 ISSN:2302-8149
e-ISSN:2540-9719

Penerapan Algoritma Klasifikasi *Naive Bayes* dan *Support Vector Machine* untuk Analisis Sentimen *Cyberbullying Bilingual* di Aplikasi *X*

Implementation of Naive Bayes and Support Vector Machine Classification Algorithms for Sentiment Analysis of Bilingual Cyberbullying on X Application

[1]**Novita Sari** *, [2]**Muhammad Jazman,** [3]**Tengku Khairil Ahsyar,** [4]**Syaifullah,** [5]**Arif Marsal**
[1,2,3,4,5]Information Systems, Faculty of Science and Technology, Universitas Islam Negeri Sultan Syarif Kasim Riau, Indonesia
*e-mail: 12050322233@students.uin-suska.ac.id

(***received*****:** 12 November 2024, ***revised*****:** 18 November 2024, ***accepted*****:** 20 November 2024)

Abstrak

Peningkatan penggunaan media sosial yang signifikan turut mempengaruhi peningkatan insiden *Cyberbullying*, khususnya dalam konteks keanekaragaman penggunaan bahasa. Penelitian ini akan melakukan analisis sentimen guna mendeteksi konten potensial *Cyberbullying* pada aplikasi *X* dengan pendekatan *bilingual* (bahasa Indonesia dan *Inggris*) menggunakan algoritma *Naive Bayes* dan *Support Vector Machine* (SVM). Data *tweet* terkumpul dan diproses melalui tahap pra-pemrosesan untuk mengekstraksi fitur yang relevan bagi analisis sentimen. Kedua algoritma kemudian diterapkan dalam mengklasifikasikan *tweet* menjadi kategori positif, negatif, atau netral serta mengidentifikasi indikasi *Cyberbullying*. Hasil uji coba menunjukkan bahwa algoritma NB memberikan kinerja yang lebih unggul dibandingkan SVM dengan tingkat akurasi sebesar 87%. Sedangkan dalam mengidentifikasi pola *Cyberbullying* dalam teks *bilingual* NB mencapai tingkat akurasi tertinggi pada bahasa Indonesia yaitu sebesar 87%. Dengan hasil tersebut, diharapkan penelitian ini dapat menjadi referensi untuk pengembangan sistem deteksi *Cyberbullying* yang lebih akurat dan responsif pada platform media sosial *bilingual*.

Kata kunci: aplikasi *X*, analisis sentimen *bilingual*, *cyberbullying*, NB, SVM

Abstract

The significant increase in social media usage has contributed to the rise in cyberbullying incidents, particularly in the context of multilingual language use. This study aims to conduct sentiment analysis to detect potential cyberbullying content on the X application using a bilingual approach (Indonesian and English) and leveraging the Naive Bayes (NB) and Support Vector Machine (SVM) algorithms. Tweets are collected and processed through a pre-processing stage to extract relevant features for sentiment analysis. Both algorithms are then applied to classify tweets into positive, negative, or neutral categories and identify indications of cyberbullying. The results of the trials indicate that the NB algorithm outperformed SVM, achieving an accuracy rate of 87%. Furthermore, in identifying cyberbullying patterns in bilingual text, NB reached the highest accuracy rate for the Indonesian language at 87%. These findings suggest that this study can serve as a reference for developing more accurate and responsive cyberbullying detection systems on bilingual social media platforms.

Keywords: *X app, bilingual sentiment analysis, cyberbullying,* NB, SVM,

1 Pendahuluan

Salah satu bentuk komunikasi yang sepenuhnya berbasis internet adalah media sosial [1]. Jejaring sosial ahli dalam menghubungkan orang-orang yang memiliki minat dan pemikiran yang sama satu sama lain secara teratur [2]. Salah satu media sosial yang cukup banyak peminat nya yaitu *X* (dulunya *Twitter*) [3]. Perkiraan analisis menyebutkan jumlah pengguna *X* pada tahun 2024 berada

 ISSN:2302-8149
e-ISSN:2540-9719

pada angka 335,7 juta pengguna dengan berada di peringkat 12 sebagai jejaring sosial media paling populer [4].

Pengguna platform *X* dapat berinteraksi dan mengekpresikan diri melalui berbagai *tweet* seperti teks, gambar dan video [5]. Platform ini memungkinkan penggunanya untuk berbagi informasi, berita, opini, dan berinteraksi dengan orang lain secara real-time [6]. Sayang sekali, komunikasi ini dapat disalahgunakan untuk *Cyberbullying*, yang merupakan bentuk pelecehan elektronik di mana orang menyerang dan menyinggung orang lain [7].

Konten *Cyberbullying* tidak hanya memicu perilaku berbahaya, tetapi juga membahayakan komunitas online dan kesehatan mental individu [8]. Perilaku ini ditandai dengan tindakan menindas yang dilakukan secara berulang, seringkali didorong oleh berbagai motif seperti kemarahan, balas dendam, atau bahkan hanya untuk mencari sensasi [9]. Fenomena *Cyberbullying* merupakan manifestasi modern dari penindasan tradisional yang memanfaatkan anonimitas dan jangkauan luas internet. Ancaman *Cyberbullying* yang bersifat non-spasial dan berkelanjutan dapat memicu kecemasan, depresi, dan bahkan ide bunuh diri pada korban [10].

Cyberbullying telah menjadi perhatian serius di masyarakat saat ini, seiring dengan meluasnya penggunaan platform media sosial dan saluran komunikasi online lainnya [11]. Dalam penelitian ini, peneliti mengeksplorasi penggunaan algoritma pembelajaran mesin untuk menganalisis konten pesan dan mengidentifikasi pesan yang berisi bahasa yang kasar atau melecehkan [12]. Studi komputasi mengenai opini yang populer dan efisien untuk menganalisis data besar, yang dapat menghasilkan pengambilan keputusan yang lebih baik yaitu menggunakan sentimen analisis[13].

Analisis sentimen merupakan teknik otomatis untuk mengidentifikasi opini positif, negatif, dan netral berdasarkan ulasan pengguna [14]. Analisis sentimen adalah metode populer untuk memanfaatkan keputusan berdasarkan data, yang dapat dibuat dengan mengekstraksi wawasan dari komentar media sosial, tanggapan survei, dan ulasan produk [15]. Teknik dan metode yang sering digunakan dalam klasifikasi sentimen diantara nya *Naive Bayes*(NB), *Support Vector Machine*(SVM), KNN berbasis *machine learning*, *Random Forest* dll [16].

Penelitian ini mengimplementasikan dua algoritma klasifikasi untuk sentimen analisis yaitu *Naive Bayes*(NB) dan *Support Vector Machine*(SVM) [17]. NB merupakan salah satu teknik untuk mengevaluasi atau memvalidasi akurasi model yang dibangun berdasarkan dataset yang digunakan [18]. Sedangkan Salah satu fungsi utama dari SVM adalah mencari garis (*hyperplane*) terbaik yang memisahkan kelas-kelas dengan jarak sejauh mungkin di dalam ruang input [19][20].

Tujuan penelitian ini adalah untuk menemukan algoritma terbaik dari NB dan SVM dalam mengklasifikasi sentimen *Cyberbullying bilingual* di *X*. Selain itu penelitian ini juga ingin melihat apakah algoritma *Naive Bayes* dan SVM masih bisa digunakan untuk analisis sentimen.

2 Tinjauan Literatur

Ada beberapa penelitian sebelumnya yang telah menerapkan algoritma NB dan SVM untuk menganalisis sentimen pada kasus *Cyberbullying* antara lain analisis perbandingan teknik Machine Learning untuk mendeteksi *Cyberbullying* di Twitter dengan membandingkan 7 model klasifikasi Machine Learning diantaranya NB, SVM, LR, LGBM, SGD, RF dan ADB, berdasarkan hasil klasifikasi diperoleh hasil eksperimen menunjukkan keunggulan LR, yang mencapai akurasi median sekitar 90,57%. Di antara pengklasifikasi, regresi logistik mencapai skor F1 terbaik (0,928), SGD mencapai presisi terbaik (0,968), dan SVM mencapai recall terbaik (1,00) [21].

Perbandingan Metode *Support Vector Machine* dan *Naive Bayes* dalam Klasifikasi *Cyberbullying* di Twitter, hasil penelitian ini adalah akurasi klasifikasi *Naïve Bayes* sebesar 97,99% dan akurasi klasifikasi SVM sebesar 99,60% [22]. Kemudian deteksi tingkat keparahan *Cyberbullying*: Pendekatan pembelajaran mesin, dalam penelitian ini kami menerapkan fitur Embedding, Sentiment, dan Le*X*icon beserta orientasi semantik PMI. Fitur yang diekstraksi diterapkan dengan algoritma Naïve Bayes, KNN, Decision Tree, Random Forest, dan *Support Vector Machine*. Hasil dari eksperimen dengan kerangka kerja yang kami usulkan dalam pengaturan multikelas menjanjikan baik berkenaan dengan Kappa, akurasi pengklasifikasi dan metrik f-measure, maupun dalam pengaturan biner [23]. Sistem deteksi *Cyberbullying* di media sosial menggunakan supervised machine learning dengan membandingkan model *Support Vector Machine* (SVM), Naïve Bayes, dan Logistic Regression (LR), didapatkan hasil jika dibandingkan dengan algoritma lain, metode SVM memiliki tingkat akurasi yang tinggi yaitu 75,5% [24].

 ISSN:2302-8149
e-ISSN:2540-9719

Berdasarkan hasil penelitian terdahulu sebelumnya yang telah saya cantumkan, maka dapat disimpulkan bahwa algoritma NB dan SVM terbukti dapat diimplementasikan untuk sentimen analisis *Cyberbullying* di aplikasi *X*. Selain itu, dalam penelitian ini juga akan melakukan penelitian sentimen dalam konteks *Cyberbullying bilingual*, yaitu perbandingan akurasi sentimen bahasa Indonesia dengan sentimen bahasa *Inggris* di *X*.

3 Metode Penelitian

Bagian 'Metode Penelitian' merupakan bagian penting dari penelitian karena memberikan penjelasan rinci tentang bagaimana penelitian dilakukan. Peneliti menguraikan pendekatan penelitian yang digunakan, populasi dan sampel yang digunakan, variabel yang diteliti, alat yang digunakan untuk pengumpulan data, dan metode yang digunakan untuk analisis data. Tujuan utama bagian ini adalah untuk memberikan gambaran yang jelas dan terbuka mengenai proses penelitian, sehingga pembaca dapat memahami dan menilai kualitas penelitian.

3.1 Pendekatan Penelitian

Penelitian ini menggunakan metode klasifikasi sentimen *bilingual* untuk menentukan *tweet* dengan kata kunci "*Cyberbullying*" di *X* sebagai positif, negatif, atau netral berdasarkan sentimen yang terkandung di dalamnya. Platform *Google Colab* dan bahasa pemrograman *Python* digunakan dalam penelitian ini untuk mengolah data *tweet* yang dikumpulkan oleh *X*. *Google Colab* menyediakan lingkungan komputasi awan di mana peneliti dapat menjalankan kode *Python* tanpa menginstal aplikasi tambahan. *Python* memiliki banyak perpustakaan kuat seperti NLTK dan spaCy untuk analisis data teks, yang membuatnya pilihan yang tepat. Untuk mendukung pengelolaan literatur dalam penelitian ini, aplikasi Zotero digunakan.

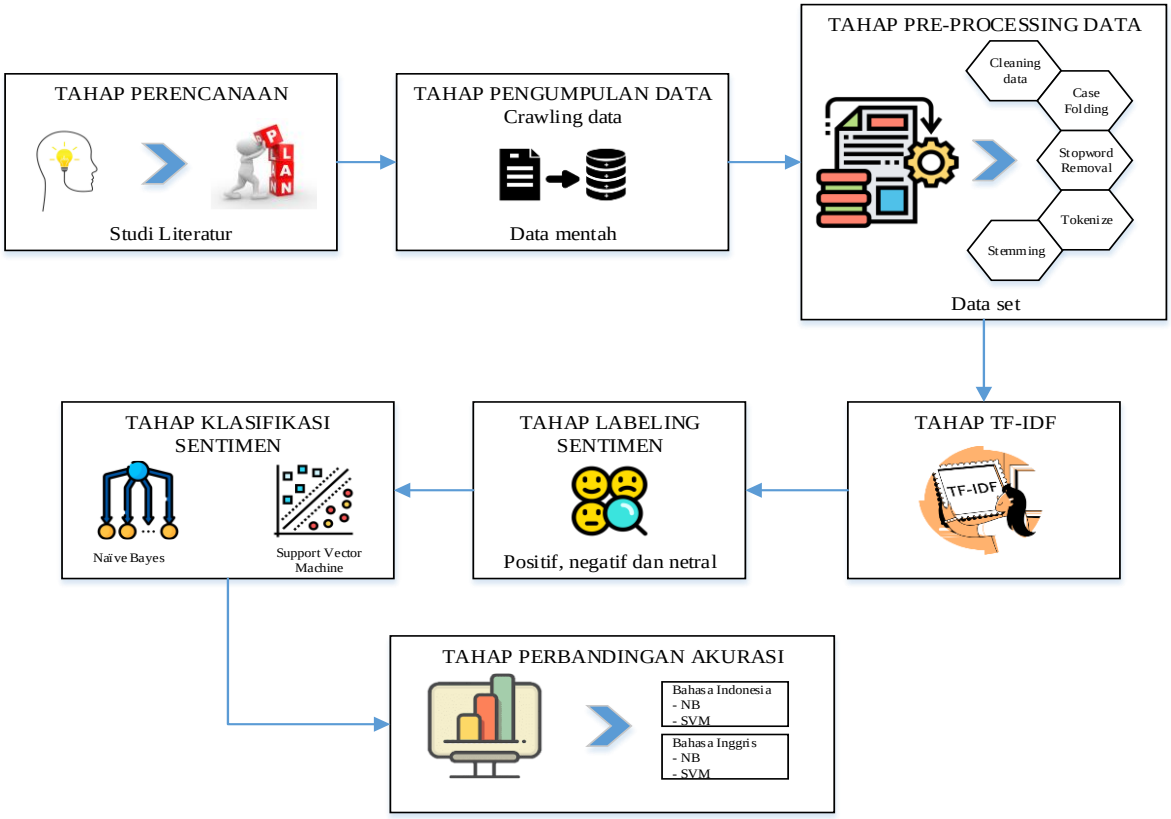
3.2 Populasi dan Sampel

Penelitian ini mengumpulkan semua *tweet* terbaru berbahasa Indonesia dan berbahasa *Inggris* yang diposting di *X* mengandung kata kunci "*Cyberbullying*". Populasi ini mewakili keseluruhan *tweet* berbahasa Indonesia dan berbahasa *Inggris* yang membahas tentang *Cyberbullying* di *X* tersebut. Metode pengambilan data atau sampel menggunakan teknik *Crawling* didapatkan 502 *tweet* data berbahasa Indonesia dan 492 *tweet* data berbahasa *Inggris* dari populasi. Ini memastikan bahwa setiap *tweet* di *X* yang memenuhi kriteria populasi memiliki peluang yang sama untuk masuk dalam sampel.

3.3 Metodologi Penelitian

Dalam penelitian ini, pertama-tama peneliti akan menentukan arah dan pondasi penelitian agar terlaksana dan efektif dan efisien, kemudian peneliti mengumpulkan data *tweet* tentang *Cyberbullying* dari *X*. Setelah itu dilakukan pra-pemrosesan data *tweet* yang telah di dapatkan. Langkah selanjutnya adalah melakukan proses *TF-IDF* untuk memberikan bobot pada kata yang penting. Kemudian kata tersebut diberikan label sentimen berupa label positif, negatif dan netral. Selanjutnya peneliti membandingkan dua model algoritma klasifikasi *Naive Bayes* dan *Support Vector Machine*. Tahap terakhir yaitu melakukan perbandingan akurasi pada kedua algoritma sekaligus perbandingan antara penggunaan algoritma untuk perbandingan dua bahasa. Gambar 1 memberikan gambaran lengkap tentang metodologi penelitian.

 ISSN:2302-8149
e-ISSN:2540-9719



Gambar 1. Metodologi penelitian

3.3.1 Tahap perencanaan

Tahapan perencanaan berisikan tahap studi literatur untuk mencari informasi mengenai referensi dari penelitian sebelumnya yang telah melakukan riset berkaitan dengan analisis sentimen *Cyberbullying bilingual* menggunakan algoritma NB dan SVM.

3.3.2 Pengumpulan Data

Tujuan pengumpulan data adalah untuk mendapatkan dataset yang dapat dianalisis. Data dikumpulkan melalui teknologi *crawling* dengan memanfaatkan bahasa pemrograman *Python* dan platform *Google Colab*. Data yang diambil berasal dari postingan pengguna pada aplikasi *X* dengan kata kunci "*Cyberbullying*".

Web *crawler* adalah program komputer yang dimaksudkan untuk mengunjungi halaman web, menyalin konten yang ada, dan menyimpannya untuk analisis kemudian [25]. Data atau sampel yang dikumpulkan dari *X* didapatkan sebanyak 349 *tweet* berbahasa Indonesia dan 349 *tweet* berbahasa *Inggris* dari populasi. Hal ini memastikan bahwa semua *tweet* yang memenuhi kriteria populasi memiliki peluang yang sama untuk dimasukkan dalam sampel.

3.3.3 *Pre-Processing* Data

Preprocessing data adalah prosedur untuk menghilangkan kebisingan atau meningkatkan kualitas teks [26]. Pra-pemrosesan data adalah metode pengolahan data yang mengubah data mentah menjadi format yang mudah dipahami. Tahap pra-pemrosesan data diperlukan untuk menyelesaikan masalah seperti data berisik, redundansi, dan nilai yang hilang [27].

Tahap ini berfokus pada manipulasi teks itu sendiri, seperti:

3.3.3.1 *Cleaning* Data

Pembersihan data melibatkan pencarian dan koreksi atau penghapusan data yang salah atau bermasalah dari kumpulan data. Proses ini biasanya digunakan untuk mencari dan mengganti data atau catatan yang tidak lengkap, tidak akurat, tidak relevan, atau salah [28].

 ISSN:2302-8149
e-ISSN:2540-9719

3.3.3.2 Case Folding

Case Folding adalah teknik yang mengubah setiap huruf dalam dokumen atau kalimat menjadi huruf yang lebih kecil yang dapat dicari dengan lebih mudah. Dalam hal penggunaan huruf kapital, tidak semua data konsisten [29].

3.3.3.3 Stopword Removal

Proses menghilangkan kata-kata seperti "di", "dari", dan "yang", serta kata-kata seperti *"a"*, *"the"*, dan *"is"* yang tidak dapat ditemukan di mesin pencari komputer. Menghapus stopwords meningkatkan efisiensi dan akurasi aplikasi penambangan teks, dan juga mengurangi kompleksitas waktu dan ruang aplikasi penambangan teks secara keseluruhan [30].

3.3.3.4 Tokenization

Tokenisasi adalah pembagian kumpulan karakter berdasarkan ruang. Pada saat yang sama, juga dapat menghapus karakter tertentu, seperti tanda baca [31]

3.3.3.5 Stemming

Periksa kata-kata yang telah diubah menjadi huruf kecil. Stemming mengurangi daftar kata dalam data pelatihan dengan membakukan kata [32].

3.3.4 TF-IDF

Frekuensi *Term* - Frekuensi Dokumen Terbalik (TF-IDF) untuk mengatasi masalah bahwa kata yang sering muncul dilebih-lebihkan oleh jumlah kata mentah yang dikaitkan dengan fitur, pendekatan TF-IDF digunakan untuk mengukur seberapa sering kata tersebut muncul dalam dokumen. Oleh karena itu, beberapa algoritme klasifikasi mungkin tidak menghitung jumlah kata secara optimal menggunakan ukuran frekuensi kata dalam dokumen [33].

Rumus untuk algoritma TF-IDF ditunjukkan pada (1) - (3), di mana TFij menunjukkan nilai TF dari kata ke-i untuk dokumen ke-j, IDFi menunjukkan nilai IDF dari kata ke-i, dan TF-IDFij menunjukkan nilai TF-IDF dari kata ke-i untuk dokumen ke-j. D adalah jumlah total semua dokumen, Dj adalah dokumen ke-j, nij adalah jumlah kemunculan kata ke-i pada dokumen ke-j, dan Ni adalah jumlah kemunculan kata ke-i pada semua dokumen.

$$TF_{ij} = \frac{n_{ij}}{D_j} \tag{1}$$

$$IDF_i = log(\frac{D}{N_i}) \tag{2}$$

$$TF - IDF_{ij} = TF_{ij}.IDF_i \tag{3}$$

Rumus TF-IDF yang telah diperbaiki ditunjukkan pada (4), dengan kata ke-i muncul pada segmen pertama dan terakhir ketika f(i) jika tidak, f(i) = 1; g(i) ketika kata ke-i adalah kata benda, dan g(i) = 1 ketika kata ke-i memiliki sifat leksikal yang berbeda. adalah angka yang lebih besar dari nol, dan suku-suku yang lain memiliki makna yang sama seperti pada persamaan (1) - (3) [34].

$$TF - IDF_{ij} = f(i).g(i).TF_{ij}.IDF_i \tag{4}$$

3.3.5 Labeling Sentimen

Nilai sentimen diberikan pada teks, yang dapat berupa nilai positif, negatif, atau netral. Nilai sentimen dapat dilabelkan dengan berbagai cara, seperti secara manual dengan memasukkan nilai sentimen pada teks, menggunakan fungsi TF-IDF, atau dengan *Python* dengan menggunakan library *TeXtBlob*. Pada penelitian ini peneliti menggunakan library *TextBlob* dalam proses pemberian labeling sentimen pada data.

 ISSN:2302-8149
e-ISSN:2540-9719

3.3.6 Klasifikasi Sentimen

Pada bagian ini dilakukan proses analisis sentimen pada data yang telah melewati tahap pemrosesan data. Tahap ini terdiri dari proses klasifikasi *Naive Bayes* dan klasifikasi *Support Vector Machine*.

3.3.6.1 *Naive Bayes* (NB)

Algoritma pembelajaran mesin *Naive Bayes* termasuk dalam kategori pembelajaran terawasi (pembelajaran mesin yang memerlukan sampel berlabel sebagai data pelatihan) [18]. Algoritme ini sangat disukai karena kesederhanaannya, efisiensinya, dan skalabilitasnya, yang membuatnya menarik untuk berbagai tugas klasifikasi [35].

Naive Bayes hanya bisa mengenali teks dan angka, tapi tidak bisa mengenali gambar. Metode ini menggunakan teorema Bayes untuk menghitung probabilitas [36]. Dalam bahasa sederhana, teorema Bayes didefinisikan sebagai kemungkinan bahwa hubungan A akan terjadi jika hubungan B telah terjadi sebelumnya dan sebaliknya [37]. Persamaan *bayes* seperti persamaan (5).

$$P(H|X) = \frac{P(X|H).P(H)}{P(X)} \quad (5)$$

Di mana :

X =Data dengan *class* yang belum diketahui
H =Hipotesis data merupakan suatu *class* spesifik
P(H|*X*) =Probabilitas hipotesis H berdasar kondisi *X* (posteriori probabilitas)
P(H) =Probabilitas hipotesis H (*prior* probabilitas)
P(*X*|H) =Probabilitas *X* berdasarkan kondisi pada hipotesis H
P(*X*) =Probabilitas *X*

3.3.6.2 *Support Vector Machine* (SVM)

Dengan menggunakan fungsi kernel, SVM dapat menangani data linear dan *nonlinear*. Ini memetakan ruang masukan ke ruang fitur berdimensi lebih besar, memungkinkan pemisahan linier [38]. Sedangkan Salah satu fungsi utama dari SVM adalah mencari garis (hyperplane) terbaik yang memisahkan kelas-kelas dengan jarak sejauh mungkin di dalam ruang input [20].

Akurasi algoritma SVM bergantung pada klasifikasi linear, yang digunakan, dan fungsi kernelnya. SVM dapat menangani data dengan karakteristik non-linier dengan baik dengan menggunakan berbagai jenis fungsi kernel, seperti kernel linier, polinomial, atau Gauss. Fungsi kernel adalah fungsi yang mengubah data ke dimensi yang lebih besar dengan tujuan meningkatkan struktur data sehingga mempermudah proses pemisahan [39].
Rumus fungsi kernel ditunjukkan pada persamaan (6)-(8).

$$Linier \;\; K(x,y) = x,y \quad (6)$$

$$Polynomial\; K(xi,xj) = \left((xi.xj) + c\right)^d \quad (7)$$

$$Gaussian\; K(xi,xj) = exp\left(-\frac{|xi-xj|^2}{2\sigma^2}\right) \quad (8)$$

Keterangan:
K(*X*i,*X*j) = Fungsi kernel
*X*i = Data ke-i
*X*j = Data ke-j
e*X*p = Operasi eksponen
d,σ = Parameter kernel

3.3.7 Perbandingan Akurasi

Perbandingan akurasi dilakukan untuk menentukan algoritma mana yang lebih baik dalam menghasilkan nilai pada data aplikasi *X* tentang *Cyberbullying* menggunakan algoritma *Naive Bayes* dan *Support Vector Machine*. Pada perbandingan akurasi juga dilakukan untuk menentukan bahasa apa yang memiliki akurasi yang lebih tinggi dalam melakukan analisis sentimen. Perbandingan yang dilakukan dengan membandingkan nilai akurasi, *precission*, *recall*, dan *f1-score* pada masing masing algoritma.

 ISSN:2302-8149
e-ISSN:2540-9719

4 Hasil dan Pembahasan

Setelah mengetahui metodologi yang akan digunakan dalam penelitian ini, langkah berikutnya adalah menjelaskan hasilnya. Mula-mula, data dikumpulkan dari *X*; kemudian dibersihkan dari *noise*, duplikat, karakter atau simbol, dan pemberian bobot; kemudian, terlepas dari kata-kata yang digunakan, diberikan label positif, negatif, atau netral sesuai dengan perasaan yang digunakan. Kemudian data diklasifikasikan menurut algoritma NB dan SVM untuk mendapatkan nilai akurasi terbaik. Selanjutnya, perbandingan akurasi dilakukan untuk menentukan algoritma mana yang paling akurat dan untuk menentukan bahasa mana yang paling akurat dari kedua algoritma tersebut. Hasil dan pembahasan lengkap dapat dilihat pada tahapan 4.1 sampai 4.6.

4.1 Pengumpulan Data

Proses pengumpulan data dilakukan dengan teknik *Crawling Data* menggunakan *Google Colab* dan bahasa pemrograman *Python* di aplikasi *X* dengan kata kunci "*Cyberbullying"*. Data *tweet* yang dikumpulkan merupakan data terbaru. Untuk dapat mengakses data *X* diperlukannya *Auth Token X* tersebut sebagai kode akses agar bisa dilakukannya *Crawling Data* oleh *Google Colab*. Proses *Crawling Data* dilakukan pada dua bahasa yang berbeda, bahasa Indonesia dan bahasa *Inggris* dengan topik yang sama yaitu *Cyberbullying.* Hasil *Crawling Data* bahasa Indonesia dan bahasa *Inggris* dari *X* ditampilkan pada Gambar 2 dan 3.

	conversation_id_str	created_at	favorite_count	full_text
0	1793148292320436356	Wed May 22 13:39:57 +0000 2024	0	@Marchfoward Masa nilah KOYAK trending. Lepas ...
1	1793270666785079302	Wed May 22 13:20:18 +0000 2024	0	Bahaya Cyberbullying! Review Video yang Menter...

✓ 0 d sel

Gambar 2. Hasil *Crawling data* bahasa Indonesia

	conversation_id_str	created_at	favorite_count	full_text
0	1793477405778305491	Thu May 23 03:01:49 +0000 2024	1	#Survivor #Survivor47 I will be cyberbullying ...
1	1793448307194544458	Thu May 23 02:59:03 +0000 2024	1	@MasteroftheTDS @Grummz Just a home address ? ...
2	1792577666794606948	Thu May 23 02:52:00 +0000 2024	7	@laylassong @acmecojim @AstroLlamaBeans sittin...

Gambar 3. Hasil *crawling data* bahasa *inggris*

Berdasarkan hasil *Crawling data* seperti terlihat pada gambar 2 dan gambar 3 diatas, didapatkan sebanyak 502 *tweet* berbahasa Indonesia dan 492 bahasa *Inggris* yang membahas tentang *Cyberbullying*.

4.2 *Pre-Processing* Data

Tahap ini bertujuan untuk membersihkan dan menstandariskan data teks agar algoritma analisis sentimen menjadi lebih terstruktur dan lebih mudah diolah. Tabel 1 menunjukkan hasil *Pre-Processing* untuk bahasa Indonesia, dan Tabel 2 menunjukkan hasil untuk bahasa *Inggris*.

Tabel 1. Hasil *pre-processing* bahasa indonesia

Alur *Pre-Processing*	Sebelum	Sesudah
Cleaning	Apa susahnya nyebut safa n si berflower itu saja alih-alih menyebut	Apa susahnya nyebut safa n si berflower itu saja alihalih menyebut nama artis

 ISSN:2302-8149
e-ISSN:2540-9719

data	nama artis secara langsung untuk dijadiin lelucon? Sangat disayangkan pelaku cyberbullying (safa) malah dijadikan seperti korban karna org hanya fokus dgn si berflower dan nama idol yg gak bersalah sllu disebut	secara langsung untuk dijadiin lelucon Sangat disayangkan pelaku cyberbullying safa malah dijadikan seperti korban karna org hanya fokus dgn si berflower dan nama idol yg gak bersalah sllu disebut
Case Folding	Apa susahnya nyebut safa n si berflower itu saja alihalih menyebut nama artis secara langsung untuk dijadiin lelucon Sangat disayangkan pelaku cyberbullying safa malah dijadikan seperti korban karna org hanya fokus dgn si berflower dan nama idol yg gak bersalah sllu disebut	apa susahnya nyebut safa n si berflower itu saja alihalih menyebut nama artis secara langsung untuk dijadiin lelucon sangat disayangkan pelaku cyberbullying safa malah dijadikan seperti korban karna org hanya fokus dgn si berflower dan nama idol yg gak bersalah sllu disebut
Stopword Removal	apa susahnya nyebut safa n si berflower itu saja alihalih menyebut nama artis secara langsung untuk dijadiin lelucon sangat disayangkan pelaku cyberbullying safa malah dijadikan seperti korban karna org hanya fokus dgn si berflower dan nama idol yg gak bersalah sllu disebut	apa susahnya nyebut safa n si berflower itu saja alihalih menyebut nama artis secara langsung untuk dijadiin lelucon sangat disayangkan pelaku cyberbullying safa malah dijadikan seperti korban karna org hanya fokus dgn si berflower dan nama idol yg gak bersalah sllu disebut
Tokenize	apa susahnya nyebut safa n si berflower itu saja alihalih menyebut nama artis secara langsung untuk dijadiin lelucon sangat disayangkan pelaku cyberbullying safa malah dijadikan seperti korban karna org hanya fokus dgn si berflower dan nama idol yg gak bersalah sllu disebut	[apa, susahnya, nyebut, safa, n, si, berflower, itu, saja, alihalih, menyebut, nama, artis, secara, langsung, untuk, dijadiin, lelucon, sangat, disayangkan, pelaku, cyberbullying, safa, malah, dijadikan, seperti, korban, karna, org, hanya, fokus, dgn, si, berflower, dan, nama, idol, yg, gak, bersalah, sllu, disebut]
Stemming	[apa, susahnya, nyebut, safa, n, si, berflower, itu, saja, alihalih, menyebut, nama, artis, secara, langsung, untuk, dijadiin, lelucon, sangat, disayangkan, pelaku, cyberbullying, safa, malah, dijadikan, seperti, korban, karna, org, hanya, fokus, dgn, si, berflower, dan, nama, idol, yg, gak, bersalah, sllu, disebut]	apa susah nyebut safa n si berflower itu saja alihalih sebut nama artis cara langsung untuk dijadiin lelucon sangat sayang laku cyberbullying safa malah jadi seperti korban karna org hanya fokus dgn si berflower dan nama idol yg gak salah sllu sebut

Tabel 2. Hasil *pre-processing* bahasa *inggris*

Alur *Pre-Processing*	Sebelum	Sesudah
Cleaning data	*@Griot2325 Most tame examples cyberbullying I've ever seen in my life Twitter reactionaries*	*Griot Most tame examples cyberbullying Ive ever seen in my life Twitter reactionaries cannot be real people*

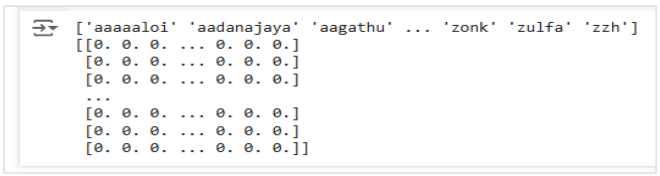
 ISSN:2302-8149
e-ISSN:2540-9719

	cannot be real people	
Case Folding	*Griot Most tame examples cyberbullying Ive ever seen in my life Twitter reactionaries cannot be real people*	*griot most tame examples cyberbullying ive ever seen in my life twitter reactionaries cannot be real people*
Stopword Removal	*griot most tame examples cyberbullying ive ever seen in my life twitter reactionaries cannot be real people*	*griot tame examples cyberbullying ive ever seen life twitter reactionaries cannot real people*
Tokenize	*griot tame examples cyberbullying ive ever seen life twitter reactionaries cannot real people*	[*griot, tame, examples, cyberbullying, ive, ever, seen, life, twitter, reactionaries, cannot, real, people*]
Stemming	[*griot, tame, examples, cyberbullying, ive, ever, seen, life, twitter, reactionaries, cannot, real, people*]	*griot tame examples cyberbullying ive ever seen life twitter reactionaries cannot real people*

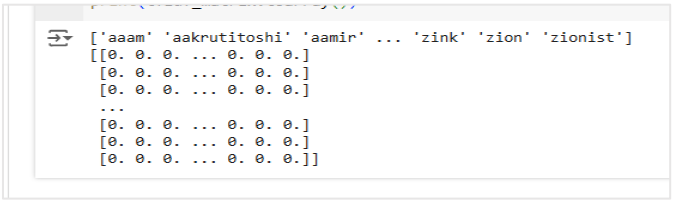
Seperti terlihat pada tabel 1 dan 2, kumpulan data yang lolos tahap prapemrosesan data bersifat seragam, terstruktur, dan tidak mengandung simbol di dalam teks. Sebab, ketika sudah sampai pada tahap selanjutnya yaitu tahap klasifikasi, komputasi menjadi lebih optimal.

4.3 TF-IDF

TF-IDF dapat mengekstrak fitur teks penting, yang membantu memahami makna dan memprediksi sentimen. Kata kunci yang sering digunakan tetapi jarang ditemukan di korpus lain akan menerima skor yang lebih tinggi, yang dapat digunakan untuk membedakan data. Sementara kata-kata umum yang sering digunakan akan dikurangi. Hasil pembobotan TF-IDF bahasa Indonesia dan bahasa *Inggris* dapat dilihat pada Gambar 4-5.



Gambar 4. Hasil TF-IDF data bahasa indonesia



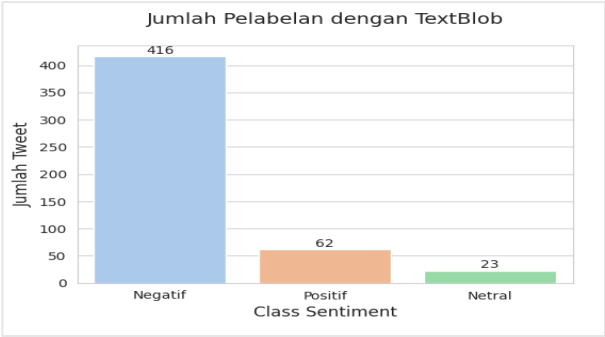
Gambar 5. Hasil TF-IDF data bahasa *inggris*

4.4 *Labeling* Data

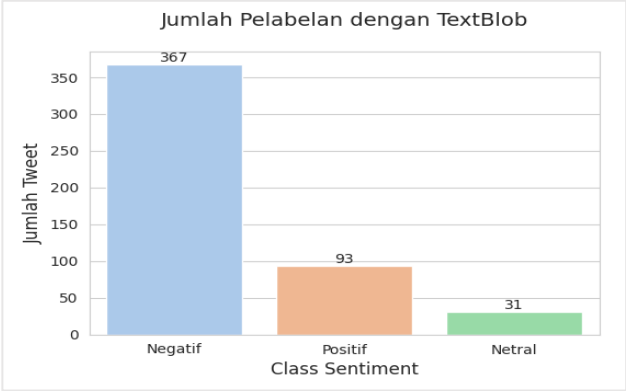
Labeling adalah proses memberikan label sentimen pada data teks. Proses labeling menggunakan library pemrograman *TextBlob*. Label ini menunjukkan apakah teks tersebut memiliki sentimen positif, negatif atau netral. Pada data yang telah didapatkan sebanyak 502 data berbahasa Indonesia dan 492 data berbahasa *Inggris*, kedua data tersebut memiliki jumlah pelabelan yang berbeda.

Berdasarkan Gambar 6 dan 7 terlihat bahwa data dengan nilai sentimen negatif lebih tinggi dibandingkan dengan sentimen positif. Ini berarti bahwa sebagian besar diskusi atau komentar tentang cyberbullying cenderung mengekspresikan perasaan negatif seperti kemarahan, kesedihan, atau ketakutan. Topik ini kemungkinan besar dianggap sebagai masalah serius yang memiliki dampak negatif pada individu dan masyarakat.

 ISSN:2302-8149
e-ISSN:2540-9719



Gambar 6. Hasil *labeling* sentimen pada data bahasa indonesia



Gambar 7. Hasil *labeling* sentimen pada data bahasa *inggris*

4.5 Klasifikasi Sentimen

Pada bagian ini akan menjelaskan bagaimana proses pengklasifikasian pada kedua algoritma yang digunakan yaitu *Naive Bayes* dan SVM. Setelah data melewati proses pra-pemrosesan dan *labeling* data, maka langkah selanjutnya mengklasifikasikan data menggunakan algoritma NB dan SVM untuk mendapatkan akurasi terbaik dari kedua algoritma tersebut. Selain itu pada bagian ini juga akan melihat akurasi algoritma pada bahasa mana yang tertinggi.

4.5.1 *Naive Bayes* (NB)

Naive Bayes adalah metode klasifikasi teks sederhana namun efektif yang digunakan untuk memprediksi kemungkinan suatu kategori berdasarkan bukti historis. Pada penelitian ini data set dibagi menjadi data latih dan data uji sebanyak 20% dari keseluruhan data. Hasil klasifikasi *Naive Bayes* dapat dilihat pada Tabel 3.

Tabel 3 Hasil klasifikasi *Naive Bayes*

Metriks klasifikasi	Bahasa Indonesia	Bahasa *Inggris*
Akurasi	69%	56%
Precision	76%	55%
Recall	69%	56%
F1-score	72%	54%

Precision merupakan proses mengukur seberapa sering prediksi positif model benar-benar positif. Tujuannya untuk meminimalkan *false positive* (memprediksi positif padahal sebenarnya negatif). Sedangkan *Recall* kebalikan dari *precision* dimana berfungsi untuk mengukur seberapa banyak dari semua contoh positif yang sebenarnya berhasil diidentifikasi oleh model. *F1-score* adalah nilai rata-rata harmonik untuk *precision* dan *recall*.

Ada sejumlah faktor yang dapat menyebabkan perbedaan hasil klasifikasi antara Bahasa Indonesia dan Bahasa Inggris di tabel tersebut. Struktur, ukuran, atau kompleksitas dataset yang digunakan dapat memengaruhi kinerja model. Aturan linguistik setiap bahasa memengaruhi kualitas

 ISSN:2302-8149
e-ISSN:2540-9719

data yang dihasilkan oleh proses preprocessing teks seperti tokenisasi dan stemming. Selain itu, mungkin model *Naive Bayes* sederhana lebih cocok untuk dataset atau bahasa tertentu.

4.5.2 *Support Vector Machine* (SVM)

Support Vector Machine (SVM) adalah salah satu algoritma yang dapat digunakan untuk klasifikasi dan regresi. Dalam konteks klasifikasi, SVM mencari *hyperplane* terbaik untuk memisahkan data menjadi dua kelas atau lebih. Pada penelitian ini data set dibagi menjadi data latih dan data uji sebanyak 20% dari keseluruhan data. Hasil klasifikasi SVM dapat dilihat pada Tabel 4.

Tabel 4. Hasil klasifikasi SVM

Metriks klasifikasi	Bahasa Indonesia	Bahasa *Inggris*
Akurasi	86%	66%
Precision	76%	43%
Recall	86%	66%
F1-score	81%	52%

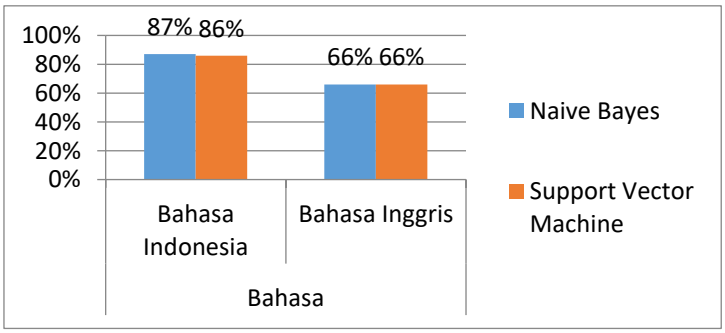
Dari tabel 4 terlihat bahwa perbedaan hasil klasifikasi antara Bahasa Indonesia dan Bahasa *Inggris* mungkin disebabkan oleh variabel seperti perbedaan struktur bahasa, kualitas, dan kuantitas data latih. Mungkin karena data Bahasa Indonesia lebih representatif atau memiliki struktur yang lebih sederhana dibandingkan dengan Bahasa *Inggris*. Model SVM yang dilatih dengan data bahasa Indonesia lebih baik dalam mengklasifikasikan sentimen dibandingkan dengan model SVM yang dilatih dengan data bahasa Inggris. Ini menunjukkan bahwa kualitas dan kuantitas data latih, serta karakteristik bahasa, sangat penting untuk keberhasilan analisis sentimen.

4.6 Perbandingan Akurasi

Perbandingan akurasi dilakukan untuk menentukan algoritma mana yang lebih baik dalam menghasilkan nilai pada data aplikasi *X* tentang *Cyberbullying* menggunakan algoritma *Naive Bayes* dan *Support Vector Machine*. Pada perbandingan akurasi juga dilakukan untuk menentukan bahasa apa yang memiliki akurasi yang lebih tinggi dalam melakukan analisis sentimen. Perbandingan yang dilakukan dengan membandingkan nilai akurasi, *precission*, *recall*, dan *f1-score* pada masing masing algoritma. Hasil perbandingan akurasi dapat dilihat pada Tabel 5 dan representasi hasil pada Gambar 8.

Tabel 5. Hasil perbandingan akurasi

Algoritma	Bahasa	
	Bahasa Indonesia	Bahasa *Inggris*
Naive Bayes	87%	66%
Support Vector Machine	86%	66%



Gambar 8. Representasi hasil perbandingan akurasi

Berdasarkan Tabel 5, *Naive Bayes* memiliki akurasi tertinggi dari SVM dan bahasa Indonesia mendapatkan nilai akurasi tertinggi dibandingkan bahasa *Inggris*. Ini bisa terjadi karena perbedaan akurasi antara bahasa Indonesia dan bahasa *Inggris* dalam analisis sentimen dapat disebabkan oleh berbagai faktor, mulai dari kompleksitas bahasa hingga kualitas data latih. Karakteristik data, tujuan analisis, dan sumber daya komputasi menentukan pilihan antara *Naive Bayes* dan SVM. Dalam

 ISSN:2302-8149
e-ISSN:2540-9719

beberapa situasi, *Naive Bayes* mungkin memberikan hasil yang lebih baik daripada SVM, terutama dalam kasus di mana data yang digunakan berkualitas tinggi dan memiliki struktur yang sederhana.

5 Kesimpulan

Penelitian ini melakukan analisis sentimen pada ulasan pengguna *X* terhadap topik *Cyberbullying* dengan membandingkan 2 Algoritma yaitu NB dan SVM. Berdasarkan hasil klasifikasi yang diperoleh melalui tahapan *Pre-Processing*, algoritma NB terbukti lebih unggul dari algoritma SVM dengan nilai akurasi sebesar 87%, sedangkan SVM 86%. Selain itu, perbandingan akurasi bahasa, tertinggi didapatkan oleh NB bahasa Indonesia yaitu sebesar 87%. Berarti hasil penelitian ini menunjukkan bahwa NB masih sangat relevan untuk algoritma klasifikasi sentimen *bilingual*. Pada penelitian berikutnya, kami sarankan untuk memperluas penggunaan algoritma terbaru dan penggunaan bahasa asing lainnya agar klasifikasi nya bisa lebih luas tidak hanya dalam ruang lingkup kecil saja.

Referensi

[1] A. Lüders, A. Dinkelberg, and M. Quayle, '*Becoming "Us" In Digital Spaces: How Online users Creatively and Strategically Exploit Social Media Affordances to Build Up Social Identity', Act Psychologica*, vol. 228, p. 103643, Aug. 2022, doi: 10.1016/j.actpsy.2022.103643.

[2] N. Sabermajidi, N. Valaei, M. S. Balaji, and S. K. Goh, *'Measuring Brand-Related Content in Social Media: a Socialization Theory Perspective', Information Technology & People*, vol. 33, no. 4, pp. 1281–1302, Jan. 2020, doi: 10.1108/ITP-10-2018-0497.

[3] L. Stracqualursi and P. Agati, '*Tweet Topics and Sentiments Relating to Distance Learning Among Italian Twitter Users', Sci Rep*, vol. 12, no. 1, p. 9163, Jun. 2022, doi: 10.1038/s41598-022-12915-w.

[4] E. S. Matsa Sarah Naseer, Jacob Liedke and Katerina Eva, '*How Americans Get News on TikTok, X, Facebook and Instagram'*, Pew Research Center. Accessed: Aug. 11, 2024. [Online]. Available: https://www.pewresearch.org/journalism/2024/06/12/how-americans-get-news-on-tiktok-x-facebook-and-instagram/

[5] R. Kullar, D. A. Goff, T. P. Gauthier, and T. C. Smith, *'To Tweet or Not to Tweet—a Review of the Viral Power of Twitter for Infectious Diseases*', *Curr Infect Dis Rep*, vol. 22, no. 6, p. 14, Jun. 2020, doi: 10.1007/s11908-020-00723-0.

[6] L. Stracqualursi and P. Agati, '*Twitter Users Perceptions of AI-based E-Learning Technologies', Scientific Reports*, vol. 14, no. 1, pp. 1–14, 2024, doi: 10.1038/s41598-024-56284-y.

[7] O. A. Alismaiel, '*Digital Media used in Education*: *The Influence on Cyberbullying Behaviors among Youth Students*', *IJERPH*, vol. 20, no. 2, p. 1370, Jan. 2023, doi: 10.3390/ijerph20021370.

[8] D. Kim, '*Cyberbullying Behaviors in Online Travel Community: Members' Perceptions and Sustainability in Online Community*', *Sustainability*, vol. 14, no. 9, p. 5220, Apr. 2022, doi: 10.3390/su14095220.

[9] Ahmad Mohamad Alomar and Hassan Sami Alabady, *'The Phenomenon of Cyber Bullying: Interpretation, Confrontation, and the Position of Islamic Law', JNS*, vol. 34, May 2023, doi: 10.59670/jns.v34i.1123.

[10] Ahmad Mohammad Alomar et Al, *'Aspect and Special Distinct Nature of Cyberbullying', Russian Law Journal*, vol. 11, no. 3, Art. no. 3, Apr. 2023, doi: 10.52783/rlj.v11i3.1817.

[11] L. H. Collantes, Y. Martafian, S. N. Khofifah, T. Kurnia Fajarwati, N. T. Lassela, and M. Khairunnisa, '*The Impact of Cyberbullying on Mental Health of the Victims', in 2020 4th International Conference on Vocational Education and Training (ICOVET)*, Malang, Indonesia: IEEE, Sep. 2020, pp. 30–35. doi: 10.1109/ICOVET50258.2020.9230008.

[12] M. P. Akhter, Z. Jiangbin, I. R. Naqvi, M. AbdelMajeed, and T. Zia, *'Correction To: Abusive Language Detection From Social Media Comments using Conventional Machine Learning and Deep Learning Approaches*', *Multimedia Systems*, vol. 29, no. 1, pp. 451–451, Feb. 2023, doi: 10.1007/s00530-021-00819-0.

 ISSN:2302-8149
e-ISSN:2540-9719

[13] I. Awajan, M. Mohamad, and A. Al-Quran, '*Sentiment Analysis Technique and Neutrosophic Set Theory for Mining and Ranking Big Data From Online Reviews'*, *IEEE Access*, vol. 9, pp. 47338–47353, 2021, doi: 10.1109/ACCESS.2021.3067844.
[14] Fathurahman Bei and Sudin Saepudin, '*Analisis Sentimen Aplikasi Tiket Online di Play Store menggunakan Metode Support Vector Machine* (SVM)', 2021.
[15] H. Hertina *et al.*, '*Data Mining Applied About Polygamy using Sentiment Analysis On Twitters In Indonesian Perception*', *Bulletin EEI*, vol. 10, no. 4, pp. 2231–2236, Aug. 2021, doi: 10.11591/eei.v10i4.2325.
[16] B. AlBadani, R. Shi, and J. Dong, '*A Novel Machine Learning Approach for Sentiment Analysis on Twitter Incorporating the Universal Language Model Fine-Tuning and SVM*', *ASI*, vol. 5, no. 1, p. 13, Jan. 2022, doi: 10.3390/asi5010013.
[17] M. R. Romadhon and F. Kurniawan, '*A Comparison of Naive Bayes Methods, Logistic Regression and KNN for Predicting Healing of Covid-19 Patients in Indonesia'*, in *2021 3rd East Indonesia Conference on Computer and Information Technology (EIConCIT)*, Surabaya, Indonesia: IEEE, Apr. 2021, pp. 41–44. doi: 10.1109/EIConCIT50028.2021.9431845.
[18] A. Roihan, P. A. Sunarya, and A. S. Rafika, 'Pemanfaatan *Machine Learning* dalam berbagai Bidang: *Review paper*', *IJCIT*, vol. 5, no. 1, May 2020, doi: 10.31294/ijcit.v5i1.7951.
[19] '*Survey on Dietary Application through Image Processing for Calorie Management*', *International Journal of Advanced Research in Science, Communication and Technology*, pp. 345–347, May 2022, doi: 10.48175/ijarsct-3666.
[20] M. Muhathir, M. H. Santoso, and D. A. Larasati, 'Wayang *Image Classification using SVM Method and GLCM Feature Extraction*', *Journal Of Informatics And Telecommunication Engineering*, vol. 4, no. 2, pp. 373–382, 2021.
[21] A. Muneer and S. M. Fati, '*A Comparative Analysis of Machine Learning Techniques For Cyberbullying Detection on Twitter*', *Future Internet*, vol. 12, no. 11, pp. 1–21, 2020, doi: 10.3390/fi12110187.
[22] N. Chamidah and R. Sahawaly, '*Comparison Support Vector Machine and Naive Bayes Methods for Classifying Cyberbullying in Twitter*', *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika*, vol. 7, no. 2, p. 338, Sep. 2021, doi: 10.26555/jiteki.v7i2.21175.
[23] B. A. Talpur and D. O'Sullivan, '*Cyberbullying Severity Detection: A Machine Learning Approach*', *PLoS ONE*, vol. 15, no. 10 October, pp. 1–19, 2020, doi: 10.1371/journal.pone.0240924.
[24] A. Perera and P. Fernando, '*Cyberbullying Detection System on Social Media using Supervised Machine Learning*', *Procedia Computer Science*, vol. 239, pp. 506–516, 2024, doi: 10.1016/j.procs.2024.06.200.
[25] C. Muehlethaler and R. A*lbert, 'Collecting Data on Textiles from the Internet using Web Crawling and Web Scraping Tools*', *Forensic Science International*, vol. 322, p. 110753, 2021, doi: 10.1016/j.forsciint.2021.110753.
[26] A. P. Natasuwarna, 'Seleksi Fitur *Support Vector Machine* pada Analisis Sentimen Keberlanjutan Pembelajaran Daring', *Techno.Com*, vol. 19, no. 4, pp. 437–448, 2020, doi: 10.33633/tc.v19i4.4044.
[27] U. Naseem, I. Razzak, and P. W. Eklund, '*A Survey Of Pre-Processing Techniques to Improve Short-Text Quality: A Case Study On Hate Speech Detection On Twitter'*, *Multimedia Tools and Applications*, vol. 80, no. 28–29, pp. 35239–35266, 2021, doi: 10.1007/s11042-020-10082-6.
[28] K. Maharana, S. Mondal, and B. Nemade, 'A review: *Data Pre-Processing and Data Augmentation Techniques*', *Global Transitions Proceedings*, vol. 3, no. 1, pp. 91–99, 2022, doi: 10.1016/j.gltp.2022.04.020.
[29] D. Alita and A. R. Isnain, 'Pendeteksian Sarkasme pada Proses Analisis Sentimen menggunakan *Random Forest Classifier*', *Jurnal Komputasi*, vol. 8, no. 2, pp. 50–58, 2020, doi: 10.23960/komputasi.v8i2.2615.
[30] D. J. Ladani and N. P. Desai, '*Stopword Identification and Removal Techniques on TC and IR Applications: A Survey', 2020 6th International Conference on Advanced Computing and Communication Systems, ICACCS 2020*, pp. 466–472, 2020, doi: 10.1109/ICACCS48705.2020.9074166.

 ISSN:2302-8149
e-ISSN:2540-9719

[31] Y. A. Singgalen, 'Analisis Sentimen Konsumen terhadap *Food, Services, and Value* di Restoran dan Rumah Makan Populer Kota Makassar Berdasarkan Rekomendasi Tripadvisor menggunakan Metode CRISP-DM dan SERVQUAL', *Building of Informatics, Technology and Science (BITS)*, vol. 4, no. 4, pp. 1899–1914, 2023, doi: 10.47065/bits.v4i4.3231.
[32] L. Hickman, S. Thapa, L. Tay, M. Cao, and P. Srinivasan, '*Text Preprocessing for Text Mining in Organizational Research: Review and Recommendations'*, *Organizational Research Methods*, vol. 25, no. 1, pp. 114–146, 2022, doi: 10.1177/1094428120971683.
[33] M. Kamyab, G. Liu, and M. Adjeisah, 'Attention-Based CNN and Bi-LSTM Model *Based on TF-IDF and GloVe Word Embedding for Sentiment Analysis'*, *Applied Sciences (Switzerland)*, vol. 11, no. 23, 2021, doi: 10.3390/app112311255.
[34] M. Liang and T. Niu, '*Research on Text Classification Techniques Based on Improved TF-IDF Algorithm and LSTM Inputs*', *Procedia Computer Science*, vol. 208, pp. 460–470, 2022, doi: 10.1016/j.procs.2022.10.064.
[35] I. Wickramasinghe and H. Kalutarage, '*Naive Bayes: Applications, Variations and Vulnerabilities: A Review of Literature with Code Snippets for Implementation'*, *Soft Comput*, vol. 25, no. 3, pp. 2277–2293, Feb. 2021, doi: 10.1007/s00500-020-05297-6.
[36] W. A. Prabowo and C. Wiguna, 'Sistem Informasi UMKM Bengkel Berbasis Web menggunakan Metode SCRUM', *mib*, vol. 5, no. 1, p. 149, Jan. 2021, doi: 10.30865/mib.v5i1.2604.
[37] O. Baines, '*Naïve Bayes: Machine Learning and Text Classification Application of Bayes*' Theorem'.
[38] J. Suzuki, '*Support Vector Machine', in Statistical Learning with Math and R: 100 Exercises for Building Logic, Singapore: Springer Nature Singapore*, 2020, pp. 171–192. doi: 10.1007/978-981-15-7568-6_9.
[39] S. Rabbani, D. Safitri, N. Rahmadhani, A. A. F. Sani, and M. K. Anam, 'Perbandingan Evaluasi Kernel SVM untuk Klasifikasi Sentimen dalam Analisis Kenaikan Harga BBM', *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 3, no. 2, pp. 153–160, 2023, doi: 10.57152/malcom.v3i2.897.