

Identifikasi Ujaran Kebencian Etnis Dan Agama Menggunakan Algoritma SVM

^{1*}Bayu Ardiansyah, ²Risky Aswi Ramadhani, ³Julian Sahertian

^{12,3} Teknik Informatika, Universitas Nusantara PGRI Kediri

E-mail: ¹bayuardiansyah965@gmail.com, ²riskyaswi@gmail.com,

³juliansahertian@unpkediri.ac.id

Penulis Korespondens : Bayu Ardiansyah

Abstrak— Perkembangan media sosial yang semakin pesat turut meningkatkan penyebaran ujaran kebencian berbasis etnis dan agama. Penelitian ini bertujuan untuk menganalisis tingkat akurasi algoritma Support Vector Machine (SVM) dalam mendeteksi ujaran kebencian pada data teks media sosial. Dataset diperoleh melalui teknik web scraping pada platform Twitter (X) dan menghasilkan 1.769 data teks. Data kemudian melalui tahap preprocessing yang meliputi case folding, tokenizing, stopword removal, dan stemming. Pelabelan data dilakukan menggunakan pendekatan lexicon-based classification, sedangkan ekstraksi fitur menggunakan metode Term Frequency-Inverse Document Frequency (TF-IDF). Model SVM selanjutnya dilatih dan dievaluasi menggunakan confusion matrix. Hasil penelitian menunjukkan bahwa model mampu mengklasifikasikan data ujaran kebencian dan non-ujaran kebencian dengan akurasi sebesar 88%. Hasil tersebut menunjukkan bahwa kombinasi preprocessing, TF-IDF, dan SVM efektif dalam mendukung deteksi otomatis ujaran kebencian berbasis etnis dan agama pada media sosial.

Kata Kunci— ujaran kebencian, Support Vector Machine, TF-IDF

Abstract— The rapid growth of social media has increased the spread of hate speech related to ethnicity and religion. This study aims to analyze the classification accuracy of the Support Vector Machine (SVM) algorithm in detecting hate speech from social media text data. The dataset was collected through web scraping on Twitter (X), resulting in 1,769 text records. The data were processed through several preprocessing stages, including case folding, tokenizing, stopword removal, and stemming. Data labeling was performed using a lexicon-based classification approach, while feature extraction employed the Term Frequency-Inverse Document Frequency (TF-IDF) method. The SVM model was then trained and evaluated using a confusion matrix. The results show that the model achieved an accuracy of 88% in classifying hate speech and non-hate speech data. These findings indicate that the combination of preprocessing techniques, TF-IDF, and SVM is effective for supporting automated detection of ethnicity- and religion-related hate speech on social media platforms..

Keywords— hate speech, Support Vector Machine, TF-IDF

This is an open access article under the CC BY-SA License.



I. PENDAHULUAN

Perkembangan teknologi informasi dan komunikasi telah mendorong perubahan signifikan dalam cara masyarakat berinteraksi dan bertukar informasi. Media sosial menjadi salah satu platform yang paling banyak digunakan untuk menyampaikan pendapat, berbagi informasi, serta berdiskusi mengenai berbagai isu sosial. Menurut laporan [1], Indonesia termasuk negara dengan jumlah pengguna media sosial terbesar di dunia, yaitu lebih dari 170 juta pengguna aktif. Tingginya intensitas penggunaan media sosial memberikan berbagai manfaat dalam penyebaran

informasi, namun di sisi lain juga meningkatkan risiko munculnya konten negatif, salah satunya adalah ujaran kebencian (*hate speech*).

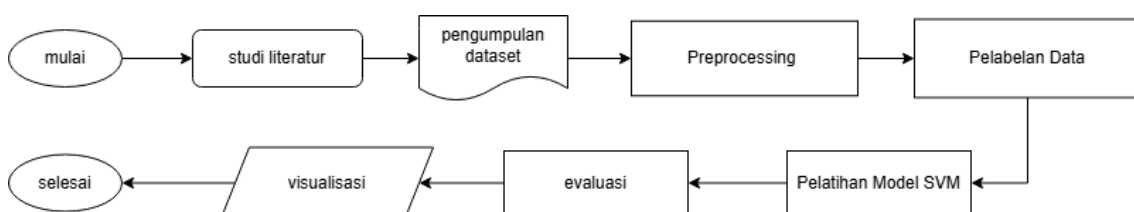
Ujaran kebencian yang berkaitan dengan etnis dan agama merupakan permasalahan yang perlu mendapat perhatian serius karena dapat memicu konflik sosial dan mengganggu kerukunan masyarakat. Di lingkungan media sosial, ujaran kebencian sering disampaikan dalam bentuk komentar, unggahan, maupun tanggapan yang mengandung unsur penghinaan, diskriminasi, atau provokasi terhadap kelompok tertentu. Penelitian yang dilakukan oleh [2] menunjukkan bahwa ujaran kebencian berbasis suku, agama, ras, dan antargolongan (SARA) masih banyak ditemukan pada berbagai platform media sosial dengan karakteristik bahasa yang bersifat provokatif dan diskriminatif. Temuan tersebut menunjukkan pentingnya pengembangan sistem otomatis yang mampu mengidentifikasi ujaran kebencian secara cepat dan akurat.

Salah satu pendekatan yang banyak digunakan untuk mengatasi permasalahan tersebut adalah penerapan teknik *machine learning* dalam klasifikasi teks. Di antara berbagai algoritma yang tersedia, *Support Vector Machine* (SVM) dikenal memiliki performa yang baik dalam menangani data teks berdimensi tinggi. Penelitian oleh [3] menunjukkan bahwa SVM mampu menghasilkan tingkat klasifikasi yang tinggi pada data teks berbahasa Indonesia. Selain itu, [4] menyatakan bahwa SVM tetap kompetitif dibandingkan beberapa algoritma klasifikasi lainnya, terutama dalam pengolahan teks pendek yang umum ditemukan pada media sosial.

Meskipun demikian, deteksi ujaran kebencian dalam Bahasa Indonesia masih menghadapi berbagai tantangan. Variasi penggunaan bahasa, singkatan, bahasa tidak baku, serta konteks percakapan yang beragam dapat memengaruhi kemampuan model dalam mengenali pola ujaran kebencian. Oleh karena itu, proses *text preprocessing* dan pemilihan representasi fitur menjadi tahapan penting untuk meningkatkan kualitas klasifikasi. Penelitian [5] menunjukkan bahwa penggunaan fitur TF-IDF dan n-gram mampu membantu model dalam menangkap pola linguistik yang relevan untuk membedakan ujaran kebencian dan non-ujaran kebencian. Hasil serupa juga ditemukan oleh [6] yang menyatakan bahwa kombinasi teknik ekstraksi fitur yang tepat dapat meningkatkan sensitivitas model terhadap konteks ujaran pada media sosial.

Berdasarkan permasalahan tersebut, penelitian ini menerapkan algoritma *Support Vector Machine* (SVM) untuk mendeteksi ujaran kebencian etnis dan agama pada data teks media sosial. Fokus utama penelitian adalah menganalisis tingkat akurasi klasifikasi yang dihasilkan oleh model SVM setelah melalui tahapan *preprocessing* dan ekstraksi fitur. Hasil penelitian diharapkan dapat memberikan gambaran mengenai efektivitas SVM dalam mengidentifikasi ujaran kebencian serta menjadi referensi bagi pengembangan sistem moderasi konten digital yang lebih akurat.

II. METODE



Gambar 1 alur penelitian

Pada gambar 1. Menggambarkan alur penelitian yang dijelaskan sebagai berikut:

1. Studi Literatur

Tahap studi literatur dilakukan dengan mengumpulkan dan menganalisis berbagai referensi yang berkaitan dengan ujaran kebencian, analisis sentimen, teknik ekstraksi fitur TF-IDF, serta algoritma *Support Vector Machine* (SVM). Selain teori dasar, penelitian terdahulu yang relevan juga dikaji untuk memperoleh pemahaman mengenai metode yang digunakan, hasil yang dicapai, serta peluang pengembangan penelitian. Kajian literatur tersebut menjadi dasar dalam penyusunan kerangka penelitian, pemilihan metode klasifikasi, serta penentuan parameter yang digunakan dalam proses deteksi ujaran kebencian berbasis teks.

2. Pengumpulan Dataset

Pengumpulan data dilakukan dengan menerapkan teknik web scraping menggunakan bahasa pemrograman Python untuk mengekstraksi data dari platform Twitter (X). Proses pengambilan data dilaksanakan pada rentang waktu tahun 2025 dengan mengumpulkan berbagai postingan yang berpotensi mengandung ujaran kebencian berbasis etnis dan agama. Data yang berhasil diperoleh kemudian digunakan sebagai sumber data utama dalam proses *preprocessing*, ekstraksi fitur, dan klasifikasi menggunakan algoritma *Support Vector Machine* (SVM).

3. Preprocessing

Sebelum memasuki tahap analisis dan klasifikasi, dataset terlebih dahulu melalui proses *preprocessing* untuk meningkatkan kualitas data serta mengurangi keberadaan *noisy* data yang dapat memengaruhi kinerja model [7]. Tahap ini bertujuan untuk membersihkan dan menormalisasi teks sehingga informasi yang terkandung dalam data dapat diolah secara lebih efektif. Proses *preprocessing* yang dilakukan meliputi beberapa tahapan, yaitu *case folding* untuk mengubah seluruh huruf menjadi huruf kecil agar format teks menjadi seragam, *tokenizing* untuk memecah kalimat menjadi kumpulan kata atau token, *filtering* dan *stopword removal* untuk menghilangkan kata-kata yang tidak memiliki kontribusi signifikan terhadap proses klasifikasi, serta *stemming* untuk mengubah kata berimbuhan menjadi bentuk kata dasar. Melalui serangkaian tahapan tersebut, data teks menjadi lebih terstruktur dan relevan sehingga dapat meningkatkan akurasi serta efektivitas algoritma *Support Vector Machine* (SVM) dalam mendeteksi ujaran kebencian berbasis etnis dan agama.

4. Pelabelan Data

Setelah melalui tahap *preprocessing*, data teks yang telah dibersihkan kemudian diberikan label menggunakan pendekatan *lexicon-based classification* [8]. Proses pelabelan dilakukan dengan memanfaatkan kamus (*lexicon*) yang berisi kumpulan kata atau frasa yang berkaitan dengan ujaran kebencian berbasis etnis dan agama.

$$Score = \sum_{i=1}^n w_i \quad (1)$$

Keterangan:

Score = skor total

w_i = bobot kata ke-*i* yang terdapat dalam kamus *lexicon*

n = jumlah kata dalam dokumen yang cocok dengan *lexicon*

Setiap dokumen dianalisis berdasarkan kemunculan kata-kata yang terdapat dalam kamus tersebut untuk menentukan kategori data. Apabila suatu teks mengandung kata atau frasa yang

terindikasi sebagai ujaran kebencian, maka data akan diberi label ujaran kebencian (*hate speech*), sedangkan teks yang tidak memenuhi kriteria tersebut akan diberi label non-ujaran kebencian (*non-hate speech*). Tahap pelabelan ini bertujuan untuk menghasilkan dataset yang terstruktur dan siap digunakan pada proses pelatihan serta pengujian model *Support Vector Machine* (SVM) dalam mendeteksi ujaran kebencian secara otomatis.

5. Model SVM

Setelah data melalui tahap *preprocessing*, ekstraksi fitur TF-IDF, dan pelabelan, data dibagi menjadi data latih dan data uji. Data latih digunakan untuk membangun model *Support Vector Machine* (SVM), sedangkan data uji digunakan untuk mengukur kemampuan model dalam mengklasifikasikan data baru.

SVM merupakan algoritma klasifikasi yang bekerja dengan mencari *hyperplane* optimal untuk memisahkan dua kelas [9], yaitu ujaran kebencian dan non-ujaran kebencian. *Hyperplane* tersebut dipilih berdasarkan margin terbesar sehingga dapat meningkatkan kemampuan model dalam melakukan klasifikasi secara akurat. Hasil prediksi dari model SVM kemudian digunakan pada tahap evaluasi untuk mengetahui performa klasifikasi yang dihasilkan.

$$f(x) = w * x + b \quad (2)$$

Keterangan

w = bobot (weight vector)

x = vector fitur

b = bias

6. Evaluasi

Evaluasi model dilakukan untuk mengukur kinerja algoritma *Support Vector Machine* (SVM) dalam mengklasifikasikan data ke dalam kategori ujaran kebencian dan non-ujaran kebencian. Proses pengujian dilakukan dengan membandingkan hasil prediksi yang dihasilkan oleh model terhadap label aktual pada data uji. Untuk menilai performa klasifikasi, digunakan *confusion matrix* yang menyajikan informasi mengenai jumlah prediksi benar maupun prediksi salah pada masing-masing kelas. Melalui *confusion matrix*, kinerja model dapat dianalisis secara lebih detail berdasarkan beberapa metrik evaluasi, seperti *accuracy*, *precision*, *recall*, dan *F1-score* [10]. Metrik-metrik tersebut digunakan untuk mengetahui tingkat ketepatan model dalam mendeteksi ujaran kebencian serta mengevaluasi kemampuan model dalam membedakan data ujaran kebencian dan non-ujaran kebencian.

$$accuracy = \frac{(TP+TN)}{(TP+FP+FN+TN)} \quad (3)$$

$$precision = \frac{(TP)}{(TP+FP)} \quad (4)$$

$$recall = \frac{(TP)}{(TP+FN)} \quad (5)$$

Keterangan:

TP = *True Positive* adalah jumlah data ujaran kebencian yang berhasil diprediksi dengan benar sebagai ujaran kebencian.

TN= *True Negative* adalah jumlah data non-ujaran kebencian yang berhasil diprediksi dengan benar sebagai non-ujaran kebencian.

FP = *False Positive* adalah jumlah data non-ujaran kebencian yang salah diprediksi sebagai ujaran kebencian.

FN= *False Negative* adalah jumlah data ujaran kebencian yang salah diprediksi sebagai non-ujaran kebencian.

7. Visualisasi

Tahap visualisasi dilakukan untuk menyajikan hasil analisis dan klasifikasi ujaran kebencian secara interaktif menggunakan *framework Streamlit*. Visualisasi bertujuan untuk memudahkan pengguna dalam memahami hasil pengolahan data, distribusi kelas, serta performa model *Support Vector Machine* (SVM) yang telah dibangun.

III. HASIL DAN PEMBAHASAN

1. Pengumpulan Data

Pengumpulan data dilakukan menggunakan teknik web *scraping* pada platform Twitter (X) dengan bantuan *library Playwright* pada bahasa pemrograman Python. Data yang dikumpulkan berupa postingan yang berkaitan dengan ujaran kebencian berbasis etnis dan agama. Dari proses pengumpulan data tersebut, diperoleh sebanyak 1.769 data teks yang selanjutnya digunakan sebagai dataset penelitian. Data yang terkumpul kemudian diproses pada tahap berikutnya.

2. Preprocessing

a. Case folding

Case folding merupakan proses mengubah seluruh huruf pada teks menjadi huruf kecil (*lowercase*) untuk menyeragamkan format penulisan. Tahap ini dilakukan agar perbedaan penggunaan huruf kapital dan huruf kecil tidak memengaruhi proses pengolahan data. Hasil proses *case folding* ditunjukkan pada Tabel 1.

Tabel 1 case folding

No	Komentar	Case folding
1	Apa yg diarepin dari islamofasis kaya mereka? Di dunia itu cuma ada islam dan non islam bagi mereka mah.	apa yg diarepin dari islamofasis kaya mereka? di dunia itu cuma ada islam dan non islam bagi mereka mah.
2	Habis dipake, mayatnya dibuang. Memang kebiasaan islam	habis dipake, mayatnya dibuang. memang kebiasaan islam

b. Tokenizing

Tokenizing merupakan proses memecah teks menjadi kata-kata tunggal (token) serta menghapus karakter yang tidak diperlukan, seperti tanda baca. Tahap ini dilakukan untuk memudahkan proses pengolahan data pada tahap selanjutnya. Hasil proses *tokenizing* ditunjukkan pada Tabel 2.

Tabel 2 tokenizing

No	Komentar	Tokenizing
1	Habis dipake, mayatnya dibuang. Memang kebiasaan islam	['habis', 'dipake', 'mayatnya', 'dibuang', 'memang', 'kebiasaan', 'islam']
2	chutiya nikamma nalayak muslim islam religion https xcomamritast amritastatus	['chutiya', 'nikamma', 'nalayak', 'muslim', 'islam', 'religion', 'https', 'xcomamritast', 'amritastatus']

c. Stopword removal

Stopword removal merupakan proses menghapus kata-kata yang tidak memiliki pengaruh signifikan terhadap proses klasifikasi, seperti kata sambung, kata depan, kata ganti, dan kata umum lainnya. Tahap ini bertujuan untuk menyisakan kata-kata yang lebih relevan sehingga dapat meningkatkan kualitas analisis teks. Hasil proses *stopword removal* ditunjukkan pada Tabel 3.

Tabel 3 stopwords removal

No	Komentar	Stopword removal
1	['habis', 'dipake', 'mayatnya', 'dibuang', 'memang', 'kebiasaan', 'islam']	['habis', 'dipake', 'mayatnya', 'dibuang', 'kebiasaan', 'islam']
2	['penyusup', 'diamdiam', 'nih', 'mereka', 'tuh', 'duaduanya', 'sender', 'dan', 'pendukung', 'yang', 'diamdiam', 'zionisntreal', 'kayak', 'setuju', 'segala', 'hal', 'kecuali', 'stance', 'islam', 'dan', 'palestina', 'apa', 'cuma', 'pengen', 'ributribut', 'doang', 'ya']	['penyusup', 'diamdiam', 'nih', 'tuh', 'duaduanya', 'sender', 'pendukung', 'diamdiam', 'zionisntreal', 'kayak', 'setuju', 'kecuali', 'stance', 'islam', 'palestina', 'pengen', 'ributribut', 'doang', 'ya']

d. Stemming

Stemming merupakan proses mengubah kata yang memiliki imbuhan menjadi bentuk kata dasar. Tahap ini bertujuan untuk menyeragamkan bentuk kata sehingga memudahkan proses analisis dan pengolahan data teks. Hasil proses *stemming* ditunjukkan pada Tabel 4.

Tabel 4 stemming

No	Komentar	stemming
1	habis dipake mayatnya dibuang memang kebiasaan islam	['habis', 'dipake', 'mayat', 'buang', 'biasa', 'islam']

2	apa yg diarepin dari islamofasis kaya mereka di dunia itu cuma ada islam dan non islam bagi mereka mah	['yg', 'diarepin', 'islamofasis', 'kaya', 'dunia', 'islam', 'non', 'islam', 'mah']
---	--	--

3. Evaluasi

Evaluasi model dilakukan untuk mengukur kinerja algoritma *Support Vector Machine* (SVM) dalam mengklasifikasikan data ujaran kebencian dan non-ujaran kebencian. Pengujian dilakukan menggunakan *confusion matrix* untuk mengetahui jumlah prediksi yang benar maupun kesalahan klasifikasi pada setiap kelas. Hasil evaluasi model menggunakan *confusion matrix* ditampilkan pada tabel 5.

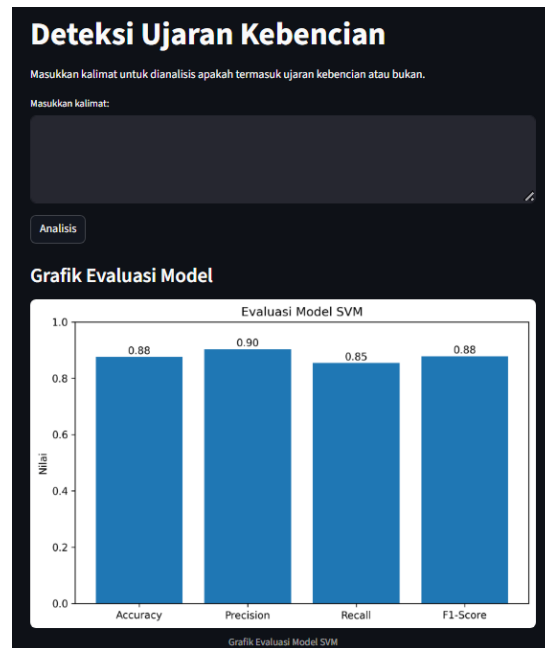
Tabel 5 confusion matrix

Confusion Matrix				
	Precision	Recall	F1-score	Support
Netral	0.85	0.90	0.87	169
Ujaran kebencian	0.90	0.85	0.88	185
Accuracy			0.88	354

Berdasarkan hasil evaluasi model pada Tabel 5, algoritma *Support Vector Machine* (SVM) memperoleh nilai akurasi sebesar 88%, yang menunjukkan bahwa model mampu mengklasifikasikan data dengan tingkat ketepatan yang baik. Pada kelas Netral, model menghasilkan nilai *precision* sebesar 0,85, *recall* 0,90, dan *F1-score* 0,87 dari total 169 data. Nilai *recall* yang tinggi menunjukkan bahwa sebagian besar data netral berhasil dikenali dengan benar oleh model.

Sementara itu, pada kelas Ujaran Kebencian, model memperoleh *precision* sebesar 0,90, *recall* 0,85, dan *F1-score* 0,88 dari total 185 data. Nilai *precision* yang lebih tinggi pada kelas ini menunjukkan bahwa sebagian besar data yang diprediksi sebagai ujaran kebencian memang termasuk ke dalam kategori tersebut.

4. Visualisasi



Gambar 2 visualisasi

Gambar 2 menampilkan antarmuka aplikasi deteksi ujaran kebencian yang dibangun menggunakan *framework streamlit*. Aplikasi ini menyediakan fitur input teks yang memungkinkan pengguna memasukkan kalimat untuk dianalisis oleh model *Support Vector Machine* (SVM). Selain itu, aplikasi juga menampilkan grafik evaluasi model sebagai bentuk visualisasi hasil pengujian yang telah dilakukan. Visualisasi tersebut bertujuan untuk memudahkan pengguna dalam memahami performa model klasifikasi yang digunakan pada penelitian ini.

IV. KESIMPULAN

Penelitian ini menunjukkan bahwa algoritma *Support Vector Machine* (SVM) dapat diterapkan secara efektif untuk mendeteksi ujaran kebencian berbasis etnis dan agama pada data teks media sosial. Melalui tahapan *preprocessing*, pelabelan menggunakan pendekatan *lexicon-based*, serta ekstraksi fitur TF-IDF, model mampu membedakan data ujaran kebencian dan non-ujaran kebencian dengan baik. Akurasi klasifikasi yang mencapai 88% menunjukkan bahwa SVM memiliki kemampuan yang baik dalam mengenali pola ujaran kebencian pada data teks berbahasa Indonesia. Dengan demikian, penelitian ini memberikan kontribusi dalam pengembangan sistem deteksi ujaran kebencian berbasis *machine learning* yang dapat mendukung proses moderasi konten digital secara lebih cepat, objektif, dan akurat.

DAFTAR PUSTAKA

- [1] W. A. Social and Meltwater, "Digital 2024: Indonesia report," 2024. [Online]. Available: <https://datareportal.com/reports/digital-2024-indonesia>
- [2] D. A. F. Sarie, Suhartono, and Mintowati, "Ujaran Kebencian Di Media Sosial (Kajian Pragmasemantik)," *J. Educ. Dev.*, vol. 9, no. 4, pp. 247–251, 2021.
- [3] E. W. Pamungkas, D. G. P. Putri, and A. Fatmawati, "Hate Speech Detection in Bahasa Indonesia: Challenges and Opportunities," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 6, pp. 1175–1181, 2023, doi: 10.14569/IJACSA.2023.01406125.
- [4] M. O. Ibrohim and I. Budi, "Hate speech and abusive language detection in Indonesian social media: Progress and challenges," *Heliyon*, vol. 9, no. 8, p. e18647, 2023, doi: 10.1016/j.heliyon.2023.e18647.
- [5] N. Aulia and I. Budi, "Hate speech detection on Indonesian long text documents using machine learning approach," *ACM Int. Conf. Proceeding Ser.*, no. April 2019, pp. 164–169, 2019, doi: 10.1145/3330482.3330491.
- [6] K. Hadi and E. Utami, "Analysis of K-NN with the Integration of Bag of Words, TF-IDF, and N-Grams for Hate Speech Classification on Twitter," *JUITA J. Inform.*, vol. 12, no. 2, p. 289, 2024, doi: 10.30595/juita.v12i2.23829.
- [7] S. A. Helmayanti, F. Hamami, and R. Y. Farifah, "Penerapan Algoritma TF-IDF dan Naïve Bayes untuk Analisis Sentimen Berbasis Aspek Ulasan Aplikasi Flip pada Google Play Store," *J. Indones. Manaj. Inform. dan Komun.*, vol. 4, no. 3, pp. 1822–1834, 2023, doi: 10.35870/jimik.v4i3.415.
- [8] N. M. Al Ghazali and Y. Sibaroni, "Sentiment Classification in E-Commerce Using Naïve Bayes and Combined Lexicon - N-Gram Features," *JIPi (Jurnal Ilm. Penelit. dan Pembelajaran Inform.*, vol. 10, no. 2, pp. 1257–1271, 2025, doi: 10.29100/jipi.v10i2.6157.
- [9] I. Riadi, A. Fadlil, and M. Murni, "Identifying Hate Speech in Tweets with Sentiment Analysis on Indonesian Twitter Utilizing Support Vector Machine Algorithm," *Khazanah Inform. J. Ilmu Komput. dan Inform.*, vol. 9, no. 2, pp. 179–191, 2023, doi: 10.23917/khif.v9i2.22470.
- [10] K. M. Suryaningrum, "Comparison of the TF-IDF Method with the Count Vectorizer to Classify Hate Speech," *Eng. Math. Comput. Sci. J.*, vol. 5, no. 2, pp. 79–83, 2023, doi: 10.21512/emacsjournal.v5i2.9978.