

Sistem Klasifikasi Komentar Media Sosial untuk Deteksi Dini Cyberbullying Menggunakan TF-IDF dan Support Vector Machine

Famela Okta Vianda^{*1}, Fajri Profesio Putra²

^{1,2}Program Studi Sarjana Terapan Rekayasa Perangkat Lunak,
Jurusan Teknik Informatika, Politeknik Negeri Bengkalis

Email: ¹famelaoktavianda79@gmail.com, ²fajri@polbeng.ac.id

Abstrak

Cyberbullying merupakan permasalahan yang terus meningkat pada berbagai platform media sosial, khususnya pada kolom komentar yang ditandai dengan penggunaan bahasa informal, singkatan, dan istilah slang yang beragam. Kondisi ini menyebabkan proses identifikasi komentar bermuatan perundungan menjadi kompleks dan kurang efektif jika dilakukan secara manual. Selain itu, metode klasifikasi berbasis deep learning yang memiliki akurasi tinggi umumnya membutuhkan sumber daya komputasi yang besar, sehingga kurang efisien untuk implementasi deteksi dini secara otomatis. Oleh karena itu, penelitian ini mengusulkan pendekatan berbasis machine learning menggunakan Support Vector Machine (SVM) dengan representasi fitur Term Frequency–Inverse Document Frequency (TF-IDF). Dataset yang digunakan terdiri dari 8.000 komentar berbahasa Indonesia yang diperoleh dari platform YouTube dan TikTok, sehingga merepresentasikan lingkungan multiplatform dengan variasi bahasa yang tinggi. Tahapan pengolahan data meliputi preprocessing teks (cleaning, normalisasi, tokenisasi, filtering, dan stemming), ekstraksi fitur menggunakan TF-IDF, serta klasifikasi menggunakan SVM dengan kernel linear. Model dievaluasi menggunakan skema pembagian data 70:30 dengan metrik accuracy, precision, recall, dan F1-score. Hasil pengujian menunjukkan bahwa model mencapai akurasi sebesar 80,75%, precision 83% pada kelas cyberbullying, serta F1-score sebesar 0,81. Hasil ini menunjukkan bahwa SVM mampu memberikan performa yang stabil dalam menangani data teks berdimensi tinggi dengan kompleksitas bahasa yang beragam. Kontribusi utama penelitian ini terletak pada penggunaan dataset multiplatform berbahasa Indonesia serta evaluasi efektivitas SVM dalam mendukung sistem deteksi dini cyberbullying yang efisien secara komputasi.

Kata kunci: Cyberbullying, Support Vector Machine, TF-IDF, TikTok, YouTube.

Abstract

Cyberbullying has become an increasingly critical issue across social media platforms, particularly within comment sections characterized by high linguistic variability, including informal expressions, abbreviations, and slang. This complexity makes manual identification of harmful comments inefficient and inconsistent. While deep learning approaches can achieve high accuracy, they typically require substantial computational resources, limiting their applicability for automatic early detection systems. Therefore, this study proposes a machine learning-based approach using Support Vector Machine (SVM) combined with Term Frequency–Inverse Document Frequency (TF-IDF) for cyberbullying detection. The dataset consists of 8,000 Indonesian-language comments collected from YouTube and TikTok, representing a multiplatform environment with diverse linguistic characteristics. The data processing pipeline includes text preprocessing (cleaning, normalization, tokenization, filtering, and stemming), feature extraction using TF-IDF, and classification using a linear kernel SVM. The model is evaluated using a 70:30 train-test split with performance metrics including accuracy, precision, recall, and F1-score. The experimental results demonstrate that the proposed model achieves an accuracy of 80.75%, a precision of 83% for the cyberbullying class, and an F1-score of 81%. These findings indicate that SVM provides a stable and computationally efficient solution for handling high-dimensional textual data with complex linguistic patterns. The main contribution of this study lies in the use of a multiplatform Indonesian dataset and the evaluation of SVM effectiveness for efficient early cyberbullying detection.

Keywords: Cyberbullying, Support Vector Machine, TF-IDF, TikTok, YouTube.

This work is an open access article and licensed under a Creative Commons Attribution-NonCommercial ShareAlike 4.0 International (CC BY-NC-SA 4.0)



1. PENDAHULUAN

Perkembangan teknologi informasi telah mendorong transformasi pola komunikasi masyarakat ke arah digital, khususnya melalui platform media sosial seperti YouTube dan TikTok. Platform tersebut menyediakan ruang interaksi terbuka melalui fitur kolom komentar yang memungkinkan pengguna menyampaikan opini secara langsung. Namun, kemudahan ini juga memicu munculnya perilaku negatif

seperti cyberbullying, yaitu tindakan perundungan yang dilakukan melalui media digital dan berpotensi memberikan dampak psikologis serius bagi korban [1]. Cyberbullying tidak hanya berdampak pada aspek sosial, tetapi juga berkontribusi terhadap gangguan kesehatan mental seperti stres, kecemasan, dan depresi [2]. Tingginya volume interaksi pengguna serta penggunaan bahasa tidak baku, singkatan, dan istilah slang pada media sosial menyebabkan proses identifikasi komentar bermuatan negatif menjadi semakin kompleks [3]. Dalam konteks ini, pendekatan manual dalam moderasi konten menjadi tidak efisien dan rentan terhadap inkonsistensi. Oleh karena itu, diperlukan pendekatan berbasis teknologi yang mampu melakukan deteksi secara otomatis untuk meningkatkan efektivitas dan efisiensi dalam proses moderasi konten digital [4].

Dalam konteks penelitian ini, deteksi dini cyberbullying diartikan sebagai kemampuan sistem untuk mengidentifikasi komentar yang mengandung unsur perundungan secara otomatis pada tahap awal kemunculannya. Pendekatan klasifikasi teks yang digunakan memungkinkan setiap komentar dianalisis dan dikategorikan secara cepat sebelum konten tersebut menyebar lebih luas, sehingga berpotensi mendukung proses moderasi konten secara lebih responsif.

Pendekatan berbasis machine learning telah banyak diadopsi dalam tugas klasifikasi teks, termasuk dalam deteksi cyberbullying. *Support Vector Machine* (SVM) merupakan salah satu algoritma yang sering digunakan karena kemampuannya dalam menangani data berdimensi tinggi serta memberikan performa klasifikasi yang stabil [5]. Penelitian sebelumnya menunjukkan bahwa kombinasi metode TF-IDF dan SVM mampu menghasilkan akurasi yang tinggi dalam mendeteksi komentar bermuatan cyberbullying pada berbagai platform media sosial [6], [7]. Selain itu, implementasi SVM pada data komentar media sosial juga terbukti efektif dalam membedakan antara konten negatif dan non-negatif [8]. Dalam konteks klasifikasi teks, representasi fitur seperti TF-IDF menghasilkan data dengan dimensi yang sangat tinggi dan bersifat sparse. Kondisi ini menuntut algoritma yang mampu membentuk batas keputusan secara optimal dalam ruang fitur berdimensi tinggi. SVM bekerja dengan membangun hyperplane optimal yang memaksimalkan margin antar kelas, sehingga memiliki kemampuan generalisasi yang baik dalam berbagai kasus klasifikasi teks [9].

Berbagai pendekatan lain juga telah digunakan dalam deteksi cyberbullying. Metode Naïve Bayes dikenal sederhana dan cepat, namun memiliki keterbatasan dalam menangani dependensi antar fitur [10]. Random Forest mampu meningkatkan performa melalui pendekatan ensemble learning, namun memerlukan proses tuning yang lebih kompleks. Sementara itu, metode berbasis deep learning seperti CNN-LSTM dan model transformer seperti BERT mampu memahami konteks bahasa secara lebih mendalam, tetapi membutuhkan sumber daya komputasi yang besar serta waktu pelatihan yang lebih lama [11], [12]. Hal ini menunjukkan bahwa pemilihan metode harus mempertimbangkan keseimbangan antara akurasi dan efisiensi komputasi.

Meskipun berbagai penelitian telah dilakukan, sebagian besar studi masih berfokus pada satu platform media sosial seperti Twitter, Instagram, atau TikTok, dan belum mempertimbangkan kompleksitas variasi bahasa lintas platform [13]. Selain itu, penelitian sebelumnya cenderung menitikberatkan pada peningkatan performa model tanpa mempertimbangkan aspek implementasi sistem secara end-to-end yang dapat digunakan secara langsung oleh pengguna [14]. Variasi bahasa informal, penggunaan slang, serta struktur kalimat yang tidak baku pada media sosial juga masih menjadi tantangan utama dalam proses klasifikasi teks yang belum sepenuhnya teratasi [15]. Dalam konteks tersebut, *Support Vector Machine* (SVM) dipilih dalam penelitian ini karena kemampuannya dalam menangani data berdimensi tinggi serta stabil terhadap representasi fitur yang bersifat sparse seperti TF-IDF. Selain itu, SVM memiliki kompleksitas komputasi yang relatif lebih rendah dibandingkan model deep learning, sehingga lebih sesuai untuk implementasi sistem deteksi dini yang membutuhkan efisiensi dan kecepatan proses. Karakteristik ini menjadikan SVM sebagai metode yang tepat untuk menangani dataset komentar media sosial yang memiliki variasi bahasa tinggi serta distribusi fitur yang kompleks.

Berdasarkan permasalahan tersebut, penelitian ini bertujuan untuk mengembangkan sistem klasifikasi komentar online untuk deteksi dini cyberbullying menggunakan algoritma *Support Vector Machine* (SVM) dengan metode TF-IDF. Berbeda dengan penelitian sebelumnya yang umumnya hanya berfokus pada aspek klasifikasi, penelitian ini mencakup seluruh tahapan pengolahan data mulai dari pengambilan data, preprocessing, hingga pemodelan dan evaluasi secara terstruktur. Kontribusi utama penelitian ini adalah: (1) penggunaan dataset multiplatform dari YouTube dan TikTok dengan karakteristik bahasa informal yang beragam, (2) analisis performa SVM dalam menangani data teks

berdimensi tinggi dengan variasi bahasa yang kompleks, serta (3) pengembangan model klasifikasi yang berpotensi diintegrasikan dalam sistem deteksi dini berbasis web.

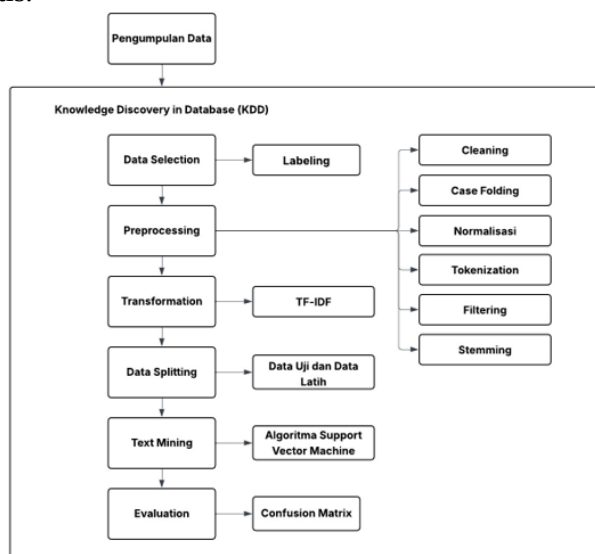
Dengan demikian, penelitian ini tidak hanya berkontribusi pada pengembangan model klasifikasi teks, tetapi juga memberikan implikasi praktis dalam mendukung sistem moderasi konten digital yang lebih responsif dan adaptif terhadap dinamika bahasa pada media sosial.

2. METODE PENELITIAN

Penelitian ini mengadopsi kerangka kerja *Knowledge Discovery in Database* (KDD) yang mencakup beberapa tahapan utama, meliputi pengumpulan data, seleksi data, preprocessing, transformasi, pemodelan, serta evaluasi. Pendekatan ini banyak digunakan dalam studi berbasis machine learning karena kemampuannya dalam mengelola data berukuran besar secara sistematis guna menghasilkan model klasifikasi yang optimal [16].

Selain itu, KDD menyediakan alur kerja yang terorganisir untuk mengekstraksi pengetahuan dari data mentah menjadi informasi yang memiliki nilai guna melalui serangkaian proses yang saling berkaitan. Setiap tahapan dalam KDD berkontribusi secara signifikan dalam menjaga kualitas data, memastikan representasi fitur yang tepat, serta meningkatkan performa dan konsistensi model yang dibangun. Dengan demikian, penerapan pendekatan ini dapat memperkuat validitas dan reliabilitas hasil penelitian secara keseluruhan. Dalam penelitian ini, data yang digunakan berupa komentar berbahasa Indonesia yang memiliki karakteristik tidak terstruktur dan mengandung variasi bahasa informal.

Dataset yang digunakan berjumlah 8.000 komentar yang diklasifikasikan ke dalam dua kategori, yaitu bullying dan non-bullying. Secara umum, alur penelitian ini mencakup proses pengolahan data teks hingga menghasilkan output berupa label klasifikasi yang digunakan sebagai dasar dalam deteksi dini cyberbullying. Proses klasifikasi yang dilakukan pada penelitian ini menjadi dasar dalam mekanisme deteksi dini, di mana setiap komentar yang masuk dapat langsung dianalisis dan dikategorikan secara otomatis.



Gambar 1. Prosedur Penelitian

Gambar 1 menunjukkan alur proses penelitian yang menggunakan pendekatan *Knowledge Discovery in Database* (KDD). Proses dimulai dari tahap pengumpulan data komentar dari media sosial, kemudian dilakukan seleksi data dan pelabelan untuk menentukan kategori *bullying* dan *non-bullying*. Tahap ini berperan penting dalam membangun fondasi dataset yang berkualitas, karena kesalahan pada tahap awal dapat berdampak langsung terhadap performa model pada tahap selanjutnya. Selanjutnya, data diproses pada tahap preprocessing yang meliputi *cleaning*, *case folding*, normalisasi, *tokenization*, *filtering*, dan *stemming* untuk meningkatkan kualitas teks. Proses ini tidak hanya bertujuan untuk menghilangkan noise, tetapi juga untuk menyederhanakan representasi bahasa sehingga pola yang relevan dapat lebih mudah dikenali oleh algoritma pembelajaran mesin.

Setelah itu, dilakukan tahap transformasi menggunakan metode TF-IDF untuk mengubah data teks menjadi representasi numerik. Representasi ini memungkinkan data tekstual yang bersifat tidak terstruktur dapat diproses dalam ruang vektor, sehingga hubungan antar kata dapat dianalisis secara

matematis. Proses ini menjadi krusial karena kualitas representasi fitur sangat memengaruhi kemampuan model dalam membedakan karakteristik antar kelas. Data kemudian dibagi menjadi data latih dan data uji sebelum masuk ke tahap *text mining* menggunakan algoritma *Support Vector Machine* (SVM) untuk proses klasifikasi. Pemilihan SVM didasarkan pada kemampuannya dalam membentuk batas keputusan yang optimal pada data berdimensi tinggi, sehingga efektif dalam menangani kompleksitas data teks yang memiliki variasi bahasa yang luas.

Tahap akhir dalam penelitian ini adalah proses evaluasi model yang dilakukan menggunakan confusion matrix untuk menilai kinerja klasifikasi yang dihasilkan. Melalui pendekatan ini, tidak hanya diperoleh informasi mengenai tingkat akurasi model, tetapi juga analisis yang lebih komprehensif terhadap jenis kesalahan klasifikasi yang terjadi, seperti false positive dan false negative. Dengan adanya evaluasi tersebut, performa model dapat dipahami secara lebih mendalam, termasuk dalam mengidentifikasi kelemahan dan potensi perbaikan. Secara keseluruhan, penerapan seluruh tahapan dalam kerangka KDD pada penelitian ini mampu menghasilkan model klasifikasi yang tidak hanya memiliki tingkat akurasi yang baik, tetapi juga menunjukkan kemampuan generalisasi yang optimal ketika dihadapkan pada data baru.

a. Pengumpulan Data

Data yang digunakan dalam penelitian ini berupa komentar dari platform media sosial YouTube dan TikTok yang dikumpulkan menggunakan teknik *web scraping*. Teknik ini memungkinkan pengambilan data secara otomatis dalam jumlah besar dan efisien dari sumber daring [17]. Dataset yang diperoleh berjumlah 8.000 komentar yang terdiri dari dua kategori, yaitu komentar *bullying* dan *non-bullying*.

Tabel 1. Contoh Dataset

Komentar	Label
Makin jelek aja anaknya, padahal ayah ibunya cakep2	Bullying
Jadi diri sendiri, jangan jadi ular apalagi dugong	Bullying
Cantik banget semoga sehat selalu kakak panutan	Non-Bullying
Yakinin dulu niatnya bg... Semoga sukses	Non-Bullying

b. Seleksi dan Pelabelan Data

Tahap seleksi dilakukan untuk menyaring data yang relevan dengan penelitian, seperti menghapus komentar duplikat dan data yang tidak memiliki makna. Setelah itu, dilakukan pelabelan data secara manual ke dalam dua kelas utama, yaitu *bullying* dan *non-bullying*. Proses pelabelan ini penting untuk memastikan kualitas data dalam proses pembelajaran model [15].

c. Preprocessing Teks

Tahap preprocessing dilakukan untuk menyiapkan data teks agar siap dianalisis secara optimal. Proses ini mencakup beberapa langkah, antara lain pembersihan data dari karakter yang tidak relevan (*cleaning*), konversi seluruh huruf menjadi bentuk kecil (*case folding*), pemecahan teks menjadi unit kata (tokenisasi), penghilangan kata-kata umum yang tidak memberikan kontribusi signifikan (*filtering*), serta reduksi kata ke bentuk dasarnya (*stemming*). Rangkaian tahapan tersebut bertujuan untuk meningkatkan kualitas dan konsistensi data, sehingga proses klasifikasi dapat dilakukan dengan lebih efektif dan akurat.

d. Transformasi Fitur (TF-IDF)

Setelah preprocessing, data teks diubah menjadi representasi numerik menggunakan metode *Term Frequency-Inverse Document Frequency* (TF-IDF). Metode ini digunakan untuk memberikan bobot pada setiap kata berdasarkan tingkat kemunculannya dalam dokumen, sehingga dapat merepresentasikan informasi penting dari teks secara lebih efektif [10]. Secara matematis, TF-IDF dirumuskan sebagai:

$$TF - IDF(t,d) = TF(t,d) \times \log\left(\frac{N}{df(t)}\right)$$

di mana $TF(t,d)$ adalah frekuensi kemunculan term t dalam dokumen d , N adalah jumlah total dokumen, dan $df(t)$ adalah jumlah dokumen yang mengandung term t .

Penggunaan TF-IDF dalam klasifikasi teks telah terbukti mampu meningkatkan kualitas representasi fitur pada berbagai penelitian [18]. Perhitungan bobot TF-IDF dalam penelitian ini mengacu pada konsep dasar term weighting, di mana bobot setiap kata ditentukan berdasarkan frekuensi kemunculan dalam dokumen serta tingkat penyebarannya dalam seluruh korpus. Secara matematis, TF-IDF dihitung dengan mengalikan nilai Term Frequency (TF) dengan Inverse Document Frequency (IDF), sehingga kata yang sering muncul dalam suatu dokumen namun jarang muncul dalam dokumen lain akan memiliki bobot yang lebih tinggi. Pendekatan ini memungkinkan model untuk lebih fokus pada kata-kata yang memiliki informasi penting dalam proses klasifikasi.

e. Pembagian Data

Dataset yang telah melalui tahap pengolahan selanjutnya dipartisi menjadi data latih dan data uji dengan perbandingan 70:30. Data latih dimanfaatkan untuk proses pembentukan model klasifikasi, sedangkan data uji digunakan untuk mengevaluasi kemampuan model dalam melakukan prediksi terhadap data yang belum pernah dikenali sebelumnya [12]. Penggunaan rasio pembagian data sebesar 70:30 dipilih karena merupakan pendekatan yang umum digunakan dalam penelitian machine learning. Rasio ini memberikan keseimbangan antara jumlah data latih yang cukup untuk membangun model serta data uji yang representatif untuk mengevaluasi performa model secara objektif terhadap data yang belum pernah dilihat sebelumnya. Pembagian data dilakukan secara acak dengan mempertahankan proporsi distribusi kelas untuk menghindari bias pada proses pelatihan dan pengujian model.

f. Pemodelan Menggunakan Support Vector Machine (SVM)

Model klasifikasi dalam penelitian ini dibangun menggunakan algoritma Support Vector Machine (SVM). Algoritma ini dipilih karena memiliki kemampuan yang baik dalam mengklasifikasikan data teks berdimensi tinggi serta mampu menghasilkan performa yang stabil [19]. Pada penelitian ini digunakan kernel linear karena lebih sederhana dan efisien secara komputasi, serta sesuai dengan karakteristik data hasil ekstraksi TF-IDF yang cenderung berdimensi tinggi. Selain itu, penggunaan kernel linear juga dapat mengurangi risiko *overfitting* dibandingkan kernel yang lebih kompleks [20].

Support Vector Machine (SVM) bekerja dengan mencari hyperplane optimal sebagai batas pemisah antar kelas dalam ruang fitur. Hyperplane tersebut direpresentasikan dalam bentuk persamaan:

$$w \cdot x + b = 0$$

di mana w merupakan vektor bobot, x adalah vektor fitur hasil ekstraksi TF-IDF, dan b adalah bias. Hyperplane ini digunakan untuk menentukan batas keputusan (*decision boundary*) dengan memaksimalkan margin terhadap titik data terdekat (*support vector*), sehingga model memiliki kemampuan generalisasi yang baik.

Parameter regularisasi (C) digunakan untuk mengontrol keseimbangan antara margin dan kesalahan klasifikasi, yang ditentukan melalui pengujian awal. Proses pelatihan model dilakukan menggunakan 70% data latih dan 30% data uji, di mana data teks yang telah melalui preprocessing dan ekstraksi TF-IDF diubah ke dalam bentuk numerik sebagai input model. Model yang dihasilkan kemudian digunakan untuk mengklasifikasikan komentar ke dalam kategori bullying dan non-bullying, sehingga dinilai efektif dalam menangani kompleksitas bahasa yang tidak terstruktur pada media sosial.

g. Evaluasi Model

Kinerja model dievaluasi menggunakan pendekatan confusion matrix untuk menilai hasil klasifikasi yang dihasilkan. Evaluasi ini melibatkan beberapa metrik utama, yaitu accuracy, precision, recall, dan F1-score. Keempat metrik tersebut digunakan untuk mengukur tingkat ketepatan model dalam mengklasifikasikan data, sekaligus menilai keseimbangan performa antara kedua kelas yang dianalisis [21]. Secara matematis, metrik evaluasi dihitung berdasarkan nilai True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN) yang diperoleh dari confusion matrix. Accuracy mengukur tingkat ketepatan keseluruhan model, precision mengukur ketepatan prediksi pada kelas tertentu, recall mengukur kemampuan model dalam mendeteksi seluruh data positif, sedangkan F1-score merupakan rata-rata harmonis antara precision dan recall. Penggunaan keempat metrik ini bertujuan untuk memberikan evaluasi yang lebih komprehensif terhadap performa model klasifikasi.

3. HASIL DAN PEMBAHASAN

a. Implementasi Sistem

Gambar 2. Halaman Input

Gambar 2 menunjukkan antarmuka input dari sistem klasifikasi yang dikembangkan. Pengguna dapat memilih platform media sosial serta memasukkan tautan konten yang akan dianalisis. Data yang dimasukkan selanjutnya diproses oleh sistem melalui tahapan preprocessing, ekstraksi fitur menggunakan TF-IDF, serta klasifikasi menggunakan SVM.

Gambar 3. Halaman Hasil Klasifikasi

Gambar 3 menunjukkan hasil klasifikasi yang dihasilkan oleh sistem. Setiap komentar dianalisis dan dikategorikan ke dalam kelas bullying atau non-bullying. Hasil ini menunjukkan bahwa sistem mampu melakukan deteksi dini secara otomatis terhadap komentar yang dianalisis.

b. Hasil Pengujian Model

Hasil klasifikasi ini menunjukkan bahwa model mampu berfungsi sebagai komponen inti dalam sistem deteksi dini, dengan kemampuan mengidentifikasi komentar bermuatan cyberbullying secara cepat berdasarkan pola bahasa yang terdeteksi. Pengujian model dilakukan dengan skenario pembagian data 70:30 antara data latih dan data uji. Hasil evaluasi ditunjukkan pada Tabel 2.

Tabel 2. Evaluasi Model

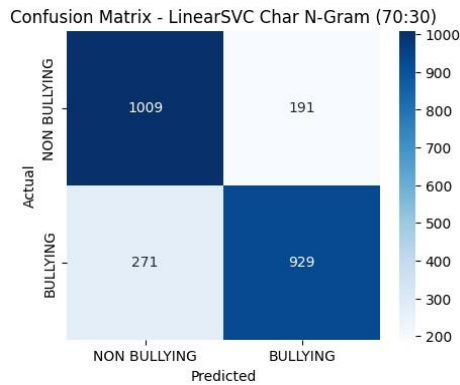
Metrik	Nilai
Accuracy	80,75%
Precision	83%
Recall	79%
F1-score	81%

Berdasarkan Tabel 2, model SVM dengan kernel linear menunjukkan performa yang kompetitif dengan akurasi sebesar 80,75%. Nilai precision sebesar 83% menunjukkan bahwa sebagian besar prediksi pada kelas cyberbullying dilakukan dengan tingkat ketepatan yang tinggi, sementara nilai recall sebesar 79% menunjukkan kemampuan model dalam mengidentifikasi data cyberbullying dari keseluruhan data yang tersedia.

Untuk memberikan gambaran yang lebih rinci terkait hasil klasifikasi, evaluasi juga dilakukan menggunakan confusion matrix yang ditunjukkan pada Gambar 4.

c. Analisis Confusion Matrix

Evaluasi lebih lanjut dilakukan menggunakan confusion matrix yang ditunjukkan pada Gambar 4.



Gambar 4. Confusion Matrix

Berdasarkan Gambar 4, model mampu mengklasifikasikan 1009 data non-bullying dan 929 data bullying dengan benar. Namun, terdapat 191 data non-bullying yang salah diklasifikasikan sebagai bullying (*false positive*) serta 271 data bullying yang tidak terdeteksi (*false negative*). Hal ini menunjukkan bahwa model masih mengalami kesulitan dalam membedakan komentar dengan makna ambigu atau konteks yang tidak eksplisit.

Distribusi kesalahan pada confusion matrix menunjukkan bahwa jumlah false negative (271) lebih tinggi dibandingkan false positive (191). Hal ini mengindikasikan bahwa model cenderung lebih sulit dalam mendeteksi komentar cyberbullying dibandingkan mengidentifikasi komentar non-bullying. Kondisi ini menunjukkan adanya kecenderungan bias model terhadap kelas non-bullying, yang dapat dipengaruhi oleh kompleksitas dan variasi bahasa pada komentar cyberbullying yang lebih kontekstual dan tidak eksplisit.

d. Analisis Performa Model

Hasil pengujian menunjukkan bahwa algoritma *Support Vector Machine* (SVM) dengan kernel linear mampu mencapai tingkat akurasi sebesar 80,75% dengan nilai precision sebesar 83% pada kelas cyberbullying serta F1-score sebesar 81%. Performa ini menunjukkan bahwa SVM memiliki kemampuan yang cukup baik dalam mengklasifikasikan data teks dengan representasi TF-IDF yang bersifat berdimensi tinggi dan sparse.

Perbedaan nilai precision (83%) dan recall (79%) menunjukkan bahwa model memiliki kecenderungan menghasilkan prediksi yang lebih konservatif terhadap kelas cyberbullying. Nilai precision yang lebih tinggi menunjukkan bahwa prediksi yang dihasilkan relatif akurat, namun nilai recall yang lebih rendah mengindikasikan bahwa masih terdapat sejumlah komentar cyberbullying yang tidak terdeteksi. Hal ini menjadi tantangan penting dalam konteks deteksi dini, karena kesalahan dalam bentuk false negative berpotensi menyebabkan konten berbahaya tidak teridentifikasi.

Keunggulan performa SVM dalam penelitian ini dapat dijelaskan melalui kemampuannya dalam membentuk hyperplane optimal yang memaksimalkan margin antar kelas. Pada dataset dengan jumlah fitur yang besar seperti hasil ekstraksi TF-IDF, SVM mampu mempertahankan stabilitas model tanpa mengalami overfitting yang signifikan. Selain itu, penggunaan kernel linear terbukti efektif karena data teks pada ruang fitur TF-IDF cenderung bersifat linearly separable.

e. Perbandingan dengan Metode Lain

Tabel 3. Perbandingan Performa Model dengan Penelitian Sebelumnya

Metode	Accuracy	Precision	Recall	F1-Score	Sumber
SVM	80,75%	83%	79%	81%	Penelitian Ini
SVM	83%	82%	83%	82%	[7]

Random Forest	82%	83%	82%	82%	[7]
Naïve Bayes	74%	72%	74%	71%	[7]

Berdasarkan Tabel 3, performa model SVM dalam penelitian ini menunjukkan hasil yang kompetitif jika dibandingkan dengan penelitian sebelumnya oleh Naufal et al. yang menggunakan dataset berskala besar dan mencakup beberapa kategori cyberbullying. Meskipun terdapat perbedaan pada nilai akurasi, hal ini dapat disebabkan oleh perbedaan karakteristik dataset, bahasa, serta jumlah data yang digunakan. Penelitian ini menggunakan dataset berbahasa Indonesia multiplatform, berbeda dengan penelitian sebelumnya yang menggunakan dataset global berbahasa Inggris. Perbandingan dilakukan berdasarkan hasil penelitian sebelumnya karena fokus penelitian ini adalah pada implementasi SVM. Namun demikian, hasil penelitian ini tetap menunjukkan bahwa SVM mampu memberikan performa yang stabil dalam klasifikasi teks berdimensi tinggi, sehingga relevan untuk digunakan dalam deteksi cyberbullying pada lingkungan media sosial.

f. Analisis Kesalahan Klasifikasi

Meskipun model menunjukkan performa yang cukup baik, masih terdapat kesalahan klasifikasi yang terjadi pada beberapa data. Kesalahan ini umumnya disebabkan oleh karakteristik bahasa pada media sosial yang bersifat tidak baku, seperti penggunaan slang, singkatan, serta ekspresi ironi dan sarkasme. Komentar yang mengandung makna ambigu atau konteks ganda sering kali sulit diklasifikasikan secara tepat oleh model berbasis fitur seperti TF-IDF. Hal ini disebabkan karena TF-IDF hanya mempertimbangkan frekuensi kemunculan kata tanpa memahami konteks semantik secara mendalam. Selain itu, penggunaan kata yang sama dalam konteks yang berbeda juga dapat menyebabkan model melakukan kesalahan prediksi.

Kesalahan klasifikasi juga dapat dipengaruhi oleh keterbatasan dataset dalam merepresentasikan seluruh variasi bahasa yang ada di media sosial. Meskipun dataset yang digunakan sudah mencakup dua platform berbeda, yaitu YouTube dan TikTok, variasi bahasa yang sangat dinamis tetap menjadi tantangan utama dalam proses klasifikasi.

g. Implikasi dan Kontribusi Penelitian

Hasil penelitian ini menunjukkan bahwa algoritma *Support Vector Machine* (SVM) memiliki potensi yang baik untuk diterapkan dalam sistem deteksi dini cyberbullying pada media sosial. Dengan performa yang cukup stabil dan efisiensi komputasi yang tinggi, model ini dapat digunakan sebagai solusi praktis dalam mendukung proses moderasi konten secara otomatis. Kontribusi utama penelitian ini tidak hanya terletak pada performa model, tetapi juga pada penggunaan dataset multiplatform yang merepresentasikan variasi bahasa yang lebih kompleks dibandingkan penelitian sebelumnya yang umumnya berfokus pada satu platform. Selain itu, model yang dihasilkan memiliki potensi untuk diintegrasikan ke dalam sistem berbasis web, sehingga mendukung implementasi deteksi dini cyberbullying secara lebih aplikatif dan efisien dalam konteks secara otomatis.

Dengan demikian, penelitian ini memberikan kontribusi baik secara akademis maupun praktis dalam pengembangan sistem deteksi cyberbullying yang adaptif terhadap dinamika bahasa pada media sosial serta dapat menjadi dasar bagi pengembangan metode yang lebih canggih di masa mendatang.

4. KESIMPULAN

Penelitian ini bertujuan untuk mengembangkan sistem klasifikasi guna mendeteksi cyberbullying pada komentar media sosial menggunakan algoritma *Support Vector Machine* (SVM) dengan representasi fitur TF-IDF. Dataset yang digunakan terdiri dari 8.000 komentar berbahasa Indonesia yang berasal dari platform YouTube dan TikTok, sehingga mampu merepresentasikan variasi bahasa yang tinggi pada lingkungan multiplatform. Hasil pengujian menunjukkan bahwa model SVM dengan kernel linear mampu mencapai akurasi sebesar 80,75%, dengan nilai precision sebesar 83% dan F1-score sebesar 0,81. Performa tersebut menunjukkan bahwa SVM memiliki kemampuan yang stabil dalam menangani data teks berdimensi tinggi, serta efektif dalam mengidentifikasi komentar yang mengandung unsur cyberbullying.

Berdasarkan analisis perbandingan dengan penelitian sebelumnya, performa SVM dalam penelitian ini tergolong kompetitif, meskipun Random Forest menunjukkan akurasi yang sedikit lebih tinggi, SVM tetap dipilih karena memberikan keseimbangan yang lebih baik antara performa dan efisiensi komputasi,

sehingga lebih sesuai untuk implementasi sistem deteksi dini berbasis real-time. Namun demikian, SVM memiliki keunggulan dalam hal efisiensi komputasi dan kesederhanaan model, sehingga lebih sesuai untuk implementasi sistem deteksi dini yang membutuhkan respons cepat dan sumber daya terbatas.

Dengan demikian, sistem klasifikasi yang dihasilkan tidak hanya berfungsi sebagai alat analisis, tetapi juga sebagai dasar dalam implementasi deteksi dini cyberbullying pada lingkungan media sosial dan juga penelitian ini memberikan kontribusi dalam bentuk penerapan sistem klasifikasi berbasis SVM pada dataset multiplatform berbahasa Indonesia, serta menunjukkan bahwa pendekatan yang digunakan mampu menjadi solusi yang efektif dan efisien dalam mendukung deteksi dini cyberbullying. Untuk penelitian selanjutnya, pengembangan dapat diarahkan pada eksplorasi metode lain dengan tetap mempertimbangkan keseimbangan antara peningkatan performa dan kompleksitas komputasi, serta perluasan dataset untuk meningkatkan kemampuan generalisasi model.

DAFTAR PUSTAKA

- [1] K. Chuluq and S. R. Nudin, "Klasifikasi Cyberbullying Pada Media Sosial Dengan Menggunakan Metode Recurrent Neural Network Dan Long Short Term Memory," *J. Inform. Comput. Sci.*, vol. 6, no. 2, pp. 501–509, 2024.
- [2] D. O. Putri and A. Qoiriah, "Klasifikasi Cyberbullying Pengguna Twitter Menggunakan Random Forest Dengan Penyematan Kata FastText," *J. Inform.*, vol. 6, pp. 265–272, 2024.
- [3] J. Lasamahu, M. G. Sopacua, and A. N. Fadillah, "Kajian Yuridis Kejahatan Cyber Bullying Di Media Sosial," *SANISA*, vol. 5, no. 1, p. 1, 2025, doi: 10.47268/sanisa.v5i1.3022.
- [4] Oryza Habibie Rahman, Gunawan Abdillah, and Agus Komarudin, "Klasifikasi Ujaran Kebencian pada Media Sosial Twitter Menggunakan Support Vector Machine," *J. RESTI Rekayasa Sist. Dan Teknol. Inf.*, vol. 5, no. 1, pp. 17–23, Feb. 2021, doi: 10.29207/resti.v5i1.2700.
- [5] S. Nauli *et al.*, "Klasifikasi Kalimat Perundungan pada Twitter Menggunakan Algoritma Support Vector Machine," *J. Inform.*, vol. 10, no. 1, pp. 107–122, 2025.
- [6] A. Ilham *et al.*, "Klasifikasi Ujaran Kebencian di Twitter Menggunakan Fitur Ekstraksi GloVe dengan Support Vector Machine," *J. Explor IT*, vol. 5, no. 36, pp. 64–72, 2023, doi: 10.35891/explorit.
- [7] M. F. Naufal *et al.*, "Analisis Perbandingan Tingkat Performa Algoritma SVM, Random Forest, dan Naive Bayes untuk Klasifikasi Cyberbullying pada Media Sosial," *Jurasik*, vol. 8, p. 82, 2023.
- [8] R. Romindo *et al.*, "Implementasi Algoritma TF-IDF dan Support Vector Machine terhadap Analisis Pendeteksi Komentar Cyberbullying di Media Sosial TikTok," *Device*, vol. 13, no. 1, pp. 124–134, 2023, doi: 10.32699/device.v13i1.5260.
- [9] F. Abdusyukur, "Penerapan Algoritma SVM untuk Klasifikasi Pencemaran Nama Baik di Twitter," *Komputa*, vol. 12, no. 1, pp. 73–82, 2023, doi: 10.34010/komputa.v12i1.9418.
- [10] H. S. Rifai *et al.*, "Analisis Sentimen Cyberbullying Menggunakan Naive Bayes," *J. IKRAITH- Inform.*, vol. 7, no. 2, pp. 183–196, 2023.
- [11] A. J. Andika *et al.*, "Deteksi Komentar Cyberbullying pada YouTube dengan CNN-LSTM," *Teknika*, vol. 12, no. 3, pp. 183–188, 2023, doi: 10.34148/teknika.v12i3.677.
- [12] S. A. Azzahra *et al.*, "Klasifikasi dan Analisis Semantik Cyberbullying Sosial Media X," *J. Inform.*, vol. 11, no. 2, pp. 353–360, 2025.
- [13] D. Nugraha *et al.*, "Media Instagram Menggunakan Metode," *J. Inform.*, vol. 8, no. 2, pp. 153–164, 2023.
- [14] K. M. Sukiakhy *et al.*, "Penggunaan Data Twitter dalam Klasifikasi dan Visualisasi Cyberbullying dengan Algoritma SVM," *Cyberspace J.*, vol. 7, no. 1, p. 1, 2023, doi: 10.22373/cj.v7i1.17358.
- [15] N. M. Farhan and B. Setiaji, "Indonesian Journal of Computer Science," *Indones. J. Comput. Sci.*, vol. 12, no. 2, pp. 284–301, 2023.
- [16] H. Santoso, R. A. Putri, and S. Sahbandi, "Deteksi Komentar Cyberbullying pada Media Sosial Instagram Menggunakan Algoritma Random Forest," *J. Manaj. Inform. JAMIKA*, vol. 13, no. 1, pp. 62–72, Mar. 2023, doi: 10.34010/jamika.v13i1.9303.
- [17] K. M. Sukiakhy, J. Junidar, and S. A. Nazhifah, "PENGUNAAN DATA PADA TWITTER DALAM KLASIFIKASI DAN VISUALISASI CYBERBULLYING DENGAN ALGORITMA SVM (SUPPORT VECTOR MACHINE)," *Cyberspace J. Pendidik. Teknol. Inf.*, vol. 7, no. 1, p. 1, Mar. 2023, doi: 10.22373/cj.v7i1.17358.

- [18] W. A. Ahmad and S. R. Wardhana, "Klasifikasi Judul Video Youtube Mengandung Clickbait Menggunakan Metode Naive Bayes," *KERNEL J. Ris. Inov. Bid. Inform. Dan Pendidik. Inform.*, vol. 4, no. 2, pp. 136–144, Dec. 2023, doi: 10.31284/j.kernel.2023.v4i2.7520.
- [19] F. Abdusyukur, "PENERAPAN ALGORITMA SUPPORT VECTOR MACHINE (SVM) UNTUK KLASIFIKASI PENCEMARAN NAMA BAIK DI MEDIA SOSIAL TWITTER," *Komputa J. Ilm. Komput. Dan Inform.*, vol. 12, no. 1, pp. 73–82, May 2023, doi: 10.34010/komputa.v12i1.9418.
- [20] A. Machmud *et al.*, "Analisis Sentimen Cyberbullying pada Komentar X Menggunakan Naive Bayes," *J. Inform.*, vol. 5, no. 1, 2025.
- [21] A. J. Andika, Y. Kristian, and E. I. Setiawan, "Deteksi Komentar Cyberbullying Pada YouTube Dengan Metode Convolutional Neural Network – Long Short-Term Memory Network (CNN-LSTM)," *Teknika*, vol. 12, no. 3, pp. 183–188, Oct. 2023, doi: 10.34148/teknika.v12i3.677.