilai tidak memberikan kontribusi penting terhadap proses klasifikasi [14].

B. Data Labeling

Pelabelan data merupakan tahap penting dalam mendukung proses klasifikasi. Setiap data percakapan (*chat*) diberikan label yang sesuai berdasarkan konten dan kata-kata yang digunakan oleh pengguna. Pelabelan ini bertujuan untuk membandingkan kategori teks berdasarkan konteks yang terkandung di dalamnya [16].

Pelabelan data dilakukan menggunakan pendekatan semi-otomatis berbasis *trigger keyword* yang diimplementasikan melalui bahasa pemrograman Python. Peneliti terlebih dahulu menyusun daftar kata kunci yang merepresentasikan masing-masing kategori toksisitas berdasarkan karakteristik bahasa toxic pada percakapan game Roblox. Selanjutnya, setiap data percakapan dianalisis untuk mendeteksi keberadaan kata kunci tertentu sehingga komentar dapat diberi label secara otomatis sesuai kategori yang telah ditentukan.

Untuk meningkatkan validitas dan konsistensi pelabelan, proses verifikasi dilakukan oleh tiga orang annotator melalui pemeriksaan ulang terhadap hasil pelabelan otomatis. Apabila terdapat perbedaan penentuan label, dilakukan diskusi hingga mencapai kesepakatan akhir. Pendekatan ini bertujuan untuk meminimalkan subjektivitas serta meningkatkan reliabilitas dataset yang digunakan dalam penelitian [17].

Berdasarkan penelitian sebelumnya, dataset yang digunakan mengklasifikasikan konten berbahaya ke dalam enam kategori, meliputi *toxic*, *severe_toxic*, *obscene*, *threat*, *insult*, dan *identity_hate* [18]. Kategori-kategori tersebut digunakan untuk mengidentifikasi berbagai bentuk ujaran berbahaya dalam teks percakapan.

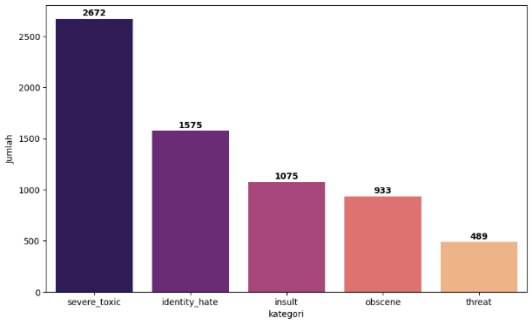
Kategori *toxic* merujuk pada teks yang mengandung unsur bahasa kasar, kalimat penuh kebencian, kata-kata tidak sopan, dan bahkan merendahkan orang lain [19]. *Severe_toxic* menunjukkan tingkat toksisitas yang lebih tinggi dan bersifat lebih agresif dibandingkan kategori *toxic* biasa. *Obscene* mencakup teks yang mengandung kata-kata vulgar atau berkonotasi seksual. *Threat* merujuk pada teks yang mengandung ancaman terhadap individu atau kelompok tertentu. *Insult* digunakan untuk mengidentifikasi teks yang berisi hinaan atau pernyataan ofensif. Sementara itu, *identity_hate* mengacu pada teks yang mengandung kebencian terhadap identitas tertentu seperti agama, ras, atau kelompok sosial [20].

Pada penelitian ini, klasifikasi tersebut digunakan sebagai dasar pelabelan data.

TABEL I
JUMLAH DATA PADA SETIAP KATEGORI

Label	Jumlah
identity_hate	1575
insult	1075
obscene	933
severe_toxic	2672
toxic	375
threat	489
Total	7119

Berdasarkan tabel tersebut diperoleh bahwa total data adalah 7.119, dengan data dengan kelas *severe_toxic* paling banyak yaitu 2.672 data.



Gambar 2. Distribusi kelas

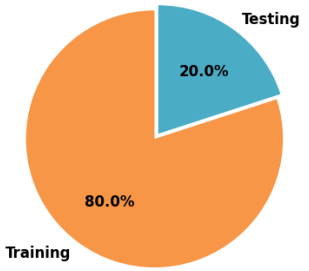
Berdasarkan distribusi kelas yang ada, terlihat bahwa jumlah data tiap kategori tidak merata. Kategori *severe toxic* mendominasi jumlah data dibamdingkan kategori lainnya, sedangkan kategori *toxic* dan *threat* memiliki jumlah data yang kecil. Ketidakseimbangan distribusi antar kelas seperti ini merupakan kondisi yang umum terjadi pada klasifikasi *multi-class* dan berpotensi mempengaruhi kinerja model, terutama dakam mengenali kelas yang memiliki representasi terbatas. Oleh sebab itu, evaluasi performa model tidak cukup hanya mengandalkan metrik akurasi saja, melainkan juga melibatkan precision, recall, dan F1-score untuk memperoleh penilaian yang lebih menyeluruh terhadap kemampuan model pada setiap kategori kelas.

C. Split Data

Setelah tahap *preprocessing*, dilakuan pemisahan data (*data splitting*). Tahap ini bertujuan untuk membagi keseluruhan dataset menjadi dua bagian utama, yaitu data pelatihan (*training data*) dan data pengujian (*testing data*) [21]. Pembagian ini dilakukan agar model dapat mempelajari pola dari sebagian data, lalu kemampuannya diuji pada data berbeda yang belum pernah dikenali selama proses pelatihan, sehingga penilaian terhadap performa model menjadi lebih objektif dan tidak bias.

Penelitian ini membagi dataset menggunakan perbandingan 80:20, di mana porsi 80% difungsikan sebagai data pelatihan guna membangun model, sementara 20% yang tersisa dialokasikan sebagai data pengujian untuk mengukur performa model yang telah dibangun.

Selain pemisahan dengan rasio 80:20, penelitian ini turut mengimplementasikan metode *k-fold cross-validation* dengan nilai k=3 untuk memperkuat reliabilitas hasil evaluasi model. Metode ini bekerja dengan membagi seluruh dataset menjadi 3 bagian (*fold*) yang berukuran sama. Pada setiap iterasi, satu *fold* berperan sebagai data pengujian, sementara dua *fold* lainnya digunakan sebagai data pelatihan. Siklus ini diulang sebanyak 3 kali hingga setiap fold mendapat giliran menjadi data pengujian tepat satu kali [21]. Nilai akurasi kemudian dirata-ratakan untuk memperoleh evaluasi yang lebih stabil dan tidak bias.



Gambar 3. Distribusi data latih dan data uji

D. Text Embedding

Text Embedding merupakan tahap penting karena algoritma klasifikasi tidak dapat langsung memproses data berbentuk teks. Oleh karena itu, data percakapan perlu ditransformasikan menjadi representasi vektor numerik agar dapat diolah oleh model [22].

Metode Bag-of-Words (BoW) digunakan untuk merepresentasikan teks berdasarkan frekuensi kemunculan kata dalam dokumen. Setiap kata unik direpresentasikan sebagai dimensi, dengan nilai yang menunjukkan jumlah kemunculannya dalam teks [23]. Meskipun sederhana, BoW cukup efektif dalam menangkap pola kata yang sering muncul pada percakapan *toxic*.

Selain BoW, digunakan juga metode *Term Frequency–Inverse Document Frequency* (TF-IDF) yang memberikan bobot lebih pada kata-kata yang lebih informatif dan menurunkan bobot kata yang terlalu umum [22]. Pendekatan ini menghasilkan representasi fitur yang lebih relevan dibandingkan BoW dalam klasifikasi teks *toxic*. Hasil representasi dari kedua metode tersebut kemudian digunakan sebagai input bagi algoritma *machine learning* dalam penelitian ini.

D. Machine Learning

Penelitian ini mengimplementasikan tiga algoritma klasifikasi dalam *machine learning*, yaitu Naive Bayes, Support Vector Machine (SVM), dan Random Forest. Ketiga algoritma tersebut dipilih karena masing-masing mewakili

pendekatan yang berbeda, mulai dari pendekatan probabilistik, *margin-based classifier*, hingga metode *ensemble* dalam klasifikasi teks.

Naive Bayes bekerja berdasarkan Teorema Bayes dengan memanfaatkan distribusi frekuensi kata, sehingga efektif untuk data berdimensi tinggi [24]. Dalam penelitian ini, algoritma ini diterapkan pada representasi BoW dan TF-IDF serta digunakan sebagai model dasar untuk mengevaluasi efektivitas fitur berbasis frekuensi.

Pemilihan SVM dalam penelitian ini, karena kemampuannya dalam membangun bidang pemisah optimal antar kelas di dalam ruang fitur berdimensi tinggi [14]. Dikombinasikan dengan representasi seperti TF-IDF, SVM mampu menangkap pola yang lebih kompleks, termasuk hubungan antar kata, sehingga berpotensi meningkatkan akurasi klasifikasi teks.

Selanjutnya, Random Forest digunakan sebagai representasi pendekatan *ensemble* yang mengombinasikan sejumlah *decision tree* untuk memperkuat kestabilan dan meningkatkan ketepatan hasil prediksi [14]. Algoritma ini mampu menangani hubungan non-linear antar fitur, sehingga diharapkan dapat menangkap pola ujaran toxic yang lebih kompleks. Dalam penelitian ini, Random Forest diterapkan pada fitur BoW dan TF-IDF untuk membandingkan kinerjanya terhadap metode lain.

E. Konfigurasi Parameter Model

Pada tahap implementasi, setiap algoritma dikonfigurasi dengan parameter tertentu untuk memastikan proses klasifikasi berjalan konsisten, dapat direplikasi, dan menghasilkan performa yang optimal. Pemilihan parameter ini berperan penting karena berpengaruh langsung terhadap kinerja dan kemampuan generalisasi model [25].

Pada representasi TF-IDF, digunakan *max_features* sebesar 10.000 dan *min_df* = 2 untuk menyaring kata yang jarang muncul. Selain itu, diterapkan *n-gram* (1,2) untuk menangkap pola kata dan frasa, serta *sublinear_tf* untuk mengurangi dominasi kata dengan frekuensi tinggi.

Model SVM diimplementasikan menggunakan LinearSVC dengan kernel linear dan parameter C = 1.0, yang dipilih karena efisien untuk data teks berdimensi tinggi [14]. Random Forest menggunakan 100 pohon (*n_estimators = 100*) dan *random_state* = 42 untuk menjaga konsistensi hasil. Sementara itu, Naive Bayes menggunakan MultinomialNB dengan parameter default.

Konfigurasi tersebut diterapkan secara konsisten pada seluruh skenario pengujian untuk memastikan validitas dan keterbandingan hasil antar model.

F. Evaluasi Kinerja Model

Evaluasi kinerja model digunakan untuk mengetahui tingkat efektivitas setiap metode dalam mengklasifikasikan percakapan *toxic* ke dalam enam kategori, yaitu *identity_hate, insult, obscene, severe_toxic, threat,* dan *toxic.*

Dalam evaluasi, *True Positive* (TP) menunjukkan prediksi benar pada suatu kategori. *False Positive* (FP) menunjukkan prediksi salah di kategori tersebut. *False Negative* (FN) menggambarkan data yang seharusnya masuk kategori tetapi tidak terdeteksi. Sementara itu, *True Negative* (TN) adalah data yang benar-benar bukan bagian dari kategori tersebut [26].

1) Metrik accuracy digunakan untuk mengukur proporsi prediksi benar secara keseluruhan, namun tidak dijadikan satu-satunya acuan karena potensi ketidakseimbangan data.

$$\text{Akurasi} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

2) Metrik precision digunakan untuk menilai tingkat ketepatan model dalam melakukan prediksi terhadap suatu kelas..

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

3) Metrik recall digunakan untuk menilai kemampuan model dalam mengidentifikasi seluruh data yang benar-benar termasuk dalam suatu kelas tertentu.

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

4) F1-score merupakan metrik yang digunakan untuk menyeimbangkan nilai precision dan recall. Pengukuran ini memberikan gambaran yang lebih komprehensif terhadap kinerja model, khususnya pada permasalahan klasifikasi multikelas.

$$\text{F1-score} = \frac{2*(Presicion*Recall)}{presicion+Recall} \quad (4)$$

Selain metrik numerik, juga digunakan confusion matrix untuk menggambarkan performa model setiap kelas secara rinci. Confusion matrix memungkinkan analisis kesalahan klasifikasi, seperti kecenderungan model salah mengklasifikasikan ujaran *toxic* sebagai *obscene* atau sebaliknya [27].

Evaluasi dilakukan dengan menggunakan skenario klasifikasi multikelas untuk memastikan bahwa model tidak hanya memiliki kinerja yang baik pada satu kelas tertentu, namun juga mampu memberikan performa yang konsisten dan seimbang pada seluruh kategori percakapan *toxic* yang dianalisis.

III. HASIL DAN PEMBAHASAN

Implementasi penelitian dilakukan Python dengan memanfaatkan library Scikit-learn, Pandas, dan NumPy untuk mendukung proses pengolahan data dan klasifikasi.

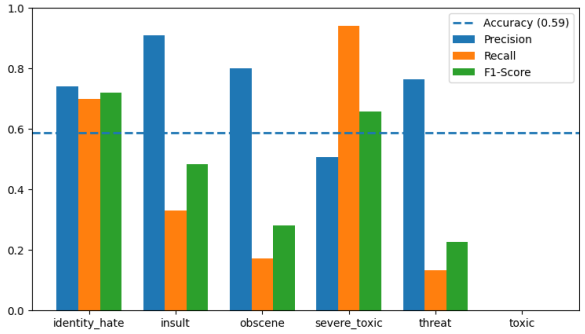
Berdasarkan tahapan metodologi yang telah dijelaskan sebelumnya, eksperimen dilakukan pada dataset percakapan yang terdiri dari lebih dari 7.000 data dengan enam kategori klasifikasi. Hasil pengujian kemudian dianalisis untuk mengevaluasi kinerja model dalam mengklasifikasikan percakapan *toxic* berdasarkan berbagai skenario representasi fitur dan algoritma yang digunakan.

Pembahasan dalam penelitian ini berfokus pada analisis perbandingan kinerja model berdasarkan dua metode representasi teks, yaitu TF-IDF dan Bag-of-Words, serta tiga

algoritma klasifikasi yang digunakan, yakni Naive Bayes, SVM, dan Random Forest. Evaluasi kinerja dilakukan dengan memanfaatkan metrik akurasi, precision, recall, dan F1-score untuk memberikan gambaran yang lebih menyeluruh terhadap kemampuan model dalam mengklasifikasikan setiap kategori percakapan toxic.

A. Model Menggunakan TF-IDF

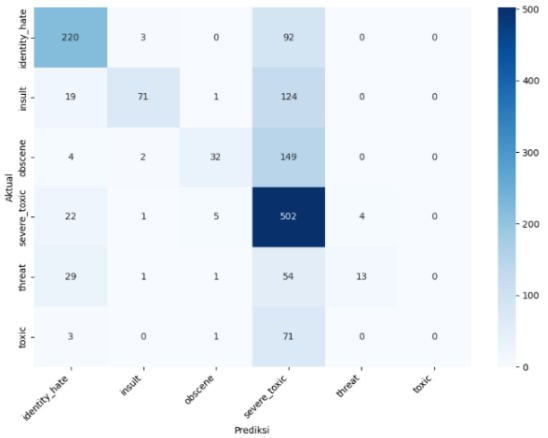
Eksperimen pertama menggunakan TF-IDF sebagai teknik konversi teks menjadi representasi numerik, yang selanjutnya diproses melalui algoritma Naive Bayes untuk keperluan klasifikasi. Model dibangun dengan memanfaatkan data latih, kemudian diuji kemampuannya terhadap data uji melalui tahap prediksi [21]. Evaluasi dilakukan dengan mengacu pada empat metrik utama, yaitu akurasi, presisi, recall, dan F1-score, selanjutnya hasil evaluasi tersebut ditampilkan dalam visual diagram batang untuk mempermudah perbandingan performa antar kelas.



Gambar 4. Performa Naive Bayes menggunakan TF-IDF

Berdasarkan hasil pengujian, kombinasi TF-IDF dan Naive Bayes menghasilkan akurasi sebesar 58,85% pada 1.424 data uji, yang menunjukkan bahwa performa model masih belum optimal untuk klasifikasi enam kategori. Secara umum, kelas *identity_hate* memiliki performa relatif baik (F1-score 0,72). Namun, kelas *insult* dan *obscene* menunjukkan ketidakseimbangan antara precision dan recall, yang mengindikasikan banyak data tidak berhasil terdeteksi. Sebaliknya, kelas *severe_toxic* memiliki recall sangat tinggi (0,94) tetapi precision rendah, yang menandakan kecenderungan model melakukan over-prediction pada kelas tersebut. Kelas *threat* memiliki performa rendah (F1-score 0,23), sementara kelas *toxic* tidak terdeteksi sama sekali.

Temuan ini mengindikasikan bahwa model mengalami bias prediksi terhadap kelas *severe_toxic*, sehingga distribusi klasifikasi menjadi tidak seimbang dan berdampak pada rendahnya kemampuan model dalam mengenali kelas lain secara akurat.

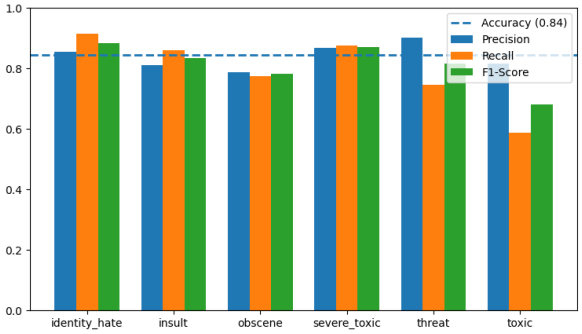


Gambar 5. Confusion matrix Naive Bayes menggunakan TF-IDF

Confusion matrix pada gambar 5 memperlihatkan bahwa kesalahan klasifikasi didominasi oleh perpindahan prediksi ke kelas *severe_toxic*. Sebagai contoh, sejumlah besar data dari kelas *identity_hate*, *insult*, dan *obscene* salah diklasifikasikan ke kelas tersebut. Pola ini konsisten dengan tingginya nilai recall namun rendahnya precision pada *severe_toxic*, yang menunjukkan bahwa model cenderung menggeneralisasi berbagai pola teks ke satu kategori dominan.

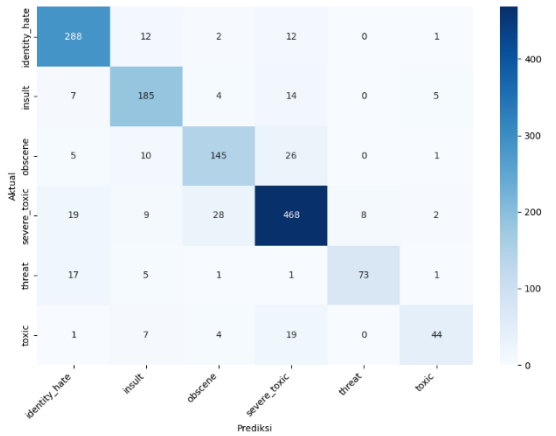
Selain itu, kelas *toxic* menjadi kategori paling bermasalah karena seluruh datanya gagal dikenali dengan benar dan sebagian besar dialihkan ke *severe_toxic*. Hal ini menunjukkan bahwa Naive Bayes dengan TF-IDF belum mampu membedakan kelas-kelas yang memiliki kemiripan leksikal tinggi. Dengan demikian, keterbatasan utama model terletak pada ketidakmampuannya membangun batas klasifikasi yang jelas antar kategori dengan karakteristik bahasa yang saling beririsan.

Selain pendekatan probabilistik menggunakan Naive Bayes, penelitian ini juga menerapkan algoritma SVM dengan representasi fitur TF-IDF. Model tersebut dilatih menggunakan dataset pelatihan yang sama dan kemudian diuji pada 1.424 data uji untuk memastikan konsistensi dalam proses perbandingan kinerja antar metode.



Gambar 6. Performa SVM menggunakan TF-IDF

Hasil pengujian menunjukkan bahwa kombinasi TF-IDF dan SVM menghasilkan performa terbaik dibandingkan model lainnya, dengan akurasi mencapai 84,48%. Kinerja ini mengindikasikan bahwa SVM mampu memanfaatkan representasi fitur berbasis TF-IDF secara lebih optimal dalam membedakan pola linguistik antar kelas. Meskipun model TF-IDF dengan SVM menunjukkan tingkat akurasi yang relatif tinggi, yaitu sebesar 84,48%, nilai tersebut tidak dapat langsung diartikan bahwa model telah bekerja optimal. Hal ini disebabkan oleh kompleksitas permasalahan klasifikasi multikelas dengan enam kategori yang memiliki karakteristik bahasa yang saling beririsan. Analisis lebih lanjut berdasarkan metrik evaluasi lainnya menunjukkan bahwa kinerja model belum merata pada seluruh kelas. Sebagai gambaran, kategori *toxic* memiliki nilai F1-score yang lebih rendah (0,68) dibandingkan kategori lainnya, yang mengindikasikan bahwa model masih mengalami kendala dalam mengidentifikasi kelas tersebut secara konsisten. Selain itu, kelas *threat* juga menunjukkan ketidakseimbangan antara precision dan recall, yang mengindikasikan adanya kesalahan dalam proses deteksi.



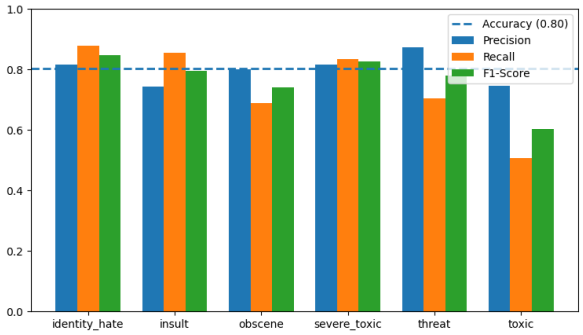
Gambar 7. Confusion matrix SVM menggunakan TF-IDF

Hasil confusion matrix pada gambar 7 turut memperkuat temuan tersebut, di mana masih terdapat kesalahan klasifikasi yang cukup dominan pada kelas-kelas dengan karakteristik serupa, seperti antara *obscene* dan *severe_toxic*, serta antara *toxic* dan *severe_toxic*. Pola ini menunjukkan bahwa model masih mengalami kesulitan dalam membedakan tingkat intensitas ujaran yang memiliki kemiripan leksikal.

Dengan demikian, meskipun nilai akurasi yang diperoleh tergolong tinggi, evaluasi model tidak dapat hanya bergantung pada satu metrik tersebut. Hal ini menunjukkan perlunya analisis yang lebih mendalam terhadap distribusi performa tiap kelas, sehingga kemampuan model dalam menangani seluruh kategori dapat dievaluasi secara lebih akurat dan tidak bias terhadap kelas tertentu.

Meskipun model SVM menunjukkan performa yang lebih baik dibandingkan Naive Bayes, hasil ini tetap mengindikasikan bahwa tantangan utama terletak pada kemampuan model dalam membedakan kelas-kelas dengan karakteristik linguistik yang mirip.

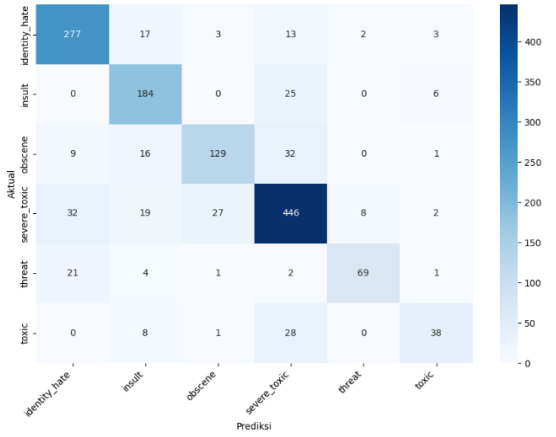
Eksperimen selanjutnya adalah menggunakan algoritma Random Forest dengan 100 pohon keputusan (n_estimators = 100) dan parameter random_state untuk menjaga konsistensi hasil. Model ini menerapkan *ensemble learning* dengan menggabungkan prediksi dari beberapa decision tree yang dibangun dari variasi data dan fitur.



Gambar 8. Performa Random Forest menggunakan TF-IDF

Hasil pengujian menunjukkan bahwa kombinasi TF-IDF dan Random Forest menghasilkan akurasi sebesar 80,27% pada 1.424 data uji. Secara umum, performa model tergolong cukup stabil, dengan sebagian besar kelas memiliki F1-score di atas 0,74, seperti *identity_hate*, *insult*, *obscene*, dan *severe_toxic*. Namun demikian, ketidakseimbangan masih terlihat pada beberapa kelas. Kategori *threat* memiliki precision tinggi (0,87) tetapi recall lebih rendah (0,70), yang menunjukkan bahwa tidak semua data berhasil terdeteksi. Sementara itu, kelas *toxic* memiliki performa terendah (F1-score 0,60), yang mengindikasikan bahwa model masih mengalami kesulitan dalam mengenali kategori tersebut secara konsisten.

Random Forest mampu memberikan performa yang relatif merata antar kelas, namun masih berada di bawah SVM baik dari sisi akurasi maupun konsistensi hasil. Hal ini menunjukkan bahwa pendekatan ensemble belum sepenuhnya mampu mengatasi kompleksitas klasifikasi multikelas pada data dengan karakteristik bahasa yang beririsan.



Gambar 9. Confusion matrix Random Forest menggunakan TF-IDF

Confusion matrix pada gambar 9 menunjukkan bahwa pola kesalahan klasifikasi cenderung serupa dengan SVM, tetapi dengan distribusi kesalahan yang lebih luas. Kesalahan terbesar terjadi pada kelas dengan kemiripan karakteristik, terutama antara *obscene* dan *severe_toxic*. Selain itu, kelas *severe_toxic* juga masih sering tertukar dengan *identity_hate* dan *obscene*, yang menunjukkan bahwa model belum mampu membedakan tingkat intensitas ujaran secara jelas.

Kelas *threat* juga menunjukkan pola kesalahan yang konsisten, di mana sebagian data diprediksi sebagai *identity_hate*. Hal ini mengindikasikan adanya kemiripan konteks antar kedua kategori yang sulit dibedakan oleh model. Sementara itu, kelas *toxic* kembali menjadi kategori paling sulit, dengan jumlah prediksi benar yang relatif rendah dan sebagian besar kesalahan mengarah ke *severe_toxic*.

Dengan demikian, meskipun Random Forest menghasilkan performa yang cukup baik, distribusi kesalahan menunjukkan bahwa model masih menghadapi kendala dalam membedakan kelas dengan karakteristik linguistik yang berdekatan. Hal ini menjelaskan mengapa akurasi keseluruhan masih lebih rendah dibandingkan SVM.

TABEL 2
PERFORMA MODEL MENGGUNAKAN TF-IDF

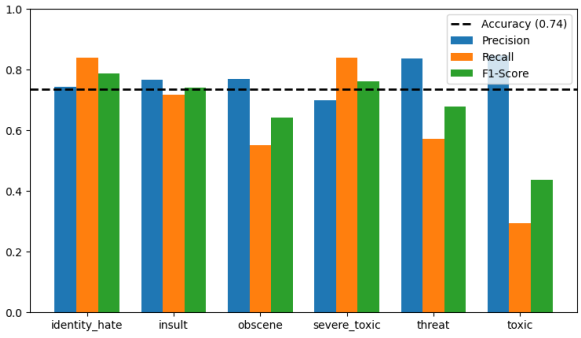
Model	***Class***	***Accuracy***	***Precision***	***Recall***	***F1-score***
Naive Bayes	*Identity_hate (0)*	0,5885	0.72	0.70	0.72
	Insult (1)		0.91	0.33	0.48
	obscene (2)		0.80	0.17	0.28
	severe_toxic (3) *Threat (4)* *Toxic (5)*		0.51 0.76 0.00	0.94 0.13 0.00	0.66 0.23 0.00
Support Vector Machine	*Identity_hate (0)*	0.8448	0.85	0.91	0.88
	Insult (1)		0.81	0.86	0.84
	obscene (2)		0.79	0.78	0.78
	severe_toxic (3) *Threat (4)* *Toxic (5)*		0.90 0.81	0.74 0.59	0.82 0.68
Random Forest	*Identity_hate (0)*	0.8026	0.82	0.88	0.85
	Insult (1)		0.74	0.86	0.79
	obscene (2)		0.80	0.69	0.74
	severe_toxic (3) *Threat (4)* *Toxic (5)*		0.82 0.87 0.75	0.84 0.70 0.51	0.83 0.78 0.60

Berdasarkan keseluruhan hasil eksperimen pada representasi TF-IDF, dapat disimpulkan bahwa:

1. SVM menunjukkan performa terbaik dengan akurasi tertinggi (0,8448) serta distribusi metrik evaluasi yang relatif lebih seimbang dibandingkan model lainnya. Meskipun demikian, performa tersebut belum sepenuhnya optimal karena masih terdapat ketidakseimbangan pada beberapa kelas, terutama pada kategori toxic yang memiliki nilai F1-score lebih rendah.
2. Naive Bayes memberikan performa kompetitif (0,5885) dan dapat dijadikan baseline yang kuat karena kesederhanaan dan efisiensinya.
3. Random Forest menunjukkan performa yang lebih rendah (0,8026), dengan penurunan paling signifikan terlihat pada kelas *toxic*

Temuan ini menunjukkan bahwa meskipun SVM memberikan performa terbaik, tantangan utama dalam klasifikasi multikelas tetap terletak pada kemiripan antar kategori, sehingga evaluasi tidak cukup hanya berdasarkan akurasi, tetapi juga perlu mempertimbangkan distribusi kesalahan pada tiap kelas.

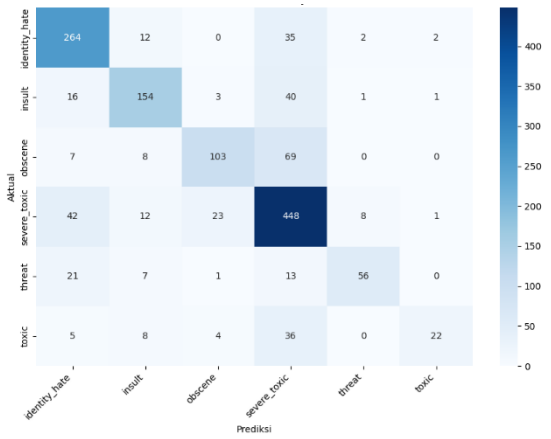
B. Model Menggunakan Bag-of-Words (BoW)



Gambar 10. Performa Naive Bayes menggunakan BoW

Model BoW dengan Naive Bayes menghasilkan akurasi sebesar 73,53% pada 1.424 data uji, yang menunjukkan peningkatan performa dibandingkan penggunaan TF-IDF pada algoritma yang sama. Hal ini mengindikasikan bahwa representasi berbasis frekuensi kata lebih sesuai untuk pendekatan probabilistik seperti Naive Bayes.

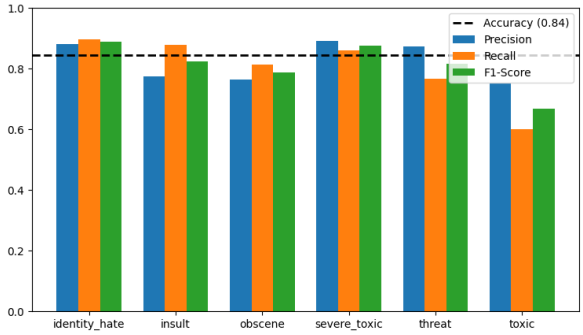
Secara umum, kelas *identity_hate* dan *severe_toxic* memiliki performa yang cukup baik (F1-score masing-masing 0,79 dan 0,76), sementara *insult* menunjukkan hasil yang relatif seimbang. Namun, beberapa ketidakseimbangan masih terlihat, seperti pada kelas *obscene* yang memiliki recall lebih rendah, serta kelas *threat* yang menunjukkan perbedaan cukup besar antara precision dan recall. Kelas *toxic* tetap menjadi kategori dengan performa terendah (F1-score 0,44), yang menandakan bahwa model masih kesulitan mengenali pola pada kelas tersebut.



Gambar 11. Confusion matrix Naive Bayes menggunakan BoW

Confusion matrix pada gambar 11 menunjukkan bahwa kesalahan klasifikasi masih didominasi oleh kelas dengan kemiripan karakteristik, terutama *obscene* yang sering diprediksi sebagai *severe_toxic*. Selain itu, kelas *threat* juga cukup sering salah diklasifikasikan sebagai *identity_hate*, yang menunjukkan adanya kedekatan konteks antar kategori.

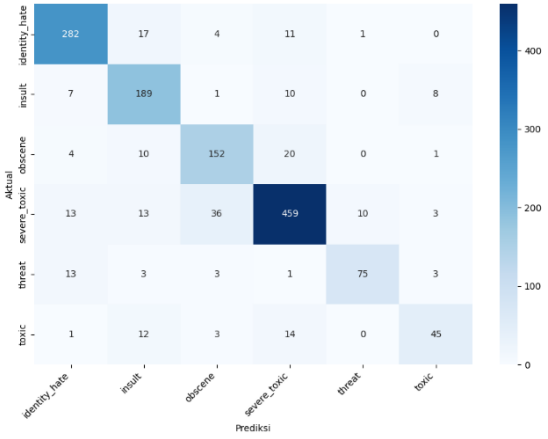
Kelas *toxic* tetap menjadi yang paling sulit dikenali, dengan sebagian besar data salah diprediksi ke *severe_toxic*. Meskipun demikian, dibandingkan dengan TF-IDF, penggunaan BoW memberikan peningkatan dalam mendeteksi kelas ini.



Gambar 12. Performa SVM menggunakan BoW

Hasil pengujian menunjukkan bahwa model BoW dengan SVM mencapai akurasi sebesar 84,41% pada 1.424 data uji, yang menandakan performa yang sangat baik dalam klasifikasi enam kategori. Sebagian besar kelas memiliki nilai F1-score yang relatif seimbang (di atas 0,79), seperti *identity_hate*, *insult*, *obscene*, dan *severe_toxic*. Kelas *threat* juga menunjukkan performa yang cukup baik, sementara *toxic* tetap menjadi kategori dengan nilai terendah (F1-score 0,67), meskipun mengalami peningkatan dibandingkan model lain.

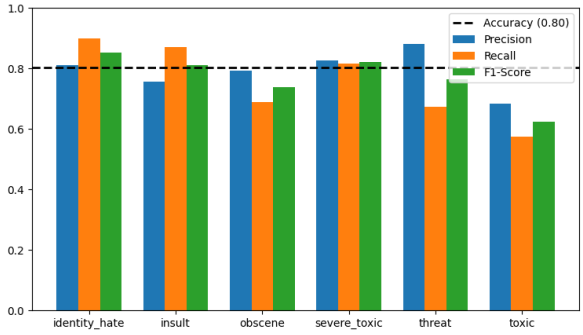
Kombinasi BoW dan SVM menunjukkan performa yang stabil, yang mengindikasikan kemampuan model dalam menangkap pola umum pada data. Namun, perbedaan performa antar kelas menunjukkan bahwa model belum sepenuhnya mampu mengatasi variasi karakteristik antar kategori.



Gambar 13. Confusion matrix SVM menggunakan BoW

Confusion matrix pada gambar 13 menunjukkan bahwa kesalahan klasifikasi relatif lebih kecil dan merata dibandingkan model BoW lainnya. Namun demikian, kesalahan masih terjadi pada kelas dengan karakteristik yang mirip, seperti *severe_toxic* yang terkadang diprediksi sebagai *obscene*. Selain itu, terdapat juga kesalahan antara *identity_hate* dan *insult*, yang mengindikasikan adanya kemiripan konteks antar kedua kelas.

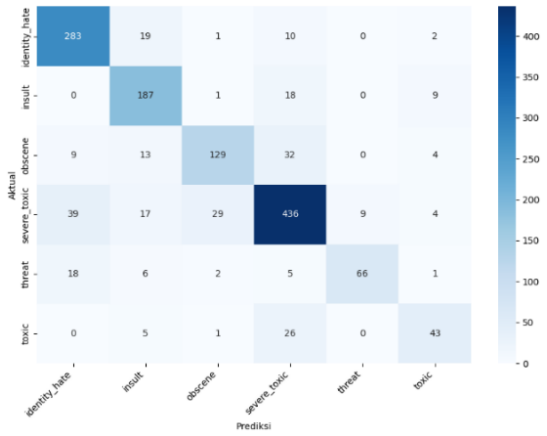
Kelas *toxic* menunjukkan peningkatan dibandingkan model lain, meskipun sebagian kesalahan masih mengarah ke *severe_toxic*. Secara keseluruhan, pola ini menunjukkan bahwa meskipun performa model sudah cukup baik, tantangan utama tetap terletak pada pembedaan kelas dengan kedekatan makna.



Gambar 14. Performa Random Forest menggunakan BoW

Hasil pengujian menunjukkan bahwa model BoW dengan Random Forest mencapai akurasi sebesar 80,34% pada 1.424 data uji, yang mencerminkan performa yang cukup baik dalam klasifikasi enam kategori. Sebagian besar kelas memiliki F1-score di atas 0,74, seperti *identity_hate*, *insult*, *obscene*, dan *severe_toxic*. Namun, ketidakseimbangan masih terlihat pada beberapa kelas. Kategori *threat* memiliki precision tinggi tetapi recall lebih rendah, sedangkan *toxic* tetap menjadi kelas dengan performa terendah (F1-score 0,62), yang menunjukkan bahwa model masih kesulitan mengenali kategori tersebut secara konsisten.

Secara umum, Random Forest menghasilkan performa yang relatif seimbang antar kelas, tetapi masih berada di bawah SVM dalam hal akurasi dan konsistensi, sehingga menunjukkan keterbatasan dalam menangani variasi karakteristik data teks.



Gambar 15. Confusion matrix Random Forest menggunakan BoW

Confusion matrix pada gambar 15 menunjukkan bahwa kesalahan klasifikasi masih didominasi oleh kelas dengan karakteristik yang mirip, terutama antara *obscene* dan *severe_toxic*. Selain itu, kelas *severe_toxic* juga cukup sering tertukar dengan *identity_hate* dan *obscene*, yang mengindikasikan bahwa model belum mampu membedakan tingkat intensitas ujaran secara jelas.

Kelas *threat* juga menunjukkan kesalahan yang konsisten dengan *identity_hate*, sementara kelas *toxic* meskipun mengalami peningkatan, masih sering salah diprediksi ke *severe_toxic*. Secara keseluruhan, distribusi kesalahan yang masih menyebar ini menjelaskan mengapa performa Random Forest belum seoptimal SVM dalam menjaga konsistensi klasifikasi antar kelas.

TABEL 3
PERFORMA MODEL MENGGUNAKAN BOW

Model	***Class***	***Accuracy***	***Precision***	***Recall***	***F1-score***
Naive Bayes	*Identity_hate (0)*	0.7352	0.74	0.84	0.79
	Insult (1)		0.77	0.72	0.74
	obscene (2)		0.77	0.55	0.64
	severe_toxic (3) *Threat (4)* *Toxic (5)*		0.70 0.84 0.85	0.84 0.57 0.29	0.76 0.68 0.44
Support Vector Machine	*Identity_hate (0)*	0.8441	0.88	0.90	0.89
	Insult (1)		0.77	0.88	0.82
	obscene (2)		0.76	0.81	0.79
	severe_toxic (3) *Threat (4)* *Toxic (5)*		0.89 0.87 0.75	0.86 0.77 0.60	0.88 0.82 0.67
Random Forest	*Identity_hate (0)*	0.7782	0.81	0.90	0.85
	Insult (1)		0.76	0.87	0.81
	obscene (2)		0.79	0.69	0.74
	severe_toxic (3) *Threat (4)* *Toxic (5)*		0.83 0.88 0.68	0.82 0.67 0.57	0.82 0.76 0.62

Berdasarkan keseluruhan hasil eksperimen pada representasi Bag-of-Words (BoW), dapat disimpulkan bahwa:

1. SVM menunjukkan performa terbaik dengan akurasi tertinggi (0,8441) serta distribusi metrik evaluasi yang relatif seimbang di hampir seluruh kelas. Hal ini menunjukkan bahwa SVM mampu menangkap pola data secara lebih konsisten, meskipun masih terdapat penurunan performa pada kelas *toxic* yang memiliki F1-score lebih rendah dibandingkan kelas lainnya.
2. Naive Bayes mengalami peningkatan performa (0,7352) dibandingkan saat menggunakan TF-IDF, yang mengindikasikan bahwa pendekatan probabilistik lebih sesuai dengan representasi berbasis frekuensi kata. Namun demikian, ketidakseimbangan antar metrik masih terlihat, terutama pada kelas *toxic* yang memiliki rescall rendah, sehingga menunjukkan keterbatasan dalam mendeteksi kategori tersebut secara konsisten.

3. Random Forest menunjukkan performa yang cukup stabil (0,7782) dengan distribusi metrik yang relatif merata, namun masih berada di bawah SVM dalam hal akurasi dan konsistensi antar kelas. Penurunan performa paling terlihat pada kelas *toxic* dan sebagian kesalahan klasifikasi masih terjadi pada kelas dengan kemiripan karakteristik.

Temuan ini menunjukkan bahwa meskipun seluruh model pada representasi BoW memberikan performa yang cukup kompetitif, SVM tetap menjadi model yang paling konsisten dalam membedakan antar kelas. Namun demikian, tantangan utama masih terletak pada kesulitan model dalam mengklasifikasikan kategori dengan kemiripan leksikal tinggi, sehingga evaluasi tidak cukup hanya berfokus pada akurasi, tetapi juga perlu mempertimbangkan distribusi performa pada setiap kelas.

Untuk memvalidasi konsistensi hasil evaluasi yang telah diperoleh, penelitian ini juga menerapkan *3-fold cross-validation* pada seluruh kombinasi model dan representasi teks. Hasil validasi silang tersebut disajikan pada Tabel 4.

TABEL 4
HASIL *3- FOLD CROSS VALIDATION* SELURUH MODEL

Model	Representasi	Fold 1	Fold 2	Fold 3	Rata-rata	Std Dev
Naive Bayes	TF-IDF	0.59	0.59	0,58	0,59	0.0026
	BoW	0.72	0.72	0.72	0.72	0.0024
Support Vector Machine	TF-IDF	0.82	0.82	0.83	0.82	0.0054
	BoW	0.83	0.83	0.84	0.83	0.0063
Random Forest	TF-IDF	0.79	0.80	0.79	0.79	0.0058
	BoW	0.79	0.80	0.79	0.79	0.0035

Hasil 3-fold cross-validation pada Tabel 4 menunjukkan bahwa SVM tetap menjadi model terbaik dengan rata-rata akurasi tertinggi, yaitu 0,83 pada representasi BoW dan 0,82 pada TF-IDF. Hasil ini sejalan dengan pengujian sebelumnya menggunakan skenario split 80:20, di mana SVM juga menunjukkan kinerja paling unggul dibandingkan model lainnya. Meskipun terdapat perbedaan nilai akurasi antara kedua metode evaluasi, hal ini merupakan kondisi yang wajar karena *cross-validation* menguji model pada beberapa variasi pembagian data yang berbeda.

Perbedaan tersebut memberikan gambaran yang lebih realistis terhadap kemampuan generalisasi model. Nilai standar deviasi yang relatif kecil juga menunjukkan bahwa performa model cenderung stabil pada setiap fold. Dengan demikian, hasil *cross-validation* menegaskan bahwa SVM merupakan model yang paling konsisten, meskipun masih terdapat variasi performa yang dipengaruhi oleh distribusi data.

C. Visualisasi Word Cloud

Visualisasi word cloud merupakan teknik representasi teks untuk menggambarkan distribusi frekuensi kata dalam korpus, yaitu kata dengan frekuensi tinggi ditampilkan dengan ukuran lebih besar [28]. Hal tersebut memudahkan pembaca dalam mengenali kata-kata dominan tanpa perlu menelaah seluruh teks secara rinci. Word cloud juga efektif untuk memberikan gambaran awal mengenai tema atau topik dalam dataset sebelum dilakukan analisis lanjutan [29].

Gambar 16. Word cloud kategori *identity_hate*

Visualisasi word cloud pada kategori *identity_hate*, kata-kata seperti “cina”, “cebong”, “kafir”, “jawa”, “aseng”, dan “papua” muncul secara dominan. Kemunculan ini menunjukkan bahwa ujaran dalam kategori tersebut banyak berorientasi pada identitas kolektif, baik yang berkaitan dengan etnis, agama, maupun afiliasi politik. Pola tersebut mengindikasikan adanya kecenderungan pelabelan kelompok yang berpotensi memperkuat polarisasi dalam interaksi digital [30].

Gambar 17. Word cloud kategori *insult*

Visualisasi *word cloud* Kategori *insult* didominasi oleh kata seperti “tolol”, “goblok”, “bego”, dan “idiot”, yang mengindikasikan bahwa bentuk ujaran dalam kelas ini berfokus pada serangan terhadap individu, terutama terkait kapasitas intelektual maupun karakter personal. Secara semantik, pola ini termasuk dalam penghinaan yang ditujukan secara langsung kepada individu [31].

Gambar 18. Word cloud kategori *obscene*

Visualisasi *word cloud* pada kategori *obscene* merepresentasikan distribusi frekuensi token dalam korpus teks yang telah melalui tahapan prapemrosesan. Berdasarkan hasil ekstraksi *top words*, kategori *obscene* didominasi oleh kata-kata seperti “banci”, “memek”, “kontol”, dan “ngews”, yang mencerminkan penggunaan bahasa vulgar serta berkonotasi seksual. Hal ini menunjukkan bahwa karakter utama kategori ini terletak pada aspek eksplisititas dan ketidaksenonohan bahasa [30]. Secara semantik, pola ujaran pada kategori ini lebih berfokus pada unsur ketidaksenonohan dan vulgaritas bahasa dibandingkan bentuk penghinaan terhadap individu maupun serangan berbasis identitas [32].

Gambar 19. Word cloud kategori *severe_toxic*

Visualisasi *word cloud* pada kategori *severe_toxic* menunjukkan distribusi token dengan tingkat intensitas yang lebih tinggi dibandingkan kategori lainnya. Kategori *severe_toxic* menunjukkan dominasi kata seperti “anjing”, “bangsat”, “kontol”, serta kemunculan referensi identitas seperti “cina”. Hal ini mengindikasikan bahwa ujaran dalam kategori ini tidak hanya mencakup penghinaan personal dan penggunaan bahasa vulgar, tetapi juga dapat melibatkan serangan berbasis identitas [30]. Dengan demikian, pola leksikal pada kelas *severe_toxic* mencerminkan eskalasi tingkat agresivitas bahasa dibandingkan kategori *toxic* maupun *insult*, yang mendukung diferensiasi kelas dalam proses klasifikasi berbasis fitur frekuensi kata.

Gambar 20. Word cloud kategori *threat*

Visualisasi *word cloud* pada kategori *threat* memperlihatkan distribusi token yang berkaitan dengan ekspresi ancaman dan ajakan kekerasan. Kategori threat ditandai oleh kemunculan kata seperti “usir”, “ganyang”, “bunuh”, dan “mati”, yang menunjukkan adanya intensi tindakan agresif. Penggunaan kata kerja tersebut mengindikasikan bahwa ujaran tidak hanya bersifat ofensif, tetapi juga mengandung ancaman langsung terhadap individu maupun kelompok [33]. Dengan demikian, kategori *threat* memperlihatkan karakteristik linguistik berupa kombinasi antara serangan berbasis identitas dan ekspresi ancaman tindakan fisik. Pola ini membedakannya secara konseptual dari kategori *identity_hate* yang berfokus pada ujaran kebencian, maupun *severe_toxic* yang menekankan intensitas penghinaan dan vulgaritas tanpa selalu mengandung unsur ancaman langsung. Distribusi kata dalam word cloud kategori *threat* mempertegas bahwa fitur leksikal yang berkaitan dengan kekerasan dan pengusiran memiliki kontribusi signifikan dalam proses identifikasi dan klasifikasi teks berbasis pembobotan frekuensi kata.

Gambar 21. Word cloud kategori *toxic*

Visualisasi word cloud pada kategori *toxic* menunjukkan dominasi kata-kata seperti “anjing”, “bacot”, dan “najis”, yang mencerminkan bentuk serangan verbal yang bersifat personal. Ujaran pada kategori ini umumnya digunakan untuk mengekspresikan emosi negatif tanpa melibatkan ancaman langsung maupun penggunaan bahasa vulgar yang eksplisit [33]. Dengan demikian, karakteristik leksikal pada kelas *toxic* menunjukkan tingkat agresivitas yang lebih rendah dibandingkan *severe_toxic*, namun tetap memperlihatkan pola bahasa yang merendahkan atau tidak sopan terhadap individu tertentu.

Selain analisis visual melalui word cloud, penelitian ini juga melakukan analisis kuantitatif dengan mengekstraksi lima kata yang paling sering muncul pada setiap kategori. Hasil ekstraksi tersebut disajikan pada Tabel 5.

TABEL 5
ANALISIS KUANTITATIF KATA KATA

Kategori	***Rank 1***	***Rank 2***	***Rank 3***	***Rank 4***	***Rank 5***
Identity_hate	Cina (422)	engak (348)	lu (217)	cebong (214)	orang (182)
Insult	lu (224)	engak (198)	gue (126)	orang (93)	tolol (85)
obscene	lu (156)	engak (128)	banci (117)	gue (112)	memek (74)
severe_toxic	lu (589)	engak (395)	gue (384)	anjing (314)	kontol (307)
Threat	cina (188)	usir (127)	ganyang (72)	lu (71)	Engak (70)
Toxic	lu (80)	engak (65)	Babi (56)	gue (55)	Anjing (52)

Secara kuantitatif, hasil ekstraksi frekuensi kata pada setiap kategori sebagaimana disajikan pada Tabel 5 mempertegas temuan visual dari word cloud. Kata "lu" dan "engak" muncul dominan di hampir seluruh kategori, yang mencerminkan karakteristik percakapan informal dalam konteks game online di mana penggunaan kata ganti orang kedua dan ekspresi penolakan sangat umum digunakan.

Pada kategori *identity_hate*, kata “cina” menjadi token paling dominan dengan 422 kemunculan, diikuti “cebong” (214), yang menunjukkan dominasi serangan berbasis etnis dan afiliasi politik. Pada kategori *insult*, kata “tolol” (85) merepresentasikan penghinaan terhadap kapasitas intelektual, sementara dominasi “lu” (224) menunjukkan bahwa hinaan bersifat langsung dan personal.

Kategori *severe_toxic* memiliki frekuensi kemunculan tertinggi, dengan “anjing” (314) dan “kontol” (307), yang mengonfirmasi tingkat agresivitas bahasa yang lebih tinggi. Pada kategori threat, kata seperti “usir” (127) dan “ganyang” (72) muncul bersama “cina” (188), menunjukkan bahwa ancaman sering diarahkan pada kelompok identitas tertentu. Sementara itu, kategori *obscene* ditandai oleh kata “banci” (117) dan “memek” (74), yang mencerminkan penggunaan diksi vulgar dan referensi seksual.

Perbedaan distribusi frekuensi ini mendukung diferensiasi kelas dalam model klasifikasi. Meskipun terdapat kata umum seperti “lu” dan “engak”, setiap kategori tetap memiliki kata khas yang berkontribusi dalam proses identifikasi label oleh model.

Perbandingan distribusi kata menunjukkan bahwa ujaran toksik tidak bersifat homogen, melainkan membentuk spektrum intensitas. Ujaran berkembang dari penghinaan personal pada kategori *toxic* dan *insult*, meningkat melalui kombinasi dengan vulgaritas pada kategori *obscene* dan *severe_toxic*, hingga menjadi ancaman melalui penggunaan verba imperatif seperti “usir”, “bunuh”, dan “ganyang” pada kategori *threat*. Pola ini menunjukkan adanya hubungan konseptual antar label sebagai bentuk eskalasi linguistik.

Berdasarkan temuan tersebut, terdapat dua implikasi penting. Pertama, diferensiasi leksikal antar kategori mendukung penggunaan fitur berbasis frekuensi kata dalam pemodelan klasifikasi. Kedua, ujaran kebencian tidak hanya mencerminkan emosi negatif, tetapi juga berpotensi berkembang menjadi ancaman kolektif ketika identitas kelompok menjadi target. Dengan memahami pola tersebut, proses deteksi tidak hanya berfungsi sebagai klasifikasi teknis, tetapi juga sebagai upaya identifikasi awal terhadap potensi eskalasi konflik dalam komunikasi daring.

Selain itu, kemunculan kata yang sama pada beberapa kategori, khususnya *toxic*, *obscene*, dan *severe_toxic*, menunjukkan adanya kemiripan fitur leksikal antar kelas. Kondisi ini berpotensi menyulitkan model dalam membedakan batas kategori dan meningkatkan kemungkinan kesalahan prediksi, yang selanjutnya dianalisis pada bagian berikutnya.

D. Analisis Kesalahan (Error Analysis)

Analisis kesalahan dilakukan untuk mengidentifikasi keterbatasan model dalam proses klasifikasi. Berdasarkan hasil evaluasi, kelas *toxic* menunjukkan performa terendah, terutama pada nilai recall dan F1-score, yang mengindikasikan bahwa model belum mampu mendeteksi kategori tersebut secara optimal. Selain itu, kelas threat menunjukkan ketidakseimbangan antara precision dan recall, sehingga performanya kurang stabil.

Berdasarkan confusion matrix, kesalahan klasifikasi didominasi pada kelas dengan kemiripan leksikal, terutama antara obscene, toxic, dan severe_toxic. Sejumlah data pada kelas *toxic* dan *obscene* sering diprediksi sebagai *severe_toxic*, yang menunjukkan kecenderungan model mengarah pada kelas dengan intensitas ujaran lebih tinggi.

Fenomena ini dipengaruhi oleh *overlap* fitur leksikal antar kategori. Kata seperti “anjing”, “kontol”, “lu”, dan “engak” muncul di berbagai kelas, sehingga menyulitkan model dalam membedakan karakteristik tiap kategori. Selain itu, kesalahan pada kelas *threat* yang diprediksi sebagai *identity_hate* menunjukkan bahwa fitur berbasis frekuensi kata belum mampu menangkap perbedaan konteks, khususnya terkait unsur ancaman.

Secara keseluruhan, tantangan utama dalam klasifikasi terletak pada kemiripan leksikal dan semantik antar kelas. Oleh karena itu, diperlukan pendekatan yang lebih kontekstual untuk meningkatkan kemampuan model dalam membedakan kategori yang memiliki karakteristik serupa.

IV. KESIMPULAN

Penelitian ini menunjukkan bahwa identifikasi percakapan toxic dalam lingkungan game daring seperti Roblox memerlukan pendekatan yang tidak hanya berfokus pada algoritma, tetapi juga pada pemahaman struktur linguistik ujaran. Analisis dilakukan menggunakan kerangka KDD yang mencakup pembersihan data, normalisasi teks, pemodelan, dan interpretasi pola bahasa.

Hasil evaluasi menunjukkan bahwa model SVM memiliki kinerja paling konsisten dalam mengklasifikasikan enam kategori ujaran toxic pada berbagai skenario pengujian. Baik menggunakan TF-IDF maupun Bag-of-Words, SVM menghasilkan performa yang relatif stabil dibandingkan algoritma lain, dengan perbedaan yang tidak signifikan dalam hal stabilitas model.

Selain itu, analisis visual dan frekuensi kata menunjukkan bahwa setiap kategori memiliki pola leksikal yang berbeda, meskipun terdapat kata umum yang muncul di berbagai kelas. Distribusi ini mencerminkan spektrum agresivitas bahasa, dari penghinaan personal hingga ancaman berbasis identitas, serta menunjukkan adanya token khas pada setiap kategori yang mendukung proses klasifikasi.

Secara Keseluruhan penelitian ini menunjukkan potensi penggunaan metode machine learning dalam mengklasifikasikan percakapan *toxic* berbahasa Indonesia pada platform permainan daring. Namun, penerapan pada sistem moderasi nyata masih memerlukan pertimbangan lebih lanjut karena keterbatasan data dan variasi bahasa. Oleh karena itu, pengujian dengan dataset yang lebih beragam dan representatif diperlukan untuk meningkatkan keandalan model.

DAFTAR PUSTAKA

[1] Y. L. Setiawan, J. Nasir, and R. hadi Putra, "Komunikasi Virtual melalui Perilaku Trash-Talking Antar Pemain Game Online Mobile Legends," *Ekasakti Jurnal Penelitian Dan Pengabdian*, vol. 05, no. 01, pp. 133–141, 2024, doi: 10.31933/ejpp.v5i1.1244.

[2] Muh. Zaad and Arni, "Perilaku Komunikasi Toxic Remaja yang Bermain Game Online Mobile Legends Pulau Barrang Lompo Kecamatan Kepulauan Sangkarrang Kota Makassar," *Jurnal Komunikasi dan Organisasi (J-KO)*, vol. 7, no. 1, pp. 15–20, 2025, doi: 10.26618/jko.v7i1.17585.

[3] Á. Zsila, R. Shabahang, M. S. Aruguete, and G. Orosz, "Toxic behaviors in online multiplayer games: Prevalence, perception, risk factors of victimization, and psychological consequences," *Aggress. Behav.*, vol. 48, no. 3, pp. 356–364, 2022, doi: 10.1002/ab.22023.

[4] T. Yuliastika and A. Fitriana Poerana, "Motif Penggunaan Game Online Roblox pada Anak Usia Sekolah," *Jurnal Ilmiah Wahana Pendidikan*, vol. 9, no. 9, pp. 364–371, 2023, doi: 10.5281/zenodo.7953027.

[5] J. Huang, "Analysis on the Young Age of Roblox Platform Audience Targeting," *Highlights in Business,Economics and Management*, vol. 11, pp. 112–117, 2023, doi: doi:10.54097/hbem.v11i.7954.

[6] H. A. Janati, E. Elvandri, S. V. Vianto, B. Agnes, and B. Cholas, "Perilaku Toxic Dalam Game Moba Dan Dampaknya Terhadap Komunitas Gamer Di Batam," *Simtek : jurnal sistem informasi dan teknik komputer*, vol. 10, no. 2, pp. 374–378, 2025, doi: 10.51876/simtek.v10i2.1585.

[7] U. Naseem, S. Shiwakoti, S. B. Shah, S. Thapa, and Q. Zhang, "GameTox: A Comprehensive Dataset and Analysis for Enhanced Toxicity Detection in Online Gaming Communities," *Proceedings of the 2025 Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies: Long Papers, NAACL-HLT 2025*, vol. 2, pp. 440–447, 2025, doi: 10.18653/v1/2025.naacl-short.37.

[8] M. D. Desriansyah, I. U. Sari, and Z. Zulfahmi, "Analisis Efektivitas Algoritma Machine Learning dalam Deteksi Hoaks: Pada Berita Digital Berbahasa Indonesia," *Jurnal Sistem Informasi Dan Informatika*, vol. 3, no. 2, pp. 63–69, 2025, doi: 10.47233/jiska.v3i1.2024.

[9] H. Ismail, A. Khalil, and A. Jasmy, "Enhancing online toxicity detection on gaming networks: a novel embeddings-based valence lexicon approach," *Int. J. Data Sci. Anal.*, vol. 20, no. 5, pp. 4489–4500, 2025, doi: 10.1007/s41060-025-00730-1.

[10] K. Andariefli, J. Leonard, V. Dewanto, and A. D. Novika, "Comparative analysis of two-class and multi-class toxicity detection using multi-source gaming chat data," *Procedia Comput. Sci.*, vol. 269, pp. 825–833, 2025, doi: 10.1016/j.procs.2025.09.025.

[11] R. P. Sidiq, B. A. Dermawan, and Y. Umaidah, "Sentimen Analisis Komentar Toxic pada Grup Facebook Game Online Menggunakan Klasifikasi Naïve Bayes," *Jurnal Informatika Universitas Pamulang*, vol. 5, no. 3, p. 356, 2020, doi: 10.32493/informatika.v5i3.6571.

[12] T. A. Alghamdi and N. Javaid, "A Survey of Preprocessing Methods Used for Analysis of Big Data Originated from Smart Grids," *IEEE Access*, vol. 10, pp. 29149–29171, 2022, doi: 10.1109/ACCESS.2022.3157941.

[13] D. Rifaldi, A. Fadlil, and Herman, "Teknik Preprocessing Pada Text Mining Menggunakan Data Tweet 'Mental Health,'" *Decode: Jurnal Pendidikan Teknologi Informasi*, vol. 3, no. 2, pp. 161–171, 2023, doi: 10.51454/decode.v3i2.131.

[14] P. Ayuningtiyas, K. Ditha Tania, and W. Kurnia Sari, "Sentiment-Based Knowledge Discovery pada Aplikasi iPusnas Menggunakan Metode Machine Learning dan Deep Learning," *Journal of Applied Informatics and Computing (JAIC)*, vol. 9, no. 5, pp. 2486–2497, 2025, [Online]. Available: http://jurnal.polibatam.ac.id/index.php/JAIC

[15] C. Suhaeni, S. A. Kamila, F. Fahira, M. Yusran, and G. A. Dito, "Exploring a Large Language Model on the ChatGPT Platform for Indonesian Text Preprocessing Tasks," *Indonesian Journal of Statistics and Its Applications*, vol. 9, no. 1, pp. 100–116, 2025, doi: 10.29244/ijsa.v9i1p100-116.

[16] A. A. Pratiwi and M. Kamayani, "Perbandingan Pelabelan Data dalam Analisis Sentimen Kurikulum Proyek di platform TikTok: Pendekatan Naïve Bayes," *Jurnal Eksplora Informatika*, vol. 14, no. 1, pp. 96–107, 2024, doi: 10.30864/eksplora.v14i1.1093.

[17] E. W. Pamungkas, C. S. Wahyuni, I. Amal, D. Purworini, and B. S. Rintyarna, "Decoding hate in memes: multimodal and multitask approaches for low-resource Indonesian social media," *PeerJ Comput. Sci.*, vol. 12, 2026, doi: 10.7717/peerj-cs.3736.

[18] R. Hayami, S. Mohnica, and Soni, "Klasifikasi multilabel komentar toxic pada sosial media twitter menggunakan convolutional neural network(CNN)," *Jurnal CoSciTech (Computer Science and Information Technology)*, vol. 4, no. 1, pp. 1–6, 2023, doi: 10.37859/coscitech.v4i1.4365.

[19] V. N. Romadina, O. Juwita, and P. Pandunata, "Analisis Komentar Toxic Terhadap Informasi COVID-19 pada YouTube Kementerian Kesehatan Menggunakan Metode Naïve Bayes Classifier," *INFORMAL: Informatics Journal*, vol. 9, no. 1, pp. 92–99, 2024, doi: 10.19184/isj.v9i1.48126.

[20] N. M. D. Sikiandani, I. M. A. Dwi Suarjaya, and Y. P. Putra, "Browser-Based Detection of Harmful Content with Deep Learning Model," *Journal of Applied Informatics and Computing*, vol. 9, no. 4, pp. 1800–1811, 2025, doi: 10.30871/jaic.v9i4.9804.

[21] R. Oktafiani, A. Hermawan, and D. Avianto, "Pengaruh Komposisi Split data Terhadap Performa Klasifikasi Penyakit Kanker Payudara Menggunakan Algoritma Machine Learning," *Jurnal Sains dan Informatika*, vol. 9, no. 1, pp. 19–28, 2023, doi: 10.34128/jsi.v9i1.622.

[22] Sutriawan, S. Mutmainnah, T. Ansyor Lorosae, and S. Ramadhan, "Model Text Embedding dan TF-IDF+Ngram untuk Meningkatkan Kinerja Algoritma Binary Classifier pada Klasifikasi SMS Palsu," vol. 4, no. 1, pp. 55–64, 2025, [Online]. Available: https://ojs.trigunadharma.ac.id/index.php/jsi

[23] A. D. M. Putri, N. Sulistianingsih, and R. Rismayati, "Pengaruh Teknik Representasi Teks Bag-of-Words dan TF-IDF terhadap Akurasi Klasifikasi Sentimen Teks Multi-Domain," *JTIM : Jurnal Teknologi Informasi dan Multimedia*, vol. 7, no. 4, pp. 675–688, 2025, doi: 10.35746/jtim.v7i4.756.

[24] S. D. Prasetyo, S. S. Hilabi, and F. Nurapriani, "Analisis Sentimen Relokasi Ibukota Nusantara Menggunakan Algoritma Naïve Bayes dan KNN," *Jurnal KomtekInfo*, vol. 10, no. 1, pp. 1–7, 2023, doi: 10.35134/komtekinfo.v10i1.330.

[25] V. Ayumi, D. Ramayanti, H. Noprisson, A. Ratnasari, and U. Salamah, "Pengaruh Tuning Parameter dan Cross Validation Pada Klasifikasi Teks Komplain Bahasa Indonesia Menggunakan Algoritma Support Vector Machine ," *JSAI : Journal Scientific and Applied Informatics* , vol. 6, no. 3, pp. 493–498, Nov. 2023.

[26] A. Hussaina and A. Aslamb, "Hate speech against women and immigrants: A comparative analysis of machine learning and text embedding techniques," *Journal of Applied Research and Technology*, vol. 22, no. 4, pp. 548–559, 2024, Accessed: Mar. 06, 2026. [Online]. Available: https://jart.icat.unam.mx/index.php/jart/article/view/2466/1129

[27] P. Ayuningtiyas, K. Ditha Tania, and W. Kurnia Sari, "Sentiment-Based Knowledge Discovery pada Aplikasi iPusnas Menggunakan Metode Machine Learning dan Deep Learning," *Journal of Applied Informatics and Computing (JAIC)*, vol. 9, no. 5, p. 2486, Oct. 2025, [Online]. Available: http://jurnal.polibatam.ac.id/index.php/JAIC

[28] T. N. Pasaribu, J. P. Tanjung, D. Hutauruk, E. S. Hutagalung, and S. Silitonga, "Study of Public Sentiment Towards Beauty Products Using A Machine Learning Approach: Random Forest Analysis On Social Media," *Sinkron: Jurnal dan Penelitian Teknik Informatika*, vol. 8, no. 3, pp. 2088–2098, 2024, doi: 10.33395/sinkron.v8i3.13969.

[29] S. Rosalin and B. F. Supriyanto, "Analisis Sentimen Program Merdeka Belajar dengan Text Analysis Wordcloud & Word Frequency," *Jurnal Minfo Polgan*, vol. 12, no. 1, pp. 25–32, 2023, doi: 10.33395/jmp.v12i1.12312.

[30] C. H. Pratama and Y. Findawati, "Klasifikasi Hate Speech dan Emosi Dalam Teks Berbahasa Indonesia Pada Pengguna Twitter Menggunakan Metode Naïve Bayes Classifier," *Indonesian Journal of Applied Technology*, vol. 1, no. 3, pp. 1–10, 2024, doi: 10.47134/ijat.v1i3.3105.

[31] Sawidiah and M. Ulfa, "Tindak Insult pada Kasus Bullying Verbal di Media Sosial Online (Kajian Linguistik Forensik)," *Stilistika: Jurnal Pendidikan Bahasa dan Sastra*, vol. 18, no. 2, pp. 347–362, 2025, doi: 10.30651/st.v18i2.25712.

[32] M. P. D. Sari, I. W. Pastika, and M. S. Satyawati, "Ujaran Kebencian Terhadap Selebgram Azizah Salsha Di Media Sosial Tiktok: Kajian Linguistik Forensik," *Kulturistik: Jurnal Bahasa & Budaya*, vol. 9, no. 2, pp. 49–57, 2025, doi: 10.22225/kulturistik.9.2.12631.

[33] Jayus, Sumariyah, A. Abdullah, and Mustafa, "YouTube, Public Discourse, and the 'Makan Siang Gratis' Program: An Analysis of Toxicity Comments on the Liputan6 Channel," *Jogjakarta Communication Conference (JCC)*, vol. 3, no. 1, pp. 367–380, 2025, [Online]. Available: https://jcc-indonesia.id/