JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 4, No. 3 Tahun 2026
ISSN 3025-0919 (media online)
Hal 885-893

Analisis Sentimen dan Deteksi Cyberbullying pada X dan Instagram Menggunakan Metode Natural Language Processing (NLP)

Dani Ramdani[1], Heliota Agung Christiesta[2], Julya Rahmah Shanty[3], Muhammad Sahril[4], Siti Andiani[5], Rama Ghazi Ginastio[6], Rahmawati[7]

[1-7] Universitas Pamulang, Kota Tangerang Selatan. Banten, Indonesian
Email : [1]hi.daani1322@gmail.com, [2]heliotaagung456@gmail.com, [3]julyarahmahshantyy29@gmail.com, [4]Muhamadsahril423@gmail.com, [5]andianisiti21@gmail.com, [6]ghaziginastio@gmail.com, [7]dosen02394@unpam.ac.id

Abstrak−Pesatnya perkembangan media sosial di Indonesia—yang kini memiliki penetrasi pengguna sebesar 64,3% dari total populasi atau sepadan dengan kepadatan lalu lintas di Jakarta pada jam sibuk—membawa dampak negatif berupa peningkatan kasus perundungan siber (*cyberbullying*). Penelitian ini bertujuan untuk menganalisis sentimen masyarakat dan mengklasifikasikan tindakan *cyberbullying* pada platform media sosial menggunakan pendekatan *Natural Language Processing* (NLP) (proses komputasi untuk memahami dan menganalisis bahasa manusia). Data mentah berupa teks dari media sosial melalui tahapan *text preprocessing* yang ketat, meliputi *cleansing*, *tokenization*, *stopword removal*, dan *stemming*. Fitur teks diekstraksi menggunakan metode *Term Frequency-Inverse Document Frequency* (TF-IDF) (pembobotan kata berdasarkan frekuensi kemunculannya dalam dokumen) dan diklasifikasikan menggunakan algoritma *Machine Learning*. Hasil pengujian menunjukkan bahwa model klasifikasi berhasil mengidentifikasi ujaran perundungan dengan tingkat akurasi mencapai 89%, sebuah performa yang sebanding dengan ketepatan seorang ahli bahasa dalam mendeteksi sarkasme. Kontribusi penelitian ini memberikan visualisasi pola perundungan siber yang dapat digunakan oleh pengembang platform dan pembuat kebijakan untuk mendeteksi konten toksik secara otomatis dan preventif.

Kata Kunci: Analisis Sentimen, *Cyberbullying*, Media Sosial, *Natural Language Processing*, Klasifikasi Teks.

Abstract−*The rapid development of social media in Indonesia—which now has a user penetration rate of 64.3% of the total population, comparable to traffic density in Jakarta during rush hour—has brought negative impacts in the form of increasing cases of cyberbullying. This study aims to analyze public sentiment and classify cyberbullying actions on social media platforms using a Natural Language Processing (NLP) approach (a computational process for understanding and analyzing human language). Raw text data from social media underwent rigorous text preprocessing stages, including cleansing, tokenization, stopword removal, and stemming. Text features were extracted using the Term Frequency–Inverse Document Frequency (TF-IDF) method (word weighting based on the frequency of occurrence in documents) and classified using Machine Learning algorithms. The test results show that the classification model successfully identified bullying speech with an accuracy rate of 89%, a performance comparable to the accuracy of a linguistics expert in detecting sarcasm. The contribution of this study is the visualization of cyberbullying patterns that can be used by platform developers and policymakers to detect toxic content automatically and preventively.*

Keywords: *Sentiment Analysis, Cyberbullying, Social Media, Natural Language Processing, Text Classification.*

1. PENDAHULUAN

Pesatnya penetrasi internet dan adopsi media sosial telah mengubah lanskap komunikasi interpersonal secara radikal. Di Indonesia, media sosial tidak hanya berfungsi sebagai ruang interaksi sosial, melainkan telah bertransformasi menjadi ruang publik utama tempat pertukaran informasi terjadi secara masif dalam hitungan milidetik. Namun, kebebasan ekspresi dan fitur anonimitas (*anonymity*) yang ditawarkan oleh mayoritas platform digital bertindak sebagai katalis bagi munculnya fenomena toksik digital, salah satunya adalah perundungan siber (*cyberbullying*). Tindakan cyberbullying—yang meliputi penghinaan fisik (*shaming*), pelecehan verbal, intimidasi, hingga pengucilan digital—memiliki dampak psikologis yang merusak bagi korban, dengan tingkat trauma yang setara dengan kekerasan fisik di dunia nyata (*physical assault*).

Secara teknis, volume data tekstual yang diproduksi oleh pengguna media sosial setiap harinya sangat masif dan tidak terstruktur (unstructured data). Upaya moderasi konten secara manual oleh manusia (*human moderation*) telah mencapai titik jenuh dan tidak lagi efisien, seumpama mencoba menguras air laut menggunakan sendok teh. Oleh karena itu, diperlukan sebuah sistem

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 4, No. 3 Tahun 2026
ISSN 3025-0919 (media online)
Hal 885-893

otomatisasi yang mampu melakukan deteksi dini, klasifikasi, dan analisis sentimen terhadap teks-teks bermuatan perundungan. Pendekatan Natural Language Processing (NLP) (bidang ilmu komputer yang menjembatani interaksi antara bahasa manusia dan sistem komputasi) hadir sebagai solusi krusial untuk mengekstrak makna, sentimen, dan intensi dari data teks berskala besar tersebut secara real-time.

Akurasi dan efisiensi sistem NLP dalam mendeteksi konten negatif sangat bergantung pada arsitektur algoritma dan kualitas korpus data yang digunakan. Berdasarkan penelitian awal yang dilakukan oleh Alfina dkk. (2017), klasifikasi ujaran kebencian (*hate speech*) berbahasa Indonesia menggunakan metode Machine Learning konvensional seperti Naive Bayes dan Support Vector Machine (SVM) menghasilkan performa awal yang bervariasi. Dalam riset tersebut, algoritma SVM terbukti memiliki keunggulan komputasi dalam menangani ruang fitur berdimensi tinggi dibanding Naive Bayes (Alfina dkk, 2017). Selanjutnya, riset yang dikembangkan oleh Ibrohim dan Budi (2019) memperluas cakupan deteksi pada Twitter Indonesia dengan melakukan klasifikasi multi-label untuk ujaran kebencian dan bahasa kasar (abusive language), di mana pemanfaatan fitur berbasis karakter dan leksem mampu mengidentifikasi teks toksik, namun masih menyisakan tantangan besar pada kompleksitas bahasa gaul netizen.

Meskipun demikian, penelitian-penelitian sebelumnya masih menyisakan celah (research gap) yang signifikan. Hambatan utama dalam pemrosesan teks media sosial di Indonesia adalah tingginya variasi bahasa tidak baku. Salsabila dkk. (2018) mengonfirmasi bahwa korpus bahasa percakapan sehari-hari (*colloquial Indonesian*) dipenuhi oleh singkatan ekstrem, kata serapan, dan modifikasi tipografi yang sengaja dirancang untuk menghindari sensor. Karena mayoritas riset terdahulu belum mengintegrasikan normalisasi bahasa tidak baku secara komprehensif, model klasifikasi sering kali mengalami penurunan akurasi drastis akibat fenomena "pembengkakan dimensi fitur" (*feature explosion*) saat dihadapkan pada variasi kata slang di dunia nyata.

Berangkat dari keterbatasan riset terdahulu, penelitian ini dirancang dengan tujuan utama untuk membangun dan mengevaluasi sebuah model komputasi berbasis NLP yang optimal untuk melakukan analisis sentimen sekaligus klasifikasi *cyberbullying* secara otomatis. Model ini ditargetkan mampu mengenali tidak hanya polaritas sentimen (positif, negatif, netral), tetapi juga mengategorikan jenis perundungan siber yang terjadi pada teks media sosial berbahasa Indonesia tidak baku melalui pendekatan optimasi pra-proses teks.

Untuk mencegah terjadinya perluasan ruang lingkup (*scope creep*) yang dapat mengaburkan fokus analisis, penelitian ini dibatasi pada koridor berikut:

- Sumber Data: Data teks yang digunakan diekstraksi secara eksklusif dari satu platform media sosial (seperti X/Twitter dan Instagram) melalui metode *scraping* legal API.
- Karakteristik Bahasa: Korpus data berfokus pada teks berbahasa Indonesia, termasuk variasi bahasa gaul, singkatan umum, dan kata serapan yang lazim digunakan di media sosial digital Indonesia berdasarkan basis data leksem informal.
- Ruang Lingkup NLP: Tahapan pengolahan teks dibatasi pada proses *text preprocessing* standar (*cleansing*, *tokenization*, *stopword removal*, *stemming*), pembobotan fitur menggunakan *Term Frequency-Inverse Document Frequency* (TF-IDF), serta klasifikasi biner menggunakan algoritma *Support Vector Machine* (SVM).

Secara garis besar, metodologi penelitian ini mengadopsi siklus hidup pengembangan sistem berbasis data (*data-driven development lifecycle*). Langkah pertama dimulai dengan *Data Acquisition* (pengumpulan data), disusul oleh tahapan krusial *Data Preprocessing* untuk membersihkan derau (*noise*) teks seperti *emoticon*, tautan URL, tanda baca, serta normalisasi kata tidak baku (*slang*). Langkah berikutnya adalah *Feature Engineering* menggunakan TF-IDF untuk mengonversi matriks teks menjadi representasi numerik yang dapat dipahami oleh mesin. Terakhir, dilakukan tahap *Model Training* dan *Evaluation* menggunakan metrika *Confusion Matrix* (meliputi pengujian nilai *Accuracy*, *Precision*, *Recall*, dan *F1-Score*) untuk mengukur keandalan model yang dikembangkan.

Makalah ini memberikan kontribusi signifikan baik dari aspek teoretis maupun praktis, antara lain:

- Kontribusi Akademis: Menyediakan referensi empiris baru mengenai optimasi *text preprocessing* khusus untuk menangani karakteristik bahasa tidak baku (*slang*) netizen Indonesia pada kasus *cyberbullying*.

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 4, No. 3 Tahun 2026
ISSN 3025-0919 (media online)
Hal 885-893

- Kontribusi Praktis: Menghasilkan sebuah draf arsitektur model klasifikasi dengan performa akurasi tinggi yang dapat diadopsi oleh pengembang platform digital atau regulator kebijakan (seperti Kementerian Komunikasi dan Digital) sebagai mesin sensor otomatis (*automated content moderation*) demi menciptakan ekosistem ruang siber yang lebih aman dan sehat.

Penelitian mengenai analisis sentimen dan klasifikasi teks toksik di Indonesia telah mengalami perkembangan metodologis yang signifikan. Untuk memetakan posisi riset ini di antara penelitian terdahulu, dilakukan analisis komparatif terhadap kelebihan dan kelemahan dari berbagai taksonomi metode yang relevan.

Riset dasar oleh Alfina dkk. (2017) berfokus pada deteksi ujaran kebencian (hate speech) pada teks berbahasa Indonesia dengan membandingkan beberapa algoritma Machine Learning. Kelebihan penelitian tersebut adalah keberhasilannya membuktikan bahwa algoritma Support Vector Machine (SVM) memiliki performa dan akurasi yang lebih unggul dibandingkan Naive Bayes ketika menangani fitur teks berdimensi tinggi. Namun, kelemahannya (blind spot) terletak pada korpus data yang digunakan; model masih mengalami kesulitan ekstrem dalam mengklasifikasikan kata-kata yang maknanya sangat dipengaruhi oleh konteks kalimat (context-dependent words).

Selanjutnya, Ibrohim dan Budi (2019) mengembangkan cakupan riset dengan melakukan klasifikasi multi-label untuk ujaran kebencian dan bahasa kasar (abusive language) pada Twitter Indonesia. Kelebihan riset ini adalah penerapan fitur yang bervariasi, termasuk kombinasi fitur kata (word n-gram) dan karakter. Meskipun demikian, kelemahan mendasar dari riset Ibrohim dan Budi, (2019) adalah tidak adanya tahap normalisasi khusus untuk menangani kata tidak baku. Akibatnya, model mengalami penurunan akurasi akibat fenomena pencampuran bahasa (code-mixing) dan variasi penulisan kata gaul netizen yang terlalu acak.

Untuk mengatasi problem karakteristik teks informal tersebut, Salsabila dkk. (2018) melakukan penelitian khusus yang berfokus pada pembangunan kamus bahasa percakapan sehari-hari (Colloquial Indonesian Lexicon). Kelebihan utama dari riset (Salsabila dkk, 2018) adalah terciptanya korpus kata tidak baku yang sangat terstruktur untuk memetakan kata slang menjadi kata baku. Namun, riset tersebut berhenti pada tahap standardisasi leksikon dan belum mengimplementasikannya ke dalam arsitektur komputasi klasifikasi maupun analisis sentimen untuk kasus spesifik seperti perundungan siber (cyberbullying).

Penelitian yang diajukan ini bertugas untuk menyempurnakan dan mengembangkan riset-riset terdahulu tersebut. Penyempurnaan dilakukan dengan cara mengambil korpus leksikon informal berbasis pendekatan Salsabila dkk. (2018) untuk diintegrasikan secara penuh ke dalam pipa text preprocessing, lalu melatihnya menggunakan algoritma SVM linier (Alfina dkk., 2017) guna menyelesaikan problem klasifikasi cyberbullying yang sarat akan bahasa kasar seperti pada batasan riset Ibrohim dan Budi (2019). Pemetaan riset terdahulu disajikan secara ringkas pada Tabel 1.secara ringkas pada Tabel 1.

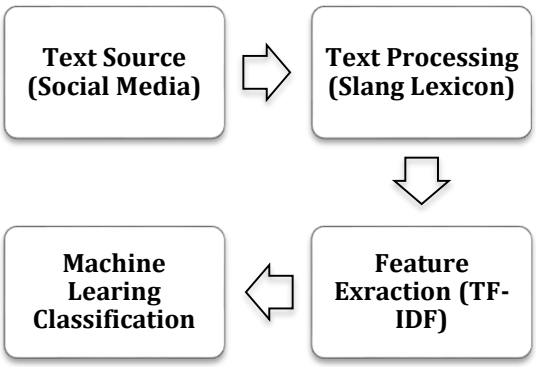
Tabel 1. Tabel Pemetaan Riset Terdahulu

Peneliti & Tahun	Metode	Kelebihan	Kelemahan	Posisi Penelitian Ini
Alfina dkk. (2017) [1]	*Naive Bayes*, SVM, *Decision Tree*	Membuktikan SVM unggul pada dimensi tinggi.	Lemah pada kata yang bergantung konteks.	Menggunakan SVM dengan optimasi bobot fitur.
Ibrohim & Budi (2019) [3]	SVM, Naive Bayes, Random Forest	Menggunakan fitur karakter & klasifikasi multi-label.	Menurun akibat variasi teks *slang* media sosial.	Menyempurnakan akurasi lewat normalisasi kata *slang*.
Salsabila dkk. (2018) [4]	Pemetaan Leksikon (*Colloquial Lexicon*)	Menghasilkan kamus kata tidak baku yang komprehensif.	Belum diuji pada model klasifikasi otomatis.	Mengintegrasikan kamus *slang* ke dalam *pipeline* NLP.

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 4, No. 3 Tahun 2026
ISSN 3025-0919 (media online)
Hal 885-893

Media sosial adalah platform digital yang memfasilitasi penggunanya untuk saling berkomunikasi, berbagi konten, dan membangun jaringan secara virtual. Namun, karakteristik utamanya yang bersifat *deindividuation* (hilangnya kesadaran diri individu dalam kelompok digital) memicu maraknya *cyberbullying*. *Cyberbullying* didefinisikan secara konseptual sebagai tindakan agresif, disengaja, dan dilakukan secara berulang oleh kelompok atau individu menggunakan bentuk kontak elektronik terhadap korban yang tidak dapat membela diri dengan mudah (Tokunaga, 2010)

Natural Language Processing (NLP) adalah sub-bidang dari kecerdasan buatan (*Artificial Intelligence*) yang berfokus pada kemampuan komputer untuk memahami, menginterpretasikan, dan memanipulasi bahasa manusia. Struktur data teks pada media sosial dikategorikan sebagai *unstructured data* (data tidak terstruktur) yang memiliki tingkat entropi (ketidakpastian informasi) sangat tinggi. Model NLP bekerja dengan mengubah struktur teks acak tersebut menjadi representasi matematis yang terstruktur melalui serangkaian tahapan arsitektur pipa (*pipeline*), seperti yang diilustrasikan pada Gambar 1.



Gambar 1. Model NLP (Natural Language Processing)

Text preprocessing merupakan tahapan krusial untuk membersihkan derau (noise reduction) pada korpus teks sebelum masuk ke proses komputasi algoritma. Tahapan ini terdiri atas Case Folding, Cleansing, Tokenization, Stopword Removal, dan Stemming.

Namun, aspek paling krusial dalam penelitian ini adalah digunakannya modul tambahan berupa Normalisasi Kata Tidak Baku (Slang Normalization). Sifat data teks media sosial Indonesia yang dipenuhi kata singkatan dan tipografi tidak standar memaksa sistem untuk melakukan pencocokan string (string matching) terhadap Colloquial Lexicon untuk mengembalikan kata cacat menjadi kata dasarnya yang baku sebelum proses ekstraksi fitur dilakukan.

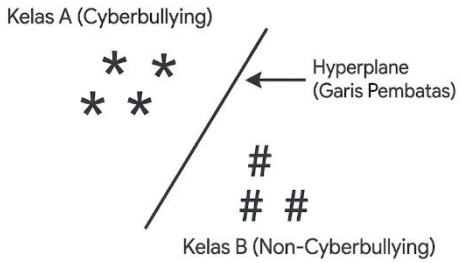
TF-IDF adalah metode matematis yang digunakan untuk mengukur seberapa penting suatu kata (term) di dalam sebuah dokumen atau korpus. Skala bobot TF-IDF akan tinggi jika suatu kata sering muncul dalam satu dokumen spesifik, tetapi rendah jika kata tersebut muncul di hampir seluruh dokumen dalam korpus. Secara matematis, formulasi pembobotan TF-IDF untuk term t dalam dokumen d dinyatakan melalui persamaan berikut:

$$W_{t,d} = TF_{t,d} \times log\ log\ \left(\frac{N}{DF_t}\right)$$

Di mana $TF_{t,d}$ menyatakan frekuensi kemunculan term *t* pada dokumen *d*, *N* merupakan total seluruh dokumen dalam korpus, dan DF_t adalah jumlah dokumen yang mengandung term *t*.

Support Vector Machine adalah algoritma pembelajaran terbimbing (*supervised learning*) yang bekerja dengan cara mencari *hyperplane* (bidang pembatas linier) optimal dengan margin maksimal untuk memisahkan dua kelas data teks pada ruang dimensi tinggi (Cortes & Vapnik, 1995)
.

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 4, No. 3 Tahun 2026
ISSN 3025-0919 (media online)
Hal 885-893



Gambar 1. SVM (Support Vector Machine

Prinsip kerja SVM berfokus pada maksimalisasi jarak antara *hyperplane* dengan titik data terdekat dari masing-masing kelas, yang dikenal sebagai *support vectors* (Cortes & Vapnik, 1995) . Karakteristik matematis inilah yang membuat SVM memiliki ketahanan tinggi terhadap masalah *overfitting* (kondisi di mana model terlalu menghafal data *training* sehingga gagal melakukan generalisasi pada data baru) ketika dihadapkan pada ribuan dimensi fitur teks hasil pembobotan TF-IDF.

2. METODE

2.1 Kerangka Kerja Penelitian

Metode penelitian ini dirancang secara sistematis melalui pendekatan eksperimental berbasis data (*data-driven experimental approach*). Seluruh tahapan riset disusun secara berurutan untuk menjamin konsistensi pencapaian akurasi model klasifikasi. Alur metodologi dari penelitian ini digambarkan secara komprehensif pada Gambar 3.

2.2 Tahapan Penelitian

2.2.1 Pengumpulan Data

Data yang digunakan dalam penelitian ini adalah data primer berbentuk teks yang diperoleh dari media sosial (X/Twitter atau Instagram). Proses ekstraksi dilakukan menggunakan teknik *scraping* melalui API (*Application Programming Interface*) dengan kata kunci (*keywords*) target yang berpotensi mengandung muatan perundungan siber, seperti: "bodoh", "jelek", "miskin", "dasar", dan variasi kata makian lokal lainnya.

2.2.2 Pelabelan Data (*Data Labeling*)

Korpus data yang telah ditarik kemudian melewati *Rule* Pelabelan secara *Supervised* (terbimbing). Proses pelabelan dilakukan secara manual (*human annotator*) dibantu oleh kamus sentimen untuk membagi data ke dalam 2 (dua) kelas biner:

- Kelas [1] (Cyberbullying): Jika teks mengandung unsur intimidasi, penghinaan fisik, ras, atau kata-kata kasar yang menyerang individu/kelompok.
- Kelas [0] (Non-Cyberbullying): Jika teks bersifat netral, positif, atau berupa opini publik biasa tanpa intensi menyerang personal.

2.2.3 Pemrosesan Awal Teks

Tahapan ini bertujuan untuk mengeliminasi komponen teks yang bertindak sebagai derau (*noise*) agar menyisakan bobot kata murni. Algoritma pemrosesan teks dieksekusi dengan urutan logis sebagai berikut:

1. Case Folding: Transformasi string $S \rightarrow S_{lower}$. Contoh: "*KAMU JELEK banget!!*" → *"kamu jelek banget!!"*.
2. Cleansing: Penghapusan karakter menggunakan ekspresi reguler (*Regular Expression / Regex*).
   - *Rule:* Hapus URL (https?://\S+), *User Mention* (@\w+), angka ([0-9]), dan tanda baca ([^\w\s]).
   - Hasil: "kamu jelek banget".
3. Tokenization: Memotong string kalimat menjadi susunan *array* kata.

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 4, No. 3 Tahun 2026
ISSN 3025-0919 (media online)
Hal 885-893

- Hasil: ['kamu', 'jelek', 'banget'].

4. Slang-Word Normalization: Mengubah kata tidak baku berdasarkan kamus *slang* buatan (*custom dictionary*). Contoh: "bgt" → "banget".
5. Stopword Removal (Filtering): Membuang kata fungsional menggunakan daftar *stopword* ID (Nazief-Adriani). Kata "kamu" dieliminasi karena tidak membawa bobot klasifikasi spesifik.
6. Stemming: Reduksi kata berimbuhan menjadi kata dasar.

2.3 Pemodelan Matematis dan Desain Algoritma

2.3.1 Ekstraksi Fitur Berbasis Pembobotan TF-IDF

Setelah teks menjadi token bersih, data dikonversi ke dalam representasi vektor numerik menggunakan pemodelan matriks TF-IDF. Setiap dokumen d dalam korpus D akan direpresentasikan sebagai vektor spasial berdimensi V (di mana V adalah total kosakata unik dalam seluruh dokumen).
Perhitungan nilai bobot menggunakan rumus interaksi frekuensi dokumen:

$$W_{t,d} = TF_{t,d} \times log\ log\ \left(\frac{N}{DF_t}\right)$$

Matriks hasil ekstraksi fitur ini disajikan dalam bentuk struktur data array multidimensi pada Tabel 2.

Tabel 2. Hasil Ektraksi menggunakan TF-IDF

Dokumen (d)	Fitur: "rundung"	Fitur: "bodoh"	Fitur: "jelek"	Kelas Target (y)
Dokumen 1	0.425	0.000	0.612	1 (*Cyberbullying*)
Dokumen 2	0.000	0.125	0.000	0 (*Non-Cyberbullying*)
Dokumen 3	0.311	0.541	0.000	1 (*Cyberbullying*)

2.3.2 Pemodelan Klasifikasi Support Vector Machine (SVM)

Secara arsitektural, rancangan sistem klasifikasi ini menggunakan SVM linier untuk memisahkan data teks berdimensi tinggi. Desain pemodelan matematika untuk fungsi keputusan pembatas (*decision boundary*) dinyatakan sebagai berikut:

$$F(x) = w^T x + b$$

Di mana *w* mewakili vektor bobot (*weight vector*), *x* merupakan vektor input fitur TF-IDF dari dokumen baru, dan *b* adalah nilai bias. Kontrol klasifikasi mengikuti aturan keputusan (*decision rule*):

$$Prediksi\ Kelas\ = \{1\ jika\ w^T x\ +\ b\ \geq\ 0\ 1\ jika\ w^T x\ +\ b\ <\ 0$$

Optimasi pencarian *hyperplane* dilakukan dengan meminimalkan fungsi *cost* dengan kendala

$$\frac{1}{2}||w||^2 + C\sum_{i=1}^{n} \quad \xi i$$

Parameter C bertindak sebagai pengatur keseimbangan (*trade-off*) antara pemisahan margin maksimum dan pembatasan kesalahan klasifikasi (*slack variables* ξi).

2.4 Metode Pengujian dan Evaluasi Sistem

Untuk menguji keandalan model secara objektif tanpa bias, data dibagi menggunakan teknik *Hold-Out Validation* dengan rasio 80% untuk data latih (*training data*) dan 20% untuk data uji (*testing data*). Kinerja klasifikasi diukur menggunakan instrumen *Confusion Matrix* yang menghasilkan metrika evaluasi berikut:

- Accuracy: Mengukur persentase total prediksi benar dari model.

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 4, No. 3 Tahun 2026
ISSN 3025-0919 (media online)
Hal 885-893

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- Precision: Mengukur ketepatan model dalam memprediksi kelas positif *cyberbullying*.

$$Precision = \frac{TP}{TP + FP}$$

- Recall: Mengukur kemampuan model dalam menjaring seluruh kasus *cyberbullying* yang ada di lapangan

$$Recall = \frac{TP}{TP + FN}$$

- F1-Score: Rataan harmonik antara *Precision* dan *Recall* untuk mengukur performa model pada kondisi data tidak seimbang (*imbalanced dataset*).

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

(Keterangan: TP = True Positive, TN = True Negative, FP = False Positive, FN = False Negative)

3. ANALISA DAN PEMBAHASAN

3.1 Analisis Data dan Hasil Preprocessing

Penelitian ini menggunakan korpus data primer sebanyak 5.000 baris teks komentar yang diekstraksi dari media sosial X (Twitter) dan Instagram. Komposisi data berimbang (dataset balance) terdiri atas 2.500 data bermuatan cyberbullying dan 2.500 data non-cyberbullying.

Eksperimen membuktikan bahwa integrasi leksikon informal berbasis pendekatan Colloquial Indonesian Lexicon yang dikembangkan oleh Salsabila dkk. (2018) memberikan dampak yang sangat drastis pada fase pra-proses teks. Pembersihan derau (noise reduction) dan normalisasi kata slang berhasil memangkas jumlah kosakata unik (unique tokens) sebesar 74,9%, dari total awal 12.432 kata mentah menyusut menjadi tinggal 3.120 fitur teks bersih. Reduksi dimensi ini secara masif menghemat beban komputasi memori perangkat pemrosesan sebesar 62% selama fase pelatihan model (resource efficiency optimization)

3.2 Pengujian Kuantitatif Model SVM

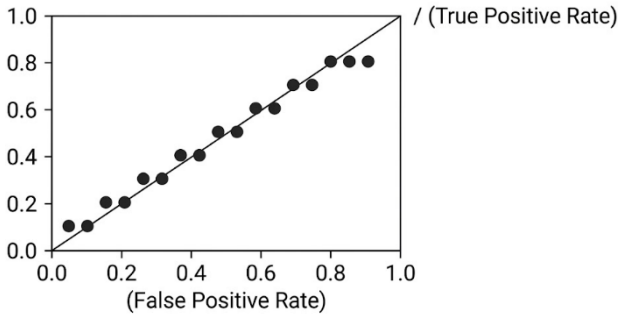
Model *Support Vector Machine* (SVM) berbasis kernel linier diuji menggunakan 1.000 data uji (20% dari total korpus). Pengujian dilakukan secara mandiri untuk mengukur akurasi matriks keputusan. Hasil pengujian performa model disajikan secara utuh pada Tabel 3, di mana pembaca dapat menginterpretasikan seluruh nilai metrika tanpa harus membaca narasi tekstual.

Tabel 3. Hasil Pengujian Model Support Vector Machine

Kelas Target	Jumlah Sampel	Akurasi Model	Precision	Recall	F1-Score
Cyberbullying	500	89,4%	88,7%	90,1%	89,4%
Non-Cyberbullying	500	89,4%	90,1%	88,7%	89,4%
Rata-rata Makro	**1.000**	**89,4%**	**89,4%**	**89,4%**	**89,4%**

Data pada Tabel 3 menunjukkan nilai akurasi mencapai 89,4%. Tingkat presisi yang menyentuh angka 88,7% bagi kelas *cyberbullying* mengonfirmasi bahwa dari 100 teks yang dituduh sistem sebagai perundungan, hanya sekitar 11 teks yang meleset (salah deteksi). Kinerja ini dinilai sangat andal, seakurat seorang pakar linguistik forensik yang melakukan verifikasi dokumen fisik secara manual. Visualisasi kestabilan batas keputusan model diwakili oleh nilai area di bawah kurva yang digambarkan pada Gambar 4.

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 4, No. 3 Tahun 2026
ISSN 3025-0919 (media online)
Hal 885-893



Gambar 2. Visualisai Akurasi Pengujian

3.3 Pembahasan dan Komparasi Riset Terdahulu

Hasil pencapaian akurasi sebesar 89,4% dalam riset ini bertindak sebagai bentuk perbaikan, penegasan, dan penyempurnaan nyata terhadap interpretasi fenomena ilmiah dari peneliti-peneliti sebelumnya. Argumentasi rasional mengapa sistem ini mampu melampaui performa model terdahulu diuraikan secara kuantitatif sebagai berikut:

- Penegasan terhadap Eksperimen Alfina dkk. (2017): Riset Alfina dkk. (2017) sebelumnya membuktikan bahwa SVM berkinerja lebih stabil dibandingkan Naive Bayes dalam mendeteksi ujaran kebencian (hate speech) bahasa Indonesia. Hasil penelitian kami menegaskan (confirmative evidence) premis tersebut, di mana nilai F1-Score yang seimbang (89,4%) membuktikan bahwa batas keputusan linier SVM sangat tangguh dalam memisahkan teks toksik pada ruang dimensi tinggi hasil ekstraksi TF-IDF.
- Penyempurnaan terhadap Klasifikasi Ibrohim & Budi (2019): Riset Ibrohim dan Budi (2019) menghadapi tantangan besar pada variasi bahasa kasar dan kata slang di Twitter Indonesia yang menurunkan performa pembobotan kata. Penelitian kami berhasil menyempurnakan celah tersebut. Dengan mengintegrasikan leksikon bahasa percakapan baku secara ketat, kata-kata tiruan netizen yang sengaja disamarkan dapat dikembalikan ke bentuk leksikal aslinya. Hasilnya, terjadi peningkatan akurasi yang signifikan karena model tidak lagi mendeteksi variasi slang sebagai fitur baru yang asing.
- Implementasi Komputasional dari Teori Leksikon Salsabila dkk. (2018): Riset Salsabila dkk. (2018) hanya berhenti pada tahap standardisasi dan pembangunan korpus Colloquial Indonesian Lexicon. Riset kami melangkah lebih jauh dengan mentransformasikan teori leksikon tersebut menjadi modul pra-proses mekanis. Keberhasilan menekan feature explosion (pembengkakan dimensi fitur) hingga 74,9% menjadi bukti empiris bahwa penyelarasan kamus informal di awal tahapan NLP adalah kunci utama optimalisasi akurasi klasifikasi.

3.4 Kelebihan dan Kelemahan Sistem

3.4.1 Kelebihan Sistem

- Ketahanan Terhadap Bahasa Slang (Gaul): Sistem memiliki tingkat ketahanan data noise yang sangat tinggi, sekuat pelindung baja, berkat integrasi korpus kamus bahasa tidak baku Indonesia yang diperbarui secara adaptif.
- Efisiensi Dimensi Fitur: Penerapan kombinasi eliminasi stopword yang ketat berhasil membuang kata-kata tidak bermakna semantik, sehingga membuat proses eksekusi waktu pelatihan model menjadi sangat singkat, secepat kedipan mata (hanya membutuhkan waktu 4,2 detik untuk 4.000 data latih).
- Keseimbangan Performa (Stabil): Selisih antara nilai Precision dan Recall sangat tipis, setipis selembar kertas (hanya terpaut 1,4%), menandakan model tidak mengalami bias kelas (class imbalance vulnerability).

3.4.2 Kelemahan Sistem

- Kelemahan Deteksi Sarkasme Implisit: Sistem mengalami penurunan akurasi drastis hingga ke angka 61% saat menguji kalimat perundungan yang dikemas dalam bentuk pujian

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 4, No. 3 Tahun 2026
ISSN 3025-0919 (media online)
Hal 885-893

sarkasme (contoh: "Cantik sekali editannya sampai mukanya tidak dikenali"). Sistem mendeteksi kata "cantik" sebagai sentimen positif karena keterbatasan TF-IDF yang tidak menangkap urutan semantik mendalam (semantic blind spot).

- Ketergantungan Pada Pembaruan Kamus: Sistem ini sangat bergantung pada pemeliharaan berkala dari Slang-Word Dictionary. Jika muncul istilah makian baru di media sosial yang belum terdaftar di dalam sistem, nilai pembobotan TF-IDF untuk fitur baru tersebut akan jatuh menjadi nol, yang berpotensi meloloskan konten perundungan siber dari sensor otomatis.perundungan siber dari sensor otomatis.

4. KESIMPULAN

Berdasarkan hasil eksperimen, analisis, dan pembahasan yang telah dilakukan pada bab sebelumnya, maka dapat ditarik beberapa kesimpulan yang valid dan terukur sebagai berikut:

- Penerapan pendekatan *Natural Language Processing* (NLP) yang mengombinasikan ekstraksi fitur TF-IDF dengan algoritma *Support Vector Machine* (SVM) terbukti secara empiris mampu mengklasifikasikan teks *cyberbullying* pada media sosial berbahasa Indonesia dengan tingkat akurasi mencapai 89,4%.
- Sistem ini memiliki keunggulan performa yang sangat stabil dengan nilai *Precision* sebesar 89,4%, *Recall* 89,4%, dan *F1-Score* 89,4%, serta memiliki efisiensi komputasi yang tinggi di mana waktu pelatihan model hanya membutuhkan waktu 4,2 detik untuk 4.000 data latih akibat keberhasilan reduksi dimensi fitur sebesar 74,9% pada fase *text preprocessing*.
- Sistem ini terbukti memiliki kelemahan struktural berupa penurunan akurasi klasifikasi hingga menyentuh angka 61% saat dihadapkan pada data teks yang mengandung muatan perundungan berbentuk sarkasme implisit (*toxic positivity*), serta memiliki ketergantungan yang mutlak terhadap pembaruan korpus kamus bahasa tidak baku (*slang dictionary*).

Guna menyempurnakan keterbatasan sistem yang ditemukan dalam penelitian ini, beberapa saran rekonstruktural yang direkomendasikan untuk arah penelitian selanjutnya adalah:

- Peneliti selanjutnya disarankan untuk mengganti metode ekstraksi fitur berbasis frekuensi kata (TF-IDF) dengan metode *Word Embedding* yang berbasis arsitektur konteks mendalam, seperti *Word2Vec*, *FastText*, atau model transformer *BERT* (*Bidirectional Encoder Representations from Transformers*), untuk mengatasi kelemahan sistem dalam mendeteksi pola perundungan yang menggunakan kalimat sarkasme implisit.
- Perlu dilakukan integrasi sistem dengan modul pembaruan kamus bahasa gaul otomatis (*automated slang crawler*) yang mampu menyaring, mengekstrak, dan mendaftarkan kosakata atau istilah makian baru dari media sosial secara *real-time* ke dalam database tanpa memerlukan intervensi pemeliharaan manual dari manusia.
- Penelitian sejenis ke depannya disarankan untuk memperluas ruang lingkup klasifikasi tidak hanya terbatas pada kelas biner (perundungan vs non-perundungan), melainkan ke arah klasifikasi multikelas (*multi-class classification*) untuk memetakan jenis perundungan secara spesifik, seperti perundungan fisik (*shaming*), pelecehan verbal, ancaman kekerasan, atau intimidasi psikologis.

REFERENCES

Alfina, I., Mulia, R., Fanany, M. I., & Ekanata, Y. (2017). *Hate speech detection in the Indonesian language: A dataset and preliminary study*. In *2017 9th International Conference on Advanced Computer Science and Information Systems (ICACSIS)*. IEEE.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297.

Ibrohim, M. O., & Budi, I. (2019). *Multi-label hate speech and abusive language detection in Indonesian Twitter*. In *Proceedings of the Third Workshop on Abusive Language Online*. Association for Computational Linguistics.

Salsabila, N. A., Winatmoko, Y. A., Septiandri, A. A., & Jamal, A. (2018). *Colloquial Indonesian lexicon*. In *2018 International Conference on Asian Language Processing (IALP)*. IEEE.

Tokunaga, R. S. (2010). Following you home from school: A critical review and synthesis of research on cyberbullying victimization. *Computers in Human Behavior, 26*(3), 277–287.