

Optimasi Ekstraksi Fitur TF-IDF Menggunakan GA Pada Metode SVM Dalam Menentukan Opini Publik Terhadap Keberlanjutan IKN

Siti Patimah¹⁾, Rudiman²⁾, Fendy Yulianto³⁾

^{1,2)} Sains Teknologi, Universitas Muhammadiyah Kalimantan Timur
Jl. Ir. H. Juanda No.15, Sidodadi, Kec. Samarinda Ulu, Kota Samarinda
Email : 2011102441235@umkt.ac.id¹⁾, rudiman@umkt.ac.id²⁾, Fy415@umkt.ac.id³⁾

Abstrak

Perubahan opini masyarakat mengenai keberlanjutan Kota Ibu Kota Negara (IKN) menjadi topik penting dalam memahami respon masyarakat terhadap kebijakan pemerintah. Penelitian ini menganalisis sentimen keberlanjutan IKN pada komentar YouTube dengan menggunakan metode *Support Vector Machine* (SVM) yang dipadukan dengan seleksi fitur *Genetic Algorithm* (GA) dan ekstraksi fitur TF-IDF. Hasil penelitian menunjukkan bahwa model dengan metode GA mencapai akurasi 78,9%, dengan seleksi 1882 fitur dari total 3776 fitur TF-IDF, menghasilkan rasio 90:10. Penggunaan GA dalam seleksi fitur terbukti efektif dalam mengurangi dimensi fitur, yang pada akhirnya meningkatkan efisiensi model dan mengurangi risiko overfitting. Dibandingkan dengan model SVM tanpa seleksi fitur yang hanya mencapai akurasi 76%, model dengan seleksi fitur menggunakan GA menunjukkan peningkatan akurasi sebesar 2,9%, serta peningkatan F1-Score dari 71% menjadi 77%. Evaluasi berdasarkan metrik akurasi, *presisi*, *recall*, dan F1-Score menunjukkan kinerja yang lebih baik dengan menggunakan GA, yang membuktikan bahwa seleksi fitur efektif dalam meningkatkan performa model SVM tanpa mengorbankan efisiensi.

Kata Kunci : *Sentimen; TF-IDF; SVM; GA; IKN*

Abstract

The change in public opinion regarding the sustainability of the National Capital City (IKN) has become an important topic in understanding the public's response to government policies. This study analyzes the sentiment of IKN sustainability in YouTube comments using the *Support Vector Machine* (SVM) method combined with *Genetic Algorithm* (GA) feature selection and TF-IDF feature extraction. The results show that the model with the GA method achieves an accuracy of 78.9%, with 1882 features selected from a total of 3776 TF-IDF features, resulting in a 90:10 ratio. The use of GA in feature selection proves effective in reducing feature dimensions, ultimately improving model efficiency and reducing the risk of overfitting. Compared to the SVM model without feature selection, which only achieved an accuracy of 76%, the model with GA feature selection shows a 2.9% accuracy improvement and an increase in F1-Score from 71% to 77%. Evaluation based on accuracy, precision, recall, and F1-Score metrics demonstrates better performance with GA, proving that feature selection is effective in enhancing the SVM model's performance without compromising efficiency.

Keywords: *Sentiment; TF-IDF; SVM; GA; IKN*

1. PENDAHULUAN

Pemindahan ibu kota negara Indonesia ke Nusantara (IKN) di Kalimantan Timur telah menjadi salah satu agenda nasional yang menyita perhatian publik sejak diumumkan pada tahun 2019. Pemerintah menyampaikan bahwa IKN tidak hanya sebagai pusat pemerintahan baru, tetapi juga sebagai simbol keberagaman dan persatuan bangsa [1]. Meskipun demikian, kebijakan ini menimbulkan berbagai reaksi, baik yang mendukung maupun yang menolak. Oleh karena itu, keterlibatan masyarakat dan partisipasi publik sangat diperlukan dalam

proses perencanaan dan pelaksanaan kebijakan, agar kebijakan tersebut memperoleh legitimasi sosial yang kuat [2]. Salah satu platform media sosial yang banyak digunakan masyarakat untuk menyuarakan pendapat mereka terkait IKN adalah YouTube. Komentar-komentar pengguna dalam video terkait pemindahan ibu kota menjadi sumber data yang potensial untuk mengetahui opini publik. Data ini dapat dimanfaatkan dalam analisis sentimen, yaitu teknik yang digunakan untuk mengidentifikasi dan mengklasifikasikan sentimen dari teks menjadi kategori seperti positif, negatif, atau netral [3]. Oleh

karena itu, riset berbasis data komentar YouTube sangat relevan untuk mengukur bagaimana masyarakat merespons keberlanjutan proyek IKN.

Dalam penelitian ini, metode klasifikasi dimanfaatkan untuk menganalisis sentimen. Proses klasifikasi dalam *Text Mining* merupakan teknik untuk mengelompokkan dokumen atau teks ke dalam kategori tertentu, berdasarkan ciri-ciri atau pola yang terkandung di dalamnya [4]. Metode klasifikasi ini banyak diterapkan untuk menganalisis teks dalam jumlah besar, seperti dalam klasifikasi sentimen, pengelompokan topik, atau deteksi spam pada email. Algoritma klasifikasi, seperti *Naïve Bayes* dan *Support Vector Machine* (SVM), sering digunakan dalam *Text Mining* untuk memprediksi kategori atau label yang sesuai untuk setiap dokumen berdasarkan fitur-fitur yang diekstraksi dari teks [5].

Jadi pada penelitian ini menggunakan metode SVM karena penelitian yang dilakukan oleh Gori [6] mengenai klasifikasi kelayakan air minum menunjukkan metode *Support Vector Machine* (SVM) menghasilkan performa terbaik dibandingkan dengan metode *KNN*. Metode *KNN* menghasilkan akurasi sebesar 69,764%, sementara metode *KNN* memperoleh akurasi tertinggi sebesar 65%. Penelitian lain yang dilakukan oleh (Ilmawan & Mude, 2020) terkait klasifikasi ulasan *Google Play Store* menunjukkan metode *Support Vector Machine* (SVM) menghasilkan akurasi sebesar 81,46%, lebih tinggi dibandingkan metode *Naïve Bayes* dengan akurasi sebesar 75,41%.

Ekstraksi fitur, yang juga dikenal sebagai penyisipan kata (*Word Embedding*), merupakan salah satu topik menarik dalam penelitian di bidang Pemrosesan Bahasa Alami (*Natural Language Processing*) [7]. Proses *word embedding* bekerja dengan memetakan setiap kata dalam dokumen ke dalam *dense vector*, dimana vektor tersebut mempresentasikan proyeksi kata di ruang vektor [8]. Beberapa metode ekstraksi fitur yang umum digunakan meliputi *Term Frequency – Inverse Document Frequency* (TF-IDF) serta *word embeddings* seperti *Word2Vec*, *GloVe*, dan *BERT*, yang bertujuan untuk merepresentasikan kata-kata ke dalam bentuk vektor numerik.

Dalam penelitian ini, penulis memilih metode TF-IDF (*Term Frequency–Inverse Document Frequency*) sebagai teknik ekstraksi fitur karena memiliki sejumlah keunggulan praktis

dibandingkan pendekatan berbasis representasi konteks [9]. TF-IDF juga memiliki keunggulan dari sisi interpretabilitas, di mana bobot tiap kata dapat dianalisis secara langsung untuk mengetahui pengaruhnya terhadap klasifikasi [10]. Penelitian yang dilakukan oleh Tri Putra [11], mengenai analisis sentimen JNE menunjukkan bahwa TF-IDF mampu menghasilkan akurasi lebih tinggi dibandingkan metode BOW dengan nilai akurasi TF-IDF 86% dan BOW 85%.

Hal ini sangat membantu dalam proses seleksi fitur menggunakan algoritma seperti *Genetic Algorithm*, karena TF-IDF menghasilkan vektor yang lebih mudah diolah [12]. Penerapan seleksi fitur dapat mempengaruhi tingkat akurasi algoritma, karena dengan seleksi fitur, kita dapat mengidentifikasi fitur-fitur yang kurang relevan serta memilih fitur-fitur yang mengandung informasi penting [13]. Beberapa algoritma yang umum digunakan untuk seleksi fitur antara lain *Principal Component Analysis* (PCA), *Recursive Feature Elimination* (RFE), dan *Genetic Algorithm* (GA).

Berdasarkan hal tersebut, penelitian ini bertujuan untuk (1) mengukur akurasi metode *Support Vector Machine* dalam klasifikasi sentimen terhadap keberlanjutan IKN dengan seleksi fitur yang relevan; (2) mengevaluasi kemampuan *genetic algorithm* dalam mereduksi dimensi hasil ekstraksi TF-IDF; dan (3) menilai sejauh mana kombinasi metode tersebut dapat meningkatkan kinerja model dalam menangkap opini publik secara lebih akurat. Dengan demikian, penelitian ini diharapkan dapat memberikan kontribusi dalam pengembangan metode analisis sentimen berbasis teks yang efisien dan akurat dalam konteks kebijakan publik.

2. METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif dengan data primer berupa komentar masyarakat di platform YouTube mengenai pemindahan dan keberlanjutan Ibu Kota Negara (IKN). Data dikumpulkan melalui teknik web scraping menggunakan Google Colab dan YouTube API, dengan mengambil komentar dari video milik channel *MetroTV* yang membahas topik IKN. Komentar tersebut kemudian disimpan dalam format CSV, dengan total sebanyak 1.274 komentar yang dijadikan bahan analisis. Untuk memastikan data dapat diolah sesuai kebutuhan penelitian, dilakukan

penyaringan awal pada kolom komentar dan nama pengguna, tanpa menyertakan data pribadi lainnya.

Dalam tahap awal analisis, komentar dikategorikan berdasarkan sentimen positif, netral, dan negatif melalui proses labeling manual, yang dikerjakan oleh tenaga profesional melalui platform freelance Projects.co.id.

Sentimen positif dalam konteks analisis teks atau komentar mengacu pada perasaan atau reaksi yang menunjukkan dukungan, apresiasi, pujian, atau optimisme terhadap suatu topik. Sentimen negatif dalam analisis teks atau komentar merujuk pada perasaan atau reaksi yang menunjukkan kritik, penolakan, ketidaksetujuan, atau kekecewaan terhadap suatu topik. Komentar dengan sentimen negatif sering kali mengungkapkan rasa kecewa, marah, atau protes terhadap kebijakan, peristiwa, atau situasi tertentu. Ciri khas dari sentimen negatif adalah penggunaan kata-kata yang menggambarkan kerugian, kesalahan, atau masalah, seperti "buruk", "tidak efektif", "merugikan", atau "gagal". Sedangkan sentimen netral dalam analisis teks atau komentar merujuk pada perasaan atau reaksi yang tidak menunjukkan kecenderungan positif maupun negatif terhadap suatu topik.

Dalam penelitian ini, alat dan bahan yang digunakan disesuaikan dengan jenis data, metode, dan tujuan penelitian. Alat yang digunakan terdiri dari perangkat keras (hardware) berupa 1 unit laptop Lenovo Idea Slim 3 dengan spesifikasi processor AMD Ryzen, RAM 8 GB, dan penyimpanan internal SSD 512 GB. Untuk perangkat lunak (software), digunakan Google Colabs, Google Chrome, dan berbagai library Python yang mendukung proses analisis. Sedangkan bahan penelitian ini berupa data komentar dari salah satu video YouTube yang diperoleh melalui teknik scraping, dengan total 1274 data yang digunakan untuk mendukung penelitian ini.

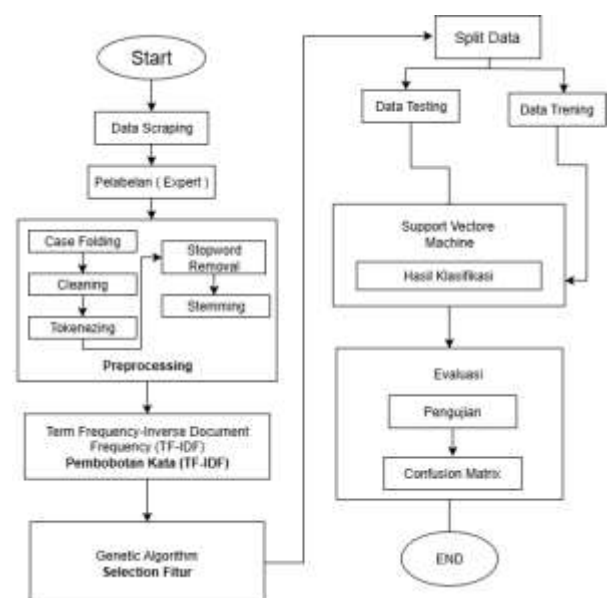
Instrumen pendukung dalam penelitian ini mencakup perangkat keras berupa laptop dengan spesifikasi memadai, serta perangkat lunak seperti Google Chrome, Google Colab, dan berbagai library Python seperti NLTK dan Scikit-learn. Metode pengujian dilakukan melalui beberapa tahapan, dimulai dari preprocessing data, ekstraksi fitur menggunakan TF-IDF, seleksi fitur menggunakan Genetic Algorithm (GA), klasifikasi dengan Support Vector Machine (SVM), dan evaluasi model menggunakan metrik akurasi, precision, recall, dan F1-score.

2.1 Tahapan Penelitian

Tahapan penelitian dimulai dari pengumpulan data komentar video YouTube yang berkaitan

dengan topik keberlanjutan IKN. Setelah data terkumpul, dilakukan proses preprocessing yang terdiri atas beberapa tahap: case folding, cleaning, tokenizing, stopwords removal, dan stemming. Proses ini bertujuan untuk membersihkan dan menstandarkan data teks agar dapat dianalisis lebih lanjut. Selanjutnya, dilakukan pembobotan kata menggunakan metode TF-IDF (Term Frequency–Inverse Document Frequency), yang memberikan nilai bobot pada setiap kata berdasarkan tingkat kepentingannya dalam dokumen. Representasi numerik ini menjadi dasar untuk klasifikasi sentimen pada tahap selanjutnya.

Setelah fitur teks diekstraksi menggunakan TF-IDF, dilakukan seleksi fitur menggunakan metode Genetic Algorithm (GA). GA digunakan untuk memilih fitur-fitur penting dan mengurangi dimensi data agar model klasifikasi menjadi lebih efisien dan akurat. Proses klasifikasi sentimen dilakukan dengan algoritma Support Vector Machine (SVM), yang dikenal handal dalam menangani data berdimensi tinggi dan memberikan hasil klasifikasi yang presisi. Terakhir, dilakukan evaluasi model dengan membandingkan hasil klasifikasi pada data latih dan data uji menggunakan metrik evaluasi seperti akurasi, precision, recall, dan F1-score. Pembagian data dilakukan dengan rasio umum 70:30, serta diuji ulang menggunakan cross-validation untuk menjamin keandalan hasil.



Gambar 1 Alur Penelitian

3. HASIL DAN PEMBAHASAN

3.1 Hasil Pengumpulan Data

Pengumpulan data dilakukan secara praktis melalui platform YouTube, di mana peneliti memanfaatkan layanan Google Colab sebagai lingkungan pemrograman berbasis cloud yang mendukung bahasa Python. Dalam proses ini, peneliti menggunakan berbagai pustaka untuk mengakses API YouTube Data v3 guna memperoleh komentar dari video tertentu. Data yang berhasil diperoleh kemudian digunakan untuk analisis sentimen. Proses crawling atau pengambilan data berlangsung selama kurang lebih 10 menit. Hasil dari proses ini ditampilkan pada gambar berikut.

No	Nama	Waktu	Komentar
1	@kubekaput...@	2025-04-25 13:21:36Z	Mantap IKN Majuterus semoga lancar pembangunannya
2	@Kubekaput...@	2025-04-27 04:16:08Z	Mantap IKN Majuterus semoga lancar pembangunannya
3	@Kubekaput...@	2025-04-21 18:05:17Z	Berhutang kemalaysia klu buat ikn.berhutang dulu.dgn negara kami malaysia klu bukan malaysia indonesia sudah gulung tikar. 🌟🌟🌟🌟🌟🌟
4	@Kubekaput...@	2025-04-20 17:04:02Z	ikn bagaikan tambang uang surga bagi para mafia infrastruktur
5	@Kubekaput...@	2025-04-20 17:59:13Z	IKN = Ikatan Korup Negara
6	@Kubekaput...@	2025-04-19 13:38:41Z	Seandainya Kita Mau Dan Mampu Mencegah Korupsi, Kita Bisa Menyelesaikan Ikn Tanpa Asing, Kita Bisa Beli 1.000 Pesawat Tempur Tercanggih Saat Ini, Bisa Menciptakan Lapangan Kerja Untuk Semua Rakyat, Bisa Memberi Modal Kpd Rakyat (Memberi, Bukan Pinjaman). Dan Masih Banyak Pembangunan Di Seluruh Indonesia Scr Merata
7	@Kubekaput...@	2025-04-19 13:38:42Z	Seandainya Kita Mau Dan Mampu Mencegah Korupsi, Kita Bisa Menyelesaikan Ikn Tanpa Asing, Kita Bisa Beli 1.000 Pesawat Tempur Tercanggih Saat Ini, Bisa Menciptakan Lapangan Kerja Untuk Semua Rakyat, Bisa Memberi Modal Kpd Rakyat (Memberi, Bukan Pinjaman). Dan Masih Banyak Pembangunan Di Seluruh Indonesia Scr Merata
8	@Kubekaput...@	2025-04-19 13:38:42Z	Seandainya Kita Mau Dan Mampu Mencegah Korupsi, Kita Bisa Menyelesaikan Ikn Tanpa Asing, Kita Bisa Beli 1.000 Pesawat Tempur Tercanggih Saat Ini, Bisa Menciptakan Lapangan Kerja Untuk Semua Rakyat, Bisa Memberi Modal Kpd Rakyat (Memberi, Bukan Pinjaman). Dan Masih Banyak Pembangunan Di Seluruh Indonesia Scr Merata
9	@Kubekaput...@	2025-04-19 13:38:42Z	Seandainya Kita Mau Dan Mampu Mencegah Korupsi, Kita Bisa Menyelesaikan Ikn Tanpa Asing, Kita Bisa Beli 1.000 Pesawat Tempur Tercanggih Saat Ini, Bisa Menciptakan Lapangan Kerja Untuk Semua Rakyat, Bisa Memberi Modal Kpd Rakyat (Memberi, Bukan Pinjaman). Dan Masih Banyak Pembangunan Di Seluruh Indonesia Scr Merata
10	@Kubekaput...@	2025-04-19 13:38:42Z	Seandainya Kita Mau Dan Mampu Mencegah Korupsi, Kita Bisa Menyelesaikan Ikn Tanpa Asing, Kita Bisa Beli 1.000 Pesawat Tempur Tercanggih Saat Ini, Bisa Menciptakan Lapangan Kerja Untuk Semua Rakyat, Bisa Memberi Modal Kpd Rakyat (Memberi, Bukan Pinjaman). Dan Masih Banyak Pembangunan Di Seluruh Indonesia Scr Merata
11	@Kubekaput...@	2025-04-19 13:38:42Z	Seandainya Kita Mau Dan Mampu Mencegah Korupsi, Kita Bisa Menyelesaikan Ikn Tanpa Asing, Kita Bisa Beli 1.000 Pesawat Tempur Tercanggih Saat Ini, Bisa Menciptakan Lapangan Kerja Untuk Semua Rakyat, Bisa Memberi Modal Kpd Rakyat (Memberi, Bukan Pinjaman). Dan Masih Banyak Pembangunan Di Seluruh Indonesia Scr Merata
12	@Kubekaput...@	2025-04-19 13:38:42Z	Seandainya Kita Mau Dan Mampu Mencegah Korupsi, Kita Bisa Menyelesaikan Ikn Tanpa Asing, Kita Bisa Beli 1.000 Pesawat Tempur Tercanggih Saat Ini, Bisa Menciptakan Lapangan Kerja Untuk Semua Rakyat, Bisa Memberi Modal Kpd Rakyat (Memberi, Bukan Pinjaman). Dan Masih Banyak Pembangunan Di Seluruh Indonesia Scr Merata
13	@Kubekaput...@	2025-04-19 13:38:42Z	Seandainya Kita Mau Dan Mampu Mencegah Korupsi, Kita Bisa Menyelesaikan Ikn Tanpa Asing, Kita Bisa Beli 1.000 Pesawat Tempur Tercanggih Saat Ini, Bisa Menciptakan Lapangan Kerja Untuk Semua Rakyat, Bisa Memberi Modal Kpd Rakyat (Memberi, Bukan Pinjaman). Dan Masih Banyak Pembangunan Di Seluruh Indonesia Scr Merata
14	@Kubekaput...@	2025-04-19 13:38:42Z	Seandainya Kita Mau Dan Mampu Mencegah Korupsi, Kita Bisa Menyelesaikan Ikn Tanpa Asing, Kita Bisa Beli 1.000 Pesawat Tempur Tercanggih Saat Ini, Bisa Menciptakan Lapangan Kerja Untuk Semua Rakyat, Bisa Memberi Modal Kpd Rakyat (Memberi, Bukan Pinjaman). Dan Masih Banyak Pembangunan Di Seluruh Indonesia Scr Merata
15	@Kubekaput...@	2025-04-19 13:38:42Z	Seandainya Kita Mau Dan Mampu Mencegah Korupsi, Kita Bisa Menyelesaikan Ikn Tanpa Asing, Kita Bisa Beli 1.000 Pesawat Tempur Tercanggih Saat Ini, Bisa Menciptakan Lapangan Kerja Untuk Semua Rakyat, Bisa Memberi Modal Kpd Rakyat (Memberi, Bukan Pinjaman). Dan Masih Banyak Pembangunan Di Seluruh Indonesia Scr Merata
16	@Kubekaput...@	2025-04-19 13:38:42Z	Seandainya Kita Mau Dan Mampu Mencegah Korupsi, Kita Bisa Menyelesaikan Ikn Tanpa Asing, Kita Bisa Beli 1.000 Pesawat Tempur Tercanggih Saat Ini, Bisa Menciptakan Lapangan Kerja Untuk Semua Rakyat, Bisa Memberi Modal Kpd Rakyat (Memberi, Bukan Pinjaman). Dan Masih Banyak Pembangunan Di Seluruh Indonesia Scr Merata
17	@Kubekaput...@	2025-04-19 13:38:42Z	Seandainya Kita Mau Dan Mampu Mencegah Korupsi, Kita Bisa Menyelesaikan Ikn Tanpa Asing, Kita Bisa Beli 1.000 Pesawat Tempur Tercanggih Saat Ini, Bisa Menciptakan Lapangan Kerja Untuk Semua Rakyat, Bisa Memberi Modal Kpd Rakyat (Memberi, Bukan Pinjaman). Dan Masih Banyak Pembangunan Di Seluruh Indonesia Scr Merata
18	@Kubekaput...@	2025-04-19 13:38:42Z	Seandainya Kita Mau Dan Mampu Mencegah Korupsi, Kita Bisa Menyelesaikan Ikn Tanpa Asing, Kita Bisa Beli 1.000 Pesawat Tempur Tercanggih Saat Ini, Bisa Menciptakan Lapangan Kerja Untuk Semua Rakyat, Bisa Memberi Modal Kpd Rakyat (Memberi, Bukan Pinjaman). Dan Masih Banyak Pembangunan Di Seluruh Indonesia Scr Merata

Gambar 2 Hasil Pengumpulan Data

Berdasarkan gambar 2 menunjukkan bahwa setelah melakukan proses pengumpulan atau crawling data maka didapatkan hasil data komentar sebanyak 1274 data dengan nama table nama, waktu dan komentar.

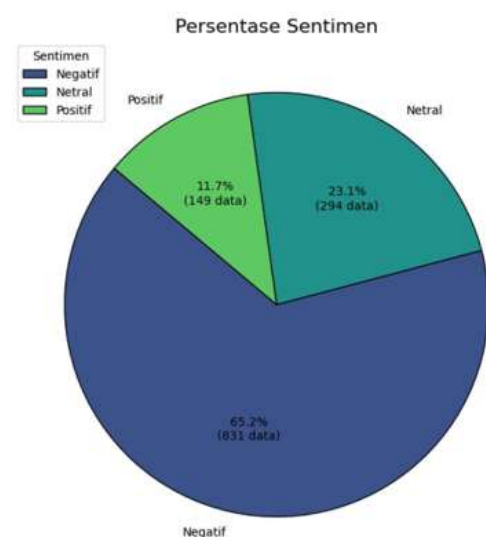
3.2 Hasil Labeling Data

Data yang diambil merupakan data yang memiliki 1 aspek didalamnya. Pelabelan dilakukan secara manual oleh ahli bahasa (expert) untuk 1274 data komentar hasil crawling pada platfrom youtube. Data komentar pada penelitian ini akan dikategorikan ke dalam tiga sentimen yaitu sentimen Negatif, Netral, dan Positif. Berikut dibawah ini sample komentar setelah diberi label.

Tabel 1 Hasil Labeling Data

Komentar	Sentimen
Mantap IKN Majuterus semoga lancar pembangunannya indonesia jaya kalimantan maju terus	Positif

pantang mundur presiden pk PRABOWO Lanjutkan demi masarakat kalimantan Merdeka	
Berhutang kemalaysia tuk buat ikn.berhutang dulu.dgn negara kami malaysia klu bukan malaysia indonesia sudah gulung tikar. 🌟🌟🌟🌟🌟🌟	Negatif
ikn bagaikan tambang uang surga bagi para mafia infrastruktur	Negatif
IKN = Ikatan Korup Negara	Negatif
Seandainya Kita Mau Dan Mampu Mencegah Korupsi, Kita Bisa Menyelesaikan Ikn Tanpa Asing, Kita Bisa Beli 1.000 Pesawat Tempur Tercanggih Saat Ini, Bisa Menciptakan Lapangan Kerja Untuk Semua Rakyat, Bisa Memberi Modal Kpd Rakyat (Memberi, Bukan Pinjaman). Dan Masih Banyak Pembangunan Di Seluruh Indonesia Scr Merata	Netral



Gambar 31 Diagram Hasil Pelabelan

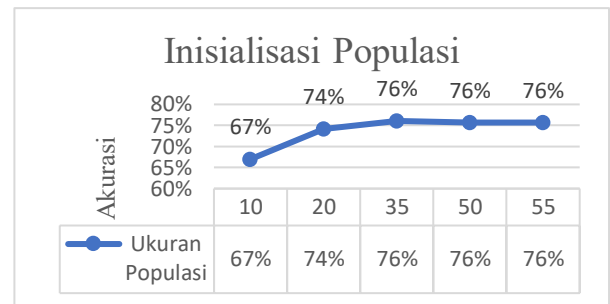
Nilai TF-IDF Tertinggi	
Token	TF-IDF
Ikn	93.136249
Yg	40.840851
Rakyat	37.641920
Bangun	31.854824
Pak	27.220096
indonesia	26.309847
Jokowi	22.278423
Negara	21.266497
Kkorupsi	19.981569
lanjut	19.802895

3.5 Seleksi Fitur *Genetic Algorithm*

Setelah tahap ekstraksi fitur menggunakan TF-IDF yang menghasilkan representasi numerik dari teks, langkah selanjutnya adalah melakukan seleksi fitur menggunakan *Genetic Algorithm* (GA). Proses seleksi fitur ini bertujuan untuk memilih subset fitur yang paling relevan dan efektif dalam membangun model, dengan harapan dapat meningkatkan akurasi model serta efisiensi komputasi. Pada tahap ini, seleksi fitur dilakukan dengan membentuk populasi awal yang terdiri dari berbagai individu atau solusi, yang masing-masing merepresentasikan subset fitur yang berbeda. Setiap individu dalam populasi dievaluasi menggunakan nilai *fitness*, yang menunjukkan seberapa baik individu tersebut dalam memenuhi kriteria yang telah ditentukan, seperti akurasi model atau performa lainnya. Proses ini berlangsung selama beberapa generasi, di mana setiap generasi menghasilkan nilai *fitness* yang berbeda untuk setiap individu, dan dilakukan seleksi untuk memilih individu terbaik.

1. Inisialisasi populasi

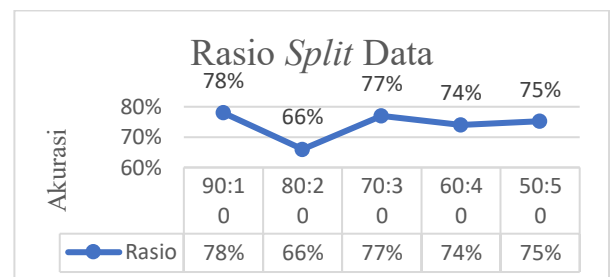
Pada tahap ini dilakukan pengujian terhadap beberapa kombinasi parameter untuk menemukan konfigurasi terbaik dalam proses optimasi GA, khususnya ukuran populasi dan jumlah generasi. Variasi ukuran populasi yang diuji meliputi (10, 20, 35, 50, 55), Hasil pengaruh variasi ukuran populasi terhadap performa klasifikasi ditampilkan pada gambar 6 sebagai berikut.



Gambar 6 Hasil Inisialisasi Populasi

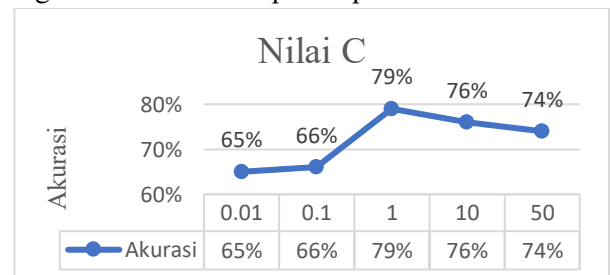
2. Evaluasi Fitness

Selanjutnya tahap evaluasi, setiap individu (kromosom) dievaluasi berdasarkan akurasi klasifikasi yang dihasilkan menggunakan algoritma *Support Vector Machine* (SVM). Penilaian dilakukan dengan menggunakan nilai akurasi tertinggi dari berbagai variasi parameter, yang terdiri atas rasio *split data* (90:10, 80:20, 70:30, 60:40, dan 50:50). Dan nilai C (0.1, 1, 10, 50, dan 100),). Setiap parameter diuji untuk memperoleh kombinasi terbaik, dan parameter dengan nilai akurasi tertinggi akan digunakan pada tahapan penelitian selanjutnya. Pengujian nilai akurasi dari hasil evaluasi *split data* dapat dilihat pada gambar 7 sebagai berikut.



Gambar 7 Hasil Rasio Terbaik

Hasil pengujian terhadap rasio *split data* menunjukkan bahwa rasio 90:10 memberikan akurasi tertinggi, yaitu sebesar 78.12%. Selanjutnya, hasil pengujian terhadap variasi nilai C pada algoritma SVM ditampilkan pada Gambar 8.



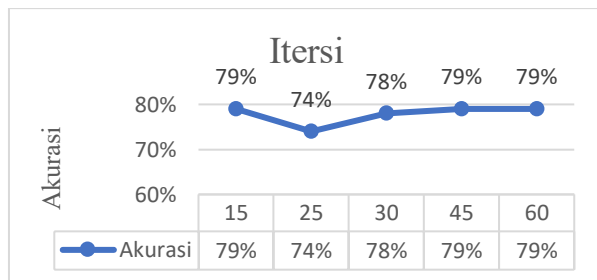
Gambar 8 Hasil Pengujian Nilai C

Hasil pengujian terhadap variasi nilai C pada algoritma *Support Vector Machine* (SVM).

Berdasarkan hasil tersebut, diperoleh bahwa nilai $C = 1$ memberikan akurasi terbaik, yaitu sebesar 79%

3. Iterasi

Iterasi atau perulangan pada algoritme genetika disebut juga dengan generasi, iterasi diperlukan sampai individu memenuhi kriteria. generasi. Variasi ukuran populasi yang diuji meliputi (15, 25, 30, 45 dan 60), Hasil pengarus variasi ukuran Iterasi terhadap performa klasifikasi ditampilkan pada gambar 9 sebagai berikut.



Gambar 9 Hasil Pengujian Iterasi

Hasil pengujian terhadap variasi iterasi pada algoritma *Support Vector Machine* (SVM). Berdasarkan hasil tersebut, diperoleh bahwa parameter terbaik adalah generasi 45 memberikan akurasi terbaik, yaitu sebesar 80%.

4. Parameter Akhir Genetic Algorithm

Parameter akhir algoritme optimasi GA adalah nilai parameter yang didapat setelah melakukan pengujian dengan mengambil yang rata-rata akurasi tertinggi. Hasil parameter ditampilkan pada tabel 3 sebagai berikut.

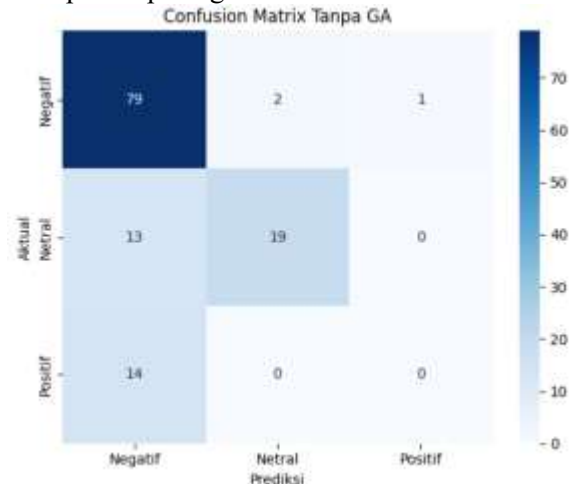
Tabel 3. 3 Parameter GA

Parameter	Nilai
Ukuran Populasi	35
Jumlah Generasi	45
Tingkat Mutasi	0.3
Tingkat <i>Crossover</i>	0.7
<i>Split</i> Data	90.10
Nilai C	1

Setelah dilakukan pengujian pada parameter, ditemukan parameter terbaik seperti yang ada pada Tabel 3.9. Populasi sebanyak 10 dan jumlah generasi sebanyak 60, dengan hasil akurasi 76.24%.

3.6 Hasil Evaluasi

Untuk tampilan *Confusion Matrix* akan ditampilkan pada gambar dibawah.



Gambar 10 Hasil Confusion Matrix Tanpa GA

Confusion matrix di atas menunjukkan hasil prediksi model untuk tiga kelas: Negatif, Netral dan Positif. Berikut adalah penjelasan dan perhitungan secara manual untuk menghitung hasil akurasi:

1 Baris 1 (Aktual Negatif)

- Kolom 1 (Prediksi Negatif): 79 (Model berhasil memprediksi 79 data dengan benar sebagai Negatif).
- Kolom 2 (Prediksi Netral) : 2 (Model berhasil memprediksi 2 data yang sebenarnya negatif menjadi netral).
- Kolom 3 (Prediksi Positif) : 1 (Model berhasil memprediksi 1 data yang sebenarnya negatif menjadi positif).

2 Baris 2 (Aktual Netral)

- Kolom 1 (Prediksi Negatif) : 13 (Model berhasil memprediksi 13 data yang sebenarnya netral menjadi negatif).
- Kolom 2 (Prediksi Netral) : 19 (Model berhasil memprediksi 19 data dengan benar sebagai netral).
- Kolom 3 (Prediksi Positif) : 0 (model berhasil memprediksi 0 data yang sebenarnya netral menjadi positif).

3 Baris 3 (Aktual Positif)

- Kolom 1 (Prediksi Negatif) : 14 (Model berhasil memprediksi 14 data yang sebenarnya positif menjadi negatif).

- Kolom 2 (Prediksi Netral) : 0 (Model berhasil memprediksi 0 data yang sebenarnya positif menjadi netral).
- Kolom 3 (Prediksi Positif) : 0 (Model berhasil memprediksi 0 data dengan benar sebagai positif).

$$Akurasi = \frac{TP_{Negatif} + TP_{Netral} + TP_{Positif}}{Total\ Jumlah\ Data} \times 100\%$$

Total jumlah data : $79 + 2 + 1 + 13 + 19 + 0 + 14 + 0 + 0 = 128$

$$Akurasi = \frac{79 + 19 + 0}{128} = \frac{98}{128} \times 100\% = 0,76 \times 100\% = 76\%$$

Hasil tersebut menunjukkan bahwasannya dalam pengujian metode *Support Vector Machine* (SVM) dan pembobotan kata menggunakan ekstraksi fitur TF-IDF dan seleksi fitur GA dalam klasifikasi sentimen perihal pemindahan Ibu Kota Indonesia. Menghasilkan akurasi yang cukup baik yaitu sebesar 76%, namun dalam penelitian ini memiliki beberapa kekurangan yang berpengaruh terhadap hasil yang didapatkan yaitu akses data komentar yang terbatas dan data yang tidak seimbang pada tahap pelabelan. Hasil Akurasi dapat dilihat pada gambar 11.

```

--- SKENARIO 1: TANPA GA (SEMUA FITUR) ---
=====
=== Evaluasi Model SVM (Tanpa GA) ===
Akurasi: 0.7656
Precision: 0.7036
Recall: 0.7656
F1 Score: 0.7176
Waktu eksekusi: 1.05 detik

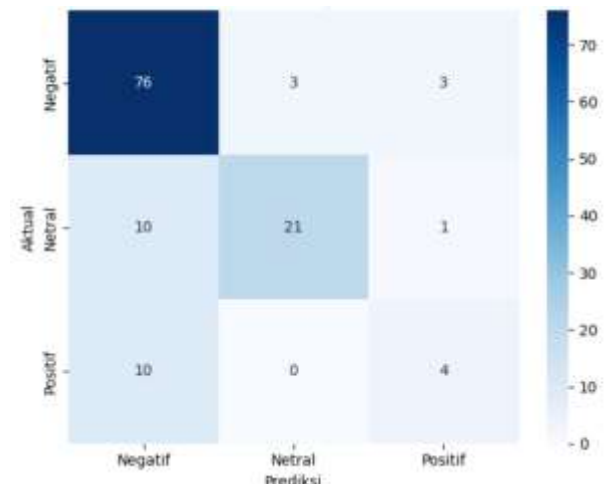
=== Classification Report ===

```

	precision	recall	f1-score	support
Negatif	0.75	0.96	0.84	82
Netral	0.90	0.59	0.72	32
Positif	0.00	0.00	0.00	14
accuracy			0.77	128
macro avg	0.55	0.52	0.52	128
weighted avg	0.70	0.77	0.72	128

Gambar 11 Hasil Klasifikasi Tanpa GA

Selanjutnya melakukan klasifikasi menggunakan GA dengan metode SVM. Untuk tampilan *Confusion Matrix* akan ditampilkan pada gambar 12 sebagai berikut.



Gambar 12 Hasil Confusion Matrix Dengan GA

Confusion matrix di atas menunjukkan hasil prediksi model untuk tiga kelas: Negatif, Netral dan Positif. Berikut adalah penjelasan dan perhitungan secara manual untuk menghitung hasil akurasi:

Confusion matrix di atas menunjukkan hasil prediksi model untuk tiga kelas: Negatif, Netral dan Positif. Berikut adalah penjelasan dan perhitungan secara manual untuk menghitung hasil akurasi:

1 Baris 1 (Aktual Negatif)

- Kolom 1 (Prediksi Negatif): 76 (Model berhasil memprediksi 76 data dengan benar sebagai Negatif).
- Kolom 2 (Prediksi Netral) : 3 (Model berhasil memprediksi 3 data yang sebenarnya negatif menjadi netral).
- Kolom 3 (Prediksi Positif) : 3 (Model berhasil memprediksi 3 data yang sebenarnya negatif menjadi positif).

2 Baris 2 (Aktual Netral)

- Kolom 1 (Prediksi Negatif) : 10 (Model berhasil memprediksi 10 data yang sebenarnya netral menjadi negatif).
- Kolom 2 (Prediksi Netral) : 20 (Model berhasil memprediksi 20 data dengan benar sebagai netral).
- Kolom 3 (Prediksi Positif) : 1 (model berhasil memprediksi 1 data yang sebenarnya netral menjadi positif).

3 Baris 3 (Aktual Positif)

- Kolom 1 (Prediksi Negatif) : 10 (Model berhasil memprediksi 10 data yang sebenarnya positif menjadi negatif).

- Kolom 2 (Prediksi Netral) : 0 (Model berhasil memprediksi 0 data yang sebenarnya positif menjadi netral).
- Kolom 3 (Prediksi Positif) : 4 (Model berhasil memprediksi 4 data dengan benar sebagai positif).

Akurasi

$$= \frac{TP_{Negatif} + TP_{Netral} + TP_{Positif}}{\text{Total Jumlah Data}} \times 100\%$$

Total jumlah data : 76 + 3 + 3 + 10 + 21 + 1 + 10 + 0 + 4 = 128

$$\text{Akurasi} = \frac{76 + 21 + 4}{128} = \frac{101}{128} \times 100\% \\ = 0,78 \times 100\% = 78\%$$

Hasil tersebut menunjukkan bahwasannya dalam pengujian metode *Support Vector Machine* (SVM) dan pembobotan kata menggunakan ekstraksi fitur TF-IDF dan seleksi fitur GA dalam klasifikasi sentimen perihal pemindahan Ibu Kota Indonesia. Menghasilkan akurasi yang cukup baik yaitu sebesar 78,9%. Hasil akurasi dapat dilihat pada gambar 13.

```

--- Hasil Algoritma Genetika ---
Jumlah Fitur yang Diseleksi: 1981 dari 3776
Nilai C Optimal yang Ditemukan oleh GA: 60
Akurasi: 0.7891
Precision: 0.7886
Recall: 0.7891
F1 Score: 0.7743
Waktu eksekusi GA (termasuk pencarian C): 227.03 detik
Waktu Evaluasi Final: 0.12 detik

=== Classification Report ===

```

	precision	recall	f1-score	support
Negatif	0.79	0.93	0.85	82
Netral	0.88	0.66	0.75	32
Positif	0.58	0.29	0.36	14
accuracy			0.79	128
macro avg	0.72	0.62	0.66	128
weighted avg	0.78	0.79	0.77	128

Gambar 13 Hasil Klasifikasi Menggunakan GA

Penelitian ini memiliki beberapa keterbatasan yang perlu diperhatikan. Pertama, masalah imbalance data tidak sepenuhnya ditangani, sehingga berisiko menyebabkan bias pada model, terutama terhadap kelas mayoritas. Teknik seperti oversampling, undersampling, atau SMOTE yang dapat mengatasi masalah ini tidak diterapkan, yang dapat mempengaruhi akurasi model. Kedua, penggunaan data komentar dari YouTube yang diperoleh melalui teknik scraping dapat memengaruhi kualitas representasi sentimen, mengingat komentar-komentar tersebut mungkin tidak selalu mencerminkan sentimen yang konsisten

atau relevan. Selain itu, penelitian ini hanya menggunakan *Support Vector Machine* (SVM) sebagai model, tanpa membandingkannya dengan model lain yang mungkin lebih efektif dalam analisis sentimen. Terakhir, meskipun akurasi model sudah cukup baik, disarankan untuk menggunakan metrik evaluasi yang lebih robust, seperti F1-score, untuk menangani ketidakseimbangan data dan memberikan hasil yang lebih representatif.

4. KESIMPULAN

Berdasar hasil penelitian, dapat disimpulkan dua hal utama: (I) Dalam analisis sentimen keberlanjutan IKN pada komentar YouTube menggunakan metode *Support Vector Machine* (SVM) dengan seleksi fitur *Genetic Algorithm* (GA) dan ekstraksi fitur TF-IDF, akurasi model mencapai 78,9%. GA berhasil menyeleksi 1882 fitur dengan rasio 90:10. (II) Penggunaan GA untuk seleksi fitur efektif dalam mereduksi dimensi 3776 fitur TF-IDF, sehingga menghasilkan model yang lebih efisien dan akurat, mengurangi risiko overfitting. Perbandingan kinerja model menunjukkan bahwa SVM tanpa seleksi fitur (3776 fitur) mencapai akurasi 76%, sementara dengan GA (1882 fitur) mencapai 78,9%. Seleksi fitur berhasil mengurangi dimensi hingga 50%, meningkatkan efisiensi tanpa mengorbankan kinerja. (III) Pada perbandingan kinerja algoritma SVM antara yang menggunakan seleksi fitur dengan yang tidak, berdasarkan hasil evaluasi metrik akurasi, *presisi*, *recall*, dan *F1-Score* memiliki perbedaan sebesar 2.9% yang lebih tinggi akurasi menggunakan GA. Perbedaan pada F1-Score menggunakan GA sebesar 77% sedangkan tanpa menggunakan GA sebesar 71%.

5. SARAN

Berdasarkan penelitian mengenai Optimasi Seleksi Fitur TF-IDF Menggunakan *Genetic Algorithm* Pada Metode *Support Vector Machine* Dalam Menentukan Opini Publik Terhadap Kebijakan Keberlanjutan IKN, beberapa saran yang dapat diberikan adalah sebagai berikut:

1. Untuk penelitian selanjutnya, disarankan untuk memperluas jumlah data yang digunakan, lebih dari 1274 data sentimen, agar dapat memperoleh hasil akurasi yang lebih optimal dan menghindari keterbatasan data pada tahap klasifikasi.

2. Disarankan untuk mengembangkan dan menerapkan teknik *preprocessing* yang lebih maju, seperti *lemmatization*, analisis ironi/sarkasme, dan analisis emosi, untuk meningkatkan kemampuan model dalam memahami konteks bahasa yang lebih kompleks. Pada tahap *stopword removal*, penting untuk mengelola dan menghapus *stopword* baik yang berasal dari daftar *default library Sastrawi* maupun yang ditambahkan secara manual, dengan mengelola daftar *stopword* secara langsung.
3. Untuk penelitian berikutnya, disarankan untuk mengeksplorasi dan membandingkan metode klasifikasi lainnya selain SVM. Peneliti dapat mencoba metode seperti *Random Forest*, *Naive Bayes*, atau algoritma berbasis neural network seperti LSTM dan BERT yang dapat meningkatkan kinerja dalam analisis opini publik terhadap kebijakan keberlanjutan IKN. Selain akurasi, disarankan untuk menggunakan metrik evaluasi lain seperti *precision*, *recall*, *F1-Score*, dan AUC-ROC untuk memberikan gambaran yang lebih komprehensif mengenai kinerja model pada berbagai aspek.

UCAPAN TERIMA KASIH

Kami ingin mengucapkan terima kasih kepada semua pihak yang mendukung penelitian ini, termasuk Dosen Pembimbing Bapak Rudiman, S.Kom., M.Sc yang telah memberi arahan kepada penulis dalam penyusunan jurnal ini. Terima Kasih pula kepada pihak lain yang belum saya sebutkan satu persatu.

DAFTAR PUSTAKA

- [1] H. Dhery, A. Assyam, and F. N. Hasan, "Analisis Sentimen Twitter Terhadap Perpindahan Ibu Kota Negara Ke IKN Nusantara Menggunakan Orange Data Mining," *KLIK: Kajian Ilmiah Informatika dan Komputer*, vol. 4, no. 1, pp. 341–349, 2023, doi: 10.30865/klik.v4i1.957.
- [2] Alfiansyah, "ANAYASA (Journal of Legal Studies) IBU KOTA NEGARA BARU DALAM PANDANGAN HUKUM LINGKUNGAN," vol. 1, no. 1, 2023.
- [3] I. Kurniawan *et al.*, "Perbandingan Algoritma Naive Bayes Dan SVM Dalam Sentimen Analisis Marketplace Pada Twitter," *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 10, no. 1, 2023, [Online]. Available: <http://jurnal.mdp.ac.id>
- [4] P. Kumala Sari and R. Randy Suryono, "KOMPARASI ALGORITMA SUPPORT VECTOR MACHINE DAN RANDOM FOREST UNTUK ANALISIS SENTIMEN METAVERSE," 2024.
- [5] Y. A. Singgalen, "Analisis Performa Algoritma NBC, DT, SVM dalam Klasifikasi Data Ulasan Pengunjung Candi Borobudur Berbasis CRISP-DM," *Building of Informatics, Technology and Science (BITS)*, vol. 4, no. 3, Dec. 2022, doi: 10.47065/bits.v4i3.2766.
- [6] T. Gori, A. Sunyoto, and H. Al Fatta, "Preprocessing Data dan Klasifikasi untuk Prediksi Kinerja Akademik Siswa," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 11, no. 1, pp. 215–224, Feb. 2024, doi: 10.25126/jtiik.20241118074.
- [7] L. Efrizoni, S. Defit, M. Tajuddin, and A. Anggrawan, "Komparasi Ekstraksi Fitur dalam Klasifikasi Teks Multilabel Menggunakan Algoritma Machine Learning," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 21, no. 3, pp. 653–666, 2022, doi: 10.30812/matrik.v21i3.1851.
- [8] H. Utama, T. Informatika, U. A. Yogyakarta, A. Masruro, T. Informatika, and U. A. Yogyakarta, "Analisis Sentimen pada Twitter menggunakan Word Embedding dengan Pendekatan Word2Vec," vol. 05, no. 02, pp. 128–134, 2022.
- [9] Regina, T. H. Saragih, and D. Kartini, "Analisis Sentimen Brand Ambassador Bts Terhadap Tokopedia Menggunakan Klasifikasi Bayesian Network Dengan Ekstraksi Fitur Tf-Idf," *Jurnal Informatika Polinema*, vol. 9, no. 4, pp. 383–390, 2023, doi: 10.33795/jip.v9i4.1333.
- [10] A. Yodi Prayoga, A. Id Hadiana, and F. Rakhmat Umbara, "Deteksi Hoax pada Berita Online Bahasa Inggris Menggunakan Bernoulli Naïve Bayes dengan Ekstraksi Fitur Tf-Idf," *Jurnal Syntax Admiration*, vol. 2, no. 10, pp. 1808–1823, Oct. 2021, doi: 10.46799/jsa.v2i10.327.
- [11] K. Tri Putra, M. Amin Hariyadi, and C. Crysdian, "Perbandingan Feature Extraction Tf-Idf Dan Bow Untuk Analisis Sentimen Berbasis Svm," *Jurnal Cahaya MAndalika*, p. 1449, 2023.
- [12] F. Septianingrum, A. Susilo, and Y. Irawan, "Metode Seleksi Fitur Untuk Klasifikasi Sentimen Menggunakan Algoritma Naive Bayes : Sebuah Literature Review," vol. 5, pp. 799–805, 2021, doi: 10.30865/mib.v5i3.2983.
- [13] G. F. Grandis and Y. Arumsari, "Seleksi Fitur Gain Ratio pada Analisis Sentimen Kebijakan Pemerintah Mengenai Pembelajaran Jarak Jauh dengan K-Nearest Neighbor," vol. 5, no. 8, pp. 3507–3514, 2021.