Hate Speech and Emotions Classification in Indonesian Language Texts on Twitter Using Naïve Bayes Classifier

[Klasifikasi *Hate Speech* dan Emosi Dalam Teks Berbahasa Indonesia Pada Pengguna Twitter Menggunakan Metode Naïve Bayes Classifier]

Chandra Hary Pratama [1)], Yulian Findawati [2)]
[1)] Program Studi Teknik Informatika, Universitas Muhammadiyah Sidoarjo, Indonesia
[2)] Program Studi Teknik Informatika, Universitas Muhammadiyah Sidoarjo, Indonesia
*Email Penulis Korespondensi: yulianfindawati@umsida.ac.id

Abstract. *Hate speech is a form of expression that incites, spreads, justifies, or encourages hatred, discrimination and violence against individuals and groups for various reasons. Hate speech is usually found on social media connected to the internet, one of which is in this study through social media twitter using the Naïve Bayes Classifier method. The dataset used in this study amounted to 1800 data labeled not hate speech and 2250 data labeled hate speech with a comparison of 60% training data and 40% test data. The results of the evaluation of test data with confusion matrix obtained measurements of matrix mean accuracy for hate speech classification 0.89 and matrix mean accuracy for emotion classification 0.59. Based on the results obtained, it can be concluded that to classify hate speech and emotions on Twitter using Naïve Bayes, the best results with the Confusion Matrix without selecting the Information Gain feature.*

Keywords - *Clasification, Hate Speech. Emotion, Naïve Bayes, Twitter*

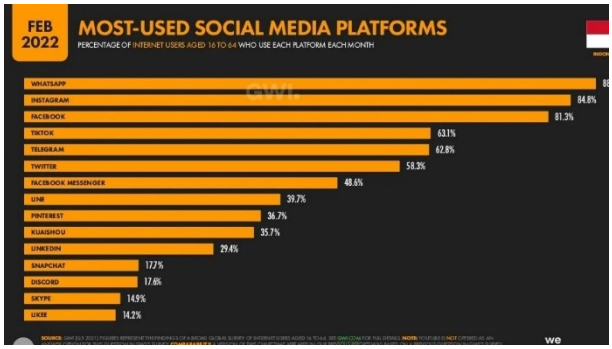
Abstrak. *Ujaran kebencian merupakan salah satu bentuk ekspresi yang menghasut, menyebarkan, membenarkan, atau mendorong kebencian, diksriminasi serta kekerasan atas individu dan kelompok sebab berbagai alasan. Hate speech biasanya ditemukan pada sosial media yang terhubung dengan internet, salah satunya pada penelitian ini melalui sosial media twitter dengan menggunakan metode Naïve Bayes Classifier. Dataset yang digunakan pada penelitian ini berjumlah 1800 data berlabel bukan ujaran kebencian dan 2250 data berlabel ujaran kebencian dengan perbandinghan 60% data latih dan 40% data uji. Hasil evaluasi data uji dengan confusion matrix diperoleh pengukuran matrix mean accuracy for hate speech classification 0,89 dan matrix mean accuracy for emotion classification 0,59. Berdasarkan hasil yang didapat tersebut dapat diambil kesimpulan bahwa untuk melakukan klasifikasi hate speech dan emosi pada Twitter menggunakan Naïve Bayes hasil paling bagus dengan Confusion Matrix tanpa melakukan seleksi fitur Information Gain.*

Kata Kunci - *Klasifikasi, Ujaran Kebencian. Emosi, Naïve Bayes, Twitter*

I. PENDAHULUAN

Hate speech atau ujaran kebencian adalah suatu bentuk ekspresi yang menghasut, menyebarkan, membenarkan, atau mendorong kebencian, diskriminasi serta kekerasan atas individu dan kelompok sebab berbagai alasan [1]. *Hate speech* atau ujaran kebencian tidak jarang kita jumpai pada kehidupan sehari-hari. Ujaran kebencian sangat sering digunakan pada status, komentar, atau postingan pada media sosial.

Linschoten [2] menjelaskan bahwa emosi orang dibagi menjadi tiga bagian menurut kategorinya, yaitu suasana hati, suasana perasaan, dan emosi. Dalam arti luas emosi adalah salah satu bagian dari perasaan. Emosi hadir dari perasaan yang bergolak, sehingga yang terlibat dapat mengalami perubahan dalam situasi emosi tersendiri. Pada penelitian [3] menunjukkan bahwa emosi tertentu seperti kemarahan dan kebencian lebih berkorelasi dengan ujaran kebencian di *Twitter*. Klasifikasi emosi terdiri dari *"Anger", "Anticipation", "Disgust", "Fear", "Joy", "Sadness", "Surprise" dan "Trust".*



Gambar 1. We Are Social Hootsuite (2022)

Survei *We Are Social* menyebutkan dalam tinjauan media sosial penduduk Indonesia yang aktif bermain media sosial mencapai 4,62 milyar orang pada tahun 2022, dan jumlah perangkat mobile yang terhubung mencapai 8,28 milyar. Media sosial merupakan sebuah wadah yang sering disalahgunakan sebagai tempat ujaran kebencian dan berekspresi. Twitter merupakan salah satu media sosial yang paling banyak digunakan. Pada survei *We Are Social*, Twitter menempati peringkat ke enam dengan presentase 58,3%. Pengguna twitter seringkali melayangkan komentar, status, bahkan postingan yang mengandung *hate speech* dan ekspresi emosi. Pengguna diberikan suatu kebebasan untuk menyalurkan ekspresi dan perasaan emosi di twitter [4].

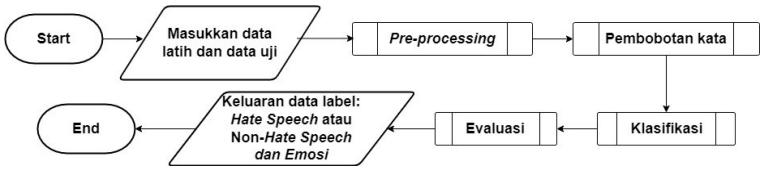
Penelitian mengenai *hate speech* dan emosi ini sebelumnya telah dilakukan dengan pembahasan mengenai deteksi *hate speech* bahasa Indonesia pada twitter [5]. Serta Deteksi sentimen di twitter menggunakan metode *Naive Bayes* [6] Hasil penelitian ini menunjukkan klasifikasi ujaran kebencian dan deteksi emosi pada akun twitter yang dimana dalam penelitian ini metode yang digunakan adalah *naïve bayes* berbasis n-gram serta seleksi fitur *Information Gain.*

Penelitian telah yang dilakukan oleh [7] memaparkan suatu koneksi negatif antara kematangan batin dan aksi ujaran kebencian (*Hate Speech*), semakin tinggi kematangan batin seseorang, semakin sedikit pula aksi ujaran kebencian (*Hate Speech*) dan juga sebaliknya. Maka dari itu pada penelitian ini akan menggabungkan kedua klasifikasi yaitu klasifikasi *hate speech* dan emosi dengan menggunakan metode *naïve bayes*.

II. METODE

Metodologi penelitian adalah suatu tahapan atau langkah-langkah yang akan dilakukan untuk memecahkan suatu permasalahan guna menemukan solusi yang tepat pada sebuah penelitian. Pada penelitian ini, penulis membahas studi kasus klasifikasi *hate speech* dan emosi terhadap pengguna twitter menggunakan *Naive Bayes Classifier* [8].

Penelitian ini dilandaskan dengan mengimplementasikan naive bayes classifier untuk mengklasifikasikan hate speech dan emosi berbahasa Indonesia terhadap pengguna twitter. Berikut merupakan langkah-langkah yang dilakukan:



Gambar 2. Diagram Alir Penelitian

Berdasarkan diagram alir pada Gambar 1dapat dilihat proses dalam tahapan penelitian yang akan menjadi acuan dalam pengerjaan penelitian ini.

Berikut adalah penjelasan dari flowchart Naïve Bayes Classifier:

1. Dataset
   Dataset pada penelitian ini diambil dari sosial media twitter. Kemudian dilakukan identifikasi kata dasar, identifikasi kata-kata yang sering muncul atau stopword serta dilakukan identifikasi kategori.
2. Text Pre-processing
   Pre-processing merupakan sebuah proses awal pengklasifikasikan dokumen dengan tujuan menyiapkan data agar data tersebut mempunyai struktur. Text pre-processing akan menghasilkan nilai yang digunakan sebagai data untuk dilakukan proses selanjutnya. Pre-processing dibagi menjadi beberapa proses antara lain case folding, tokenizing, filtering, stemming, dan penghitungan bobot kata [9].

3. Pembobotan Kata
   Pembobotan kata (Term Weighting) yaitu suatu mekanisme pemberian nilai pada setiap kata berdasarkan indeks [10].
4. Klasifikasi
   Tujuan dari proses klasifikasi ini yaitu untuk memperoleh hasil dari data uji untuk mendapatkan luaran berupa label hate speech atau non-hate speech dan emosi yang terkandung pada data yang telah dimasukkan. Semua proses atau langkah selesai ketika diperoleh hasil luaran berupa label kelas tersebut.
5. Evaluasi
   Tahapan akhir yang dilakukan adalah proses evaluasi, proses evaluasi yang bertujuan menguji hasil dari klasifikasi dengan cara pengukuran nilai kebenaran dari sistem tersebut.

III. HASIL DAN PEMBAHASAN

A. Dataset

Data dari media sosial Twitter yang berhasil dikumpulkan sebanyak 3.972 tweet. Memiliki 15 atribu tantara lain URL, Tanggal, Tweet, ID, Username, Likes, Quotes dan sebagainya yang disimpan dalam format .csv [11]. Data selanjutnya diolah ke tahapan preprocessinguntuk meningkatkan struktur dari data. Tabel 1 menunjukkan hasil crawling tweet.

Tabel 1. Sampel tweet hasil *Crawling*

NO	Tweet
1	@arunariftan makin gila lihat LBP mewarnai Indonesia tercinta, Barisan Saku Hati Kadrun2 sich ya mampus saja ha ha ha racain.
2	Ha ha ha sigundul penguasa ancol karena selama ini taunya hanya jilat2 gabenar mulai dibuka jeroannya, KPK Kejaksaan Agung Mabes Polri tolong segera turun/selidiki MERDEKA. https://t.co/Zzp3e2IZ67
.....	.....
3972	Buat KADRUN2 nih

B. Pre-processing

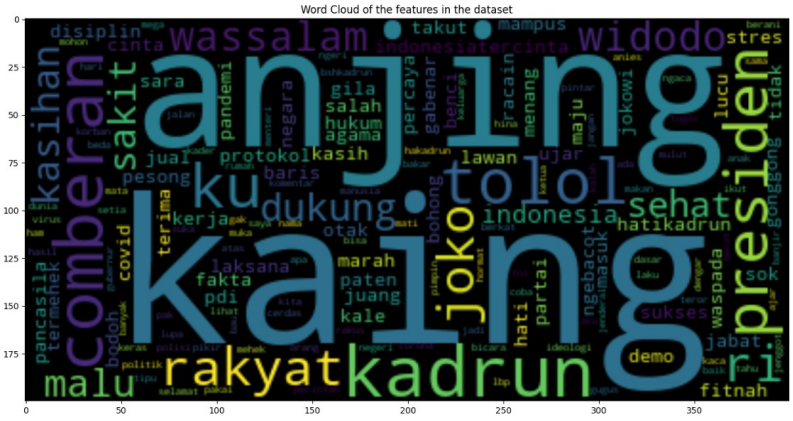
Tahapan preprocessingdata dilakukan untuk meningkatkan performa masing-masing algoritma klasifikasi dalam melakukan prediksi sehingga didapatkan data yang lebih presisi. Tahapan ini meliputi cleaningdata, Setelah dilakukan proses preprocessingdata jumlah data yang dihasilkan sejumlah 3.972 tweet.

Tabel 2. Proses *Preprocessing* Data

Proses	Tweet
Cleaning data	@arunariftan makin gila lihat LBP mewarnai Indonesia tercinta, Barisan Saku Hati Kadrun2 sich ya mampus saja ha ha ha racain.
Tokenization	[‘makin’, ‘gila’, ‘lihat’, ‘lbp’, ‘mewarnai’, ‘indonesia’, ‘tercinta’, ‘barisan’, ‘saku’, ‘hati’, ‘kadrun’, ’sich’, ‘ya’, ‘mampus’, ‘saja’, ‘ha’, ‘ha’, ‘ha’, ‘racain’]
Stopword Removal	[‘makin’, ‘gila’, ‘lihat’, ‘lbp’, ‘mewarnai’, ‘indonesia’, ‘tercinta’, ‘barisan’, ‘saku’, ‘hati’, ‘kadrun’, ’sich’, ‘ya’, ‘mampus’, ‘saja’, ‘ha’, ‘ha’, ‘ha’, ‘racain’]
Stemming	makin gila lihat lbp mewarnai indonesia tercinta barisan saku hati kadrun2 sich ya mampus saja ha ha ha racain

C. Pembobotan Kata

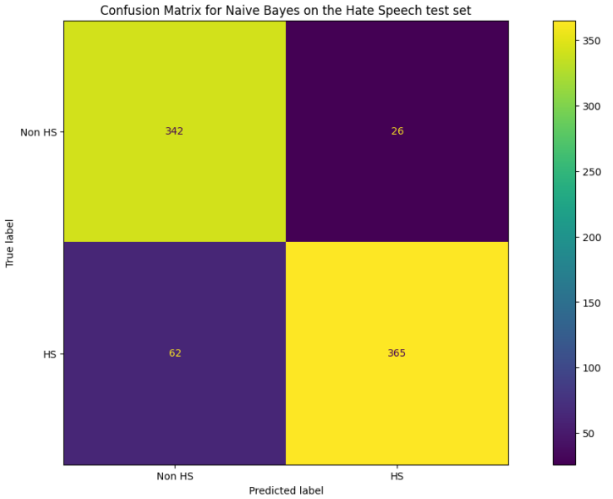
Pembobotan kata (Term Weighting) yaitu suatu mekanisme pemberian nilai pada setiap kata berdasarkan indeks. Proses pembobotan kata ini bertujuan untuk memperoleh jumlah kemunculan kata setelah dilakukan proses pre-processing pada dataset. Pada penelitian ini metode pembobotan kata yang digunakan adalah TF-IDF (Term Frequency Invers Document Frequency)[12]. Metode inidigunakan untuk menentukan keterhubungan kata terhadap dokumen dengan cara memberikan bobot pada setiap kata. Pada perhitungan TF-IDF terlebih dahulu menghitung nilai TF pada setiap kata, dengan bobot satu kata adalah 1.



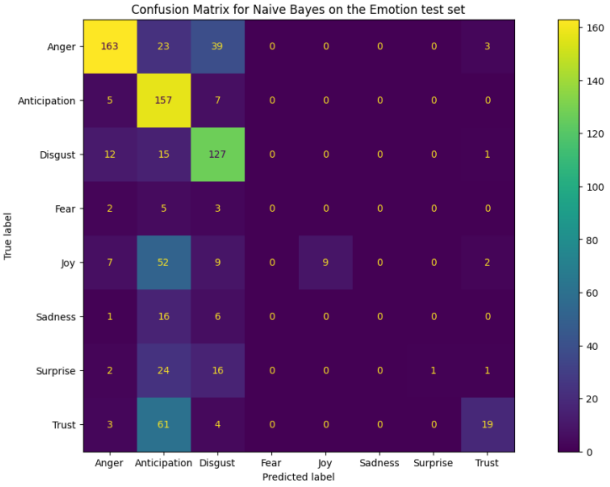
Gambar 3. *Word Cloud*

D. Klasifikasi

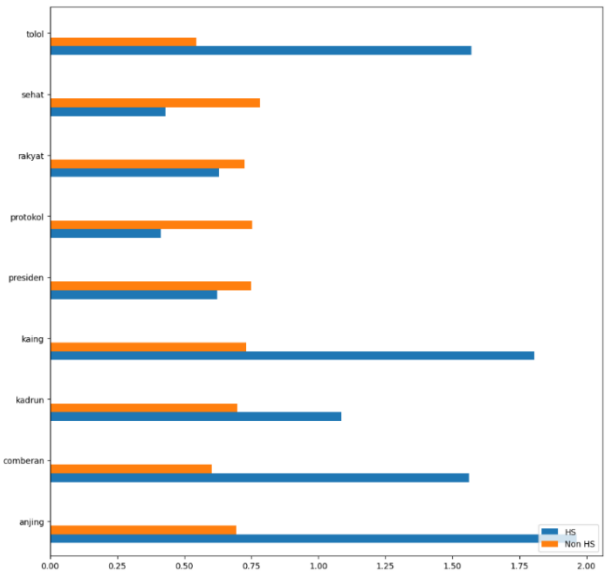
Setelah melalui tahapan proses pembobotan kata, langkah berikutnya adalah memasukkan data ke dalam fase pemodelan klasifikasi. Sebelumnya, data akan dibagi menjadi dua bagian, yakni data latih dan data uji, dengan rasio 60:40, di mana 60% digunakan untuk data latih dan 40% untuk data uji dalam eksperimen ini [13]. Data uji berperan sebagai alat evaluasi model, sementara data latih berfungsi untuk pembangunan model dan identifikasi pola. Setelah pembagian data diselesaikan, langkah selanjutnya melibatkan proses klasifikasi model. Algoritma pertama yang akan dilakukan percobaan adalah Naive Bayes [14].



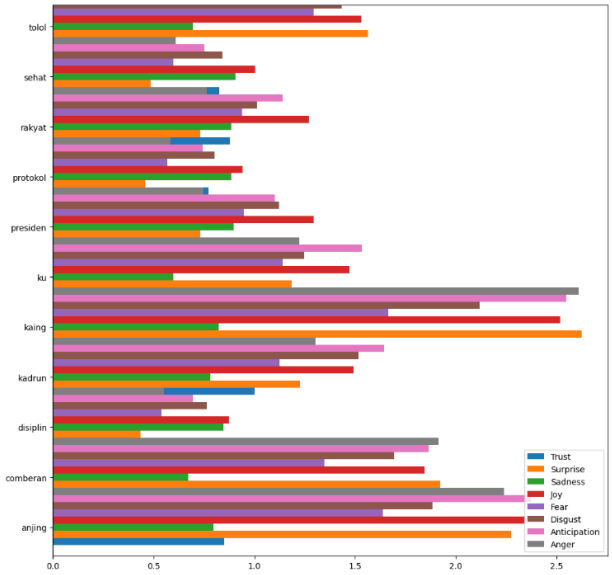
Gambar 4. *Confusion Matrix Hate Speech Set*



Gambar 5. *Confusion Matrix* Emosi



Gambar 6. Plot *Hate Speech*



Gambar 7. Plot Emosi

E. Evaluasi

Tahapan akhir yang dilakukan adalah proses evaluasi, proses evaluasi yang bertujuan menguji hasil dari klasifikasi dengan cara pengukuran nilai kebenaran dari sistem tersebut [15]. Tolok ukur yang digunakan sebagai acuan dalam mengukur adalah accuracy. Pada penelitian ini ekstraksi fitur yang digunakan adalah precision dengan persamaan sebagai berikut:

$$\text{Accuracy} = \left(\frac{TP + TN}{TP + TN + FP + FN}\right) \times 100\% \quad (1)$$

Tabel 3. *Mean Accuracy*

Model Algoritma	***Hate Speech***	**Emosi**
Mean Accuracy	0.889308176100629	0.5987421383647799

F. Output

Output yang dihasilkan dari program ini adalah untuk mendeteksi hate speech dan emosi dari sebuah kalimat random. Ada dua jenis output dari program ini yatu berupa keterangan HS/Non HS dan emosi.

Tabel 4. Output klasifikasi *Hate Speech* dan Emosi

Text	***Hate Speech***	**Emosi**
gin ngebacot tunjukin muka tampang ditwitter bawa islam islam islam sadar	True	Anger
ideologi indonesia pancasila	False	Anticipation
komentar dukung setia joko widodo presiden ri cinta rakyat kerja keras	True	Anger
makin gila lihat LBP mewarnai Indonesia tercinta, Barisan Saku Hati Kadrun2	True	Anticipation
sich ya mampus saja ha ha ha racain.		
kasihan korban akibat cuci otak kadrun tanggung terima kasih polisi paspampres		
waspada waspada waspada		

IV. SIMPULAN

Dengan Output dari program ini adalah hasil klasifikasi teks tweet berbahasa Indonesia yang telah dianalisis menggunakan algoritma Naïve Bayes Classifier. Program ini mengeluarkan dua jenis klasifikasi utama:

1. Klasifikasi Hate Speech: Program akan memberikan label pada setiap tweet apakah termasuk dalam kategori ujaran kebencian atau tidak. Ini dilakukan dengan memproses teks tweet dan membandingkannya dengan pola-pola yang telah dipelajari dari dataset pelatihan.
2. Klasifikasi Emosi: Selain mendeteksi hate speech, program juga mengidentifikasi emosi yang terkandung dalam teks tweet. Emosi yang dapat dideteksi misalnya marah, sedih, bahagia, dan lain-lain. Setiap tweet akan diberi label sesuai dengan emosi yang paling dominan berdasarkan analisis fitur linguistiknya.

Output akhir yang dihasilkan berupa data terstruktur yang menunjukkan tweet, label hate speech *("hate speech = true" atau " hate speech = false"*), dan label emosi (*"Anger", "Anticipation", "Disgust", "Fear", "Joy", "Sadness", "Surprise" dan "Trust"*). Data ini dapat digunakan untuk analisis lebih lanjut atau untuk tindakan moderasi konten di platform Twitter.

UCAPAN TERIMA KASIH

Penulis Semoga isi artikel ini memberikan manfaat dan pemahaman yang berharga. Terima kasih juga kepada semua pihak yang telah turut serta dalam proses penulisan artikel ini. Dukungan dan kontribusi dari berbagai pihak sangat berarti, dan tanpa mereka, artikel ini tidak akan terwujud. Teruslah mendukung dan membagikan pengetahuan, karena bersama-sama kita dapat menciptakan ruang diskusi yang bermanfaat.hasil.

REFERENSI

[1] I. Liu and Y. A. Sari, "Klasifikasi Hate Speech Berbahasa Indonesia di Twitter Menggunakan Naive Bayes dan Seleksi Fitur Information Gain dengan Normalisasi Kata," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 3, no. 5, pp. 4914–4922, 2019.

[2] S. Al Baqi, "Ekspresi Emosi Marah," *Bul. Psikol.*, vol. 23, no. 1, p. 22, 2015, doi: 10.22146/bpsi.10574.

[3] B. Martins, G. Sheppes, J. J. Gross, and M. Mather, "Age Differences in Emotion Regulation Choice: Older Adults Use Distraction Less Than Younger Adults in High-Intensity Positive Contexts," *Journals Gerontol. - Ser. B Psychol. Sci. Soc. Sci.*, vol. 73, no. 4, pp. 603–611, 2018, doi: 10.1093/geronb/gbw028.

[4] H. Ahmad Gozali and M. Alfan Rosid, "Classification of Student Complaints with Naive Bayes and Literary Methods Klasifikasi Keluhan Mahasiswa dengan Metode Naive Bayes dan Sastrawi," *Network, Comput. Sci. |*, vol. 3, no. 1, pp. 22–26, 2020.

[5] M. Hakiem, M. A. Fauzi, and Indriati, "Klasifikasi Ujaran Kebencian pada Twitter Menggunakan Metode Naïve Bayes Berbasis N-Gram Dengan Seleksi Fitur Information Gain," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 3, no. 3, pp. 2443–2451, 2019, [Online]. Available: http://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/4682

[6] F. Fanesya, R. C. Wihandika, and Indriati, "Deteksi Emosi pada Twitter Menggunakan Metode Naive Bayes dan Kombinasi Fitur," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 3, no. 7, p. 3, 2019.

[7] M. F. A. Afif, Y. Nurhamidah, and M. F. Mashuri, "Kematangan emosi dalam perilaku ujaran kebencian pada kebijakan politik," *Cognicia*, vol. 9, no. 1, pp. 25–30, 2021, doi: 10.22219/cognicia.v9i1.14234.

[8] Ahmad Wildan Attabi, Lailil Muflikhah, and Mochammad Ali Fauzi, "Penerapan Analisis Sentimen untuk Menilai Suatu Produk pada Twitter Berbahasa Indonesia dengan Metode Naïve Bayes Classifier dan Information Gain," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 11, pp. 4548–4554, 2018.

[9] I. G. A. Socrates, A. L. Akbar, M. S. Akbar, A. Z. Arifin, and D. Herumurti, "Optimasi Naive Bayes Dengan Pemilihan Fitur Dan Pembobotan Gain Ratio," *Lontar Komput. J. Ilm. Teknol. Inf.*, vol. 7, no. 1, p. 22, 2016, doi: 10.24843/lkjiti.2016.v07.i01.p03.

[10] T. E. Hidayat and A. Rosid, "Analysis of Community Sentiments Regarding Plans to Relocate National Capital Using the Naïve Bayes Method Analisa Sentimen Masyarakat Tentang Rencana Pemindahan Ibukota Negara Dengan Metode Naïve Bayes," *Network, Comput. Sci. |*, vol. 3, no. 2, pp. 43–49, 2020.

[11] A. Deolika, K. Kusrini, and E. T. Luthfi, "Analisis Pembobotan Kata Pada Klasifikasi Text Mining," *J. Teknol. Inf.*, vol. 3, no. 2, p. 179, 2019, doi: 10.36294/jurti.v3i2.1077.

[12] A. Kumari, "Study on Naive Bayesian Classifier and its relation to Information Gain," *Int. J. Recent Innov. Trends Comput. Commun.*, vol. 2, no. 3, pp. 601–602, 2014.

[13] A. P. J. Dwitama, "Deteksi Ujaran Kebencian Pada Twitter Bahasa Indonesia Menggunakan Machine Learning: Reviu Literatur," *J. Sains, Nalar, dan Apl. Teknol. Inf.*, vol. 1, no. 1, pp. 31–39, 2021, doi: 10.20885/snati.v1i1.5.

[14] T. Ghassani Saskia, "Klasifikasi Hate Speech Dan Abusive LanguagePada Twitter Bahasa Indonesia Dengan MetodeNaive Bayes Classifier," 2021.

[15] N. M. S. Hadna, P. I. Santosa, and W. W. Winarno, "Studi Literatur Tentang Perbandingan Metode Untuk Proses Analisis Sentimen Di Twitter," *Semin. Nas. Teknol. Inf. dan Komun.*, vol. 2016, no. March, pp. 57–64, 2016.

Conflict of Interest Statement:
The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.