Journal of Informatics Engineering and Software Applications

Cyberbullying Detection Model Using IndoBERT Representation and Support Vector Machine

Alda Mauliza[1], Mahdi[2]*, Musta'inul Abdi[3]

[123] Jurusan Teknologi Informasi dan Komputer, Politeknik Negeri Lhokseumawe, Kota Lhokseumawe, 24301 INDONESIA
*Penulis Korespondensi : mahdi@pnl.ac.id

INFORMASI ARTIKEL

Riwayat artikel:
Diajukan pada 10 November 2025
Direvisi pada 26 November 2025
Publikasi pada 20 Desember 2025

Kata kunci:
Perundungan
Psikologis
Support Vector Machine
Confusion Matrix
Machine Learning

Keywords:
Bullying
Psychological
Support Vector Machine
Confusion Matrix
Machine Learning

ABSTRAK

Peningkatan penggunaan media sosial telah menyebabkan banyak individu belum menyadari bahwa komentar atau ulasan yang mereka sampaikan dapat dikategorikan sebagai tindakan *cyberbullying*. *Cyberbullying* merupakan bentuk perundungan yang umum terjadi di ruang digital dan dapat menimbulkan dampak psikologis serius bagi korban, seperti depresi, gangguan tidur, serta penurunan produktivitas kerja. Penelitian ini mengembangkan model machine learning untuk mengklasifikasikan teks sebagai *bullying* atau *non-bullying*, dengan memanfaatkan dataset Hugging Face. Model ini dibangun melalui analisis teks menggunakan representasi IndoBERT serta metode klasifikasi Support Vector Machine. Evaluasi kinerja model dilakukan menggunakan Confusion Matrix yang menghasilkan tingkat akurasi sebesar 89,59%. Hasil tersebut menunjukkan bahwa kombinasi IndoBERT dan Support Vector Machine merupakan pendekatan yang efektif dalam mendeteksi *cyberbullying*.

ABSTRACT

The increasing use of social media has led many individuals to remain unaware that their comments or reviews may be categorized as cyberbullying. Cyberbullying is a common form of harassment in digital spaces and can cause serious psychological impacts on victims, such as depression, sleep disturbances, and reduced work productivity. This study develops a machine learning model to classify text as bullying or non-bullying, utilizing a dataset from Hugging Face. The model is built through text analysis using IndoBERT representation and the Support Vector Machine classification method. Model performance evaluation was conducted using a Confusion Matrix, resulting in an accuracy rate of 89.59%. These results indicate that the combination of IndoBERT and Support Vector Machine is an effective approach for detecting cyberbullying.

1. Pendahuluan

Perundungan atau *bullying*, merupakan perilaku agresif yang dilakukan oleh individu maupun kelompok terhadap seseorang yang dianggap memiliki posisi lebih lemah. Pelaku, yang merasa memiliki kekuasaan atau keunggulan tertentu, cenderung melakukan tindakan yang tidak sesuai dengan norma sosial dan etika yang berlaku. Perilaku ini bersifat berulang dan berlangsung dalam jangka waktu tertentu, dengan tujuan untuk melemahkan kondisi fisik maupun mental korban hingga menimbulkan rasa tidak berdaya [1].

Perundungan yang berlangsung melalui jaringan internet dikenal sebagai *cyberbullying* atau pelecehan siber. Saat ini, terdapat berbagai bentuk *cyberbullying* yang dapat dikenali. Contohnya termasuk penulisan konten teks yang bersifat tidak pantas serta penyebaran konten visual yang tidak sesuai, seperti meme yang mengandung unsur penghinaan atau ejekan [2]. Di era digital saat ini, penggunaan kata-kata sering kali tidak terkontrol, terutama ketika media sosial tidak dimanfaatkan dengan bijak. Selain itu, penghinaan di dunia maya menjadi lebih mudah dilakukan dengan memanfaatkan identitas palse, sehingga pelaku sulit dilacak [3]. *Cyberbullying* berpotensi menimbulkan berbagai gangguan kesehatan mental, termasuk depresi, kecemasan, tekanan emosional, hingga gejala stres pascatrauma. Dampak dari insiden ini kerap meninggalkan luka psikologis yang mendalam bagi korban. Mereka dapat mengalami gangguan tidur seperti insomnia, penurunan produktivitas kerja, serta menunjukkan perilaku yang tidak adaptif di lingkungan profesional [4].

Berdasarkan latar belakang yang telah dipaparkan, dapat disimpulkan bahwa permasalahan *cyberbullying* merupakan isu yang sangat serius. Perkembangan teknologi digital telah mempermudah pelaku melakukan tindakan *bullying* secara daring dengan menggunakan identitas palsu, sehingga menyulitkan proses identifikasi dan penegakan hukum. Selain itu, *cyberbullying* juga berpotensi menimbulkan gangguan mental seperti depresi, kecemasan, dan stres berkepanjangan, yang berdampak negatif terhadap kualitas hidup serta produktivitas korban.

Langkah yang dipilih untuk menyelesaikan permasalahan tindak *cyberbullying* adalah dengan membangun sebuah model untuk mendeteksi *cyberbullying*. Model ini dapat digunakan sebagai filtering komentar pada media sosial dan diharapkan *cyberbullying* dapat di deteksi sedini mungkin. Penelitian sebelumnya [5] mengembangkan sistem klasifikasi komentar berbasis analisis sentimen untuk mendeteksi indikasi *cyberbullying*. Metode yang digunakan meliputi TF-IDF sebagai teknik ekstraksi fitur dan Support Vector Machine sebagai algoritma klasifikasi. Hasil evaluasi menunjukkan performa sistem yang tinggi, dengan akurasi sebesar 93%, precision 95%, dan recall 97% yang mencerminkan efektivitas pendekatan tersebut dalam mengidentifikasi potensi *cyberbullying* melalui analisis teks. Selanjutnya pada penelitian [6] mengembangkan sistem deteksi *cyberbullying* berbasis pendekatan deep learning dengan menggabungkan model BERT dan Bi-LSTM. Kombinasi kedua model ini memungkinkan pemahaman konteks bahasa serta identifikasi pola perundungan dalam teks berbahasa Indonesia. Evaluasi sistem menunjukkan akurasi sebesar 90%, yang menandakan efektivitas metode tersebut dalam mendeteksi berbagai bentuk *cyberbullying* secara otomatis.

Berdasarkan tinjauan pustaka yang dipaparkan, solusi yang ditawarkan pada penelitian ini berupa membangun model deteksi *cyberbullying* menggunakan Support Vector Machine sebagai algoritma utama dan IndoBERT digunakan sebagai model ektraksi fitur.

Support Vector Machine (SVM) merupakan metode suppervised classifier yang terbukti sangat efektif dalam menyelesaikan berbagai permasalahan pengenalan pola dan visi komputer. Pelatihan SVM dilakukan dengan menentukan hyperplane yang memisahkan data pelatihan dari dua class. Posisi hyperplane ini ditentukan oleh sejumlah kecil vektor dari data pelatihan, yang disebut sebagai support vector [7]. Jika suatu data dapat dipisahkan secara sempurna menggunakan hyperplane linear, maka metode SVM yang digunakan disebut sebagai SVM linier. Sementara itu, fungsi kernel berperan dalam

mentransformasikan data ke ruang berdimensi lebih tinggi guna memperjelas pola struktur data, sehingga proses pemisahan antar class menjadi lebih efektif [8]. Fungsi ini bersifat sederhana karena tidak melakukan transformasi non-linier terhadap data, melainkan hanya memetakan data ke dalam ruang fitur yang sama. Kernel linier ideal digunakan ketika data dapat dipisahkan secara linier dalam dimensi fitur aslinya. Walaupun tergolong sederhana, kernel ini tetap menunjukkan efektivitas tinggi dalam berbagai kasus klasifikasi data yang bersifat linier [9].

Model word embedding yang telah dilatih sebelumnya dikenal sebagai pre-trained word embedding yang dibangun dari korpus besar agar mampu memahami makna dan sintaksis bahasa secara lebih baik. Pada tahun 2018, diperkenalkan BERT (Bidirectional Encoder Representations from Transformer) yang menunjukkan performa unggul dalam berbagai studi NLP. BERT menggunakan arsitektur Transformer dengan mekanisme self-attention untuk mempelajari konteks antar kata dalam teks. Khusus Bahasa Indonesia, versi pre-trained BERT yang disebut IndoBERT dikembangkan pada tahun 2020 [10]. Model ini dilatih menggunakan dataset besar bernama Indo4B, yang terdiri dari sekitar 4 miliar kata yang dikumpulkan dari berbagai sumber publik seperti media sosial, blog, dan portal berita [11]. IndoBERT memiliki 12 lapisan tersembunyi dan telah dilatih menggunakan kata-kata dari Wikipedia Bahasa Indonesia (74 juta), artikel berita dari Kompas, Tempo, dan Liputan6 (55 juta), serta korpus web Bahasa Indonesia (90 juta) [12].

Penelitian ini menawarkan pendekatan baru dalam deteksi *cyberbullying* dengan menggabungkan model IndoBERT sebagai ekstraksi fitur dan algoritma Support Vector Machine (SVM) sebagai klasifikator. Penggunaan IndoBERT memungkinkan model memahami konteks bahasa Indonesia secara lebih mendalam, sementara SVM memberikan efisiensi dalam klasifikasi biner. Pendekatan ini melanjutkan tren penelitian sebelumnya yang telah menggunakan TF-IDF dan deep learning, namun dengan kombinasi yang lebih adaptif terhadap karakteristik lokal. Model yang dibangun diharapkan mampu meningkatkan akurasi dan efektivitas dalam menyaring komentar bermuatan perundungan di media sosial.

2. Metode

Bagian ini menjelaskan secara rinci mengenai rancangan penelitian yang dilakukan. Meliputi data, *flowchart* dan juga pengujian model.

2.1 Data dan Pengumpulan Data

Dataset yang digunakan dalam penelitian ini berupa komentar Instagram yang diperoleh dari platform Hugging Face. Hugging Face sendiri merupakan platform open-source yang menyediakan beragam liblary dan tools untuk mendukung pengembangan serta penerapan model NLP tingkat lanjut [13]. Dataset yang digunakan memiliki 1103 data dengan 2 buah label, yaitu label positif mewakili *non-bullying* dan label negatif mewakili *bullying*. Pada Tabel 1 merupakan sampel dari dataset *cyberbullying*, label positif pada dataset diubah menjadi label *Non-bullying* Ún label negatif diubah menjadi *bullying*. Jumlah data berlabel *bullying* sebanyak 627 data dan data berlabel *non-bullying* sebanyak 476 data. Nantinya dataset ini dibagi menjadi 2, yaitu 80% data latih dan 20% data uji.

Tabel 1. Sample Dataset *Cyberbullying*

No	Teks	Label
1	JUAL GROSIR MAKANAN ANJING MAKANAN KUCING UNTUK PETSHOP	*Non-bullying*
2	Cokelat merupakan racun bagi anjing dan kucing.	*Non-bullying*
3	sok geulis anjing	*Bullying*
4	MUKAA LO SAMA DIA PASSS SAMA SMAA BULUKKK JD JADIANN DEH SONO ANJING SAMA BABIII	*Bullying*

2.2 *Preprocessing* Data

Preprocessing data merupakan serangkaian teknik yang bertujuan untuk membersihkan data mentah sebelum dilakukan analisis lebih lanjut. Dengan melakukan *preprocessing* data, hasil analisis menjadi lebih valid dan andal, serta memungkinkan model bekerja secara optimal [14]. *Preprocessing* data terdiri dari beberapa tahapan diantara lain:

1. Case folding

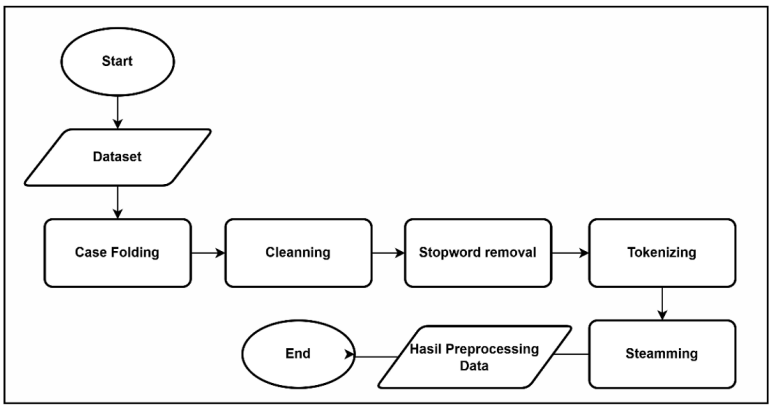
   Case folding adalah tahapan untuk menyamaratakan seluruh data menjadi huruf kecil (lowercase) ataupun huruf besar (uppercase) [15]
2. Cleanning

   Cleanning merupakan tahapan untuk membersihkan data dari karakter-karakter yang tidak diperlukan, seperti tanda baca, simbol ataupun karakter lainnya [15].
3. Tokenizing

   Tokenizing merupakan tahapan membagi teks menjadi token kata, pembagian ini bertujuan untuk mempermudah dilakukannya analisis data [16].
4. Stopword removal

   Stopword removal adalah tahapan untuk menghapus kata-kata umum yang tidak memiliki makna signifikan dalam analisis teks. Seperti konjungsi ataupun kata bantu [16].
5. Steamming

   Steamming mengubah kata imbuhan menjadi kata dasarnya kembali, ini dilakukan untuk menjaga konsistensi agar mengurangi variasi kata [16].

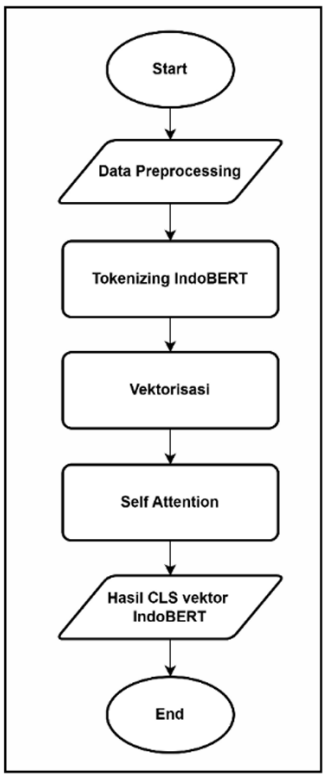


Gambar 2. *Flowchart Preprocessing* Data

Gambar 2 menggambarkan tahapan *preprocessing* data sebagai langkah awal sebelum pelatihan model. Proses dimulai dari pengambilan dataset teks mentah, kemudian dilakukan case folding untuk menyeragamkan huruf, cleaning untuk menghapus karakter yang tidak relevan, stopword removal agar kata bersifat umum yang tidak penting dihilangkan, tokenizing untuk memecah teks menjadi kata, serta stemming berfungsi untuk mengubah kata ke bentuk dasarnya. Setelah tahap ini selesai, data siap digunakan pada proses berikutnya.

2.3 Embedding IndoBERT

IndoBERT memanfaatkan tokenizer WordPiece dengan jumlah kosakata 31.923 token, sehingga mampu mengakomodasi ciri khas bahasa Indonesia seperti kata majemuk, reduplikasi serta slang. Dengan arsitektur 12 lapisan transformer dan representasi vektor berdimensi 768, IndoBERT dapat menghasilkan representasi kata yang kontekstual, di mana makna kata disesuaikan dengan konteks kalimat [17]. Proses representasi input pada BERT, termasuk IndoBERT diawali dengan memecah teks menjadi token. Setiap token kemudian dipetakan ke dalam vektor numerik melalui tiga embedding, token embedding (arti kata), position embedding (urutan kata), dan segment embedding (pembeda segmen). Ketiganya digabung menjadi satu vektor berdimensi tetap. Selain itu, token khusus [CLS] digunakan sebagai representasi keseluruhan kalimat untuk klasifikasi, sedangkan [SEP] berfungsi sebagai pemisah antar segmen [18].



Gambar 3. *Flowchart* IndoBERT

Pada gambar 3 menunjukkan *flowchart* IndoBERT yang menggambarkan alur representasi teks. Proses dimulai dari data hasil *preprocessing*, kemudian melalui tahap tokenizing dengan menambahkan token khusus [CLS] dan [SEP]. Setiap token selanjutnya diubah menjadi vektor numerik melalui embedding

IndoBERT dan diproses oleh mekanisme self-attention untuk menangkap konteks antar kata. Hasil akhirnya berupa vektor [CLS] yang merepresentasikan keseluruhan teks dan digunakan dalam klasifikasi SVM.

2.4 Metode Support Vector Machine

Pelatihan SVM dilakukan dengan menentukan hyperplane yang memisahkan data pelatihan dari dua class. Posisi hyperplane ini ditentukan oleh sejumlah kecil vektor dari data pelatihan, yang disebut sebagai support vector [7]. Fungsi kernel adalah mentransformasi data ke ruang fitur berdimensi tinggi. Transformasi ini membantu model menemukan hyperplane yang lebih optimal antara class yang sulit dipisahkan pada dimensi awal.[19]. Kernel linear dipilih pada penelitian ini karena karakteristik data memungkinkan pemisahan antar class dengan menggunakan hyperplane garis lurus.

Klasifikasi SVM linear kernal dapat digambarkan sebagai berikut. Misalkan ada m sampel pengamatan (set pelatihan), (xi, yi), i = 1, 2, ...,m di mana:

$$x^T_i = (xi1, \dots xid) \in R^d \quad (1)$$

Dimana x_i^T merepresentasikan fitur berdimensi-d dari sampel ke-i dan y ∈ {-1, +1} merupakan label class biner. Jika sampel xi berada di class positif, maka yi adalah +1, sedangkan jika berada pada class negatif maka yi adalah -1. Himpunan data latih tersebut dapat dipisahkan oleh sebuah hyperplane dengan persamaan $x^T xi + b = 0$, di mana w adalah vektor bobot dan b merupakan bias. Adapun persamaan hyperplane H_1 dan H_2 dapat dilihat pada

$$H_1: (w^T xi + b) = 1 \quad (2)$$

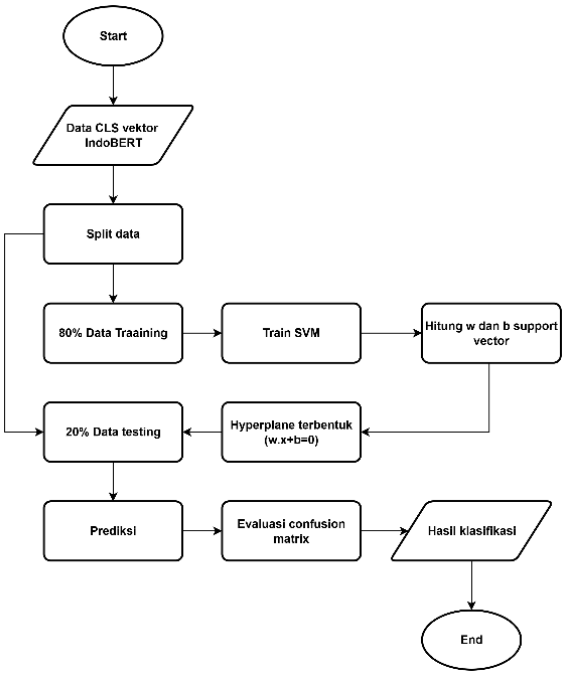
$$H_2: (w^T xi + b) = -1 \quad (3)$$

Sehingga , titik-titik yang diklasifikasikan dengan benar memenuhi pertidaksamaan.

$$y_i: (w^T xi + b) = \geq 1 \quad (4)$$

Untuk xi, i = 1,2,…,m. Jarak antara hyperplanes marginal yaitu sama dengan $\frac{2}{||w||}$ [20].

Pada Gambar 4 menggambarkan tahapan model klasifikasi teks algoritma Support Vector Machine (SVM). Proses dimulai dengan mengambil data vektor CLS IndoBERT yang sudah dihasilkan pada proses sebelumnya. Vektor ini merepresentasikan makna keseluruhan kalimat secara kontekstual. Selanjutnya, data tersebut dibagi menjadi dua bagian, yaitu 80% data latih dan 20% data uji. Data latih yang telah diseleksi kemudian digunakan untuk melatih model SVM. Pada tahap ini model SVM dilatih dari data latih untuk kemudian menghitung nilai w yang merupakan vektor bobot untuk menentukan arah atau orientasi dari hyperplane dan b yang merupakan nilai skalar yang menggeser hyperplane menjauh dari titik asal kedua nilai tersebut digunakan untuk mencari nilai pemisah agar SVM dapat melakukan klasifikasi.



Gambar 4. *Flowchart* Metode SVM

2.5 Pengujian Confusion Matrix

Confusion Matrix merupakan penyajian kinerja pengklasifikasi yang berguna dan komprehensif. Matrix ini biasanya digunakan dalam evaluasi model klasifikasi multiclass dan berlabel tunggal, di mana setiap instans data dapat menjadi bagian dari satu class pada titik waktu tertentu [21]. Keakuratan hasil diukur menggunakan nilai recall, precision, dan accuracy. Recall mengacu pada rasio identifikasi benar positif terhadap total data yang benar positif. Precision adalah rasio identifikasi benar positif dibandingkan dengan semua hasil identifikasi positif. Sementara itu accuracy adalah rasio identifikasi benar positif terhadap seluruh data yang ada [22].

Tabel 2. Pembagian Data Latih dan Data Uji

Data Latih	Data Uji
882	221

Dataset berjumlah 1103 terbagi ke dalam 2 label, yaitu *bullying* dan *non-bullying*. Jumlah data dari setiap labelnya adalah sebagai berikut.

Tabel 3. Data Per Label

No	Label	Jumlah Data
1	*Bullying*	627
2	*Non-bullying*	476

Pada Table 3 dapat dilihat bahwa pembagian jumlah data tidak seimbang, hal ini tentu akan mempengaruhi akurasi model nantinya. Namun penggabungan antara model IndoBERT yang sudah dilatih menggunakan jutaan kalimat serta metode klasifikasi SVM diharapkan dapat memberikan akurasi yang baik pada model meski jumlah data per label tidaklah seimbang. Evaluasi model menggunakan Confusion Matrix memetakan segala hasil dalam 4 bagian utama, yaitu sebagai berikut:

a. True Positive (TP), model memprediksi bahwa sampel adalah positif, dan kenyataannya memang positif.
b. False Positive (FP)), model memprediksi bahwa sampel adalah positif, padahal sebenarnya negatif.
c. True Negative (TN), model memprediksi bahwa sampel adalah negatif, dan kenyataannya memang negatif.
d. False Negative (FN), model memprediksi bahwa sampel adalah negatif, padahal sebenarnya positif [23].

3. Hasil Dan Pembahasan

Penelitian ini akan mennjelaskan secara sistematis prosedur serta tahapan yang dilalui, yaitu di awali dengan mengambil dataset, kemudian dilakukan *preprocessing* data, lalu data diubah menjadi vector IndoBERT sehingga dapat dilatih menggunakan metode SVM untuk membangun model dan terakhir dilakukan pengujian menggunakan Confusion Matrix.

3.1 Dataset

Penelitian ini menggunakan dataset berisi 1.103 teks berbahasa Indonesia yang diklasifikasikan ke dalam dua kategori: *bullying* dan *non-bullying*. Dataset diperoleh dari platform Hugging Face dan telah melalui proses normalisasi label dari positif-negatif menjadi dua class utama. Tujuan normalisasi ini adalah untuk memperjelas kategori dan meningkatkan akurasi model dalam mendeteksi *cyberbullying* secara otomatis.

no	teks	label
1	JUAL MAKANAN ANJING (DOG FC	non-bullying
2	JUAL GROSIR MAKANAN ANJING	non-bullying
3	Jangan mentang-mentang lu anak	non-bullying
4	Males itu kalo kerja pagi trus gak	bullying
5	Pagi ini cuma mau panggil anjing	bullying
6	Mau tanya, hukumya org Islam pe	non-bullying
7	Cuma karena anjing... okay	non-bullying
8	sok geulis anjing	bullying
9	Cokelat merupakan racun bagi an	non-bullying
10	anjing jg lo ler	bullying
…		
1098	monyet itu bisa membuka kaleng	non-bullying
1099	babi itu sangat kotor dan bau.	bullying
1100	anjing, hebat banget cara lu nyele	bullying
1101	anjing itu sangat loyal pada pemi	non-bullying
1102	monyet, lu selalu bikin ribet!	bullying
1103	gue liat monyet bermain di hutan	non-bullying

Gambar 5. Dataset

Pada Gambar 5 merupakan dataset yang telah melewati proses normalisasi pada label yang sebelumnya ialah positif dan negatif berganti menjadi *bullying* dan *non-bullying*. Proses ini bertujuan untuk memperjelas kategori data serta meningkatkan akurasi dalam tahap pelatihan dan evaluasi model deteksi *cyberbullying*.

3.2. Case Folding

Case folding merupakan tahap awal dalam *preprocessing* yang bertujuan untuk menyamakan bentuk huruf dengan mengubah seluruh karakter menjadi huruf kecil (lowercase).

teks	case_folding
JUAL MAKANAN ANJIN	jual makanan anjing (dog food)
JUAL GROSIR MAKANA	jual grosir makanan anjing mak
Jangan mentang-ment	jangan mentang-mentang lu an
Males itu kalo kerja pa	males itu kalo kerja pagi trus ga
Pagi ini cuma mau pan	pagi ini cuma mau panggil anjir
Mau tanya, hukumya o	mau tanya, hukumya org islam
Cuma karena anjing... c	cuma karena anjing... okay
sok geulis anjing	sok geulis anjing
Cokelat merupakan rac	cokelat merupakan racun bagi a
anjing jg lo ler	anjing jg lo ler
Wah aku memelihara a	wah aku memelihara anjing n k
So Sweet, Pemain Ini G	so sweet, pemain ini gendong a
Anak anjing jantan sen	anak anjing jantan sengaja mer
Sesama anjing putih be	sesama anjing putih berantem
kesel deh pagi2 anjing	kesel deh pagi2 anjing ku masa

Gambar 6. Hasil Case Folding

Pada Gambar 6 merupakan hasil Case folding yang bertujuan untuk menyamakan bentuk huruf agar analisis teks menjadi lebih konsisten. Pada tahap ini, seluruh karakter dalam teks yang semula terdiri dari huruf besar atau kapital diubah menjadi huruf kecil (lowercase). Proses ini penting untuk menghindari redundansi dalam pemrosesan kata dan meningkatkan akurasi dalam tahap tokenisasi serta klasifikasi.

3.3 Cleanning dan Stopwords Removal

Tahap kedua *preprocessing*, yaitu cleaning untuk menghapus karakter tidak relevan dan tahap ketiga adalah stopwords removal, yaitu penghapusan kata-kata umum.

cleaning	stopwords_removed
jual makanan anjing dog food ha	jual makanan anjing dog food ha
jual grosir makanan anjing makar	jual grosir makanan anjing maka
jangan mentangmentang lu anak	mentangmentang lu anak gaul pa
males itu kalo kerja pagi trus gak	males kalo kerja pagi trus gak yg
pagi ini cuma mau panggil anjing	pagi panggil anjing aja elo motor
mau tanya hukumya org islam pe	hukumya org islam pelihara anjir
cuma karena anjing okay	anjing okay
sok geulis anjing	sok geulis anjing
cokelat merupakan racun bagi an	cokelat racun anjing kucing
anjing jg lo ler	anjing jg lo ler
wah aku memelihara anjing n kuc	memelihara anjing n kucing nih
so sweet pemain ini gendong anji	so sweet pemain gendong anjing
anak anjing jantan sengaja memb	anak anjing jantan sengaja anak
sesama anjing putih berantem	anjing putih berantem

Gambar 7. Hasil Cleanning dan Stopwords Removal

Pada Gambar 7 merupakan hasil cleanning dan stopwords removal, yaitu tahapan penting dalam pra-pemrosesan data teks. Proses cleaning dilakukan untuk menghapus elemen-elemen yang tidak relevan

seperti URL, angka, simbol, dan tanda baca yang tidak diperlukan dalam analisis. Selanjutnya, dilakukan stopwords removal untuk menghilangkan kata-kata umum yang tidak memiliki makna signifikan dalam konteks klasifikasi, seperti “dan”, “yang”, atau “di”. Tahapan ini menggunakan pustaka NLTK yang menyediakan daftar stopwords dalam Bahasa Indonesia, sehingga membantu meningkatkan kualitas data yang akan digunakan dalam pelatihan model.

3.4 Tokenizing dan Steamming

Tahapan keempat dalam preprocessinh adalah tokenizing, yaitu proses memecah teks menjadi unit-unit kata atau token agar dapat dianalisis secara individual oleh model. Selanjutnya, tahap terakhir adalah stemming yang bertujuan untuk mengubah kata ke bentuk dasarnya.

tokens	stemmed
['jual', 'makanan', 'anjing', 'dog', 'food'	['jual', 'makan', 'anjing', 'dog', 'food'
['jual', 'grosir', 'makanan', 'anjing', 'ma	['jual', 'grosir', 'makan', 'anjing', 'ma
['mentangmentang', 'lu', 'anak', 'gaul',	['mentangmentang', 'lu', 'anak', 'gau
['males', 'kalo', 'kerja', 'pagi', 'trus', 'ga	['males', 'kalo', 'kerja', 'pagi', 'trus', '
['pagi', 'panggil', 'anjing', 'aja', 'elo', 'm	['pagi', 'panggil', 'anjing', 'aja', 'elo',
['hukumya', 'org', 'islam', 'pelihara', 'ar	['hukumya', 'org', 'islam', 'pelihara',
['anjing', 'okay']	['anjing', 'okay']
['sok', 'geulis', 'anjing']	['sok', 'geulis', 'anjing']
['cokelat', 'racun', 'anjing', 'kucing']	['cokelat', 'racun', 'anjing', 'kucing']
['anjing', 'jg', 'lo', 'ler']	['anjing', 'jg', 'lo', 'ler']
['memelihara', 'anjing', 'n', 'kucing', 'ni	['pelihara', 'anjing', 'n', 'kucing', 'nih
['so', 'sweet', 'pemain', 'gendong', 'anji	['so', 'sweet', 'main', 'gendong', 'anji

Gambar 8. Hasil Tokenizing dan Steamming

Pada Gambar 8 menampilkan hasil dari tahapan tokenizing dan stemming yang merupakan bagian penting dalam proses *preprocessing*. Tokenizing berfungsi memecah kalimat menjadi unit kata (tokens) sehingga dapat dianalisis secara terpisah. Selanjutnya, stemming dilakukan untuk mengembalikan kata ke bentuk dasarnya guna menjaga konsistensi serta mengurangi variasi kata dengan makna serupa. Pada penelitian ini, proses stemming memanfaatkan liblary Sastrawi, yaitu library natural language processing yang dikembangkan khusus untuk Bahasa Indonesia.

3.5 Implementasi Embedding IndoBERT

Tahapan selanjutnya adalah proses embedding menggunakan model IndoBERT yang bertujuan untuk mengubah teks menjadi vector numerik yang dapat di proses oleh metode Support Vector Machine. Melalui Gambar 9 memperlihatkan hasil embedding dari model IndoBERT, di mana setiap token dalam teks direpresentasikan sebagai vektor berdimensi 768. Representasi ini tidak hanya memuat informasi dari token, tetapi juga siap diproses lebih lanjut untuk menangkap relasi semantik. Selanjutnya, vektor-vektor tersebut diproses melalui mekanisme self-attention yang memungkinkan setiap token mempertimbangkan hubungan dengan token lain di sekitarnya, baik sebelum maupun sesudahnya. Sehingga makna kata dapat dipahami sesuai konteks kalimat secara menyeluruh. Proses ini menghasilkan representasi kontekstual yang lebih kaya. Hasil akhirnya dirangkum ke dalam token khusus [CLS] yang juga berupa vektor 768 dimensi.

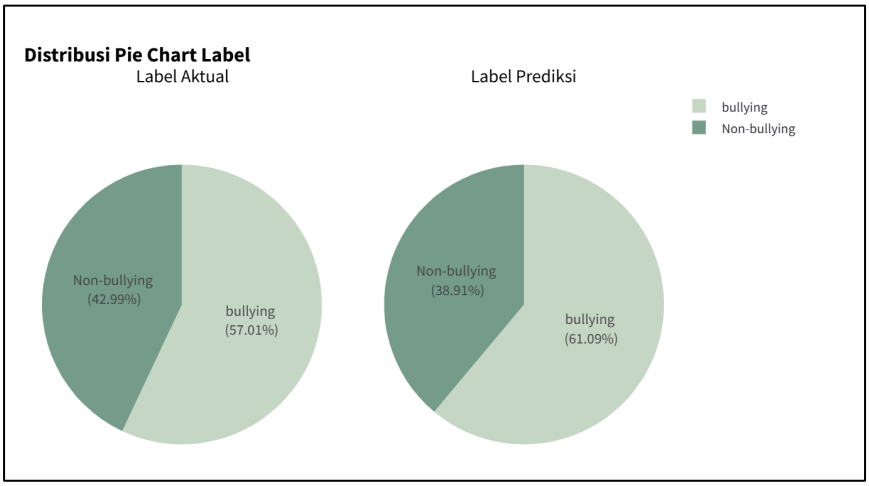
Token [CLS] ini berfungsi sebagai representasi teks atau kalimat, sehingga dapat digunakan sebagai input utama pada tahap klasifikasi.

id	0	1	2	3	4	5
10501	0.686967	1.322128	0.827438	0.534569	1.803968	-1.68781
10502	0.38328	1.189288	0.57313	0.921233	1.355692	-1.13503
10503	1.70402	1.810656	-0.36822	0.801798	1.721097	-0.84086
10504	0.634085	1.83912	-0.20012	-0.33991	1.418288	-0.88102
10505	0.956517	1.517083	0.004919	0.651136	2.327857	-0.92179
10506	0.681802	1.510966	0.003994	0.24837	2.043717	-1.618
10507	0.135785	1.125732	0.820886	1.338695	1.881329	-1.64051
10508	0.003503	0.578484	-0.2614	1.265945	1.300164	0.782119
10509	0.409322	1.486792	0.007486	0.069273	2.13176	-2.00312
10510	0.768888	1.339789	0.281932	0.699499	1.68627	-1.2496
10511	0.652298	1.514297	0.60481	0.89891	1.42965	-1.47952
10512	0.304687	2.12805	0.025222	0.041176	2.053119	-1.49289
10513	1.143246	1.852064	0.339957	0.56184	1.664197	-1.78802
10514	0.664203	1.317115	0.34986	0.582213	1.537315	-1.77733

Gambar 9. Hasil Representasi IndoBERT

3.6 Implementasi Metode Support Vector Machine

Pada tahap ini, dilakukan proses pelatihan dan klasifikasi menggunakan metode SVM. Model dilatih menggunakan 80% data latih dari total dataset yang telah melalui tahapan *preprocessing* dan embedding IndoBERT. Sisa 20% data digunakan sebagai data uji untuk mengevaluasi performa model dalam mengklasifikasikan teks.



Gambar 10. Pie Chart Klasifikasi

Pada Gambar 10 menampilkan dua pie chart yang membandingkan distribusi label aktual dan label prediksi dalam klasifikasi *cyberbullying*. Pie chart di sebelah kiri menunjukkan 'Label Aktual' di mana kategori *bullying* mencakup 57,01% dari data, sedangkan *Non-bullying* sebesar 42,99%. Sementara itu, pie chart di sebelah kanan menggambarkan 'Label Prediksi' dari model yang dibangun, dengan proporsi *bullying* sebesar 61,09% dan *Non-bullying* sebesar 38,91%.

Perbedaan distribusi antara label aktual dan prediksi ini memberikan gambaran awal mengenai kecenderungan model dalam mengklasifikasikan data. Terlihat bahwa model cenderung sedikit overpredict pada kategori *bullying* yang dapat berdampak pada nilai False Positive dalam evaluasi performa.

3.7 Evaluasi Pengujian Confusion Matrix

Tahapan ini merupakan evaluasi performa model menggunakan confusion matrix. Pengujian dilakukan terhadap data uji sebanyak 20% dari total dataset dengan tujuan untuk mengukur kemampuan model dalam membedakan teks *bullying* dan *non-bullying*.

```
Confusion matrix:
 [[119   7]
 [ 16  79]]
Classification report:
              precision    recall  f1-score   support

    bullying     0.8815    0.9444    0.9119       126
Non-bullying     0.9186    0.8316    0.8729        95

    accuracy                         0.8959       221
   macro avg     0.9000    0.8880    0.8924       221
weighted avg     0.8974    0.8959    0.8951       221
```

Gambar 11. Hasil Pengujian Confusion Matrix

Berdasarkan classification report, model menunjukkan performa yang cukup baik dalam membedakan teks *bullying* dan *non-bullying*. Nilai precision dan recall pada kedua class relatif tinggi, dengan F1-score yang seimbang, menandakan bahwa model mampu menjaga trade-off antara ketepatan dan kelengkapan prediksi.

a. Precision *bullying* (88.15%) menunjukkan bahwa sebagian besar teks yang diprediksi sebagai *bullying* memang benar mengandung unsur *bullying*. Begitupun sebaliknya pada precision *non-bullying* (91.86%) yang memperlihatkan bahwa teks yang diprediksi sebegai *non-bullying* dan benar memiliki class *non-bullying* juga tinggi.
b. Recall *bullying* (94.44%) menandakan bahwa model berhasil menangkap sebagian besar teks *bullying* yang sebenarnya ada dalam data.
c. Sebaliknya, recall *non-bullying* (83.16%) sedikit lebih rendah, yang berarti masih terdapat teks netral yang salah diklasifikasikan sebagai *bullying* (false positive).

Nilai akurasi keseluruhan sebesar 89.59% menunjukkan bahwa model cukup andal dalam melakukan klasifikasi, namun masih ada ruang untuk perbaikan, terutama dalam mengurangi kesalahan klasifikasi pada class *non-bullying*.

4. Kesimpulan

Setelah melalui tahapan perancangan, implementasi, dan pengujian model deteksi *cyberbullying*, diperoleh hasil evaluasi dari kombinasi representasi IndoBERT dan metode Support Vector Machine (SVM) dengan tingkat akurasi sebesar 89,59%. Model dilatih dan diuji menggunakan total 1103 data teks berbahasa Indonesia yang telah diberi label *bullying* dan *non-bullying*. Berdasarkan classification report, pada class *bullying* diperoleh nilai precision sebesar 88.15%, recall 94.44%, dan f1-score sebesar 91.19%. Sementara itu, pada class *non-bullying*, model menghasilkan precision sebesar 91.86%, recall 83.16%, dan f1-score sebesar 87.29%.

Hasil tersebut menunjukkan bahwa model mampu melakukan klasifikasi teks dengan baik pada kedua class. Penggabungan antara IndoBERT yang telah dilatih sebelumnya untuk memahami konteks

bahasa Indonesia dengan metode SVM memungkinkan model untuk bekerja secara efektif dengan mempertimbangkan hubungan antar kata dalam kalimat. Pendekatan ini memungkinkan model untuk mengenali fitur tekstual yang relevan dalam mendeteksi komentar yang mengandung unsur *bullying* maupun *non-bullying*.

Referensi

[1] A. Diannita, F. Salsabela, L. Wijiati, and A. M. S. Putri, "Pengaruh *Bullying* terhadap Pelajar pada Tingkat Sekolah Menengah Pertama," *J. Educ. Res.*, vol. 4, no. 1, pp. 297–301, 2023, doi: 10.37985/jer.v4i1.117.

[2] M. T. Hasan, M. A. E. Hossain, M. S. H. Mukta, A. Akter, M. Ahmed, and S. Islam, "A Review on Deep-Learning-Based *Cyberbullying* Detection," *Futur. Internet*, vol. 15, no. 5, pp. 1–47, 2023, doi: 10.3390/fi15050179.

[3] M. D. Ihkam and I. G. N. Parwata, "Tindak Pidana Cyber *Bullying* Dalam Perspektif Huku Pidaa di Indonesia," *J. Kertha Wicara*, vol. 9, no. 11, pp. 1–10, 2016.

[4] S. Bansal, N. Garg, J. Singh, and F. Van Der Walt, "*Cyberbullying* and mental health: past, present and future," *Front. Psychol.*, 2023, doi: 10.3389/fpsyg.2023.1279234.

[5] W. A. Prabowo and F. Azizah, "Sentiment Analysis for Detecting *Cyberbullying* Using TF-IDF and SVM," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 4, no. 6, pp. 11–12, 2020, doi: 10.29207/resti.v4i6.2753.

[6] F. Farasalsabila, E. Utami, and H. Hanafi, "Deteksi *Cyberbullying* Menggunakan BERT dan Bi-LSTM," *J. Teknol.*, vol. 17, no. 1, pp. 1–6, 2024, doi: 10.34151/jurtek.v17i1.4636.

[7] J. Nalepa and M. Kawulok, "Selecting training sets for support vector machines: a review," *Artificial Intelligence Review*. 2019. doi: 10.1007/s10462-017-9611-1.

[8] J. Ipmawati, S. Saifulloh, and K. Kusnawi, "Analisis Sentimen Tempat Wisata Berdasarkan Ulasan pada Google Maps Menggunakan Algoritma Support Vector Machine," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 4, no. 1, pp. 247–256, 2024, doi: 10.57152/malcom.v4i1.1066.

[9] S. D. Wahyuni and R. H. Kusumodestoni, "Optimalisasi Algoritma Support Vector Machine (SVM) Dalam Klasifikasi Kejadian Data Stunting," *Bull. Inf. Technol.*, vol. 5, no. 2, pp. 56–64, 2024, doi: 10.47065/bit.v5i2.1247.

[10] R. Merdiansah, S. Siska, and A. Ali Ridha, "Analisis Sentimen Pengguna X Indonesia Terkait Kendaraan Listrik Menggunakan IndoBERT," *J. Ilmu Komput. dan Sist. Inf.*, vol. 7, no. 1, pp. 221–228, 2024, doi: 10.55338/jikomsi.v7i1.2895.

[11] B. Wilie *et al.*, "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," 2020, [Online]. Available: http://arxiv.org/abs/2009.05387

[12] G. Z. Nabiilah, I. N. Alam, E. S. Purwanto, and M. F. Hidayat, "Indonesian multilabel classification using IndoBERT embedding and MBERT classification," *Int. J. Electr. Comput. Eng.*, vol. 14, no. 1, pp. 1071–1078, 2024, doi: 10.11591/ijece.v14i1.pp1071-1078.

[13] U. R. Pol, "Hugging Face: Revolutionizing AI and NLP," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 12, no. 8, pp. 1121–1124, 2024, doi: 10.22214/ijraset.2024.64023.

[14] C. Fan, M. Chen, X. Wang, J. Wang, and B. Huang, "A Review on Data *Preprocessing* Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data," *Frontiers in Energy Research*. 2021. doi: 10.3389/fenrg.2021.652801.

[15] A. Zahra, R. Mayasari, and I. Pernamasari, "Analisis Sentimen pada Aplikasi M-Paspor Menggunakan Algoritma Naïve Bayes Classifier," *Action Res. Lit.*, vol. 8, no. 8, pp. 2365–2371, 2024, doi: 10.46799/arl.v8i8.466.

[16] Y. Wulandari, E. Haerani, S. K. Gusti, and S. Ramadhani, "Klasifikasi Berita Menggunakan Algoritma C4.5," *J. Nas. Komputasi dan Teknol. Inf.*, vol. 5, no. 2, pp. 279–289, 2022, doi: 10.32672/jnkti.v5i2.4194.

[17] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," in *COLING 2020 - 28th International Conference on Computational Linguistics, Proceedings of the Conference*, 2020. doi: 10.18653/v1/2020.coling-main.66.

[18] A. Rogers, O. Kovaleva, and A. Rumshisky, "A primer in bertology: What we know about how bert works," *Trans. Assoc. Comput. Linguist.*, vol. 8, pp. 842–866, 2020, doi: 10.1162/tacl_a_00349.

[19] S. Rabbani, D. Safitri, N. Rahmadhani, A. A. F. Sani, and M. K. Anam, "Comparative Evaluation of SVM Kernels for Sentiment Classification in Fuel Price Increase Analysis," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 3, no. 2, pp. 153–160, 2023.

[20] N. G. Ramadhan and A. Khoirunnisa, "Klasifikasi Data Malaria Menggunakan Metode Support Vector Machine," *J. Media Inform. Budidarma*, vol. 5, no. 4, p. 1580, 2021, doi: 10.30865/mib.v5i4.3347.

[21] G. Rininda, I. Hartami Santi, and S. Kirom, "PENERAPAN SVM DALAM ANALISIS SENTIMEN PADA EDLINK MENGGUNAKAN PENGUJIAN CONFUSION MATRIX," *JATI (Jurnal Mhs. Tek. Inform.*, 2024, doi: 10.36040/jati.v7i5.7420.

[22] D. Krstinic, L. Seric, and I. Slapnicar, "Comments on 'MLCM: Multi-Label Confusion Matrix,'" *IEEE Access*. 2023. doi: 10.1109/ACCESS.2023.3267672.

[23] S. Sathyanarayanan, "Confusion Matrix-Based Performance Evaluation Metrics," *African J. Biomed. Res.*, vol. 27, no. 4, pp. 4023–4031, 2024, doi: 10.53555/ajbr.v27i4s.4345.