JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 10, No. 1, Maret 2025, Pp. 107-122

KLASIFIKASI KALIMAT PERUNDUNGAN PADA TWITTER MENGGUNAKAN ALGORITMA SUPPORT VECTOR MACHINE

Sumitta Nauli*[1)], Sunneng Sandino Berutu[2)], Haeni Budiati[3)], Febe Maedjaja[4)]
1. Informatika, Fakultas Sains dan Komputer, Universitas Kristen Immanuel, Indonesia
2. Informatika, Fakultas Sains dan Komputer, Universitas Kristen Immanuel, Indonesia
3. Informatika, Fakultas Sains dan Komputer, Universitas Kristen Immanuel, Indonesia

Article Info
Kata Kunci: *Klasifikasi; Perundungan siber; Support Vector Machine; Twitter;*

Keywords: *Classification; Cyberbullying; Support Vector Machine; Twitter*

Article history:
Received 15 Oktober 2024
Revised 17 November 2024
Accepted 1 Maret 2025
Available online 1 Maret 2025

DOI :
https://doi.org/10.29100/jipi.v10i1.5749

* Corresponding author.
Sumitta Nauli
E-mail address:
miitaastl@gmail.com

ABSTRAKSI
Cyberbullying menjadi masalah yang semakin serius, terutama dengan meningkatnya media sosial dan teknologi. Twitter adalah sebagai salah satu media digital yang kerap dijadikan ajang untuk memunculkan tekanan dari sesama pengguna. Dalam riset yang dilakukan, teknik Support Vector Machine (SVM) akan diterapkan untuk membuat klasifikasi terhadap tweet sebagai tujuannya. Crawling data digunakan untuk mengumpulkan data pelatihan, yang kemudian diproses dengan melakukan tokenisasi, pembersihan data, dan TF-IDF untuk pembobotan kata. Kata-kata yang membentuk sebuah frasa memiliki fungsi sebagai fitur. Untuk menentukan model klasifikasi yang ideal, teknik SVM dikembangkan dengan memanfaatkan beberapa jenis kernel dan parameter yang berbeda. Klasifikasi tweet dilakukan berdasarkan aspek fisik dan non-fisik. Dataset yang digunakan terdiri dari 2752 data, dengan 568 kategori bullying, 2183 kategori non bullying, dan 428 untuk aspek fisik, 132 untuk aspek non-fisik. Klasifikasi terbaik ditunjukkan melalui kernel Linear yang memiliki perbandingan 90:10, menghasilkan 89,47% akurasi, 57,14% recall, 100% presisi, serta 72,72% f1-score. Perolehan riset membuktikan yaitu nilai parameter tertentu dan model teknik SVM oleh esensi linear mahir mengklasifikasikan kalimat perundungan pada Twitter dengan akurasi yang tinggi. Penelitian ini memberikan kontribusi dalam upaya mendeteksi dan menangani perundungan siber pada platform media sosial.

ABSTRACT
Cyberbullying now represents an increasingly serious problem, especially with the rise of social media and technology. One frequently used digital media platform for generating peer pressure is Twitter. During this research, to make classification of tweets as the goal will be applied utilizing the *Support Vector Machine* technique. Data crawling is used to collect training data, which is then processed by performing tokenization, data cleaning, and TF-IDF for word weighting. The words that make up a phrase function as features. To determine the ideal classification model, the SVM technique was developed by utilizing several different kernel types and parameters. Classification of tweets was done based on physical and non-physical aspects. The dataset used consists of 2752 tweets, with 568 bullying categories, 2183 non-bullying categories, and 428 for physical aspects, 132 for non-physical aspects. The best classification is shown through the Linear kernel which has a ratio of 90:10, resulting in 89.47% accuracy, 57.14% recall, 100% precision, and 72.72% f1-score. The results show that SVM models with linear kernels and certain parameter values can classify bullying sentences on Twitter with high akurasi. This research contributes to efforts to detect and deal with cyber bullying on social media platforms.

I. PENDAHULUAN

FENOMENA penggunaan media sosial khususnya Twitter, sebagai sarana berekspresi publik berkembang pesat dalam satu dekade terakhir. Media sosial menyediakan platform untuk berkomunikasi dan berbagi informasi, namun pengguna sering menjadi korban pelecehan dalam bentuk perundungan di platform tersebut [1]. Penindasan di media sosial dapat berupa kata-kata kasar, ancaman, dan perilaku berbahaya lainnya,

yang dapat berdampak negatif pada kesehatan mental korban [2].

Pelecehan online yang sering dikenal sebagai cyberbullying, adalah jenis perundungan yang umum terjadi dimasa kini. Kebanyakan pengguna tidak menyadari bahwa postingan atau ulasan yang dilihat merupakan contoh cyberbullying. Masalah ini mungkin menjadi salah satu perhatian utama bagi pengguna, masyarakat, pejabat, dan pemerintah untuk meminimalisir terjadinya insiden tersebut [3].

Dalam penelitian ini [4] menurut ketentuan hukum yang ada tertulis, terdapat suatu bagian memaparkan larangan penyebaran informasi yang dengan sengaja menimbulkan kebencian atau kebencian terhadap orang atau kelompok tertentu atas dasar SARA (ras, agama, kasta, dan golongan). Ketentuan spesifik dalam regulasi terkait informasi dan transaksi elektronik sering dikaitkan dengan masalah legislasi atau pelecehan online.

Riset ini berfokus pada kategorisasi kasus perundungan di Twitter dengan memanfaatkan teknologi algoritma Support Vector Machine. Dari masalah tersebut peneliti ingin mengukur tingkat akurasi menggunakan metode tersebut. SVM terbukti efektif untuk tugas klasifikasi teks dan membantu mengidentifikasi kalimat yang mengandung unsur perundungan dengan akurasi tinggi [5]. SVM bekerja dengan menemukan hyperplane optimal yang paling baik memisahkan titik-titik sampel yang tergabung ke dalam kelas-kelas tertentu di lingkup ruang dimensi tinggi dengan menggunakan teknik yang disebut trik kernel. *Kernel* tersebut adalah *linear*, *polynomial*, RBF dan *sigmoid*. SVM diterapkan dalam mendeteksi perundungan karena beberapa alasan. Ini termasuk kemampuan untuk menangani fitur teks yang sangat besar, ketahanan terhadap overfitting, kinerja yang baik dalam klasifikasi kalimat, dan kemampuan untuk menggunakan berbagai jenis kernel (linier, RBF, polinomial, sigmoid) yang dapat dicocokkan dengan karakteristik data. SVM sering kali memperlihatkan performa yang lebih baik dalam tugas klasifikasi kalimat perundungan dibandingkan dengan metode lain seperti Naive Bayes atau Decision Tree. Dataset pada penelitian ini berupa komentar atau cuitan [6].

Sebelum diolah, data terlebih dahulu melalui tahap preprocessing [7]. Tujuannya adalah untuk mendapatkan data yang baik dan mengurangi noise pada data. Sumber datanya mengumpulkan dataset dari komentar-komentar di jejaring sosial Twitter [8]. Setelah dilatih, model SVM dapat digunakan secara otomatis untuk mengklasifikasikan data kalimat ke dalam kategori perundungan atau non-perundungan, lalu diidentifikasi lagi kategori perundungan tersebut berdasarkan aspek menjadi aspek fisik atau non-fisik. Hasil klasifikasi ini kemudian dapat dievaluasi menggunakan metrik seperti akurasi, presisi, recall, dan skor F1 untuk mengetahui kinerja model. Selain itu, teknologi memungkinkan visualisasi dan interpretasi hasil klasifikasi, seperti dengan menampilkan kata-kata atau fitur yang paling berkontribusi terhadap klasifikasi perundungan. Hal ini dapat memberikan wawasan berguna tentang pola bahasa atau karakteristik yang terkait dengan perundungan. Melalui penerapan Support Vector Machine pada data di twitter terhadap kalimat perundungan, dapat membantu untuk meningkatkan kualitas aplikasi, memahami perspektif pengguna serta menangkap atau memfilter suatu postingan yang ada di aplikasi twitter.

Metode kategorisasi teks yang sering digunakan antara lain CNN, Classification Trees, Support Vector Machines, dan Naive Bayes. SVM memiliki hasil akurasi terbaik dari semua algoritma ini. Banyak penelitian sebelumnya yang membahas klasifikasi teks telah menunjukkan hal ini. Sebuah penelitian tentang analisis komponen sentimen internet atas ulasan Instagram dengan menggunakan kaidah klasifikasi Support Vector Machine menunjukkan hitungan yang 90% akurat [3]. Pada penelitian lain yang terdiri dari penelitian tentang pengelompokan tipe pantun memakai Support Vector Machine, akurasi yang ditemukan sangat tinggi yaitu 81,91% [9]. Di samping itu, metode SVM juga digunakan untuk mendeteksi cuitan di media sosial seperti Twitter dengan akurasi sebesar 75% [10]. Dalam penelitian lainnya [11] mengenai analisis kinerja SVM dengan identifikasi komentar bullying pada jejaring sosial yang menggunakan dataset sebanyak 1000 data dan perbandingan 60:40 menghasilkan tingkat akurasi sebesar 87,75%. Berdasarkan penelitian tersebut, dapat disimpulkan bahwa peneliti sebelumnya lebih berfokus pada analisis sentimen pada data ulasan perundungan terhadap media sosial dengan penerapan metode Support Vector Machine. Namun, penelitian ini memberikan beberapa inovasi dan nilai tambahan yang membedakan dengan penelitian sebelumnya. Jumlah dataset yang digunakan lebih besar, data ini berbeda dengan yang digunakan dalam penelitian sebelumnya karena karakteristiknya dan fitur untuk mengklasifikasikan kalimat perundungan berdasarkan aspek ini belum banyak digunakan dalam mendeteksi kalimat perundungan. Penelitian ini menyempurnakan upaya-upaya sebelumnya dalam mengkategorikan kalimat perundungan secara otomatis.

II. METODE PENELITIAN

Perancangan sistem yang hendak dipasang di penelitian dibagi kedalam beberapa bagian yaitu crawling data, pra-pemrosesan data, pengelompokan, kompilasi fitur, pemisahan data latih dan uji, proses klasifikasi menggunakan SVM, dan evaluasi terakhir. Gambar 1 menunjukkan cara *Support Vector Machine* digunakan.

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 10, No. 1, Maret 2025, Pp. 107-122

Gambar 1 Rancangan Sistem

A. Crawling Data

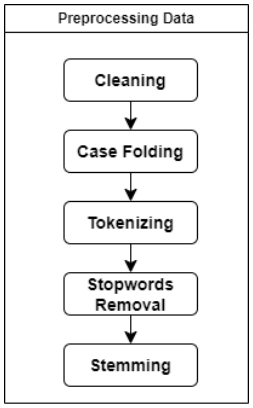
Crawling data Twitter adalah proses pengumpulan data dari Twitter menggunakan program komputer atau bot, terkadang dikenal sebagai crawler. Proses ini dilakukan dengan mengumpulkan data yang tersedia untuk umum dari Twitter dan mentransfernya ke sebuah data dasar atau file [12]. Data yang dikumpulkan meliputi tweet, akun pengguna, lokasi, dan informasi lain yang tersedia di Twitter.

Metode untuk crawling data dari media sosial Twitter adalah dengan menggunakan program crawler yang mengekstrak data dari Twitter tanpa memerlukan API key. Data yang diperoleh dari Twitter dikategorikan berdasarkan kata kunci, jumlah data yang diperoleh dan limitnya kemudian akan diekspor dalam format file CSV yang dapat dibuka dengan aplikasi Excel tanpa ada gangguan [13].

B. Preprocessing Data

Data prapemrosesan melibatkan transformasi, integrasi, dan pembersihan data untuk mempersiapkannya untuk analisis lebih lanjut. Teknik-teknik yang umum dilakukan meliputi normalisasi, deteksi data outlier, dan koreksi nilai yang hilang [14]. Preprocessing data bertujuan untuk meningkatkan kualitas data, mengurangi kesalahan dalam analisis data, dan meningkatkan efisiensi waktu dan volume data yang akan dianalisis.

Fase ini dibagi menjadi beberapa tahap, antara lain pembersihan, case folding, tokenisasi, penghapusan kata, dan pemblokiran. Kategori yang dibuat dalam kumpulan data ditunjukkan dalam gambar 2.



Gambar 2 Flowchart Preprocessing Data

Menurut gambar 2, tahap pertama dari prapemrosesan data adalah pembersihan yang bertujuan untuk membersihkan komponen yang tidak berarti dari data ulasan yang telah diperoleh. Komponen yang dibersihkan merupakan komponen yang tidak berarti atau tidak relevan untuk proses analisis data seperti URL, mention, hashtag, tanda baca, emotikon, angka, karakter tunggal dan white space. Tujuannya adalah untuk mengurangi kesalahan (noise) pada data. Langkah kedua adalah file yang mengubah semua kata di dokumen , kecuali huruf yang dihapus, menjadi huruf kecil dan huruf lainnya. Tahap ketiga adalah tokenisasi, di mana kalimat dipecah menjadi token atau kata-kata individual. Jawaban ini bergantung pada waktu. Langkah keempat adalah menghapus kata atau kata yang akan diblokir. Stopwords merupakan kumpulan kata yang tidak dianggap mempunyai arti. Selanjutnya, langkah terakhir diperkuat, realisasi menjawab kata dengan kata dasar [9].

C. Pengelompokan

Dengan demikian tahap preprocessing data dijalankan, tahap selanjutnya adalah tahap pengelompokan. Proses pengelompokan bisa juga disebut sebagai klasifikasi, yang berarti mengelompokkan benda/bagian yang sama dan memisahkan benda/bagian yang tidak sama. Secara umum, dapat mengatakan bahwa perbatasan diterapkan sebagai upaya menata dunia informasi secara sistematis [15]. Pengelompokan dapat memudahkan untuk memahami pendapat atau perasaan yang terkandung dalam teks tersebut [16].

D. Ekstraksi Fitur

Ekstraksi fitur adalah proses mengidentifikasi dan mengekstrak fitur atau karakteristik yang relevan dari data mentah yang dapat digunakan untuk representasi data yang lebih efisien [17]. Tujuannya adalah untuk mengurangi dimensionalitas data dan meningkatkan akurasi model pembelajaran mesin. Ekstraksi fitur dimanfaatkan untuk mencari informasi yang memungkinkan dan menampilkan kata-kata seperti properti komponen. Data vektor ini akan dijadikan input untuk pengklasifikasi selanjutnya. Salah satu cara untuk melakukan proses pengambilan informasi penting dari suatu data adalah dengan menggabungkan frekuensi kata dalam dokumen tertentu dan kebalikan dari frekuensi kata dalam kumpulan dokumen secara keseluruhan. Frekuensi kata dalam setiap dokumen dihitung untuk mengetahui seberapa sering suatu kata muncul, dan invers frekuensi kata dalam keseluruhan kumpulan dokumen dihitung untuk mengetahui seberapa sering suatu kata muncul. [9].

E. Data Training & Testing

Semua data harus dibagi menjadi dua kelompok yang berbeda pada tahap awal proses pemodelan pembelajaran mesin [18]. Kelompok pertama akan digunakan untuk memberikan pelatihan pada model yang akan dibangun, dan kelompok kedua akan digunakan untuk menilai seberapa baik model tersebut bekerja. Pemisahan data menjadi dua kelompok ini sangat penting untuk menghindari situasi di mana model terlalu disesuaikan dengan data pelatihan sehingga tidak dapat berfungsi dengan baik dengan data baru. Setelah pemisahan, percobaan dilanjutkan dengan membagi data menjadi tiga rasio yang berbeda antara porsi data untuk pelatihan dan pengujian, masing-masing 70:30, 80:20, dan 90:10 [19]. Hasil dari proses pemisahan data dengan ketiga rasio tersebut direpresentasikan dalam bentuk tabel.

Tabel 1 Hasil Komparatif

Data	Komparasi
Rasio 1	70:30
Rasio 2	80:20
Rasio 3	90:10

Pembagian data latih dan uji ini dilakukan dengan menggunakan ketiga rasio diatas. Untuk rasio 70:30 dimana 70% dari keseluruhan data akan digunakan sebagai data latih, dan 30% sisanya dijadikan data uji. Lalu rasio 80:20 dimana 80% data digunakan untuk pelatihan model, sedangkan 20% data digunakan untuk pengujian. Dan untuk rasio 90:10 dimana 90% data dipakai untuk data latih dan 10% untuk pengujian. Dalam penelitian ini, rasio yang dipilih nantinya adalah rasio yang memiliki evaluasi kinerja hasil akurat yang terbaik.

F. Klasifikasi dengan SVM

Metode SVM adalah sebuah algoritma pembelajaran yang populer untuk klasifikasi dan prediksi. Tujuannya adalah untuk mengetahui titik-titik terbaik yang mendistribusikan data berdasarkan jangkauan maksimum. SVM banyak digunakan karena kemampuannya untuk menangani data dalam jumlah besar dan menawarkan hasil yang baik dalam berbagai situasi [20].

Algoritma mesin vektor dukungan dikategorikan sebagai diawasi. Keunikannya adalah data yang digunakan dalam pembelajaran mesin adalah data yang telah dihasilkan sebelumnya. Hasilnya, mesin akan menerapkan hasil pengujian berdasarkan berbagai faktor dalam proses pengambilan keputusan [21]. Pada penelitian ini, SVM dipakai untuk mengkategorikan kalimat kedalam dua kelas yaitu perundungan dan non-perundungan, lalu diidentifikasi kembali kalimat perundungan tersebut kedalam dua aspek yaitu aspek fisik dan non-fisik.

Pengoperasian metode Support Vector Machine melibatkan proyeksi kernel dalam ruang yang besar, terutama untuk persoalan nonlinier. Cara kerja SVM ini cukup unik, targetnya ialah untuk mencari garis batas (hyperplane) atau pembagi yang dapat mewakili selisih (margin) antar kumpulan data. Agar dapat menemukan hyperplane yang sesuai, dapat menetapkan batas dan mendapatkan nilai maksimum. Metode representasi hyperplane yang terbaik didasarkan pada Support Vector Machine [22]. Selain itu, SVM dapat memilih jenis kernel yang umum digunakan. Dengan menggunakan kernel ini, data dapat dipetakan ke dalam fitur yang lebih tinggi. Ini memungkinkan pemisahan data yang sebelumnya tidak dapat dipisahkan secara linear. Kernel linear, kernel polinomial, kernel RBF, dan kernel sigmoid adalah beberapa jenis kernel yang biasa digunakan dalam SVM. SVM sangat baik untuk

menangani data berukuran besar. Dalam klasifikasi kalimat, setiap kalimat akan digambarkan sebagai vektor fitur yang memiliki dimensi yang sangat besar, tergantung pada jumlah kata unik yang ada dalam dataset. SVM bekerja dengan baik dan akurat meskipun menangani data berukuran besar seperti ini, tidak seperti metode klasifikasi lain yang mungkin akan menghadapi kesulitan.

Model SVM menerapkan fitur kernel linier, polinomial, RBF, dan sigmoid terhadap kelas yang berbeda. Fungsi "fit" digunakan untuk melatih model dengan data latih. Setelah model dilatih, dilakukan prediksi pada data uji menggunakan fungsi "predict". Setelah prediksi dilakukan kemudian dilakukan perhintungan akurasi menggunakan fungsi "accurary_score" dari library scikit-learn. Fungsi ini akan membandingkan nilai prediksi dengan nilai yang sebenarnya (data test) untuk menghitung akurasi dari setiap kernel [23].

Penelitian ini menggunakan Python sebagai alat utama untuk membangun dan melatih model SVM. Python dipilih karena mudah digunakan dan didukung banyak library yang kuat terkait dalam pengolahan data dan machine learning. Salah satu library yang dimanfaatkan adalah scikit-learn, yaitu machine learning untuk Python yang dibangun diatas Numpy atau matplotlib. Untuk mengimplementasikan model SVM dalam scikit-learn dengan menggunakan modul SVC (Support Vector Classifier). Modul ini menawarkan metode yang sederhana untuk membangun dan melatih model SVM dengan berbagai pengaturan parameter yang dapat disesuaikan. Selain scikit-learn, penelitian ini juga memanfaatkan library Python lain untuk membantu proses preprocessing data seperti spaCy. Untuk memastikan kinerja terbaik dari model dalam mengklasifikasikan kalimat perundungan dan non-perundungan, pengaturan parameter dan pemilihan library pendukung dilakukan dengan hati-hati.

G. Evaluasi

Untuk mengukur kinerja model yang dihasilkan dari proses SVM, tahap evaluasi dilakukan [24]. Pada tahap ini, metrik seperti tingkat keakuratan, ketepatan, recall, dan f1-skor [25]. Tujuan utama dari proses penilaian ini adalah untuk mendapatkan pemahaman tentang seberapa baik kinerja model dalam menangani masalah yang dihadapi dan untuk menemukan bagian mana yang perlu diperbaiki. Akurasi ini yang mengukur persentase prediksi yang benar dari keseluruhan data yang diuji. Selanjutnya presisi ini menghitung seberapa banyak prediksi positif, seperti kalimat perundungan yang benar-benar positif dari prediksi positif secara keseluruhan. Lalu, ada recall yang menghitung seberapa banyak prediksi positif yang benar dari total kasus positif yang nyata. Terakhir, f1-skor yang merupakan nilai rata-rata antara presisi dan recall. F1-skor ini berguna untuk memberikan gambaran tentang bagaimana kinerja model dengan mempertimbangkan kedua metrik tersebut.

Keempat metrik ini digunakan dalam penelitian untuk mengevaluasi kinerja model SVM dalam mengklasifikasikan kalimat perundungan dan non-perundungan. Presisi dan recall memberikan wawasan lebih detail tentang kinerja dalam memprediksi positif yaitu perundungan dan negatif yaitu non-perundungan, sementara akurasi memberikan gambaran umum seberapa baik model. Setelah itu, skor F1 menggabungkan semua metrik tersebut menjadi satu nilai. Dengan adanya metrik ini dapat menganalisis kekuatan dan kelemahan model dengan lebih baik, serta membandingkannya dengan metode lain yang mungkin digunakan dalam penelitian serupa. Semakin tinggi nilai metrik tersebut, semakin baik kinerja model dalam melakukan klasifikasi teks perundungan.

III. HASIL DAN PEMBAHASAN

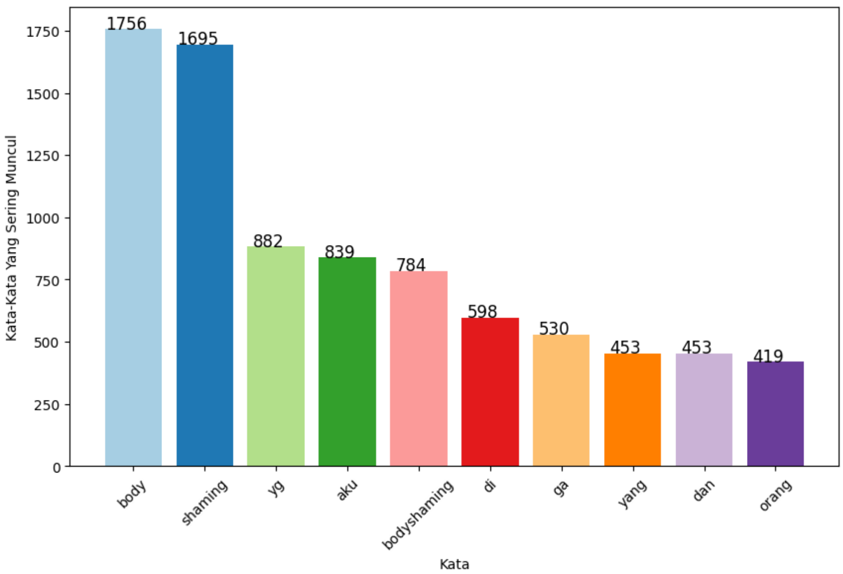
A. Hasil Crawling Data

Proses *crawling* data memerlukan beberapa filter untuk mengambil data ulasan. Filternya yaitu bahasa tweet adalah bahasa indonesia, data paling relevan (*most relevant*), jumlah maksimum data 5000 dengan kata kunci 'body shaming'. Dari hasil proses *crawling* tersebut, maka didapat jumlah tweet sebanyak 2757 data dengan 15 kolom yang terdiri dari kolom *conversation id str, created at, favorite count, full text, id str, image url, in reply to screen name, lang, location, quote count, reply count, retweet count, tweet url, user id str, username*. Hasil dari tweet akan divisualisasikan menggunakan library *WordCloud* dan *Plot*.

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 10, No. 1, Maret 2025, Pp. 107-122

Gambar 3 Hasil Visualisasi *Wordcloud Crawling* Data

Gambar 3 menunjukkan bagaimana word cloud menggambarkan hasil grup tweet. Word cloud ini berisi berbagai kata dan frasa yang berhubungan dengan binaraga. Kata “body shaming” sendiri muncul di tengah-tengah kata yang sangat besar, menandakan bahwa inilah topik perbincangan sebenarnya.



Gambar 4 Hasil Visualisasi *Plot Frekuensi Kata*

Untuk menunjukkan kata-kata yang paling sering digunakan dalam konteks perundungan dalam kumpulan data, representasi visual diberikan dalam bentuk grafik. Grafik ini menunjukkan kata-kata yang paling sering digunakan dalam konteks perundungan dan menunjukkan frekuensi kemunculan masing-masing kata yang terkait. Sebagai contoh, grafik dengan jumlah 1756 menggunakan kata “body” paling sering. Ini dapat merujuk pada istilah kasar, cercaan, atau hinaan yang biasa digunakan untuk mengintimidasi.

B. Hasil Preprocessing Data

1. Cleaning

Cleaning adalah proses menganalisis kualitas data secara akurat dengan menggunakan metode seperti memodifikasi, mengganti, dan menghilangkan data yang tidak sesuai dengan format atau file yang paling sering ditemukan pada dataset untuk menghasilkan data dengan kualitas tinggi. Hasil proses cleaning pada tabel 2 dapat dilihat bahwa data telah bersih dari komponen yang tidak berarti diantarannya URL, mention, hashtag, tanda baca, emotikon, angka, karakter tunggal dan white space.

Tabel 2 Hasil Cleaning Data

full_text	Cleaning
kenapa lo tangisin? seorang pembully dan penipu kalian tangisin? seorang leader ga becus yg malah biarin membernya bodyshaming dan colorshaming member lain? dan terang an nyuruh membernya diet? org itu yg kalian tangisin kalo gue sih malah gue ketawain karma soalnya	kenapa lo tangisin seorang pembully dan penipu kalian tangisin seorang leader ga becus yg malah biarin membernya bodyshaming dan colorshaming member lain dan terang an nyuruh membernya diet org itu yg kalian tangisin kalo gue sih malah gue ketawain karma soalnya
Berat badanku awalnya 55 terus sering kena bodyshaming dibilang penyakitan cacingan dan lainnya. Sekarang udah jadi 70an kg masih tetep aja ada yang bilang terlalu kurus dan kurang ideal. Awalnya aku ngga ambil pusing eh lama2 sedih juga digituin terus.	Berat badanku awalnya terus sering kena bodyshaming dibilang penyakitan cacingan dan lainnya Sekarang udah jadi an kg masih tetep aja ada yang bilang terlalu kurus dan kurang ideal Awalnya aku ngga ambil pusing eh lama sedih juga digituin terus
@putheerow @kochengfs : jgn body shaming kak.. aku badannya ideal begini https://t.co/qYodASC7EG	putheerow kochengfs jgn body shaming kak aku badannya ideal begini
@tanyakanrl UP ⬆ 🆙 !!!!!!!! gue juga nge fans HSH tapi itu ngeliat die harder nya nyebokkin delusional n kasar sampe body shaming ga perlu itu ewwww banget	tanyakanrl UP gue juga nge fans HSH tapi itu ngeliat die harder nya nyebokkin delusional n kasar sampe body shaming ga perlu itu ewwww banget

2. Case Folding

Teks tweet harus dibulatkan atau diubah untuk mengurangi karakter huruf kecil dan non-alfanumerik.

Tabel 3 Hasil Case Folding Data

Cleaning	Case Folding
kenapa lo tangisin seorang pembully dan penipu kalian tangisin seorang leader ga becus yg malah biarin membernya bodyshaming dan colorshaming member lain dan terang an nyuruh membernya diet org itu yg kalian tangisin kalo gue sih malah gue ketawain karma soalnya	kenapa lo tangisin seorang pembully dan penipu kalian tangisin seorang leader ga becus yg malah biarin membernya bodyshaming dan colorshaming member lain dan terang an nyuruh membernya diet org itu yg kalian tangisin kalo gue sih malah gue ketawain karma soalnya
Berat badanku awalnya terus sering kena bodyshaming dibilang penyakitan cacingan dan lainnya Sekarang udah jadi an kg masih tetep aja ada yang bilang terlalu kurus dan kurang ideal Awalnya aku ngga ambil pusing eh lama sedih juga digituin terus	berat badanku awalnya terus sering kena bodyshaming dibilang penyakitan cacingan dan lainnya sekarang udah jadi an kg masih tetep aja ada yang bilang terlalu kurus dan kurang ideal awalnya aku ngga ambil pusing eh lama sedih juga digituin terus
putheerow kochengfs jgn body shaming kak aku badannya ideal begini	putheerow kochengfs jgn body shaming kak aku badannya ideal begini
tanyakanrl UP gue juga nge fans HSH tapi itu ngeliat die harder nya nyebokkin delusional n kasar sampe body shaming ga perlu itu ewwww banget	tanyakanrl up gue juga nge fans hsh tapi itu ngeliat die harder nya nyebokkin delusional n kasar sampe body shaming ga perlu itu ewwww banget

3. Tokenizing

Proses mengubah teks menjadi kode disebut tokenisasi. Token bisa berbentuk kalimat, ungkapan, angka, atau elemen-elemen tertulis lainnya. Tokenisasi bertujuan untuk membuat dokumen lebih mudah dibaca dan diubah. Tabel 4 menunjukkan bahwa metode identifikasi yang berbeda memiliki hasil yang baik dalam identifikasi (kata).

Tabel 4 Hasil Tokenizing

Case Folding	Tokenisasi
kenapa lo tangisin seorang pembully dan penipu kalian tangisin seorang leader ga becus yg malah biarin membernya bodyshaming dan colorshaming member lain dan terang an nyuruh membernya diet org itu yg kalian tangisin kalo gue sih malah gue ketawain karma soalnya	['kenapa', 'lo', 'tangisin', 'seorang', 'pembully', 'dan', 'penipu', 'kalian', 'tangisin', 'seorang', 'leader', 'ga', 'becus', 'yg', 'malah', 'biarin', 'membernya', 'bodyshaming', 'dan', 'colorshaming', 'member', 'lain', 'dan', 'terang', 'an', 'nyuruh', 'membernya', 'diet', 'org', 'itu', 'yg', 'kalian', 'tangisin', 'kalo', 'gue', 'sih', 'malah', 'gue', 'ketawain', 'karma', 'soalnya']
berat badanku awalnya terus sering kena bodyshaming dibilang penyakitan cacingan dan lainnya sekarang udah jadi an kg masih tetep aja ada yang bilang terlalu kurus dan kurang ideal awalnya aku ngga ambil pusing eh lama sedih juga digituin terus	['berat', 'badanku', 'awalnya', 'terus', 'sering', 'kena', 'bodyshaming', 'dibilang', 'penyakitan', 'cacingan', 'dan', 'lainnya', 'sekarang', 'udah', 'jadi', 'an', 'kg', 'masih', 'tetep', 'aja', 'ada', 'yang', 'bilang', 'terlalu', 'kurus', 'dan', 'kurang', 'ideal', 'awalnya', 'aku', 'ngga', 'ambil', 'pusing', 'eh', 'lama', 'sedih', 'juga', 'digituin', 'terus']
putheerow kochengfs jgn body shaming kak aku badannya ideal begini	['putheerow', 'kochengfs', 'jgn', 'body', 'shaming', 'kak', 'aku', 'badannya', 'ideal', 'begini']
tanyakanrl up gue juga nge fans hsh tapi itu ngeliat die harder nya nyebokkin delusional n kasar sampe body shaming ga perlu itu ewwww banget	['tanyakanrl', 'up', 'gue', 'juga', 'nge', 'fans', 'hsh', 'tapi', 'itu', 'ngeliat', 'die', 'harder', 'nya', 'nyebokkin', 'delusional', 'n', 'kasar', 'sampe', 'body', 'shaming', 'ga', 'perlu', 'itu', 'ewwww', 'banget']

4. Stopwords Removal

Pemfilteran disebut juga pemblokiran kata, yang merupakan proses menghapus kalimat yang banyak tampil dalam bacaan namun tidak mengandung informasi berharga ketika disortir. Stopword tidak memiliki dampak yang signifikan terhadap kualitas teks dan dapat memengaruhi hasil evaluasi. Namun, penting untuk dipahami bahwa pilihan stopword dapat beragam, tergantung pada tujuan analisis dan bahasa yang digunakan.

Tabel 5 Hasil Stopwords Removal

Tokenisasi	Stopwords Removal
['kenapa', 'lo', 'tangisin', 'seorang', 'pembully', 'dan', 'penipu', 'kalian', 'tangisin', 'seorang', 'leader', 'ga', 'becus', 'yg', 'malah', 'biarin', 'membernya', 'bodyshaming', 'dan', 'colorshaming', 'member', 'lain', 'dan', 'terang', 'an', 'nyuruh', 'membernya', 'diet', 'org', 'itu', 'yg', 'kalian', 'tangisin', 'kalo', 'gue', 'sih', 'malah', 'gue', 'ketawain', 'karma', 'soalnya']	['lo', 'tangisin', 'pembully', 'penipu', 'tangisin', 'leader', 'ga', 'becus', 'yg', 'biarin', 'membernya', 'bodyshaming', 'colorshaming', 'member', 'terang', 'an', 'nyuruh', 'membernya', 'diet', 'org', 'yg', 'tangisin', 'kalo', 'gue', 'sih', 'gue', 'ketawain', 'karma']
['berat', 'badanku', 'awalnya', 'terus', 'sering', 'kena', 'bodyshaming', 'dibilang', 'penyakitan', 'cacingan', 'dan', 'lainnya', 'sekarang', 'udah', 'jadi', 'an', 'kg', 'masih', 'tetep', 'aja', 'ada', 'yang', 'bilang', 'terlalu', 'kurus', 'dan', 'kurang', 'ideal', 'awalnya', 'aku', 'ngga', 'ambil', 'pusing', 'eh', 'lama', 'sedih', 'juga', 'digituin', 'terus']	['berat', 'badanku', 'kena', 'bodyshaming', 'dibilang', 'penyakitan', 'cacingan', 'udah', 'an', 'kg', 'tetep', 'aja', 'bilang', 'kurus', 'ideal', 'ngga', 'ambil', 'pusing', 'eh', 'sedih', 'digituin']
['putheerow', 'kochengfs', 'jgn', 'body', 'shaming', 'kak', 'aku', 'badannya', 'ideal', 'begini']	['putheerow', 'kochengfs', 'jgn', 'body', 'shaming', 'kak', 'badannya', 'ideal']
['tanyakanrl', 'up', 'gue', 'juga', 'nge', 'fans', 'hsh', 'tapi', 'itu', 'ngeliat', 'die', 'harder', 'nya', 'nyebokkin', 'delusional', 'n', 'kasar', 'sampe', 'body', 'shaming', 'ga', 'perlu', 'itu', 'ewwww', 'banget']	['tanyakanrl', 'up', 'gue', 'nge', 'fans', 'hsh', 'ngeliat', 'die', 'harder', 'nya', 'nyebokkin', 'delusional', 'n', 'kasar', 'sampe', 'body', 'shaming', 'ga', 'ewwww', 'banget']

5. Stemming

Stemming adalah proses mengkonversi kata dari bentuk aslinya ke bentuk yang lebih simpel. Proses stemming digunakan untuk menemukan kata-kata yang telah ditranskripsikan sebelumnya. Salah satu contoh proses stemming adalah mengubah kata “dikembangkan” menjadi kata dasar “kembang”. Teks kata-kata yang telah di-stemming akan dianggap sebagai satu kata, yang dapat meningkatkan akurasi dalam pengolahan data.

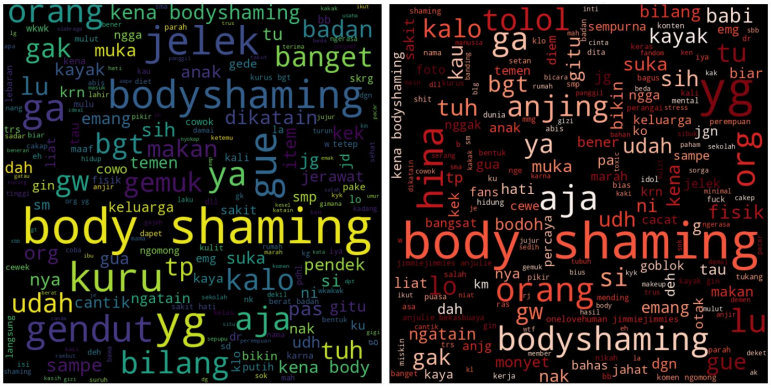
Tabel 6 Hasil Stemming Data

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 10, No. 1, Maret 2025, Pp. 107-122

Stopwords Removal	Stemming Data
['lo', 'tangisin', 'pembully', 'penipu', 'tangisin', 'leader', 'ga', 'becus', 'yg', 'biarin', 'membernya', 'bodyshaming', 'colorshaming', 'member', 'terang', 'an', 'nyuruh', 'membernya', 'diet', 'org', 'yg', 'tangisin', 'kalo', 'gue', 'sih', 'gue', 'ketawain', 'karma']	lo tangisin pembully tipu tangisin leader ga becus yg biarin membernya bodyshaming colorshaming member terang an nyuruh membernya diet org yg tangisin kalo gue sih gue ketawain karma
['berat', 'badanku', 'kena', 'bodyshaming', 'dibilang', 'penyakitan', 'cacingan', 'udah', 'an', 'kg', 'tetep', 'aja', 'bilang', 'kurus', 'ideal', 'ngga', 'ambil', 'pusing', 'eh', 'sedih', 'digituin']	berat badan kena bodyshaming bilang sakit cacing udah an kg tetep aja bilang kurus ideal ngga ambil pusing eh sedih digituin
['putheerow', 'kochengfs', 'jgn', 'body', 'shaming', 'kak', 'badannya', 'ideal']	putheerow kochengfs jgn body shaming kak badan ideal
['tanyakanrl', 'up', 'gue', 'nge', 'fans', 'hsh', 'ngeliat', 'die', 'harder', 'nya', 'nyebokkin', 'delusional', 'n', 'kasar', 'sampe', 'body', 'shaming', 'ga', 'ewwww', 'banget']	tanyakanrl up gue nge fans hsh liat die harder nya nyebokkin delusional n kasar sampe body shaming ga ewwww banget

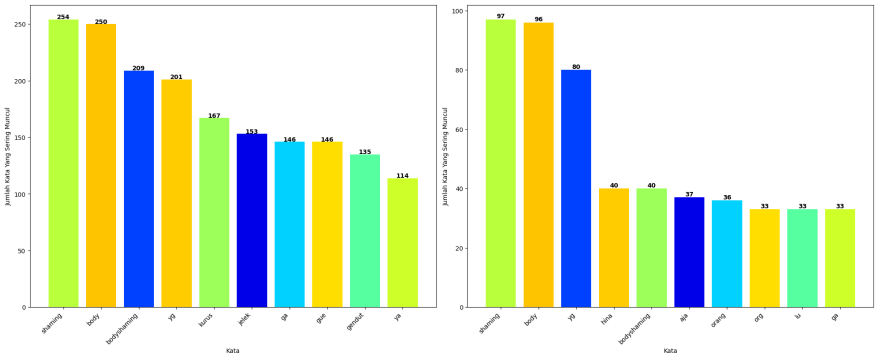
C. *Hasil Pengelompokan*

Pengelompokan dilakukan dengan mengelompokkan tweet kedalam kategori Bullying atau Non-bullying, lalu di indentifikasi lagi jika tweet masuk kedalam kategori Bullying maka akan dikategorikan lagi berdasarkan aspeknya yaitu Fisik dan Non-fisik. Dalam proses pengelompokan dilakukan pengecekan kondisi menggunakan tool spaCy dengan pernyataan if-elif-else untuk menentukan label berdasarkan kata kunci yang ada pada kolom Dataframe. Merujuk pada visualisasi yang disajikan dalam representasi grafis sebelumnya, word cloud dan plot digunakan untuk menggambarkan hasil grup tweet.



a. Aspek Fisik b. Aspek Non-fisik

Gambar 5 Wordcloud Pengelompokan Kategori Bullying pada Aspek a) Fisik dan b) Non-fisik



a. Aspek Fisik b. Aspek Non-fisik

Gambar 6 Plot Pengelompokan Kategori Bullying pada Aspek a) Fisik dan b) Non-fisik

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 10, No. 1, Maret 2025, Pp. 107-122

Ilustrasi yang disajikan menunjukkan dua konsep yang berbeda satu sama lain. Jumlah kata dalam kategori fisik ditampilkan di sebelah kiri, dan jumlah kata dalam kategori non-fisik ditampilkan di sebelah kanan. Aspek fisik mencakup istilah yang berhubungan dengan objek fisik, sedangkan aspek non-fisik mencakup istilah yang berhubungan dengan pikiran, emosi, dan pemikiran abstrak. "orang", "tubuh", dan "badan" adalah terminologi dengan frekuensi kemunculan tertinggi dalam kategori fisik, sedangkan "kata", "frekuensi", dan "jumlah" adalah kosakata yang paling banyak dimanfaatkan dalam kategori non-fisik.

Pada hasil crawling sebelumnya didapatkan akumulasi jumlah kalimat bullying sebanyak 568 data. Namun setelah dilakukan pengelompokan berdasarkan aspek fisik dan non-fisik ditemukan 428 data untuk aspek fisik, 132 data untuk aspek non-fisik, dan 8 data untuk aspek yang tidak dapat teridentifikasi.

D. Ekstraksi Fitur

Untuk ekstraksi faktor pada penelitian ini, pembobotan TF-IDF diterapkan sebagai langkah selanjutnya. Dalam tahapan ini, data teks akan diubah menjadi angka. Gambar 7 menunjukkan hasil dari fungsi TF-IDF.

```
(0, 1367)     0.16612016093544024
(0, 2232)     0.16612016093544024
(0, 190)      0.18095119613703958
(0, 239)      0.5639775733675453
(0, 2994)     0.1223754338946239
(0, 1841)     0.17519801199426827
(0, 1565)     0.15079988879612602
(0, 3016)     0.15607130555983575
(0, 3233)     0.06700813838138932
(0, 48)       0.0804053596230067
(0, 2250)     0.1953506904611606
(0, 2686)     0.04855534455649223
(0, 438)      0.04884177754031376
```

Gambar 7 Hasil Pembobotan TF-IDF

E. Data Training & Testing

Setelah melakukan pembobotan kata, metode yang menggabungkan frekuensi kata yang muncul dalam dokumen tertentu dan invers frekuensi kata yang muncul dalam dokumen secara keseluruhan, langkah selanjutnya adalah membagi semua data yang tersedia menjadi dua subset. Subset pertama akan digunakan untuk proses pelatihan model, dan subset kedua akan digunakan untuk menilai kinerja model yang telah dilatih. Pembahagian data menjadi tiga rasio yang berbeda antara ukuran subset pelatihan dan subset evaluasi digunakan untuk melakukan eksperimen, yaitu 0,7:0,3; 0,8:0,2; 0,9:0,1.

Tabel 7 Hasil Data Latih dan Data Uji dalam 3 Rasio

Rasio	Aspek	Data Training	Data Testing
70:30	Fisik	299	129
	Non-fisik	92	40
80:20	Fisik	342	86
	Non-fisik	105	27
90:10	Fisik	385	43
	Non-fisik	118	14

F. Klasifikasi SVM dan Evaluasi

Pemodelan melibatkan kernel linear, polinomial, RBF, dan sigmoid dengan teknik klasifikasi yang memanfaatkan prinsip pemisahan bidang vektor dalam ruang fitur diaplikasikan pada penelitian ini. Setelah model diterapkan, hasil nyata dievaluasi menggunakan matriks konfusi. Berdasarkan subjek, indeks, recall, dan skor f1 dicari.

Setelah dilakukan pemodelan dengan 4 kernel svm dan ke 3 rasio data, di dapatkan hasil *akurasi* tertinggi di rasio 3 dengan pembagian data 90:10 sebesar 89,47% untuk kernel Linear. Dari semua akurasi yang didapatkan,

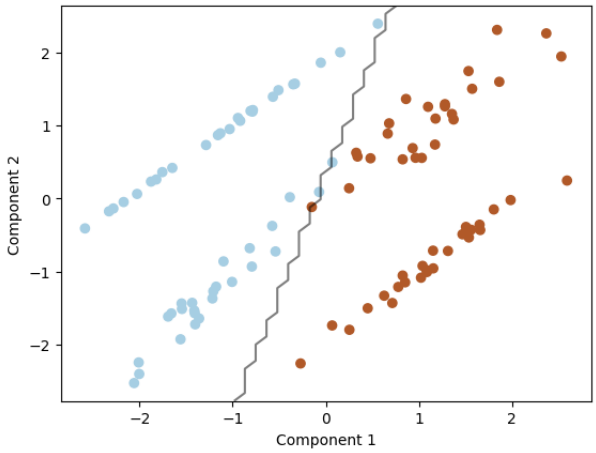
JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 10, No. 1, Maret 2025, Pp. 107-122

disimpulkan kernel *linear* memiliki *akurasi* yang lebih tinggi dari kernel lainnya. Tabel 8 menunjukkan hasil perbandingan.

Tabel 8 Hasil Klasifikasi SVM

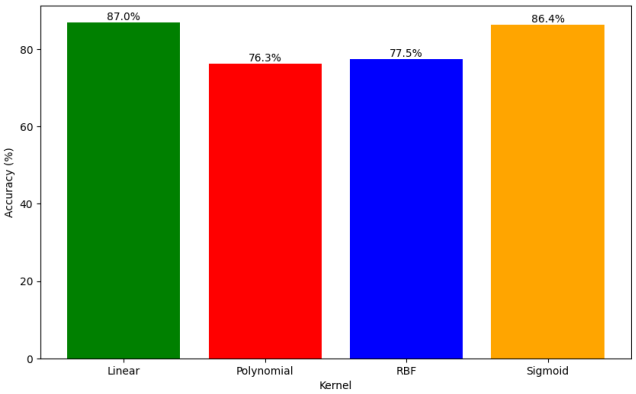
Rasio	Kernel	Akurasi
70:30	Linear	86,98%
	Polynomial	76,33%
	RBF	77,51%
	Sigmoid	86,39%
80:20	Linear	84,95%
	Polynomial	76,10%
	RBF	77,87%
	Sigmoid	84,07%
90:10	**Linear**	**89,47%**
	Polynomial	75,43%
	RBF	77,19%
	Sigmoid	87,71%

Selanjutnya, garis visi inti hyperplane dibuat dan data 3 diberi definisi yang benar. Gambar 8 menunjukkan hasilnya. Hyperplane untuk SVM dengan data kernel linear 3 adalah garis lurus yang memisahkan dua kumpulan data yang berbeda. Dalam kasus ini, hyperplane adalah garis lurus yang membagi data menjadi dua lapisan. Titik biru dan merah menunjukkan dua kelas data yang berbeda, dan sumbu x dan y menunjukkan dua elemen yang digunakan untuk mengurutkan data.



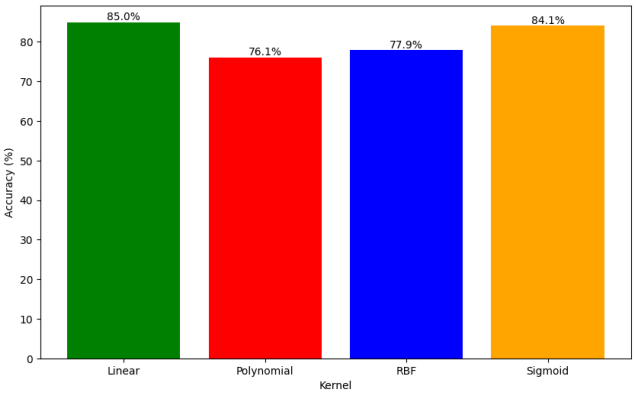
Gambar 8 Hyperplane Kernel Linear dengan Rasio Data 3

Berikut hasil *akurasi* dari setiap kernel berdasarkan rasio data 1, data 2, data 3 dengan nilai persentasenya. Kernel linier disorot dengan warna hijau, kernel polinomial berwarna merah, kernel rbf berwarna biru, dan kernel sigmoid berwarna merah-oranye. Temuan yang didapatkan diilustrasikan melalui visualisasi pada ketiga gambar berikut, yakni Gambar 9, Gambar 10, dan Gambar 11.



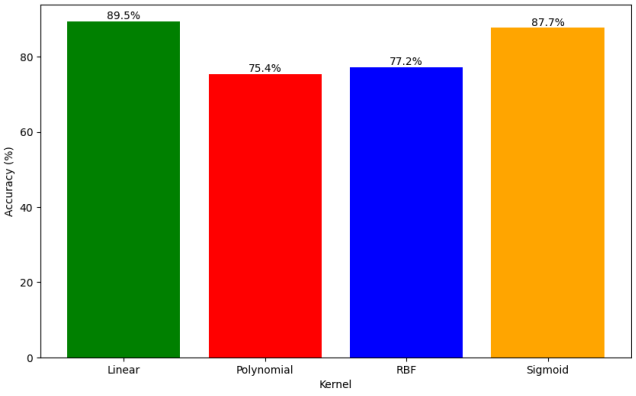
Gambar 9 Perbandingan *Akurasi Kernel* dengan Rasio Data 1

Gambar 9 ini menunjukkan grafik yang membandingkan akurasi dari empat algoritma SVM linier, SVM polinomial, SVM RBF, dan SVM Sigmoid. Perbandingannya adalah dari 1 hingga 70 persen, dengan 30% data yang digunakan untuk pengujian dan 70% lainnya untuk pelatihan. Akurasi SVM linier adalah yang tertinggi (87,0%) dibandingkan dengan algoritma SVM lainnya, seperti yang ditunjukkan pada grafik. SVM linear memiliki akurasi polynomial sebesar 76,3%, akurasi SVM RBF sebesar 77,5%, dan akurasi SVM Sigmoid sebesar 86,4%. Linear SVM adalah algoritma terbaik untuk mengklasifikasikan data dalam dataset ini. Namun, perlu diingat bahwa akurasi algoritma dapat bervariasi tergantung pada jenis data, ukuran data, dan algoritma.



Gambar 10 Perbandingan *Akurasi Kernel* dengan Rasio Data 2

Kemudian, grafik yang membandingkan akurasi dari keempat algoritma (SVM linier, SVM polinomial, SVM RBF, dan SVM sigmoid) ditunjukkan pada Gambar 10. Model ini menggunakan rasio 80:20, yang berarti 20 persen data digunakan untuk pengujian dan 80 persen untuk pelatihan. Akurasi SVM linier jelas lebih tinggi (85,0%) dibandingkan dengan algoritma SVM lainnya. Akurasi SVM Polynomial adalah 76,1%, akurasi SVM RBF adalah 77,9%, dan akurasi SVM Sigmoid adalah 84,1%, yang mengindikasikan bahwa SVM linier adalah algoritma SVM yang paling akurat untuk klasifikasi data pada dataset ini. Namun, perlu diingat bahwa akurasi algoritma dapat bervariasi tergantung pada jenis data, ukuran data, dan algoritma.

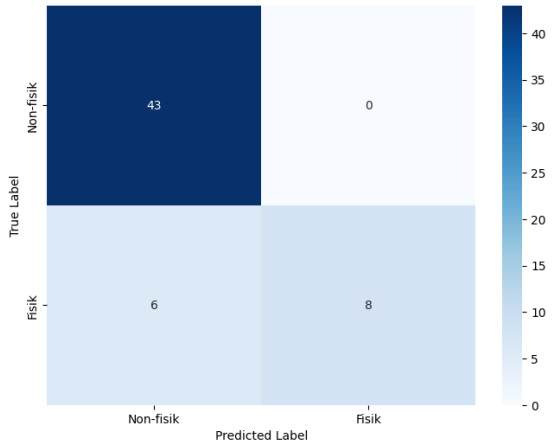


Gambar 11 Perbandingan *Akurasi Kernel* dengan Rasio Data 3

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 10, No. 1, Maret 2025, Pp. 107-122

Terakhir, grafik yang membandingkan akurasi dari keempat algoritma SVM linier, SVM polinomial, SVM RBF, dan SVM sigmoid ditunjukkan pada Gambar 11. Visualisasi tersebut dibuat dengan menggunakan sebaran data yang terbagi atas perbandingan 10:90. Sembilan puluh persen dari sebaran data digunakan untuk proses pembelajaran, dan sepuluh persen lainnya digunakan untuk pengujian. Berdasarkan grafik, dapat dilihat bahwa Linear SVM memiliki akurasi tertinggi (89.5%) dibandingkan dengan algoritma SVM lainnya. Polynomial SVM memiliki akurasi 75.4%, RBF SVM memiliki akurasi 77.2%, dan Sigmoid SVM memiliki akurasi 87.7%. Hasilnya menunjukkan bahwa metode SVM linier adalah yang paling efektif untuk mengkategorisasi kumpulan data yang digunakan. Dengan demikian, dapat disimpulkan bahwa, dibandingkan dengan simulasi 1, tingkat ketepatan klasifikasi yang dihasilkan oleh SVM linier lebih tinggi pada simulasi 3 dan 2. Hal ini menunjukkan bahwa Linear SVM cocok digunakan pada dataset yang berisi data dalam jumlah besar. SVM Presisi SVM Polinomial dan SVM Sigmoid lebih tinggi untuk data rate 1 dibandingkan data 3 dan 2. Hal ini menunjukkan bahwa Polynomial SVM dan Sigmoid SVM cocok untuk data dengan data yang lebih sedikit. SVM RBF yang sebenarnya stabil untuk ketiga kumpulan data. Temuan ini mengindikasikan bahwa pendekatan RBF SVM memiliki karakteristik yang tidak kompleks serta memiliki sifat fleksibilitas dalam pengaplikasiannya pada berbagai ragam data.

Visualisasi dilakukan dengan menggunakan kurva linear untuk mendapatkan hasil konfusi matriks yang akurat. Gambar 12 menunjukkan gambar yang dihasilkan.



Gambar 12 Visualisasi *Confusion Matrix* Kernel Linear

Perhitungan manual dilakukan dengan hasil klasifikasi yang tertinggi. Untuk mendapatkan nilai *presisi, recall,* dan *f1-score*, matriks dibagi menjadi 2 kelas. Peneliti memilih hasil kernel Linear dengan rasio data 3 yang memiliki nilai akurasi tertinggi. Berdasarkan data pada Gambar 12, ada 43 data non-fisik yang benar diprediksi yang termasuk dalam kelas aspek non-fisik dan 6 data non-fisik yang benar diprediksi yang termasuk dalam kelas aspek fisik. Selain itu, ada 8 data fisik yang benar diprediksi yang termasuk dalam kelas aspek fisik.

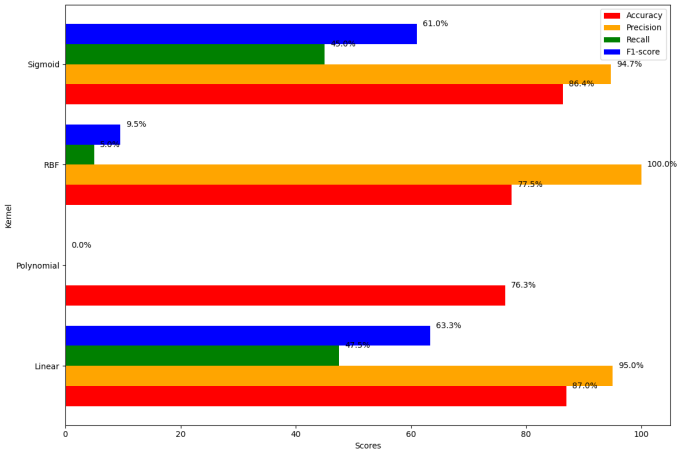
Setelah semua proses implementasi dan evaluasi metode *svm* dengan semua kernel, maka didapatkan tingkat ketepatan klasifikasi, ketelitian, keberhasilan pemanggilan kembali, dan skor harmonis, yang menggabungkan presisi dan recall, adalah metrik yang digunakan untuk menilai. Tabel 9 memperlihatkan yaitu hasil perbandingan masing-masing kernel dengan data rasio yang berbeda dimana kernel Linear pada rasio 3 memiliki persentase akurasi yang paling tinggi dengan nilai 89,47%, presisi 100%, recall 57,14% dan f1-score dengan nilai 72,72%. Kernel yang memiliki akurasi terendah dari antara kernel lainnya dengan testing 3 rasio data adalah kernel polynomial.

Tabel 9 Hasil Perbandingan Evaluasi Rasio

Rasio	Kernel	Akurasi	Presisi	Recall	F1-score
70:30	Linear	86,98%	95,0%	47,5%	63,33%
	Polynomial	76,33%	nan	0,0%	nan
	RBF	77,51%	100,0%	5,0%	9,52%
	Sigmoid	86,39%	94,73%	45,0%	61,01%

80:20	Linear	84,95%	100,0%	37,03%	54,05%
	Polynomial	76,10%	nan	0,0%	nan
	RBF	77,87%	100,0%	7,40%	13,79%
	Sigmoid	84,07%	100,0%	33,33%	50,0%
90:10	**Linear**	**89,47%**	**100,0%**	**57,14%**	**72,72%**
	Polynomial	75,43%	nan	0,0%	nan
	RBF	77,19%	100,0%	7,14%	13,33%
	Sigmoid	87,71%	100,0%	50,0%	66,66%

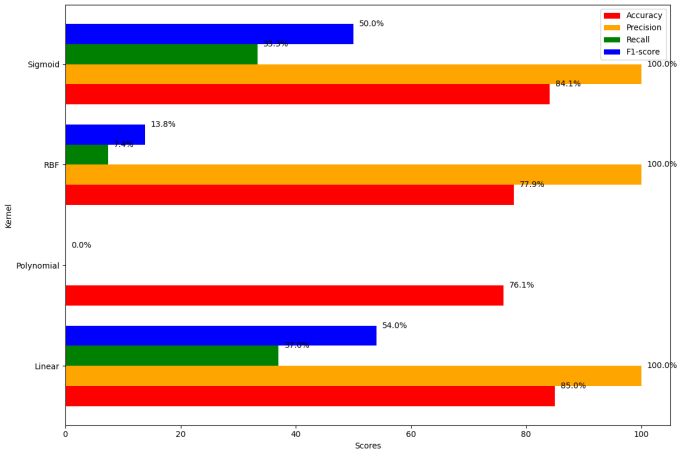
Hasil visualisasi statistik kinerja setiap core terhadap data pertama, data kedua, dan data ketiga, beserta persentasenya, disajikan di bawah ini. Warna merah menandakan akurasi, warna oranye menandakan interpretasi, warna hijau menandakan pengingat, dan warna biru menandakan skor f1. Hasil terperinci ditunjukkan pada Gambar 13.



Gambar 13 Performa *Metrics Kernel* Rasio Data 1

Pada proses looping semua kernel terhadap data 1, didapatkan akurasi tertinggi dan terendah untuk setiap kernel. Berdasarkan data pada gambar diatas kernel Linear mempunyai hasil akurasi tertinggi yaitu 86.98%, sedangkan kernel Polynomial mendapatkan akurasi yang terendah dengan 76.33%. Sementara untuk hasil presisi tertinggi ada pada kernel RBF dengan nilai yang sempurna 100%. Kernel Linear memperoleh recall 47.5% dan 63.33% nilai f1 tertinggi, lalu presisi, recall, dan nilai f1 yang terendah adalah kernel Polynomial.

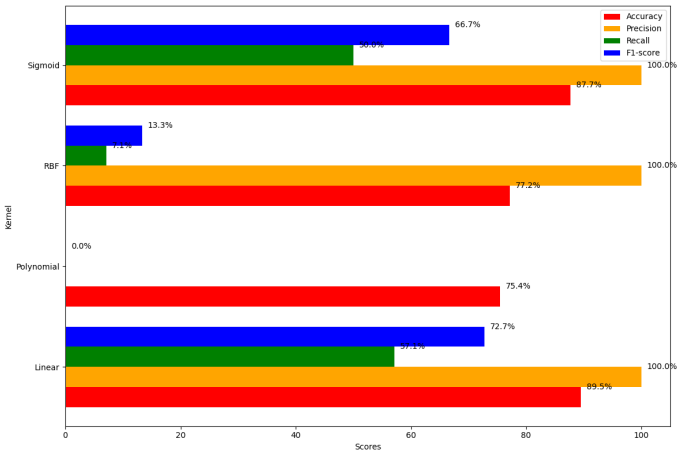
Hasil visualisasi metrik kinerja untuk setiap core berdasarkan rasio data 2 dengan persentasenya ditunjukkan di bawah ini. Pewarnaan biru menunjukkan akurasi, perwarnaan hijau menunjukkan presisi, perwarnaan merah menunjukkan recall, dan perwarnaan oranye menunjukkan nilai f1. Gambar 14 menunjukkan hasil yang lebih detail.



Gambar 14 Performa *Metrics Kernel* Rasio Data 2

Dalam proses looping semua kernel untuk data 2, diperoleh akurasi tertinggi dan terendah untuk setiap kernel. Berdasarkan data di atas, kernel Linear memiliki akurasi tertinggi sebesar 84.95%, sedangkan kernel Polynomial memiliki akurasi terendah sebesar 76.10%. Kernel Linear, RBF, Sigmoid memiliki nilai sempurna 100% untuk hasil presisi, tetapi kernel Linear juga memiliki nilai recall 37.03% dan nilai f1 54.05% tertinggi, kemudian untuk hasil recall, nilai f1 dan presisi yang terendah ada pada kernel Polynomial.

Di bawah ini adalah hasil visualisasi dari metrik kinerja untuk setiap core berdasarkan rasio data 3 dengan persentase yang ditampilkan. Kode warna biru mengindikasikan akurasi, hijau mengindikasikan presisi, merah mengindikasikan recall, dan oranye mengindikasikan nilai f1. Untuk hasil yang lebih spesifik dapat dilihat pada Gambar 15.



Gambar 15 Performa *Metrics Kernel* Rasio Data 3

Selama proses looping semua kernel dengan data 3, diperoleh akurasi tertinggi dan terendah untuk setiap kernel. Berdasarkan data pada Gambar 15, akurasi tertinggi sebesar 89,47% didapatkan kernel linear, sedangkan kernel polynomial memiliki akurasi terendah sebesar 75,43%. Kernel linier, RBF, dan Sigmoid masih sama dengan data diatas yang memiliki nilai sempurna 100% untuk hasil presisi, tetapi kernel linier juga memiliki nilai recall tertinggi sebesar 57,14% dan nilai f1 tertinggi sebesar 72,72%. Selanjutnya kernel polynomial memperoleh nilai presisi, recall, dan f1-skor terendah.

Penelitian ini menggunakan metode SVM untuk mengklasifikasikan kalimat perundungan. Didapatkan hasil yang menunjukkan bahwa kernel Linear memberikan akurasi tertinggi sebesar 89,47% dalam mengklasifikasikan kalimat perundungan berdasarkan aspek fisik dan non-fisik pada dataset yang digunakan. Penemuan ini sejalan dengan penelitian sebelumnya yang juga menggunakan SVM untuk mengklasifikasikan kalimat perundungan. Penelitian yang pernah dilakukan [6] menggunakan SVM dengan kernel RBF untuk menganalisis komentar perundungan pada Facebook yang menemukan bahwa dibandingkan dengan kernel Linear, polynomial atau sigmoid, kernel RBF lebih akurat.

Namun terdapat perbedaan dalam nilai akurasi yang diperoleh penelitian sebelumnya dengan penelitian yang sedang dilakukan. Penelitian ini [6] mendapatkan akurasi tertinggi sebesar 75%. Sementara penelitian lainnya [9] menunjukkan akurasi sebesar 81,91%. Perbedaan ini dapat disebabkan oleh beberapa faktor, seperti teknik preprocessing dan ekstraksi fitur, parameter SVM, serta perbedaan dalam dataset yang digunakan. Secara keseluruhan, penelitian ini mendukung penelitian sebelumnya bahwa SVM menjadi metode yang efektif untuk klasifikasi kalimat perundungan. Penelitian juga memberikan kontribusi baru dengan penggunaan dataset yang lebih besar.

IV. KESIMPULAN

Secara keseluruhan, penelitian ini telah membuat model yang mampu dalam mengenali kalimat-kalimat bullying berdasarkan teks dengan menggunakan SVM. Hasil dari implementasi TF-IDF didapatkan 428 data aspek fisik dengan persentase 75.4%, 132 data aspek non-fisik dengan persentase 23.2% dan 8 data aspek yang tidak dapat teridentifikasi dengan persentase 1.4%. Kinerja terbaik didapat oleh kernel Linear dengan menggunakan pembagian data rasio 90:10, dimana parameter mendapatkan akurasi sebesar 89.47%, presisi sebesar 100%, recall sebesar 57.14%, dan f1-score sebesar 72.72%. Untuk penelitian selanjutnya, dapat ditambahkan lebih banyak informasi dan fitur, dapat ditambahkan komentar dalam bahasa Inggris, dan dapat menambahkan lebih dari satu aspek.

DAFTAR PUSTAKA

[1] D. Shabira *et al.*, "MIND (Multimedia Artificial Intelligent Networking Database Deteksi Seksisme Online menggunakan Support Vector Machine dan Naïve Bayes," *J. MIND J. | ISSN*, vol. 8, no. 2, pp. 254–266, 2023, [Online]. Available: https://doi.org/10.26760/mindjournal.v8i2.254-266

[2] J. Pardede, "Deteksi Komentar Cyberbullying Pada Media Sosial Berbahasa Inggris Menggunakan Naïve Bayes Classification," *J. Inform.*, vol. 7, no. 1, pp. 46–54, 2020, doi: 10.31311/ji.v7i1.6920.

[3] W. Athira Luqyana, I. Cholissodin, and R. S. Perdana, "Analisis Sentimen Cyberbullying pada Komentar Instagram dengan Metode Klasifikasi Support Vector Machine," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 11, pp. 4704–4713, 2018, [Online]. Available: http://j-ptiik.ub.ac.id

[4] R. Indonesia, "Undang-Undang Republik Indonesia Nomor 19 Tahun 2016 Tentang Perubahan Atas Undang-Undang Nomor 11 Tahun 2008 Tentang Informasi Dan Transaksi Elektronik," *UU No. 19 tahun 2016*, no. 1, pp. 1–31, 2016.

[5] N. W. Sri wahyuni, "Penerapan Metode Klasifikasi Support Vector Machine (Svm) Untuk Menentukan Karyawan Putus Kontrak Pada Pt. Tae Hang Indonesia," *J. Inform. SIMANTIK*, vol. 4, no. September, pp. 10–15, 2019.

[6] R. M. Kamal and E. Rainarli, "Analisis Sentimen Cyberbullying Pada Komentar Facebook Dengan Metode Klasifikasi Support Vector Machine," *Univ. Komput. Indones.*, 2019.

[7] L. Afina, H. Raudhoti, A. Herdiani, and D. A. Romadhony, "Identifikasi Cyberbullying pada Kolom Komentar Instagram dengan Metode Support Vector Machine dan Semantic Similarity (Cyberbullying Identification on Instagram Comment Using Support Vector Machine and Semantic Similarity)," *J-Cosine*, vol. 4, no. 1, pp. 1–8, 2020, [Online]. Available: http://jcosine.if.unram.ac.id/

[8] M. . Imelda A.Muis & Muhammad Affandes, "Penerapan Metode Support Vector Machine (SVM) Menggunakan Kernel Radial Basis Function (RBF) Pada Klasifikasi Tweet," *Sains, Teknol. dan Ind. Sultan Syarif Kasim Riau*, vol. 12, no. 2, pp. 189–197, 2015.

[9] H. N. Irmanda and Ria Astriratma, "Klasifikasi Jenis Pantun Dengan Metode Support Vector Machines (SVM)," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 4, no. 5, pp. 915–922, 2020, doi: 10.29207/resti.v4i5.2313.

[10] N. M. G. D. Purnamasari, M. A. Fauzi, Indriarti, and L. S. Dewi, "Identifikasi Tweet Cyberbullying pada Aplikasi Twitter menggunakan Metode Support Vector Machine (SVM) dan Information Gain (IG) sebagai Seleksi Fitur," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 11, pp. 5326–5332, 2018.

[11] A. C. Sitepu, W. Wanayumini, and Z. Situmorang, "Analisis Kinerja Support Vector Machine dalam Mengidentifikasi Komentar Perundungan pada Jejaring Sosial," *J. Media Inform. Budidarma*, vol. 5, no. 2, p. 475, 2021, doi: 10.30865/mib.v5i2.2923.

[12] H. P. P. Zuriel and A. Fahrurozi, "Implementasi Algoritma Klasifikasi Support Vector Machine Untuk Analisa Sentimen Pengguna Twitter Terhadap Kebijakan Psbb," *J. Ilm. Inform. Komput.*, vol. 26, no. 2, pp. 149–162, 2021, doi: 10.35760/ik.2021.v26i2.4289.

[13] T. M. Dista and F. F. Abdulloh, "Clustering Pengunjung Mall Menggunakan Metode K-Means dan Particle Swarm Optimization," *J. Media Inform. Budidarma*, vol. 6, no. 3, p. 1339, 2022, doi: 10.30865/mib.v6i3.4172.

[14] D. Deb and V. E. Balas, "Intelligent Systems Reference Library 153," vol. 72, 2019.

[15] G. Subrata, "Klasifikasi Bahan Pustaka," *Pustak. Perpust.*, vol. 1, no. Ddc, pp. 1–13, 2019.

[16] D. Rizfinanda, S. Ningrum, R. A. Yaksa, and U. Jambi, "Identifikasi Perilaku Bullying Verbal Dalam Hubungan Pertemanan Di Desa Simpang Terusan Kabupaten Batang Hari," vol. 3, pp. 10330–10343, 2023.

[17] W. Agastya and Aripin, "Pemetaan Emosi Dominan pada Kalimat Majemuk Bahasa Indonesia Menggunakan Multinomial Naïve Bayes," *J. Nas. Tek. Elektro dan Teknol. Inf.*, vol. 9, no. 2, pp. 171–179, 2020, doi: 10.22146/jnteti.v9i2.157.

[18] R. M. Candra and A. Nanda Rozana, "Klasifikasi Komentar Bullying pada Instagram Menggunakan Metode K-Nearest Neighbor," *IT J. Res. Dev.*, vol. 5, no. 1, pp. 45–52, 2020, doi: 10.25299/itjrd.2020.vol5(1).4962.

[19] R. Dhian Syarif, A. Herdiani, W. Astuti, and M. Kom, "Identifikasi Cyberbullying pada Komentar Instagram menggunakan Metode Lexicon-Based dan Naïve Bayes Classifier (Studi kasus: Pemilihan Presiden Indonesia Tahun 2019)," *e-Proceeding Eng.*, vol. 6, no. 2, p. 8838, 2019.

[20] R. VINDO, *Named Entity Recognition (Ner) Bahasa Indonesia Berbasis Multi Class Classification*. 2023. [Online]. Available: http://digilib.unila.ac.id/70458/%0Ahttp://digilib.unila.ac.id/70458/3/SKRIPSI TANPA BAB PEMBAHASAN.pdf

[21] E. Agisyaputri, N. A. Nadhirah, and I. Saripah, "Identifikasi fenomena perilaku bullying pada remaja," *J. Bimbing. dan Konseling*, vol. 3, pp. 19–30, 2023.

[22] N. A. Sinaga, A. Setiawan, and A. Noertjahyana, "Sistem Deteksi Reputasi Akun Seller Pada Steam Community Menggunakan Metode Klasifikasi Support Vector Machine," *J. Infra*, 2022, [Online]. Available: https://publication.petra.ac.id/index.php/teknik-informatika/article/view/12820%0Ahttps://publication.petra.ac.id/index.php/teknik-informatika/article/download/12820/11120

[23] R. P. I. Putra, M. Akbar, and R. Amalia, "Analisis Sentimen Masyarakat Terhadap Kinerja Persatuan Sepakbola Seluruh Indonesia Menggunakan Metode Backpropagation," *J. Inf. Technol. Ampera*, vol. 1, no. 2, pp. 106–118, 2020, doi: 10.51519/journalita.volume1.isssue2.year2020.page106-118.

[24] S. S. Berutu, H. Budiati, J. Jatmika, and F. Gulo, "Data preprocessing approach for machine learning-based sentiment classification," *J. Infotel*, vol. 15, no. 4, pp. 317–325, 2023, doi: 10.20895/infotel.v15i4.1030.

[25] Z. Adhari, F. Informatika, and U. Telkom, "Identifikasi Ujaran Kebencian pada Twitter Menggunakan Metode Convolutional Neural Network (CNN)," vol. 10, no. 3, pp. 3464–3474, 2023.