SKANIKA: Sistem Komputer dan Teknik Informatika
Volume 9, Nomor 2, Juli 2026, Halaman 234-247 E-ISSN: 2721-4788

Komparasi MNB, CNB, dan SVM untuk Deteksi Ujaran Kebencian Bahasa Indonesia

Fisco Maulana Ikhwan[1], Edi Sugiarto[2*]
[1,2]Fakultas Ilmu Komputer, D3 Teknik Informatika, Universitas Dian Nuswantoro, Semarang, Indonesia
E-mail: [1]fiscomaulana@gmail.com, [2*]edi.sugiarto@dsn.dinus.ac.id
(*: corresponding author)

Abstrak

Indonesia memiliki 235,26 juta pengguna internet dengan tingkat penetrasi 81,72% pada 2026, sementara ujaran kebencian mengancam kohesi sosial masyarakat majemuk. Berdasarkan literatur yang ditelaah dalam penelitian ini, belum ditemukan studi yang melakukan *benchmark* tiga-arah MNB, CNB, dan SVM dengan *ablation study preprocessing*, uji McNemar, dan analisis *error* Wilson CI pada konteks *hate speech* bahasa Indonesia. Penelitian ini mengusulkan kerangka komparatif menggunakan *dataset* 13.169 tweet, penanganan *imbalanced class* per-algoritma (*SMOTE* untuk MNB/CNB; *CSL* dan *Lexicon-Based Coefficient Override* untuk SVM), serta ablasi enam konfigurasi *preprocessing*. Hasil: MNB mencapai *F1 Macro* 72,10% dan *Recall Abusive* tertinggi (84,30%), sementara SVM mencapai *Recall Abusive* 80,20% dengan latensi inferensi 0,10 ms pada *CPU*; perbandingan langsung dengan IndoBERT pada *GPU* T4 tidak sepenuhnya setara lintas-perangkat-keras, namun pengukuran tambahan pada *CPU* yang sama menunjukkan IndoBERT berlatensi ~102,4 ms (rasio ~1.020×). Uji McNemar mengkonfirmasi perbedaan SVM vs MNB/CNB signifikan ($p<0,0001$), CNB vs MNB tidak signifikan ($p=0,0704$), mengindikasikan interaksi negatif *SMOTE*×CNB pada *dataset* dan konfigurasi penelitian ini. SVM dipilih sebagai model *deployment* berdasarkan *Utility Score* (81,39) pada skenario bobot yang diuji, dan dapat menjadi alat bantu penyaringan awal (bukan pengganti verifikasi manusia) pada platform yang beroperasi di bawah regulasi UU ITE No. 1/2024.

Kata kunci: deteksi ujaran kebencian, Complement Naive Bayes, TF-IDF, imbalanced class, lexicon-based calibration

Abstract

Indonesia had 235.26 million internet users and 81.72% internet penetration in 2026, while hate speech threatens social cohesion. Based on the literature reviewed in this study, no prior study has been found that systematically conducted a three-way benchmark of MNB, CNB, and SVM with preprocessing ablation, McNemar testing, and Wilson CI error analysis in Indonesian hate speech detection, using a dataset of 13,169 tweets, with per-algorithm imbalanced-class handling (SMOTE for MNB/CNB; CSL and Lexicon-Based Coefficient Override for SVM), and ablation of six preprocessing configurations. MNB achieves F1 Macro 72.10% and highest Abusive Recall (84.30%), while SVM achieves Abusive Recall 80.20% with inference latency 0.10 ms on CPU; a direct comparison with IndoBERT on GPU T4 is not fully equivalent across hardware, though an additional CPU-only measurement shows IndoBERT at ~102.4 ms (~1,020× ratio). McNemar tests confirm SVM differs significantly from MNB/CNB ($p<0.0001$), CNB vs MNB non-significant ($p=0.0704$), indicating a negative SMOTE×CNB interaction specific to this study's dataset and configuration. SVM is selected as the deployment model based on Utility Score (81.39) under the weighting scenarios tested, and may serve as a first-pass screening aid (not a substitute for human verification) for platforms operating under Indonesian ITE Law No. 1/2024.

Keywords: *hate speech detection, Complement Naive Bayes, TF-IDF, imbalanced class, lexicon-based calibration*

1. PENDAHULUAN

Indonesia memiliki 235,26 juta pengguna internet dengan penetrasi 81,72% pada 2026 [1]. Tingginya penetrasi media sosial di tengah masyarakat yang terdiri dari lebih dari 300 suku bangsa menciptakan ruang ekspresi yang rentan terhadap ujaran kebencian (*hate speech*) dan konten abusif, yang dapat mengancam kohesi sosial dan kerukunan antarkelompok. Regulasi Undang-Undang Informasi dan Transaksi Elektronik (UU ITE) Nomor 1 Tahun 2024 [2] menegaskan kebutuhan mendesak akan sistem deteksi otomatis yang efektif, akurat, dan dapat dioperasikan secara efisien pada infrastruktur yang tersedia luas. Penelitian terdahulu telah meletakkan fondasi penting dalam deteksi *hate speech* otomatis. Survei komprehensif yang

dilakukan oleh [3] mengidentifikasi bahwa tantangan utama deteksi otomatis mencakup ambiguitas bahasa, variasi lintas-budaya, dan ketidakseimbangan kelas pada berbagai kerangka klasifikasi. [4] membangun *dataset* anotasi *hate speech* berbahasa Indonesia berbasis Twitter yang menjadi acuan komunitas peneliti. [5] mengeksplorasi pendekatan berbasis TF-IDF dan augmentasi data untuk konteks hate speech berbahasa Indonesia. Lebih lanjut, oleh [6] menginvestigasi perbandingan metode klasikal pada data tidak seimbang, namun tidak menyertakan CNB maupun pengujian signifikansi statistik.

Terdapat beberapa gap penelitian yang belum terpenuhi: (a) berdasarkan literatur yang ditelaah dalam penelitian ini, tidak ditemukan *benchmark* tiga-arah yang menyertakan CNB bersama MNB dan SVM secara sistematis pada domain *hate speech* bahasa Indonesia; (b) tidak ada ablation study *preprocessing* yang komprehensif pada domain *hate speech* bahasa Indonesia; (c) tidak ada penggunaan McNemar *test* dengan koreksi Bonferroni; dan (d) tidak ada analisis *error* kuantitatif dengan *Wilson Confidence Interval*. Tabel 1 merangkum penelitian terdahulu dan analisis gap secara sistematis.

Tabel 1. Ringkasan Penelitian Terdahulu dan Analisis Gap

No	Penulis (Tahun)	*Dataset*	Metode	Gap / Keterbatasan
1	Widiarta et al. (2024) [5]	Twitter	SVM + TF-IDF + augmentasi	Tanpa CNB; tanpa ablation; tanpa McNemar
2	Zikrina & Fitriyani (2025) [7]	Twitter ID	GNN + TF-IDF	Tanpa CNB; tanpa ablation; tanpa McNemar
3	Putri et al. (2021) [8]	Twitter ID	SVM, NB, RFDT	Multi-label; tanpa CNB; tanpa ablation
4	Difandana et al. (2026) [6]	Twitter ID + Twitter API	MNB, SVM (linear kernel)	Imbalanced data; tanpa CNB; tanpa ablation
5	**Penelitian ini**	Twitter ID 13.169 / 12.024 valid	MNB + CNB + SVM + IndoBERT + TF-IDF	Ekstensi lengkap: CNB, imbalanced per-algoritma, ablation, McNemar, IndoBERT

Pendekatan klasifikasi teks menggunakan Naïve Bayes dan SVM telah terbukti efektif pada berbagai domain, termasuk klasifikasi ulasan aplikasi berbahasa Indonesia [9]. Menurut [8] menggunakan *dataset* sama (13.169 tweet, [4]) dengan pendekatan Graph Neural Network (GCN) yang membentuk graf antar-tweet via *cosine similarity* TF-IDF (3.000 fitur), dibanding *baseline* RNN. *Preprocessing* 4 tahap (*cleaning*, *case folding*, *stopword removal*, *stemming*) tanpa normalisasi slang. Kedua label (Abusive, HS) dievaluasi biner terpisah (bukan 3-kelas tunggal), *split* 80:20 (10.525/2.634) tanpa penyeimbangan kelas eksplisit. GNN terbaik: Accuracy 92,90% (Abusive), 89,78% (HS); namun tanpa CNB, ablation *preprocessing*, maupun uji McNemar. Difandana et al. (2026) [7] membandingkan MNB dan SVM (linear, one-vs-rest) pada *dataset* diperluas dari Ibrohim & Budi dengan tambahan data via Twitter API (13.169 tweet, komposisi berbeda: HS 27,52%, Abusive 34,25%, Netral 38,23% jauh berbeda dari *dataset* penelitian ini). *Preprocessing*: *case folding*, tokenisasi, *stopword removal*, *stemming* Nazief-Adriani tanpa normalisasi slang dan tanpa penanganan imbalanced class (tanpa SMOTE/CSL/*override*). TF-IDF unigram, min_df=3. Evaluasi 10-Fold CV saja tanpa *holdout* terpisah. SVM unggul (Accuracy 93,28% vs 84,71%), namun tanpa CNB, ablation, maupun uji McNemar - konsisten dengan gap Tabel 1.

Berdasarkan analisis gap tersebut, penelitian ini merumuskan kontribusi sebagai berikut: (1) Kontribusi Empiris: *benchmark* sistematis tiga-arah MNB vs CNB vs SVM dengan strategi imbalanced handling per-algoritma; (2) Kontribusi Analitis: ablation study *preprocessing* sistematis pada domain hate speech bahasa Indonesia; (3) Kontribusi Metodologis: *error analysis* kuantitatif dengan Wilson 95% CI; dan (4) Kontribusi Teknis: analisis komparatif SVM vs IndoBERT fine-tuned beserta implementasi sistem web FastAPI + Vue.js 3.

Hipotesis penelitian dirumuskan sebagai: H_0: tidak ada perbedaan signifikan pola prediksi antar model (McNemar, α=0,0167 setelah koreksi Bonferroni); H_{1a}: terdapat perbedaan signifikan antara SVM dan MNB; $H_{1\beta}$: terdapat perbedaan signifikan antara SVM dan CNB; H_{1c}: terdapat perbedaan signifikan antara CNB dan MNB.

2. METODE PENELITIAN

Penelitian ini menggunakan pendekatan eksperimental komparatif dalam dua jalur paralel: (1) *pipeline* klasikal MNB/CNB/SVM dengan fitur TF-IDF dan strategi penanganan *imbalanced class* yang disesuaikan per algoritma; dan (2) eksplorasi IndoBERT *fine-tuned* sebagai ekstensi komparatif berbasis *deep learning*.

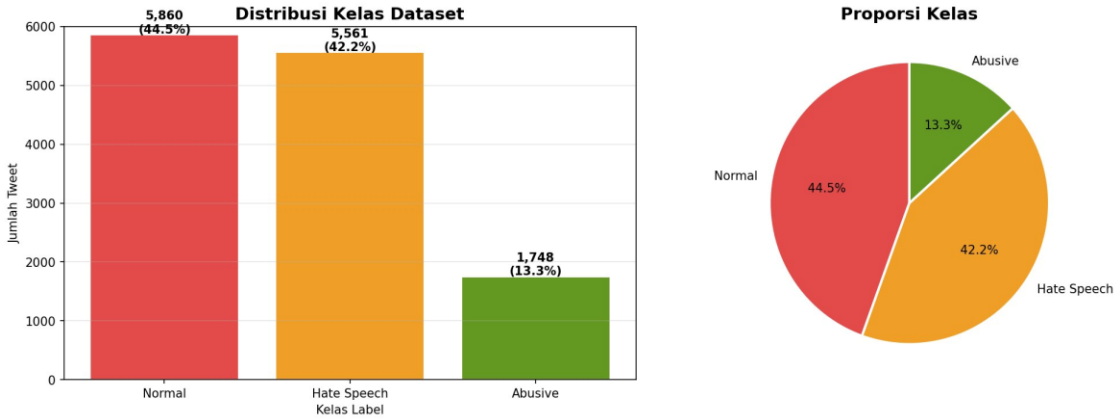
2.1. *Dataset*

Penelitian ini menggunakan *dataset* Indonesian *Hate Speech* yang dibangun oleh [4] dan dibahas dalam kajian perkembangan riset selanjutnya [10], terdiri dari 13.169 tweet dari platform Twitter dipilih karena merupakan *benchmark* yang paling banyak diadopsi dalam penelitian hate speech bahasa Indonesia, memungkinkan perbandingan langsung dengan state-of-the-art. Karena dataset awal bersifat multi-label, penelitian ini melakukan konversi ke single-label dengan skema prioritas HS > Abusive > Normal untuk menyelesaikan konflik anotasi. Urutan prioritas ini didasarkan pada tingkat keparahan: Hate Speech yang menargetkan kelompok identitas tertentu memiliki dampak sosial lebih besar dan konsekuensi hukum lebih berat berdasarkan UU ITE [2] dibandingkan konten Abusive yang bersifat kasar tanpa target spesifik, sehingga label HS diprioritaskan apabila suatu tweet memenuhi kedua kriteria secara bersamaan. Distribusi kelas setelah konversi ditunjukkan pada Tabel 2 dan divisualisasikan pada Gambar 1.

Ketiga kelas mengikuti skema anotasi [4]: (1) Normal tanpa unsur ujaran kebencian, contoh: "deklarasi pilkada 2018 aman dan anti hoax warga dukuh sari jabon"; (2) *Hate Speech* menyerang individu/kelompok berdasarkan identitas (etnis, agama, ras, *gender*), contoh (disunting): "USER USER Kaum c*bong k*pir udah keliatan d*ngoknya dari awal tambah d*ngok lagi hahahah"; (3) Abusive, bahasa kasar generik tanpa target identitas, contoh: "Gue baru aja kelar re-watch Aldnoah Zero!!! paling kampret emang endingnya!". Perbedaan utama HS vs Abusive: ada-tidaknya target identitas kelompok, sesuai skema prioritas HS > Abusive > Normal.

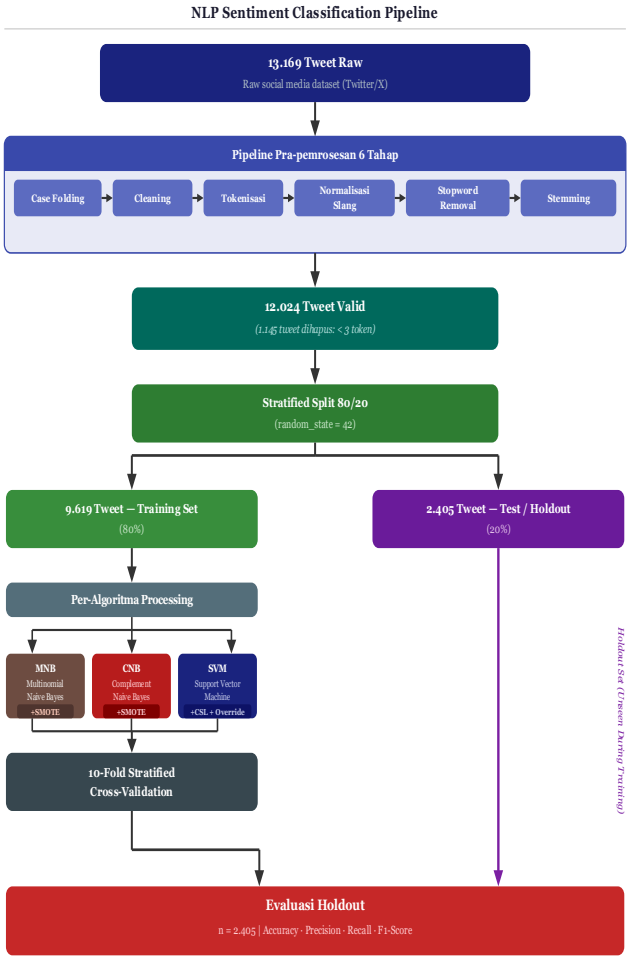
Tabel 2. Distribusi Kelas *Dataset* (N = 13.169)

Kelas	Label	Jumlah Tweet	Persentase	Keterangan
Normal	0	5.860	44,50%	Konten netral
Hate Speech	1	5.561	42,23%	Tweet bermuatan kebencian terhadap kelompok identitas
Abusive	2	1.748	13,27%	Kata kasar tanpa target identitas spesifik
TOTAL	—	**13.169**	**100%**	Sumber: Ibrohim & Budi (2019) [4]



Gambar 1. Distribusi Kelas *Dataset*: Normal (44,50%), Hate Speech (42,23%), Abusive (13,27%)

Alur pemrosesan data secara keseluruhan ditunjukkan pada Gambar 2. *Dataset* awal 13.169 tweet melewati *pipeline* pra-pemrosesan enam tahap (dijelaskan pada Bagian 2.2), menghasilkan 12.024 *tweet valid* (menghapus 1.145 *tweet* dengan kurang dari tiga token). *Dataset* valid kemudian dibagi secara *stratified* 80/20 menjadi 9.619 *tweet training* dan 2.405 *tweet test* (random_state=42). Seluruh eksperimen *cross-validation* dan evaluasi *holdout* dilakukan pada partisi ini.



Gambar 2. Alur Pemrosesan Data: dari 13.169 tweet mentah hingga partisi *training* (n=9.619) dan test (n=2.405)

2.2. Pra-pemrosesan Teks

Pipeline pra-pemrosesan mencakup enam tahap berurutan: (1) *Case folding*, konversi seluruh teks ke huruf kecil; (2) *Cleaning*, penghapusan URL, @mention, hashtag, emoji, dan angka; (3) Tokenisasi, segmentasi teks menjadi token kata; (4) Normalisasi slang, pemetaan kata tidak baku menggunakan KamusAlay v2018 [11] dengan lebih dari 12.000 entri; (5) *Stopword removal*, penghapusan kata-kata umum tidak informatif; (6) *Stemming* menggunakan algoritma Sastrawi/Nazief-Adriani. Tweet dengan kurang dari tiga token setelah proses eliminasi dihapus, menghasilkan 12.024 tweet valid. Contoh: tweet Abusive "Gue baru aja kelar re-watch Aldnoah Zero!!! paling kampret emang endingnya! 2 karakter utama cowonya kena friendzone bray! XD URL" menjadi "gue selesai re watch aldnoah zero kampret karakter utama cowok kena friendzone bro xd" setelah *cleaning* (hapus URL/angka), normalisasi slang (aja→saja, kelar→selesai, bray→bro, dst.), *stopword removal*, dan *stemming* (cowoknya→cowok).

2.3. Ekstraksi Fitur TF-IDF

Pembobotan kata menggunakan formula TF-IDF: $w(t,d) = tf(t,d) \times \log(N/df(t))$ [12], di mana $tf(t,d)$ adalah frekuensi term t dalam dokumen d, N adalah jumlah total dokumen, dan $df(t)$

adalah jumlah dokumen yang mengandung term t. Konfigurasi vectorizer: Unigram+Bigram dengan min_df=2 menghasilkan 7.316 fitur untuk MNB/CNB, dan min_df=3 menghasilkan 6.518 fitur untuk SVM. Vectorizer di-fit hanya pada *training set* untuk mencegah data *leakage* [13]. Sebagai ilustrasi, untuk tweet pada Bagian 2.2, bobot TF-IDF tertinggi: xd (0,4538), karakter (0,4007), utama (0,3620), bro (0,3213), cowok (0,3163) — term seperti "re" dan "aldnoah" tidak muncul karena di bawah ambang min_df training.

2.4. Algoritma dan Penanganan Imbalanced Class

Ketiga algoritma klasikal dipilih untuk merepresentasikan spektrum pendekatan berbeda: MNB sebagai *baseline* probabilistik standar; CNB sebagai pembanding pada distribusi kelas tidak seimbang yang ditemui pada dataset ini; dan SVM karena keunggulannya pada fitur TF-IDF berdimensi tinggi dan *sparse* serta kemampuannya dimodifikasi manual melalui intervensi koefisien (Lexicon-Based *Coefficient Override*), yang tidak dimungkinkan pada model probabilistik. *Dataset* dibagi stratified 80/20 (random_state=42): 9.619 tweet training, 2.405 test. Seluruh komponen stokastik memakai seed tetap (SGDClassifier & SMOTE: random_state=42) untuk *reproducibility*.

Strategi penanganan *imbalanced class* dirancang berbeda per-algoritma. MNB [14] dan CNB dikonfigurasi dengan *Laplace smoothing* α=1,0 dan SMOTE [15] (menyeimbangkan ke 13.158 *sampel*). SVM final [16] diimplementasikan dengan SGDClassifier(*loss*='hinge') — bukan LinearSVC — karena *solver* ini mengekspos atribut coef_ yang dapat dimodifikasi pasca-fitting, sesuai kebutuhan *Coefficient Override*. Sebagai pembanding, SVM *Baseline* pada Tabel 3 dan Tabel 9 memakai Linear SVC tanpa CSL/*Oversampling*/*Override*, merepresentasikan SVM tanpa penanganan *imbalanced class*. SVM final menerapkan tiga mekanisme: (1) *Cost-Sensitive Learning* via class_weight={Abusive:10, HS:1, Normal:1} [17]; (2) Partial *Oversampling* 3× khusus kelas Abusive; (3) Lexicon-Based *Coefficient Override* — modifikasi post-hoc pada matriks coef_ untuk 24 kata makian, dikompilasi dari analisis frekuensi term Abusive di *training set* (bukan seluruh *dataset*) dan divalidasi manual oleh penulis tanpa validator eksternal/IAA (keterbatasan dinyatakan di Bagian 3.7). Satu kata ("pelacur") tidak muncul di korpus training/test, sehingga leksikon efektif berjumlah 23 kata. Koefisien kelas Abusive di-*override* +2,0, kelas *Hate Speech* −5,0 (koefisien Normal tidak diubah); nilai ini dievaluasi pada *validation set* (bukan *test set*, menghindari *leakage*) dari perspektif F1-Macro, konfigurasi tanpa *override* (0,0) sebenarnya lebih tinggi (75,76% vs 69,91%), namun (+2,0/−5,0) tetap dipilih karena tujuan eksplisit penelitian adalah memaksimalkan *Recall Abusive* sebagai kelas minoritas kritis (lihat *Deployment Utility Score*, Bagian 3.2) keputusan domain-driven, bukan hasil *grid search* murni. Ruang pencarian terbatas (24 kata × 2 parameter) memungkinkan evaluasi exhaustive manual; Bayesian *optimization* pada ruang lebih luas direkomendasikan untuk penelitian lanjutan. Kode implementasi lengkap tersedia di github.com/feastco/hate-speech-detection.

Tabel 3. Perbandingan Tahapan *Pipeline*: SVM *Baseline* vs SVM Final.

Tahap	SVM *Baseline* (LinearSVC)	SVM Final (SGDClassifier)
1	Fit TF-IDF (*training set*)	Fit TF-IDF (*training set*)
2	—	Partial *Oversampling* 3× (kelas Abusive)
3	LinearSVC, C=1,0, tanpa class_weight	SGDClassifier(loss='hinge') + Cost-Sensitive Learning (class_weight custom)
4	Fit langsung ke data	Fit model
5	Prediksi *holdout*/CV — Recall Abusive <10%	Lexicon *Override* post-hoc pada coef_ (24 kata) → Recall Abusive 80,20%

2.5. IndoBERT *Fine-tuning*

IndoBERT disertakan sebagai pembanding *contextual embedding* untuk menguji apakah pretraining bahasa berskala besar unggul dibanding pendekatan leksikal-statistik (TF-IDF) pada deteksi ujaran kebencian Indonesia yang kaya slang. Model pretrained yang digunakan adalah indobenchmark/indobert-base-p2 [18]. IndoBERT merupakan model pralatih bahasa Indonesia berbasis arsitektur BERT. *Fine-tuning* dilakukan dengan konfigurasi: batch_size=16, learning rate=5×10^{-5}, optimizer AdamW, dan *training* selama 5 *epoch* dengan pemilihan *checkpoint*

terbaik pada *Epoch* 3 berdasarkan F1 Macro pada *validation set*. *Class weighting* diterapkan dengan bobot Normal=0,749, HS=0,789, dan Abusive=2,511. Infrastruktur yang digunakan adalah Google Colab dengan GPU Tesla T4.

2.6. Protokol Evaluasi

Metrik evaluasi meliputi *Accuracy*, *Precision*, *Recall*, F1 Macro [19], dan ROC-AUC. Validasi silang menggunakan 10-*Fold Stratified Cross-validation* untuk model klasikal, sejalan dengan pendekatan yang diterapkan pada klasifikasi teks berbahasa Indonesia [20]. SMOTE diterapkan secara independen di dalam setiap *fold* pelatihan (*in-fold resampling*), sehingga data sintetis tidak pernah bocor ke *fold* validasi. Pendekatan ini mengikuti rekomendasi [13] untuk menghindari bias estimasi akibat data *leakage* pada *pipeline oversampling*. Uji signifikansi menggunakan McNemar *test* [19] dengan koreksi Bonferroni ($\alpha\_adj=0{,}05/3\approx0{,}0167$) [19] pada *holdout test set* (n=2.405). *Error analysis* dilakukan dengan Wilson 95% CI [21] pada n=50 sampel *misclassified* per model (inter-rater $\kappa\_avg=0{,}743$, substantial agreement [22]). Ablation study mencakup 6 konfigurasi *preprocessing* pada SVM *baseline*. Latensi dirata-ratakan atas 1.000 iterasi inferensi murni dan n=500 HTTP request end-to-end via FastAPI. Implementasi menggunakan scikit-learn 1.3.0 [23] dan Python 3.10. Kode sumber tersedia di github.com/feastco/hate-speech-detection.

3. HASIL DAN PEMBAHASAN

3.1. Hasil *Cross-validation*

Evaluasi 10-*Fold Stratified Cross-validation* menunjukkan performa yang relatif stabil pada ketiga model klasikal. Hasil lengkap disajikan pada Tabel 4. Seluruh model menunjukkan stabilitas tinggi antar *fold* dengan standar deviasi ≤1,00%, mengindikasikan tidak adanya *overfitting* pada level *cross-validation*.

Tabel 4. Hasil 10-Fold Stratified *Cross-validation*

Algoritma	Mean Accuracy	Std Dev	Setting
MNB	81,62%	±1,00%	SMOTE
CNB	81,05%	±0,89%	SMOTE
SVM *Baseline* (LinearSVC)	80,21%	±0,79%	*Baseline*
SVM Final (SGD+CSL+*Override*)	79,39%	±0,88%	*Deployment*

Gap ~5,8pp antara CV (79,39%) dan *Holdout* (73,56%) disebabkan Lexicon *Override* yang tidak terukur dalam CV (pasca-fitting) dan Partial *Oversampling* bukan *overfitting*. Perlu ditegaskan bahwa Partial *Oversampling* dan *Coefficient Override* tidak diimplementasikan pada setiap *fold cross-validation* dalam penelitian ini, karena keduanya memerlukan modifikasi *loop* CV kustom di luar fungsi *cross*_val_score standar *scikit-learn*; hal ini diakui sebagai keterbatasan metodologis (lihat Bagian 3.7).

3.2. Performa *Holdout Test set*

Evaluasi pada *holdout test set* (n=2.405) memberikan gambaran performa model pada data baru. Ringkasan performa keseluruhan disajikan pada Tabel 5, evaluasi per-kelas pada Tabel 6, dan visualisasi *confusion matrix* pada Gambar 3.

Tabel 5. Performa Keseluruhan *Holdout Test set* (n=2.405)

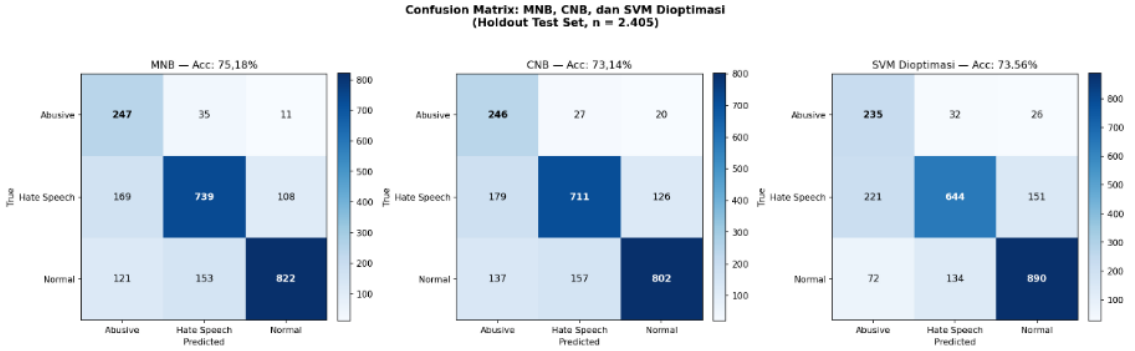
Model	Accuracy	F1 Macro	Setting
MNB	75,18%	72,10%	TF-IDF Bigram + SMOTE
CNB	73,14%	70,14%	TF-IDF Bigram + SMOTE
SVM	73,56%	70,03%	SGD+CSL+*Override* (AB=+2,0; HS=−5,0)

Model	Accuracy	F1 Macro	Setting
IndoBERT	68,69%	63,03%	Evaluasi pada *split* identik n=2.405

Keterangan: IndoBERT dievaluasi pada *split* identik (n=2.405); hasil *split* asli (Acc=78,32%, F1=74,59%) tidak dapat dibandingkan langsung. Uji McNemar mengkonfirmasi superioritas SVM (χ^2=246,0, p=5,45×10^{-15}).

Tabel 6. Evaluasi Per-Kelas: Precision, Recall, F1 (MNB, CNB, SVM) — *Holdout Test set* (n=2.405)

Kelas	MNB-P	MNB-R	MNB-F1	CNB-P	CNB-R	CNB-F1	SVM-P	SVM-R	SVM-F1
Abusive	46,00%	84,30%	59,52%	43,77%	83,96%	57,54%	44,51%	80,20%	57,25%
Hate Speech	79,72%	72,74%	76,07%	79,44%	69,98%	74,41%	79,51%	63,39%	70,54%
Normal	87,35%	75,00%	80,71%	84,60%	73,18%	78,47%	83,41%	81,20%	82,29%
Macro Avg	71,02%	77,35%	72,10%	69,27%	75,70%	70,14%	69,14%	74,93%	70,03%



Gambar 3. *Confusion Matrix* Tiga Model pada *Holdout Test set* (n=2.405): (kiri) MNB, (tengah) CNB, (kanan) SVM

Fenomena *accuracy* paradox terlihat jelas pada SVM: model *baseline* memiliki accuracy 77,96% namun Recall kelas Abusive di bawah 10%. Setelah penerapan CSL dan Lexicon *Override*, Recall Abusive meningkat drastis menjadi 80,20% (+70 persentase poin). Perlu dicatat secara eksplisit bahwa MNB mencapai Recall Abusive tertinggi (84,30%) dan F1 Macro tertinggi (72,10%) di antara model klasikal — sehingga MNB merupakan pilihan optimal untuk use-case yang memprioritaskan akurasi per-kelas di atas latensi. Precision Abusive yang relatif rendah pada seluruh model (43,77%–46,00%) dijelaskan oleh tumpang tindih linguistik antara kelas Abusive dan Hate Speech: banyak kata makian generik (mis. "anjing", "kampret") juga muncul pada tweet Hate Speech yang menyertakan target identitas, sehingga model TF-IDF berbasis kata cenderung salah mengklasifikasikan tweet HS bermuatan makian sebagai Abusive murni (false positive). Sebaliknya, *Precision Hate Speech* antar model relatif konsisten (79,44%–79,72%) karena kelas ini memiliki dukungan sampel lebih besar (n=1.016) dan pola linguistik penargetan identitas yang lebih terpisah secara leksikal dari sekadar kata kasar generik.

Untuk mengkuantifikasi pemilihan model secara formal dan menghindari keputusan berbasis single metric yang dapat menyesatkan pada *dataset* imbalanced [19], didefinisikan *Deployment* Utility Score sebagai fungsi tertimbang empat dimensi operasional:

U = 0,40 × Recall_Abusive + 0,25 × Recall_Normal + 0,20 × F1_Macro + 0,15 × Latensi_Score

Bobot ditetapkan berdasarkan prioritas operasional sistem moderasi konten near real-time: (1) Recall_Abusive diberi bobot tertinggi (0,40) karena kegagalan mendeteksi konten abusif memiliki konsekuensi regulatoris langsung di bawah UU ITE [2]; (2) Recall_Normal (0,25) meminimalkan *false alarm* pada kelas mayoritas; (3) F1_Macro (0,20) memastikan keseimbangan performa lintas kelas [18]; dan (4) Latensi_Score (0,15) merefleksikan kebutuhan throughput produksi — pada volume tinggi, latensi rendah (SVM, 0,10 ms) mencegah backlog, sementara latensi tinggi (IndoBERT, ~102,4 ms CPU) berisiko antrian [23]. Latensi_Score dinormalisasi min-max (SVM=100, MNB=62,5); perhitungan lengkap pada Tabel 7. Bobot keempat komponen

ditetapkan peneliti berdasarkan prioritas operasional yang diasumsikan, belum divalidasi melalui elisitasi pakar domain/pengguna akhir; kesimpulan superioritas SVM dibatasi pada skenario bobot yang diuji (lihat Tabel 8), bukan kesimpulan mutlak.

Tabel 7. Analisis Komponen *Deployment Utility Score*

Komponen	Bobot	MNB	SVM	Nilai Tertimbang MNB	Nilai Tertimbang SVM
Recall Abusive (%)	0,40	84,30	80,20	33,72	32,08
Recall Normal (%)	0,25	75,00	81,20	18,75	20,30
F1 Macro (%)	0,20	72,10	70,03	14,42	14,01
Latensi Score	0,15	62,50	100,00	9,38	15,00
U_*deployment*	—	—	—	76,27	81,39

Catatan: Latensi Score = normalisasi min-max (lebih cepat = lebih tinggi). MNB: 62,50; SVM: 100,00.

Untuk memvalidasi robustitas keputusan terhadap variasi preferensi bobot, dilakukan analisis sensitivitas pada tiga skenario alternatif sebagaimana disajikan pada Tabel 7.
Recall kelas *Hate Speech* tidak disertakan sebagai komponen independen dalam *Utility Score* karena dua alasan. Pertama, *Hate Speech* merupakan kelas terbesar kedua (42,23%) dan seluruh model menunjukkan Recall HS yang relatif tinggi (63,39%–78,35%) tanpa optimasi khusus, sehingga tidak menjadi faktor pembeda utama antar model. Kedua, kontribusi Recall HS sudah direpresentasikan secara implisit melalui komponen F1 Macro. Analisis sensitivitas (Tabel 8) mengkonfirmasi robustitas formula terhadap variasi pembobotan.

Tabel 8. Analisis Sensitivitas Utility Score pada Empat Skenario Bobot

Skenario	w_Ab	w_N	w_F1	w_Lat	U_MNB	U_SVM	Winner
Default (paper ini)	0,40	0,25	0,20	0,15	76,27	81,39	SVM
Prioritas Keamanan	0,60	0,20	0,15	0,05	79,52	79,86	SVM
Prioritas Kecepatan	0,20	0,20	0,20	0,40	71,28	86,29	SVM
Tanpa Latensi	0,47	0,29	0,24	0,00	78,67	78,05	MNB

Keterangan:
w_Ab = bobot Recall Abusive;
w_N = bobot Recall Normal;
w_F1 = bobot F1 Macro;
w_Lat = bobot Latensi.

Analisis sensitivitas mengungkapkan bahwa pemilihan model optimal bergantung pada prioritas kontekstual: SVM unggul pada tiga dari empat skenario (Default, Prioritas Keamanan, dan Prioritas Kecepatan), sedangkan MNB menjadi pilihan optimal pada skenario Tanpa Latensi (U_MNB=78,67 vs U_SVM=78,05). Perlu dicatat bahwa pada skenario Prioritas Keamanan, selisih antara kedua model sangat tipis (U_SVM=79,86 vs U_MNB=79,52, Δ=0,34), mengindikasikan keduanya praktis ekuivalen ketika bobot keamanan tinggi. Untuk konteks platform moderasi konten *near real-time* yang beroperasi di bawah UU ITE [2], di mana kecepatan respons dan minimasi *false alarm* sama-sama kritis. SVM dengan U_*deployment* = 81,39 merupakan pilihan yang paling seimbang secara operasional. Meskipun demikian, untuk platform yang memprioritaskan cakupan deteksi konten berbahaya di atas segalanya tanpa batasan latensi, MNB dengan *Recall Abusive* 84,30% tetap merupakan alternatif yang lebih optimal.

3.3. Uji Signifikansi McNemar

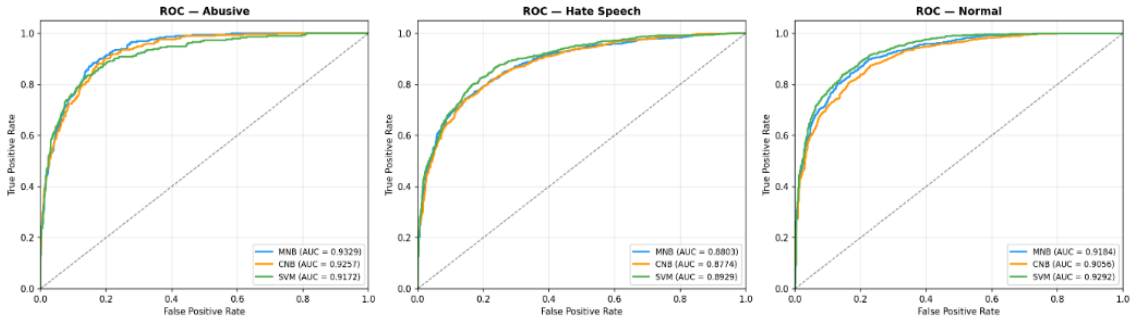
Uji McNemar dengan koreksi Bonferroni (α_adj=0,0167) dilakukan pada *holdout test set* untuk memverifikasi signifikansi perbedaan pola prediksi berpasangan antar model lebih tepat dibanding uji independen karena model dievaluasi pada sampel identik. Hasil disajikan pada Tabel 9, kurva ROC-AUC pada Gambar 4. Dampaknya: superioritas SVM/MNB terhadap CNB terbukti signifikan secara statistik (bukan sekadar variasi acak) pada α=0,0167 setelah koreksi Bonferroni untuk 3 perbandingan berpasangan.

Tabel 9. Hasil Uji McNemar dengan Koreksi Bonferroni (α_adj=0,0167, n=2.405)

Pasangan Model	b	c	χ^2 (Yates)	p-value	Signifikan (α=0,0167)
SVM vs MNB	220	346	27,6060	<0,0001	✓ **Ya**
SVM vs CNB	164	259	20,8889	0,000005	✓ **Ya**
CNB vs MNB	122	153	3,2727	0,0704	✗ Tidak
CNB dgn SMOTE vs CNB tanpa SMOTE	58	164	49,6622	$1,83\times10^{-12}$	✓ Ya

Keterangan: CNB tanpa SMOTE, *pipeline* utama (bukan *pipeline* ablation), *split* identik (n=2.405): Acc=77,55%, Recall Abusive=69,62%, F1 Abusive=61,72%.

* Uji konfirmasi *ablation*; di luar kerangka H_{1a}–H_{1c}, tidak dikenakan koreksi Bonferroni.



Gambar 4. Kurva ROC-AUC per Kelas untuk MNB, CNB, dan SVM pada *Holdout Test set*

Berdasarkan hasil uji McNemar, H_0 berhasil ditolak. H_{1a} (SVM≠MNB) dan $H_{1\beta}$ (SVM≠CNB) terkonfirmasi pada α=0,0167. H_{1c} (CNB≠MNB) tidak terkonfirmasi (p=0,0704 > α_adj), mengindikasikan interaksi negatif SMOTE×CNB pada *dataset* dan konfigurasi penelitian ini, diverifikasi terpisah pada *pipeline* utama: CNB tanpa SMOTE mencapai akurasi 77,55% dan Recall Abusive 69,62%, vs CNB dengan SMOTE (akurasi 73,14%, Recall Abusive 83,96%; χ^2=49,6622, $p=1,83\times10^{-12}$, sangat signifikan). Interaksi ini dijelaskan dua mekanisme hipotesis: (1) Dalam konfigurasi penelitian ini, penambahan SMOTE pada CNB berpotensi menciptakan koreksi ganda yang mendistorsi estimasi kelas mayoritas; (2) sampel sintetis SMOTE adalah interpolasi linear dalam ruang fitur TF-IDF [14] yang dapat melanggar asumsi distribusi multinomial CNB. Trade-off: CNB tanpa SMOTE unggul di akurasi global dan F1 Abusive tertinggi antar model klasikal (61,72%), namun Recall Abusive lebih rendah; CNB dengan SMOTE memaksimalkan Recall Abusive yang diprioritaskan dalam konteks UU ITE [2], dengan konsekuensi akurasi turun 4,41pp. Panduan: gunakan CNB+SMOTE bila recall minoritas prioritas absolut; CNB tanpa SMOTE bila keseimbangan presisi lintas kelas diutamakan. ROC-AUC: tidak ada model dominan di semua kelas MNB unggul Abusive (AUC=0,9329), CNB unggul Normal (0,9350), SVM unggul *Hate Speech* (0,8929).

Koreksi Bonferroni merupakan metode paling konservatif; alternatif Holm-Bonferroni [19] memberikan power lebih tinggi. Namun seluruh keputusan signifikansi tetap konsisten bahkan dengan koreksi ketat ini, sehingga pilihan metode tidak memengaruhi kesimpulan.

3.4. *Ablation Study Preprocessing*

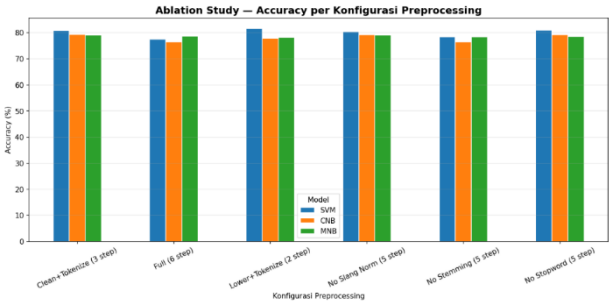
Ablation study atas enam konfigurasi *preprocessing* dilakukan untuk mengidentifikasi kontribusi setiap tahap terhadap performa model. Catatan: pada kolom SVM, seluruh konfigurasi dievaluasi menggunakan LinearSVC sebagai SVM *Baseline* (tanpa CSL, Partial *Oversampling*, maupun Lexicon *Override*); kolom MNB dan CNB dievaluasi menggunakan konfigurasi standar masing-masing (α=1,0, tanpa SMOTE), sehingga *ablation study* ini murni mengisolasi efek tahapan *preprocessing*, terpisah dari mekanisme penanganan *imbalanced class*. Hasil disajikan pada Tabel 10 dan divisualisasikan pada Gambar 5.

Tabel 10. Hasil Ablation Study *Preprocessing* (SVM *Baseline*, *Holdout Test set*, n=2.405)

Konfigurasi *Preprocessing*	Acc SVM (%)	Acc CNB (%)	Acc MNB (%)
Full (6 tahap)	77,55	76,51	78,67
Tanpa *Stemming* (5 tahap)	78,50	76,51	78,50
Tanpa *Stopword removal* (5 tahap)	80,91	79,19	78,60
Tanpa Normalisasi Slang (5 tahap)	80,43	79,19	79,07
Cleaning + Tokenisasi (3 tahap)	80,79	79,34	79,11
Lowercase + Tokenisasi (2 tahap)*	81,60	77,86	78,21
SVM CNB MNB Final ★	73,56	73,14	75,18

*Recall Abusive <10% pada konfigurasi ini meski accuracy tinggi.

★Metrik relevan model final adalah F1 Macro, bukan accuracy semata.



Gambar 5. Hasil Ablation Study *Preprocessing*: Perbandingan Accuracy SVM, CNB, dan MNB pada 6 Konfigurasi

Tiga temuan kunci dari ablation study: (1) *Stopword removal* merupakan tahap prapemrosesan yang paling memengaruhi performa secara negatif — implementasi tahap ini mereduksi accuracy SVM sebesar 3,36 pp (dari 80,91% menjadi 77,55%), mengindikasikan adanya penghapusan term yang secara semantik penting untuk klasifikasi ujaran kebencian; (2) Normalisasi slang memberikan kontribusi signifikan dengan penurunan 2,88 pp jika dihilangkan; (3) Konfigurasi Lowercase+Tokenisasi menghasilkan accuracy tertinggi (81,60%) namun menunjukkan Recall Abusive di bawah 10%, mempertegas bahwa accuracy bukan metrik relevan untuk *deployment* pada *dataset* imbalanced. Secara spesifik, konfigurasi Clean+Tokenize (3 tahap, 80,79%) lebih tinggi dari Full 6-tahap (77,55%) karena tidak menyertakan *stopword removal*/*stemming* yang paling merusak sinyal kata kasar, namun sedikit lebih rendah dari 2-tahap (81,60%) karena tokenisasi eksplisit memecah entitas gabungan (hashtag/kata majemuk) yang mengurangi term diskriminatif pada *vocabulary* TF-IDF.

3.5. *Error analysis* Linguistik

Analisis *error* kuantitatif dilakukan pada n=50 sampel *misclassified* per model (total 150 sampel). Anotasi kategori *error* dilakukan secara independen oleh dua annotator: penulis dan satu annotator eksternal yang berasal dari Fakultas Bahasa dan Seni untuk memahami tweet dari bahasa Indonesia terutama dan Inggris. Inter-rater agreement diukur menggunakan Cohen's Kappa [22] secara terpisah per model: κ_MNB=0,697, κ_CNB=0,739, dan κ_SVM=0,794, dengan rata-rata κ=0,743 yang diinterpretasikan sebagai substantial agreement berdasarkan interpretasi skala kappa [22]. Sampel dengan ketidaksepakatan (MNB: 11 sampel; CNB: 10 sampel; SVM: 8 sampel) diselesaikan melalui diskusi konsensus antara kedua annotator sebelum analisis distribusi dilakukan.

Tabel 11. Distribusi Pola *Error* dengan Wilson 95% CI (n=50 sampel per model). Wilson interval dipilih (bukan normal *approximation*) karena lebih akurat untuk proporsi pada sampel kecil/ekstrem, mencegah interval melampaui rentang 0–100%, sehingga interpretasi overlap antar-model lebih dapat dipercaya meski tetap eksploratif.

Tabel 11. Distribusi Pola Error dengan Wilson 95% CI (n=50 sampel per model)

Pola Error	MNB [95% CI]	CNB [95% CI]	SVM [95% CI]
Code-switching / Ironi	12 (24%) [14.3%, 37.4%]	10 (20%) [11.2%, 33.0%]	12 (24%) [14.3%, 37.4%]
Ambiguitas HS-Abusive	15 (30%) [19.1%, 43.8%]	13 (26%) [15.9%, 39.6%]	11 (22%) [12.8%, 35.2%]
Ambiguitas Target	16 (32%) [20.8%, 45.8%]	14 (28%) [17.5%, 41.7%]	7 (14%) [7.0%, 26.2%]
Ekspresi Implisit	4 (8%) [3.2%, 18.8%]	4 (8%) [3.2%, 18.8%]	6 (12%) [5.6%, 23.8%]
Tidak Terkategorikan	3 (6%) [2.1%, 16.2%]	9 (18%) [9.8%, 30.8%]	14 (28%) [17.5%, 41.7%]

Distribusi pola error berdasarkan hasil konsensus disajikan pada Tabel 11. Tiga temuan utama dapat diidentifikasi. Pertama, MNB didominasi oleh Ambiguitas Target (32,0%) dan Ambiguitas HS-Abusive (30,0%), mengindikasikan bahwa kesalahan MNB terutama terjadi pada tweet dengan target ujaran yang ambigu atau yang berada di batas antara konten Abusive dan *Hate Speech*. Kedua, SVM menunjukkan profil error yang berbeda secara substansial dibandingkan MNB: proporsi "Tidak Terkategorikan" pada SVM (28,0%) jauh lebih tinggi dibandingkan MNB (6,0%), dan merupakan satu-satunya pasangan dengan confidence interval yang tidak saling tumpang tindih (MNB: [2,1–16,2] vs SVM: [17,5–41,7]), mengindikasikan SVM cenderung membuat kesalahan pada tweet berkarakteristik noise tinggi yang tidak masuk kategori error linguistik standar. Ketiga, profil error MNB dan CNB relatif serupa di seluruh kategori dengan confidence interval yang saling tumpang tindih, konsisten dengan temuan McNemar bahwa perbedaan pola prediksi keduanya tidak signifikan secara statistik (p=0,0704). Perlu dicatat bahwa dengan n=50 per model, lebar Wilson CI berkisar antara 13–25 persentase poin, sehingga interpretasi perbedaan antar-model bersifat eksploratif dan perlu dikonfirmasi pada sampel yang lebih besar.

3.6. Komparasi SVM vs IndoBERT

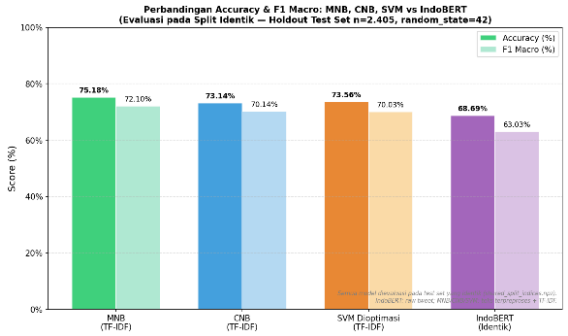
Perbandingan final seluruh model pada *split* identik (n=2.405) disajikan pada Tabel 12 dan Gambar 6. Uji McNemar mengkonfirmasi superioritas SVM terhadap IndoBERT secara statistik (χ^2=246,0, p=5,45×10^{-15}).

Gambar 6 menyajikan accuracy dan F1 Macro seluruh model pada *split* identik (n=2.405), *apple-to-apples*; IndoBERT (Acc=68,69%, F1=63,03%) di bawah ketiga model klasikal. Tabel 13 menyajikan performa IndoBERT pada *split* asli per *epoch* (terbaik *Epoch* 3, F1=74,53%) sebagai konteks tambahan, bukan pembanding langsung. Superioritas model klasikal kemungkinan karena: (1) IndoBERT hanya di-fine-tune sekali tanpa *cross-validation* (Bagian 3.7); (2) *Coefficient Override* memberi model klasikal keunggulan langsung mengenali kata makian eksplisit; (3) data training (9.619 tweet) relatif kecil untuk *fine-tuning* transformer optimal. Perbandingan SVM vs IndoBERT juga bukan sepenuhnya *apple-to-apples* secara arsitektur pembanding lebih setara adalah IndoBERT vs SBERT, direkomendasikan untuk penelitian lanjutan.

Tabel 12. Perbandingan Final Semua Model: Akurasi, F1 Macro, Latensi, dan *Hardware*

Model	Acc (%)	F1 Macro	Latensi Inf.	End-to-End	HW	Keterangan
MNB	75,18	72,10%	0,16 ms	4,65 ms avg	CPU	Bigram+SMOTE
CNB	73,14	70,14%	0,17 ms	4,65 ms avg	CPU	Bigram+SMOTE
SVM	73,56	70,03%	0,10 ms	4,62 ms avg	CPU	SGD+CSL+*Override*
IndoBERT†	68,69	63,03%	4,89 ms	93,11 ms avg	GPU T4	*Split* identik n=2.405

Keterangan: *Split* identik (n=2.405). Perbandingan SVM (CPU) vs IndoBERT (GPU T4) lintas-hardware tidak sepenuhnya setara; sebagai pembanding tambahan pada *hardware* yang sama (CPU), IndoBERT berlatensi ~102,4 ms sehingga rasio menjadi ~1.020× (inferensi murni, CPU-to-CPU).



Gambar 6. Perbandingan *Accuracy* dan F1 Macro Seluruh Model pada *Split* Identik (*Holdout Test set*, n=2.405). Seluruh model, termasuk IndoBERT, dievaluasi pada *split* data uji yang sama.

Tabel 13. Performa IndoBERT pada *Split* Asli *Fine-tuning* per *Epoch* (bukan *split* identik, tidak dapat dibandingkan langsung dengan Tabel 12).

Epoch	Val Accuracy	Val F1 Macro
1	68,22%	65,25%
2	75,82%	72,50%
3 (best)	77,71%	74,53%
4	78,28%	74,26%
5	78,44%	74,23%

Panduan *deployment* berdasarkan Tabel 13: (1) CPU-only atau volume tinggi → SVM (0,10ms, Recall Abusive 80,20%); (2) Prioritas deteksi Abusive → MNB (Recall Abusive 84,30%); (3) GPU tersedia → IndoBERT (F1=74,59%). Dalam konteks implementasi di bawah UU ITE Nomor 1 Tahun 2024 Pasal 28 ayat (2) [2], terdapat beberapa pertimbangan operasional yang perlu diperhatikan. Pertama, dengan Precision kelas Abusive sebesar 44,51% pada SVM, sekitar 55% konten yang ditandai sebagai abusif merupakan *false positive* implikasi ini signifikan terhadap kebebasan berekspresi pengguna. Oleh karena itu, sistem ini direkomendasikan sebagai alat bantu triase (*first-pass* filter) yang memerlukan verifikasi manusia sebelum tindakan penghapusan atau sanksi. Kedua, threshold confidence untuk flagging konten perlu dikalibrasi berdasarkan toleransi risiko platform: *threshold* rendah memaksimalkan *recall* namun meningkatkan beban moderator manusia, sementara *threshold* tinggi mengurangi *false alarm* namun meningkatkan risiko konten berbahaya yang lolos. Kalibrasi optimal ini merupakan keputusan kebijakan yang memerlukan konsultasi dengan tim hukum dan moderasi platform.

3.7. Keterbatasan Penelitian

Penelitian ini memiliki enam keterbatasan. Pertama, seluruh eksperimen menggunakan satu *dataset* [4] dari Twitter pada periode tertentu belum divalidasi pada platform lain (Instagram, TikTok) atau data temporal lebih baru, sehingga generalisasi dan robustitas terhadap concept drift belum dapat diklaim. Kedua, *fine-tuning* IndoBERT hanya satu seed tanpa *cross-validation*, sehingga variance-nya tidak terukur; perbandingan SVM vs IndoBERT perlu diinterpretasikan dengan keterbatasan ini. Ketiga, bobot *Deployment* Utility Score (w_Ab=0,40; w_N=0,25; w_F1=0,20; w_Lat=0,15) ditetapkan peneliti, belum divalidasi survei stakeholder meski analisis sensitivitas dilakukan (Tabel 8), pemilihan bobot default tetap subjektif. Keempat, penelitian ini tidak membandingkan dengan LLM (GPT-4, Llama) fokus model klasikal untuk solusi deployable pada infrastruktur CPU minimal. Kelima, leksikon 24 kata *Coefficient Override* bersifat statis tanpa pembaruan otomatis terhadap evolusi kosakata. Keenam, gap ~5,8pp antara CV (79,39%) dan *holdout* (73,56%) pada SVM Final adalah konsekuensi metodologis (Bagian 3.1), bukan *overfitting*.

4. KESIMPULAN DAN SARAN

Penelitian ini menghasilkan empat temuan utama. Pertama, SVM (CSL+Lexicon-Based *Coefficient Override*) memperoleh *Utility Score* tertinggi (81,39) pada skenario bobot yang diuji, latensi 0,10 ms CPU; berpotensi sebagai alat bantu penyaringan awal (bukan dasar otomatis

sanksi) di bawah UU ITE [2], tetap memerlukan verifikasi manusia mengingat *Precision Abusive* 44,51%. Kedua, MNB mencapai Recall Abusive tertinggi (84,30%) dan F1 Macro tertinggi (72,10%) di antara model klasikal optimal ketika cakupan deteksi diprioritaskan di atas latensi. Ketiga, uji McNemar mengkonfirmasi H_{1a} dan $H_{1\beta}$ ($p<0,0001$), namun H_{1c} tidak terkonfirmasi ($p=0,0704$), mengindikasikan interaksi negatif SMOTE×CNB pada *dataset* dan konfigurasi ini bukan kesimpulan umum. Keempat, ablation study mengidentifikasi *stopword removal* sebagai tahap terkritis (−3,36pp); konfigurasi minimal tanpa penanganan imbalanced class gagal mendeteksi Abusive (*Recall* <10%).

Kode implementasi tersedia di github.com/feastco/hate-speech-detection. Penelitian lanjutan yang direkomendasikan: (1) *cross*-temporal *validation* pada *dataset* IndoToxic2024 [24] untuk mengukur concept drift; (2) *fine-tuning* IndoBERTweet [25] yang spesifik untuk korpus Twitter Indonesia; (3) eksplorasi ensemble SVM+IndoBERT dengan meta-learner; serta (4) *multi-seed evaluation* IndoBERT untuk menghasilkan confidence interval yang valid secara statistik.

DAFTAR PUSTAKA

[1] Asosiasi Penyelenggara Jasa Internet Indonesia (APJII), "*Survei Profil Internet Indonesia 2026*," 2026. [Online]. Available: https://apjii.or.id

[2] Republik Indonesia, "*Undang-Undang Nomor 1 Tahun 2024 tentang Perubahan Kedua atas Undang-Undang Nomor 11 Tahun 2008 tentang Informasi dan Transaksi Elektronik*," 2024. [Online]. Available: https://peraturan.bpk.go.id/details/274494/uu-no-1-tahun2024

[3] A. Gandhi, P. Ahir, K. Adhvaryu, P. Shah, R. Lohiya, E. Cambria, S. Poria, and A. Hussain, "Hate speech detection: A comprehensive review of recent works," *Expert Systems*, vol. 41, no. 8, Art. e13562, 2024, doi: 10.1111/exsy.13562.

[4] M. O. Ibrohim and I. Budi, "Multi-label Hate Speech and Abusive Language Detection in Indonesian Twitter," in *Proceedings of the 3rd Workshop on Abusive Language Online (ALW3)*, 2019, pp. 46–57, doi: 10.18653/v1/w19-3506.

[5] I. Putu Widiarta Nandana Githa, A. Syananda, R. Faustine, I. S. Edbert, and D. Suhartono, "Hate Speech Classification in Indonesian Tweets Using TF-IDF and Data Augmentation," in *2024 International Conference on Green Energy, Computing and Sustainable Technology, GECOST 2024*, 2024, pp. 61–65, doi: 10.1109/GECOST60902.2024.10474781.

[6] I. I. R. Difandana and I. Imaduddin, "Comparative Analysis of Naive Bayes and Support Vector Machine for Hate Speech Classification," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 6, no. 1, pp. 414–422, Feb. 2026, doi: 10.57152/MALCOM.V6I1.2571.

[7] S. A. Zikrina and F. Fitriyani, "Advancing Hate Speech Detection in Indonesian Language Using Graph Neural Networks and TF-IDF," *J. RESTI*, vol. 9, no. 1, pp. 137–145, 2025, doi: 10.29207/resti.v9i1.6179.

[8] S. D. A. Putri, M. O. Ibrohim, and I. Budi, "Abusive Language and Hate Speech Detection for Indonesian-Local Language in Social Media Text," in *Lecture Notes in Networks and Systems*, 2021, pp. 88–98, doi: 10.1007/978-3-030-79757-7_9.

[9] N. Khoirunnisaa, K. Nabila Nastiti Kesuma, S. Setiawan, and A. Yunizar Pratama Yusuf, "Klasifikasi teks ulasan aplikasi Netflix pada Google Play Store menggunakan algoritma Naïve Bayes dan SVM," *SKANIKA Sist. Komput. dan Tek. Inform.*, vol. 7, no. 1, pp. 64–73, 2024, doi: 10.36080/skanika.v7i1.3138.

[10] M. O. Ibrohim and I. Budi, "Hate speech and abusive language detection in Indonesian social media: Progress and challenges," *Heliyon*, vol. 9, no. 8, Art. e18647, 2023, doi: 10.1016/j.heliyon.2023.e18647.

[11] A. N. A. Saputra, R. E. Saputro, and D. I. S. Saputra, "Labeling Optimization and Hybrid CNN Model in Sentiment Analysis of Movie Reviews with Slang Handling," *Jurnal Teknik Informatika (JUTIF)*, vol. 6, no. 6, 2025, doi: 10.52436/1.jutif.2025.6.6.4465.

[12] X. Xiang, "Application of an Improved TF-IDF Method in Literary Text Classification," *Advances in Multimedia*, 2022, doi: 10.1155/2022/9285324.

[13] D. Rosadi and S. Rakasiwi, "Anti-Data Leakage Pipeline for Differentiated Thyroid Cancer Recurrence Prediction: Integrating SMOTE, Optuna-based Optimization, and Bootstrap BCa Validation," *Journal of Applied Informatics and Computing*, vol. 10, no. 3, 2026, doi: 10.30871/jaic.v10i3.12922.

[14] M. Koren, O. Peretz, and O. Koren, "Naive Bayes classifier – An ensemble procedure for recall and precision enrichment," *Engineering Applications of Artificial Intelligence*, vol. 136, Part B, Art. 108972, 2024, doi: 10.1016/j.engappai.2024.108972.

[15] D. Elreedy, A. F. Atiya, and F. Kamalov, "A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning," *Machine Learning*, vol. 113, pp. 4903–4923, 2024, doi: 10.1007/s10994-022-06296-4.

[16] A. M. Mardiana, I. F. Rozi, and R. Arianto, "Support Vector Machine with FastText Word Embedding for Hate Speech Aspect Categorization," *Paradigma - Jurnal Komputer dan Informatika*, vol. 27, no. 2, pp. 92–98, 2025, doi: 10.31294/p.v27i2.5127.

[17] W. Chen, K. Yang, Z. Yu, Y. Shi, and C. L. P. Chen, "A survey on imbalanced learning: latest research, applications and future directions," *Artif. Intell. Rev.*, vol. 57, no. 6, p. 137, 2024, doi: 10.1007/s10462-024-10759-6.

[18] A. S. Aribowo, Y. Fauziah, Y. Bantulu, S. Saifullah, and A. M. A. Fubalo, "Hate Speech Analysis Using IndoBERT in YouTube Comments on the 2024 Indonesian Presidential Debate Video," *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, vol. 11, no. 3, 2026, doi: 10.22219/kinetik.v11i3.2604.

[19] O. Rainio, J. Teuho, and R. Klén, "Evaluation metrics and statistical tests for machine learning," *Scientific Reports*, vol. 14, Art. 6086, 2024, doi: 10.1038/s41598-024-56706-x.

[20] T. Sugihartono and R. R. C. Putra, "Penerapan metode Support Vector Machine dalam klasifikasi ulasan pengguna aplikasi Mobile JKN," *SKANIKA Sist. Komput. dan Tek. Inform.*, vol. 7, no. 2, pp. 144–153, 2024, doi: 10.36080/skanika.v7i2.3193.

[21] L. Andersson and A. Nerman, "A Note on Confidence Intervals for a Binomial p: Andersson–Nerman vs. Wilson," *Stat*, vol. 13, Art. e70027, 2024, doi: 10.1002/sta4.70027.

[22] G. Rau and Y.-S. Shih, "Evaluation of Cohen's kappa and other measures of inter-rater agreement for genre analysis and other nominal data," *Journal of English for Academic Purposes*, vol. 53, Art. 101026, 2021, doi: 10.1016/j.jeap.2021.101026.

[23] scikit-learn 1.3.0, *Zenodo*, 30 June 2023, doi: 10.5281/zenodo.8098905. [Online]. Available: https://zenodo.org/record/8098905

[24] M. I. Wijanarko, L. Susanto, P. A. Pratama, I. Idris, T. Hong, and D. Wijaya, "Monitoring Hate Speech in Indonesia: An NLP-based Classification of Social Media Texts," in *Proceedings of EMNLP 2024: System Demonstrations*, pp. 142–152, 2024, doi: 10.18653/v1/2024.emnlp-demo.15.

[25] F. Koto, J. H. Lau, and T. Baldwin, "INDOBERTWEET: A Pretrained Language Model for Indonesian Twitter with Effective Domain-Specific Vocabulary Initialization," in *EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings*, 2021, pp. 10660–10668, doi: 10.18653/v1/2021.emnlp-main.833.