Using K-Nearest Neighbor to Classify Hate Speech and Emotions on Twitter

[Penggunaan K-Nearest Neighbor untuk Klasifikasi Hate Speech dan Emosi pada Twitter]

Mochamad Yanuar Kusdianto [1)], Uce Indahyanti [2)]
[1)] Program Studi Teknik Informatika, Universitas Muhammadiyah Sidoarjo, Indonesia
[2)] Program Studi Teknik Informatika, Universitas Muhammadiyah Sidoarjo, Indonesia
*Email Penulis Korespondensi: uceindahyanti@umsida.ac.id

Abstract. *Hate speech is a form of expression that has a purpose against an individual or group. The expression can be in the form of inciting and spreading slander, justifying, or encouraging hatred, even discrimination and violence based on various reasons. Usually, hate speech is often found on social media that is connected to the internet. This study focuses on the Twitter platform using the K-Nearest Neighbor method. In this study, the dataset used consisted of 1842 data with the description "not hate speech" and 2130 data with the description "hate speech," which was divided into 80% training data and 20% test data. The results of the evaluation of the test data using the confusion matrix showed an average accuracy value for hate speech classification of 0.855, while for emotion classification it was 0.534. Based on these results, it can be concluded that the K-Nearest Neighbor algorithm is quite effective in analyzing and classifying hate speech and emotions in Twitter text with certain datasets.*

Keywords *- Clasification, Hate Speech. Emotion, K-Nearest Neighbor, Twitter*

Abstrak. *Hate speech ialah sebuah ungkapan yang memiliki tujuan terhadap individu atau kelompok. Ekspresi tersebut dapat berupa menghasut dan menyebarkan fitnah, membenarkan, atau mendorong kebencian, bahkan diskriminasi serta kekerasan dengan berdasarkan berbagai alasan. Biasanya, hate speech atau ujaran kebencian ini banyak ditemukan di media sosial yang terhubung dengan internet. Penelitian ini berfokus pada platform twitter dengan menggunakan metode k-nearest neighbor. Pada penelitian ini dataset yang digunakan terdiri dari 1842 data dengan keterangan "bukan ujaran kebencian" dan 2130 data dengan keteranganl "ujaran kebencian," yang dibagi menjadi 80% data latih dan 20% data uji. Hasil evaluasi data uji menggunakan confusion matrix menunjukkan nilai rata-rata akurasi untuk klasifikasi ujaran kebencian sebesar 0,855, sedangkan untuk klasifikasi emosi sebesar 0,534. Berdasarkan hasil tersebut, dapat disimpulkan bahwa algoritma k-nearest neighbor cukup efektif dalam menganalisis dan mengklasifikasikan ujaran kebencian serta emosi pada teks Twitter dengan dataset tertentu.*

Kata Kunci *- Klasifikasi, Ujaran Kebencian. Emosi, K-Nearest Neighbor, Twitter*

I. PENDAHULUAN

Perkembangan teknologi internet dan media sosial telah mengubah cara komunikasi dan interaksi manusia secara signifikan. Sebagai salah satu platform media sosial terbesar, twitter memungkinkan penggunanya untuk berbagi pendapat dan pandangan dalam bentuk tweet singkat yang dapat menjangkau audiens luas. Namun, kemudahan akses ini juga memicu penyebaran konten negatif, termasuk ujaran kebencian (*hate speech*). Hate speech di media sosial tidak hanya merusak tatanan sosial, tetapi juga dapat menyebabkan konflik antarindividu maupun kelompok [1].

Hate speech dan emosi adalah dua aspek penting dalam analisis teks di media sosial. Ujaran kebencian didefinisikan sebagai ujaran menyudutkan atau merendahkan berdasarkan atribut seperti agama, ras, atau gender [2]. Di sisi lain, analisis emosi bertujuan untuk mengidentifikasi perasaan atau ekspresi emosional dalam teks, seperti kemarahan, kebahagiaan, atau ketakutan. Kedua jenis analisis ini semakin penting, mengingat pengaruh signifikan media sosial terhadap perilaku dan persepsi masyarakat. Penelitian yang dilakukan oleh Martins membuktikan, kemarahan dan kebencian merupakan sebuah emosi yang lebih berkonotasi dengan ujaran kebencian. Klasifikasi emosi dapat di bagi menjadi 8 yaitu terdiri dari *"Fear", "Joy", "Trust", "Anticipation", "Surprise", "Sadness", "Disgust" dan "Anger"* [3].

Klasifikasi hate speech dan emosi pada Twitter menjadi semakin relevan menjelang Pilpres, karena platform ini sering digunakan sebagai arena perdebatan politik yang intens. Analisis teks menggunakan metode seperti *K-Nearest Neighbor* (KNN) dapat membantu mengidentifikasi ujaran kebencian (*hate speech*) serta emosi seperti marah, benci, atau dukungan dalam percakapan daring. Dengan meningkatnya polarisasi selama Pilpres, teknologi ini dapat digunakan untuk mendeteksi dan mengelola konten yang berpotensi memicu konflik, sehingga menciptakan ruang diskusi yang lebih sehat dan aman di media sosial.

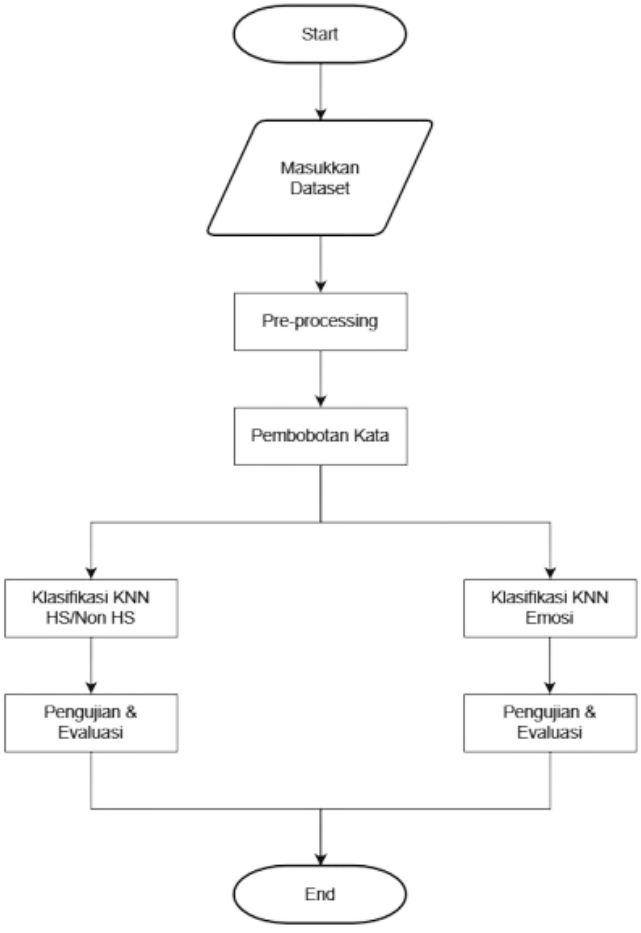
Penelitian tentang *hate speech* ini sebelumnya sudah pernah dilakukan dengan pembahasan judul analisis sentimen *hate speech* kepada pengguna layanan twitter menggunakan metode *naïve bayes classifier* [4]. Serta penerapan metode *k-means clustering* dalam menganalisis sentimen masyarakat terhadap *k-popers* pada twitter [5]. Juga penerapan analisis sentimen ujaran kebencian terhadap vaksinasi covid-19 pada tweet berbahasa indonesia menggunakan algoritme *k-nearest neighbor* [6]. Pada penelitian ini nantinya akan menunjukkan hasil dari klasifikasi ujaran kebencian dan deteksi emosi di sosial media twitter yang mana dalam penelitian ini metode yang dipakai ialah *k-nearest neighbor*.

Oleh karena itu, untuk mengurangi efek negatif dari interaksi di media sosial, teknologi machine learning seperti *K-Nearest Neighbor* (KNN) menjadi semakin penting untuk mengklasifikasikan dan mengidentifikasi hate speech serta emosi.

II. METODE

Metodologi penelitian mencakup tahapan proses yang akan diambil untuk menyelesaikan masalah dan menemukan solusi. Penulis melakukan studi kasus dengan menggunakan *K-Nearest Neighbor* untuk mengkategorikan sentimen dan hate speech pada Twitter.

Penelitian ini didasarkan pada penggunaan *K-Nearest Neighbor* untuk mengkategorikan sentimen dan ujaran pelecehan di Twitter. Ini adalah tindakan yang dilakukan:



Gambar 1. Diagram Alir Penelitian

Menurut Gambar 1 yang berisi diagram alir, dapat diketahui apa saja yang menjadi acuan pengerjaan pada penelitian ini.

Berdasarkan *flowchart K-Nearest Neighbor :*

1. Dataset
   Dataset dalam penelitian ini diperoleh dari media sosial Twitter. Selanjutnya mengkaji kata dasar, menganalisis kata yang kerap nampak atau dikenal stopword, serta pengelompokan ke dalam golongan tertentu.
2. Text Pre-processing
   Tahapan selajutnya merupakan tahapan pre-processing teks, pada tahap harus melalui beberapa tahapan yaitu diantaranya ada setidaknya 4 tahapan yang harus dilakukan, pertama tahap *transform case,* pada tahap ini semua kata dan huruf kapital yang terdapat dalam data diubah menjadi huruf kecil semua. Tahap kedua, *tokenization,* pada tahap ini teks akan dibersihkan dari tanda baca, spasi yang berulang-ulang juga baris baru akan diubah menjadi spasi, dan kata per kata nantinya akan dipisahkan dari kalimatnya. Tahapan yang ketiga yaitu *stemming,* pada tahapan ini imbuhan yang terdapat pada kata atau teks akan dihilangkan. Dan tahap yang terakhir yaitu *stopword,* tahap ini kata yang tidak dibutuhkan akan dihapus [7].
3. Pembobotan Kata
   Pemberian bobot pada setiap kata guna melihat sejauh mana kata tersebut penting dengan acuan pada tingkat kemunculan kata dalam dokumen tersebut [8].
4. Klasifikasi
   Klasifikasi adalah sebuah proses yang dimana label kelas nantinya akan diprediksi dari sampel yang telah disajikan berdasarkan karakteristik atau sekumpulan fitur [9].
5. Evaluasi
   Confusion matrix ialah sebuah metode yang digunakan sebagai alat ukur untuk model evaluasi performansi yang diimplementasikan guna mengukur performa dari suatu algoritma [10].

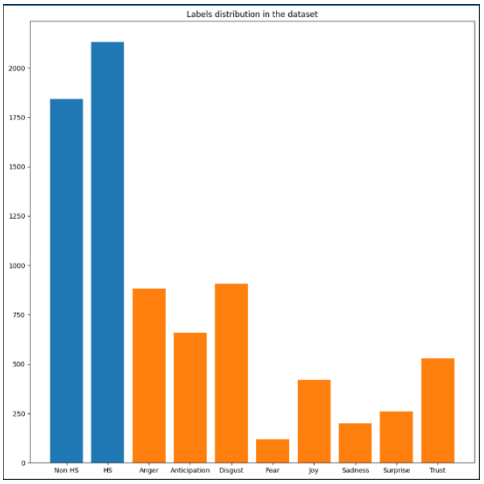
III. HASIL DAN PEMBAHASAN

A. Dataset

Menurut penelitian yang diambil melalui media sosial yaitu twitter diperoleh Dataset periode Desember 2023 sampai Februari 2024 pada isu Pilpres 2024. Selanjutnya kata dasar, kata-kata yang sering tampak atau stopword tersebut diidentifikasi dan dipilah menurut golongan [11]. Hasil Crawling tweet dapat dilihat dari tabel 1 berikut.

Tabel 1. Sampel tweet hasil *Crawling*

No	Tweet
1.	Hidup penuh perjuangan,tidak semudah membalikkan telapak tangan,tapi gunakanlah telapak tangan untuk hidup yang jauh lebih baik.
2.	Paduka @jokowi dan si @prabowo bisa gak menjelaskan ini?
.....	.....
3972	KASIANNYA PROF GILA DEMO SENDIRIAN GA MALU PARTAI @PDemokrat????



Gambar 2. *Labels Distribution in the Dataset*

B. Pre-processing

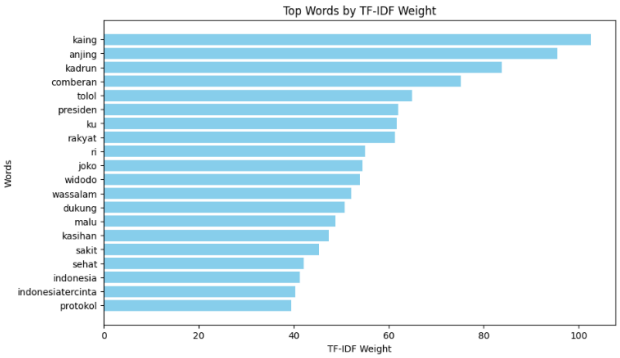
Pemrosesan teks yaitu membersihkan dan menyederhanakan teks sehingga dapat di proses lebih lanjut [12]. Text Pre-processing atau bisa disebut pemrosesan teks terdiri atas beberapa tahap, antara lain Transform case, Tokenize, Filter token, Stemming, dan Stopword.

Tabel 2. Proses *Preprocessing* Data

Proses	Tweet
Cleaning data	KASIANNYA PROF GILA DEMO SENDIRIAN GA MALU PARTAI @PDemokrat????
Tokenization	[‘kasiannya’, ‘prof’, ‘gila’, ‘demo’, ‘sendirian’, ‘ga’, ‘malu’, ‘partai’]
Stopword Removal	[‘kasiannya’, ‘prof’, ‘gila’, ‘demo’, ‘sendirian’, ‘ga’, ‘malu’, ‘partai’]
Stemming	kasian prof gila demo sendirian ga malu partai

C. Pembobotan Kata

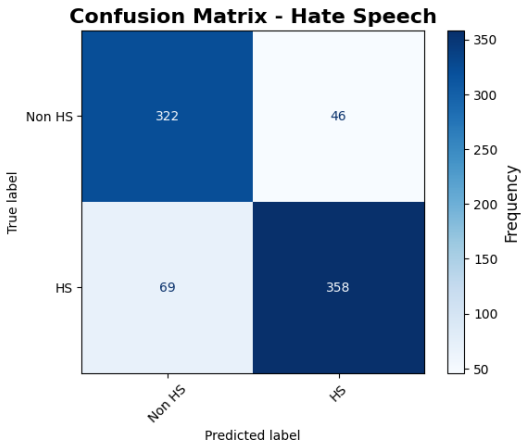
Proses berikutnya, ekspansi kata selanjutnya, dilakukan proses pembobotan kata dengan memakai metode TF-IDF yaitu dengan kepanjangan Term Frequency-Inverse Document Frequency. Tahap ini bertujuan mengubah data berbentuk teks menjadi format numerik. TF-IDF berfungsi untuk menemukan kata yang paling signifikan dalam dokumen yang sudah ditentukan.



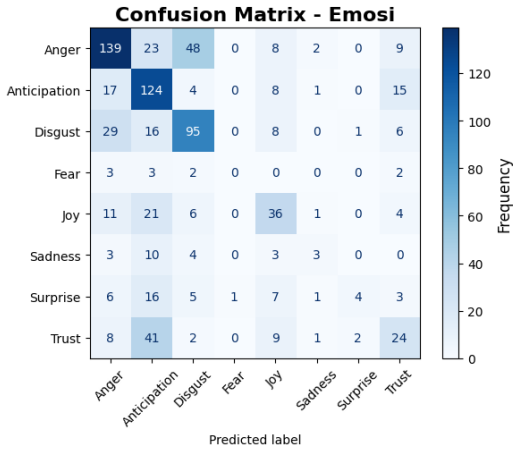
Gambar 3. *Word Cloud*

D. Klasifikasi

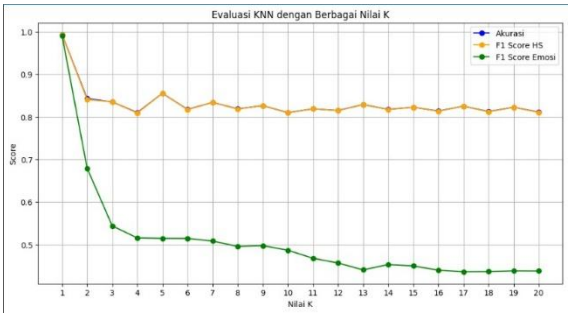
Setelah melakukan tahapan pembobotan kata, selanjutnya dilakukan pemodelan klasifikasi. Pada tahapan ini data dibagi menjadi dua yaitu data latih dengan rasio 80% dan data uji dengan raiso 20%. *K-Nearest Neighbor* (KNN) ialah metode klasifikasi dalam supervised learning yang berdasarkan pada perhitungan jarak. Data uji dan data latih akan dibanding jarang antara keduanya, dengan cara tersebut algoritma *k-nearest neighbor* ini bekerja [13].



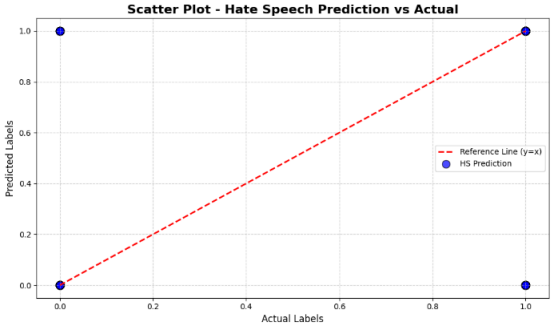
Gambar 4. *Confusion Matrix Hate Speech Set*



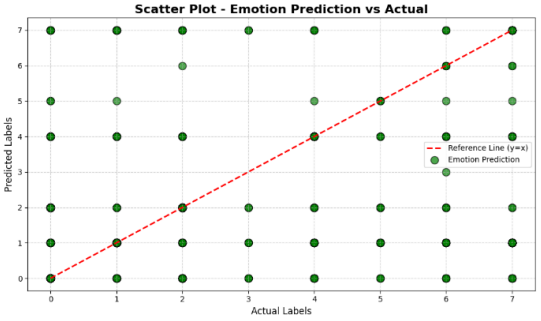
Gambar 5. *Confusion Matrix* Emosi



Gambar 6. Evaluasi KNN Dengan Berbagi Nilai K



Gambar 7. Plot *Hate Speech*



Gambar 8. Plot Emosi

E. Evaluasi

Tabel yang mengungkapkan klasifikasi jumlah data uji yang benar dan yang salah ialah confusion matrix [14]. F1 *Score* digunakan sebagai acuan tolak ukur dalam penelitian ini. F1 score auto muncul ketika hasil klasifikasi dari presision dan recall saling bertentangan. Jika meningkatkan precision, maka kemungkinan recall akan menurun, begitupun sebaliknya F1 *score* digunakan untuk menemukan keseimbangan antara precision dan recall. F1 *Score* yaitu perbandingan rata-rata presisi dan recall dengan pemberian bobot [15].

Tabel 3. *Mean Accuracy*

Model Algoritma	***Hate Speech***	**Emosi**
Mean Accuracy	0.8553459119496856	0.5345911949685535

C. Output

Program ini dirancang untuk mendeteksi hate speech serta emosi dari sebuah kalimat acak. Hasilnya terdiri dari dua jenis output, yaitu klasifikasi HS/Non-HS dan identifikasi emosi.

Tabel 4. Output klasifikasi *Hate Speech* dan Emosi

Text	***Hate Speech***	**Emosi**
Karena sembrono gabenar tetap hanya Bakal Calon Presiden, karena koalisinya rapuh jadi pasti peca kongsi setelah calonnya gagal jadi Bakal Calon Wakil Presiden gabenar dan sekarang partainya sudah dibujuk bergabung di KIB MERDEKA.	True	Fear
Setelah berjilid jilid demo dan tdk didukung Rakyat Indonesiatercinta, Ormas Buruh yg Ketua Umumnya Kadrun mengancam mau demo lage 10 November ha ha ha jualannya nggak laku nie ye mau diletakkan kemana itu muka badaknya MERDEKA.	True	Disgust
Aku juga sangat setujuuuuu tangkap dan borgol kadrun ini sudah benar2 melanggar hukum di Negara Indonesia terCinta yg ber Ideologi Pancasila MERDEKA.	False	Anticipation
Sama banget Komentar ini dgn para pendukung setia Bpk Joko Widodo Presiden RI ke 7, yg sangat dicintai Rakyatnya krn Kerja Keras MERDEKA.	False	Anticipation

IV. SIMPULAN

Penelitian ini bertujuan unutk mengklasifikasikan ujaran kebencian (*hate speech*) dan emosi pada platform twitter menggunakan metode K-Nearest Neighbor (KNN). Adapun tahapan utama dari penelitan ini yaitu meliputi pengumpulan data melalui proses crawling, pre-processing data untuk meningkatkan kualitas data, pembobotan data menggunakan Term Frequency-Inverse Document Frequency (TF-IDF), dan klasifikasi menggunakan algoritma KNN. Berdasarkan hasil evaluasi, algoritma KNN menunjukkan performa yang signifikan dalam mendeteksi *hate speech* dan emosi dengan akurasi rata-rata 0,855 untuk klasifikasi hate speech dan 0,534 untuk klasifikasi emosi. Penelitian ini berhasil membuktikan bahwa metode KNN mampu memberikan hasil yang cukup akurat dalam menganalisis teks dengan dataset tertentu.

hasil klasifikasi ini diharapkan dapat digunakan untuk mendukung moderasi konten di media sosial, khususnya dalam mengidentifikasi dan mengurangi penyebaran ujaran kebencian serta analisis emosi pengguna. Temuan ini juga membuka peluang bagi penelitian lanjutan untuk meningkatkan akurasi klasifikasi emosi dengan mengeksplorasi metode atau kombinasi algoritma yang lebih canggih.

Penelitian ini memiliki keterbatasan, antara lain pada jumlah dan variasi dataset yang digunakan serta rasio akurasi klasifikasi emosi yang lebih rendah dibandingkan klasifikasi hate speech. Oleh karena itu, penelitian di masa depan dapat difokuskan pada peningkatan akurasi klasifikasi emosi dengan memanfaatkan teknik seperti deep learning atau dengan memperluas cakupan dataset.

UCAPAN TERIMA KASIH

Artikel yang telah disusun oleh penulis bertujuan agar memberi kontribusi dalam ilmu pengetahuan. Terima kasih kepada Tuhan Yang Maha Esa dan seluruh pihak yang mendukung tersusunnya artikel. Penulis mempunyai harapan agar artikel ini dapat bermanfaat dan memberikan pemahaman yang bernilai. Terima kasih kepada Tuhan Yang Maha Esa dan seluruh pihak yang telah berkontribusi dan mendukung tersusunnya artikel ini, sehingga artikel ini dapat diselesaikan.

REFERENSI

[1] "PENDETEKSIAN HATE SPEECH PADA SOSIAL MEDIA INDONESIA DENGAN ALGORITMA LOGISTIC REGRESSION," pp. 1–13.

[2] F. Poletto, V. Basile, and M. Sanguinetti, "Resources and benchmark corpora for hate speech detection : a systematic review," *Lang. Resour. Eval.*, vol. 55, no. 2, pp. 477–523, 2021, doi: 10.1007/s10579-020-09502-8.

[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," vol. 73, no. 4, pp. 603–611, 2018, doi: 10.1093/geronb/gbw028.

[4] I. Riadi and A. Fadlil, "Analisis Sentimen Hate Speech pada Pengguna Layanan Twitter dengan Metode Naïve Bayes Classifier (NBC)," vol. 10, no. 2, 2023, doi: 10.30865/jurikom.v10i2.5984.

[5] M. Alysha, Z. Larasati, N. Anisa, S. Winarsih, and M. S. Rohman, "Penerapan Metode K-Means Clustering Dalam Menganalisis Sentimen Masyarakat Terhadap K-Popers Pada Twitter," pp. 201–210, 2016.

[6] J. Kalyzta, M. A. Willdan, and S. Halfiani, "PENERAPAN ANALISIS SENTIMEN UJARAN KEBENCIAN TERHADAP VAKSINASI COVID-19 PADA TWEET BERBAHASA INDONESIA MENGGUNAKAN ALGORITME K-NEAREST NEIGHBOR," vol. 5, pp. 87–97, 2022.

[7] R. Sistem, P. Metode, T. T. Pada, K. Teks, and H. M. K. Neighbor, "JURNAL RESTI," vol. 5, no. 10, pp. 911–918, 2021.

[8] P. Simposium, N. Multidisiplin, and U. M. Tangerang, "Analisis Sentimen Kinerja Pemerintahan Menggunakan Algoritma," vol. 4, pp. 114–121, 2022.

[9] A. Naïve, "Jurnal KomtekInfo Analisis Sentimen Relokasi Ibukota Nusantara Menggunakan," vol. 10, pp. 1–7, 2023, doi: 10.35134/komtekinfo.v10i1.330.

[10] R. Ariandi, U. Telkom, O. N. Pratiwi, U. Telkom, and U. Telkom, "Klasifikasi Soal Sejarah Tingkat SMA Berdasarkan Level Kognitif Revised Bloom ' s Taxonomy Menggunakan Algoritma K-Nearest Neighbour Manhattan," vol. 10, no. 2, pp. 1549–1555, 2023.

[11] U. M. Sidoarjo, "PENGGUNAAN K-NEAREST NEIGHBOR UNTUK Oleh : Tahun 2024," 2024.

[12] N. Aula, M. Ula, and L. Rosnita, "ANALISIS SENTIMEN REVIEW CUSTOMER TERHADAP PERUSAHAAN EKSPEDISI JNE , J & T EXPRESS DAN POS INDONESIA MENGGUNAKAN METODE SUPPORT VECTOR MACHINE (SVM) ANALYSIS OF CUSTOMER REVIEW SENTIMENT TO JNE , J & T EXPRESS AND POS INDONESIA EXPEDITION COMPANIES USING SVM METHOD," vol. 9, no. 1, pp. 81–86, 2023.

[13] L. Handayani, "IMPLEMENTASI K-NEAREST NEIGHBOR DALAM MENGKLASIFIKASIKAN HATE TWEET K-POPERS PADA TWITTER Skripsi," 2023.

[14] J. S. Komputer, "Implementasi Naïve Bayes Classifier Dan Confusion Matrix Pada Analisis Sentimen Berbasis Teks Pada Twitter," vol. 5, no. September, pp. 697–711, 2021.

[15] P. Romadloni, B. Adhi Kusuma, and W. Maulana Baihaqi, "Komparasi Metode Pembelajaran Mesin Untuk Implementasi Pengambilan Keputusan Dalam Menentukan Promosi Jabatan Karyawan," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 6, no. 2, pp. 622–628, 2022, doi: 10.36040/jati.v6i2.5238.

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.