JURNAL BUFFER INFORMATIKA
Volume 12 Nomor 1, April 2026

p-ISSN : 2527-4856, e-ISSN : 2614-5413
https://journal.fkom.uniku.ac.id/index.php/buffer

Optimasi Seleksi Fitur BERT Menggunakan GA Pada Metode KNN Dalam Menentukan Opini Publik Terkait Keberlanjutan IKN

Augie Sugiarto Nunka[1)] , Rudiman[2)] , Fendy Yulianto[3)]

[1,2,3,] *Program Studi Teknik Informatika, Fakultas Sains dan Teknologi, Universitas Muhammadiyah Kalimantan Timur*
Jl. Ir. H. Juanda No.15, Sidodadi, Kec. Samarinda Ulu, Kota Samarinda
Email :2111102441146@umkt.ac.id[1)], rudiman@umkt.ac.id [2)],Fy415@umkt.ac.id [3)]

Abstrak

Penelitian ini berfokus pada optimasi klasifikasi opini publik terkait keberlanjutan Ibu Kota Nusantara (IKN) yang beragam di media sosial. Tujuan utama dari penelitian ini adalah untuk meningkatkan performa klasifikasi sentimen dengan mengintegrasikan model *Bidirectional Encoder Representations from Transformers* (BERT) untuk ekstraksi fitur, Algoritma Genetika (*Genetic Algorithm*) untuk seleksi fitur, dan K-Nearest Neighbors (KNN) sebagai metode klasifikasi. Metode penelitian diawali dengan pengumpulan 1.274 data komentar dari *YouTube*, diikuti oleh pelabelan pakar, pra-pemrosesan data, dan ekstraksi fitur menggunakan IndoBERT yang menghasilkan 768 fitur. Algoritma Genetika kemudian diterapkan untuk menyeleksi fitur-fitur paling relevan. Hasil penelitian menunjukkan bahwa model tanpa seleksi fitur mencapai akurasi sebesar 76,56%. Sementara itu, model yang menggunakan seleksi fitur Algoritma Genetika berhasil mereduksi jumlah fitur menjadi 371 dan memperoleh akurasi sebesar 75,00%. Meskipun terjadi sedikit penurunan akurasi sebesar 1,56%, seleksi fitur terbukti mampu meningkatkan efisiensi komputasi secara signifikan dengan mengurangi dimensi fitur hingga 51,7% tanpa mengorbankan kinerja secara drastis, meskipun kedua model gagal dalam mengklasifikasikan kelas netral secara efektif.
Kata Kunci Algoritma Genetika; Analisis Sentimen; BERT;IKN; K-Nearest Neighbors

Abstract

This research focuses on optimizing the classification of public opinion from social media regarding the sustainability policy of Indonesia's new capital, Nusantara (IKN). The main objective is to enhance sentiment classification performance by integrating a Bidirectional Encoder Representations from Transformers (BERT) model for feature extraction, Genetic Algorithm for feature selection, and the K-Nearest Neighbors (KNN) algorithm as the classification method. The research method began with the collection of 1,274 comments from YouTube, followed by expert labeling, data preprocessing, and feature extraction using IndoBERT, which yielded 768 features. The Genetic Algorithm was then applied to select the most relevant features. The results show that the model without feature selection achieved an accuracy of 76.56%, while the model with Genetic Algorithm-based feature selection successfully reduced the feature count to 371 and obtained an accuracy of 75.00%. Although there was a slight accuracy decrease of 1.56%, feature selection proved to significantly improve computational efficiency by reducing feature dimensionality by 51.7% without drastically sacrificing performance; however, both models failed to effectively classify the neutral class.
Keywords: BERT; Genetic Algorithm; IKN; K-Nearest Neighbor; Sentiment Analysis

1. PENDAHULUAN

Pembangunan Ibu Kota Nusantara (IKN) merupakan salah satu Rencana Strategis Nasional yang berlokasi di Kecamatan Sepaku, Kabupaten Penajam Paser Utara, Provinsi Kalimantan Timur[1]. Pemindahan Ibu Kota Nusantara (IKN) memunculkan berbagai pro dan kontra, undang-undang yang mengatur pemindahan tersebut dinilai terlalu terburu-buru, meskipun pemerintah telah mengajak masyarakat Indonesia untuk turut mendukung pembangunan IKN [2]. Di sisi lain, terdapat sejumlah risiko dalam pemindahan ibu kota ke luar Pulau Jawa, terutama yang berkaitan dengan kesiapan wilayah tujuan dalam menyediakan infrastruktur yang memadai guna mendukung penyelenggaraan pemerintahan [1]. Untuk memahami opini masyarakat di media sosial terkait isu ekonomi, politik, dan hukum, dapat digunakan

pendekatan metode tradisional seperti *Lexicon-Based* [3]. Selain itu, beberapa metode modern juga telah dikembangkan, salah satunya adalah *Machine Learning*. *Machine Learning* merupakan bagian dari *Artificial Intelligence* (AI), didefinisikan sebagai metode pembelajaran dari pengalaman yang mampu menyesuaikan diri terhadap masukan baru untuk bertindak secara otomatis serta menghasilkan keputusan tepat dan sesuai [4]. *Machine* Learning memiliki beberapa pendekatan utama, antara lain *Clustering, Deep Learning,* dan Klasifikasi.

Klasifikasi digunakan untuk mengetahui seberapa besar sentimen yang berkembang di masyarakat terhadap suatu isu tertentu dengan membangun model yang mampu mengidentifikasi atau memetakan data ke dalam kategori berdasarkan karakteristiknya. Dalam konteks analisis sentimen, klasifikasi berfungsi untuk menentukan polaritas opini, yaitu apakah bersifat positif, negatif, atau netral [5]. Beberapa algoritma klasifikasi yang umum digunakan dalam bidang data *mining* antara lain *K-Nearest Neighbors* (KNN), *Decision Tree*, dan *Naive Bayes* [6]. Penelitian yang dilakukan oleh Louis & Dian, (2022) mengenai klasifikasi promosi karyawan menunjukkan metode *K-Nearest Neighbors* (KNN) memiliki kinerja terbaik dibandingkan dengan metode *Decision Tree* dan *Random Forest*, metode *Decision Tree* menghasilkan akurasi sebesar 85,29%, *Random Forest* sebesar 86,37%, sementara metode KNN memperoleh akurasi tertinggi sebesar 86,57%. Penelitian lain yang dilakukan oleh Mulyani et al., (2025) terkait klasifikasi penyakit diabetes menunjukkan metode KNN menghasilkan akurasi sebesar 83,33%, lebih tinggi dibandingkan metode *Support Vector Machine* (SVM) dengan akurasi sebesar 81,67%.

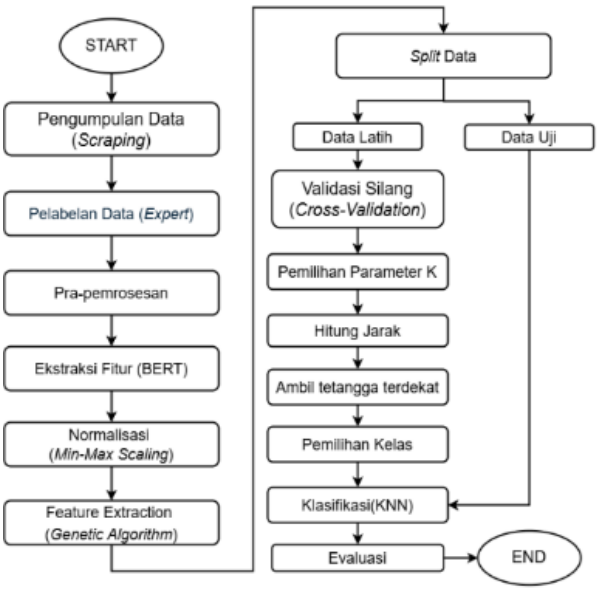
Tahap awal dalam klasifikasi teks adalah ekstraksi fitur (*feature extraction*), yaitu proses mengubah dokumen teks menjadi representasi fitur numerik yang dapat diumpankan ke sebuah *classifier* [9]. Proses *Word Embedding* bekerja dengan memetakan setiap kata dalam dokumen ke dalam *dense vector*, di mana vektor tersebut merepresentasikan proyeksi kata di ruang vektor [10]. Beberapa metode ekstraksi fitur yang umum digunakan meliputi TF-IDF (*Term Frequency–Inverse Document Frequency*) serta *Word Embeddings* seperti *Word2Vec*, *GloVe*, dan *BERT*, yang bertujuan untuk merepresentasikan kata-kata ke dalam bentuk vektor numerik [11]. Penelitian ini menerapkan metode *Bidirectional Encoder Representations from Transformers* (BERT) karena keunggulannya dalam memahami konteks kalimat secara dua arah (*bidirectional*), sehingga mampu mendeteksi nuansa kompleks dan makna tersirat dalam teks [8].

Proses *ekstraksi fitur* memiliki kelemahan karena dapat menghasilkan dimensi data yang sangat tinggi, sehingga memicu *overfitting* dan meningkatkan waktu komputasi. *Seleksi fitur* merupakan metode untuk menghilangkan atau mengurangi fitur yang kurang relevan [12]. Penggunaan seleksi fitur juga dapat meningkatkan akurasi algoritma dengan mempertahankan fitur-fitur yang memiliki 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).Algoritma Genetika (GA) adalah sebuah metode optimasi yang terinspirasi dari proses evolusi biologis dan seleksi alam. Penelitian terkait metode seleksi fitur yang dilakukan oleh Abdullah & Haddi Hassan, (2021), ditemukan bahwa *Genetic Algorithm* (GA) menghasilkan akurasi tertinggi sebesar 98,40%, dibandingkan metode *Particle Swarm Optimization* (PSO) dengan akurasi 90,10%.

Menurut Akbar, (2024) algoritma *K-Nearest Neighbor* (KNN) adalah teknik *Lazy Learning* yang termasuk dalam kategori *Instance-Based Learning*, di mana prinsip kerjanya adalah mencari jarak terdekat antara data yang akan dievaluasi dengan k tetangga terdekatnya dalam data pelatihan. KNN memiliki keunggulan dalam menghasilkan klasifikasi yang andal serta efektif diterapkan pada kumpulan data berukuran besar [16]. Penelitian ini bertujuan mengoptimalkan klasifikasi opini publik terkait keberlanjutan Ibu Kota Nusantara (IKN) yang tersebar di media sosial. Tujuan utamanya adalah meningkatkan kinerja klasifikasi sentimen melalui integrasi model *Bidirectional Encoder Representations* from *Transformers* (BERT) untuk ekstraksi fitur, Algoritma Genetika (*Genetic Algorithm*) untuk seleksi fitur, dan *K-Nearest Neighbors* (KNN) sebagai metode klasifikasi.

2. METODE PENELITIAN

Metode penelitian dilakukan melalui langkah-langkah atau alur yang dirancang agar penelitian dapat berjalan dengan baik. Alur metode penelitian dapat dilihat pada Gambar 1.



Gambar 1 Alur penelitian

Penelitian ini terdiri atas beberapa tahapan sebagaimana ditunjukkan pada Gambar 1. Tahapan tersebut meliputi proses ekstraksi fitur menggunakan *Bidirectional Encoder Representations from Transformers* (BERT), dilanjutkan dengan seleksi fitur menggunakan algoritma *Genetic Algorithm* (GA), serta klasifikasi dengan algoritma *K-Nearest* Neighbors (KNN). Seluruh data diproses melalui tahapan-tahapan yang dijelaskan pada bagian berikut.

2.1 Pengumpulan Data (*Scraping*)

Pengumpulan data dilakukan menggunakan teknik *web scraping* yang bersumber dari kolom komentar pada video *YouTube* yang relevan dengan topik IKN. Web *Scraping* adalah metode untuk mengekstraksi data dari *world wide web* (WWW) dan menyimpannya ke sistem *file* atau *database* untuk pengambilan atau analisis selanjutnya [17].

2.2 Pelabelan Data (*Expert*)

Dalam penelitian ini, proses pelabelan data dilakukan melalui layanan projects.co.id oleh H. Irfan Abdul Hakim, S.Sos., M.A., yang memiliki latar belakang sosiologi dan analisis data. Pelabelan sentimen dibagi ke dalam tiga kategori, yaitu positif (mengandung makna positif), netral (bersifat deskriptif atau informatif tanpa kecenderungan emosional), dan negatif (berisi konotasi negatif seperti kritik, sindiran, atau penolakan).

2.3 Pra-pemrosesan (*Pre-processing*)

Setelah tahap pelabelan data, langkah berikutnya adalah pra-pemrosesan dataset yang mencakup beberapa proses, yaitu *case folding* (mengubah seluruh karakter menjadi huruf kecil), *cleaning* (menghapus elemen yang tidak diperlukan seperti kode HTML, emotikon, *hashtag*, *mention*, dan URL), *tokenizing* (memecah teks menjadi token berupa kata atau karakter dengan menghilangkan tanda baca, angka, dan simbol yang tidak relevan), *stopword removal* (menghapus kata-kata yang tidak memiliki makna signifikan), *stemming* (mengubah kata ke bentuk dasar dengan menghilangkan imbuhan), serta normalisasi (mentransformasi kata tidak baku, slang, singkatan, atau bentuk penulisan tidak sesuai kaidah bahasa menjadi bentuk standar).

2.4 Ekstraksi Fitur

Bidirectional Encoder Representations from Transformers (BERT) merupakan model bahasa berbasis arsitektur *transformer* yang dikembangkan oleh Google. Dalam penelitian ini, digunakan varian BERT yang telah disesuaikan untuk bahasa Indonesia, yaitu IndoBERT. IndoBERT merupakan hasil adaptasi dari BERT yang dilatih menggunakan korpus teks berbahasa Indonesia dalam jumlah besar.

2.5 Normalisasi (*Min-Max Scaling*)

Normalisasi data merupakan salah satu langkah penting dalam pra-pemrosesan untuk mengurangi redundansi dan menyetarakan skala fitur. Pada penelitian ini digunakan metode *Min-Max Normalization*, di mana data asli ditransformasikan secara linier ke dalam rentang 0 hingga 1 agar skala data lebih kecil dan akurasi pelatihan model meningkat [18]. Proses normalisasi dilakukan dengan rumus berikut.

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Dengan keterangan: X_{scaled} adalah nilai fitur setelah dinormalisasi, X adalah nilai asli, X_{min} adalah nilai minimum pada fitur, dan X_{max} adalah nilai maksimum pada fitur.

2.6 Seleksi Fitur

Pada tahap ini dilakukan seleksi fitur menggunakan *Genetic Algorithm* (*GA*) terhadap hasil ekstraksi fitur dari *Bidirectional Encoder Representations from Transformers* (*BERT*). Proses ini bertujuan memperoleh subset atau kombinasi fitur yang paling relevan dan optimal sehingga akurasi serta efisiensi algoritma *K-Nearest Neighbors* (*KNN*) dalam mengklasifikasikan opini publik terkait kebijakan keberlanjutan Ibu Kota Nusantara (IKN) dapat meningkat.

2.7 Rasio Data dalam Split Data

Pada tahap ini dilakukan proses *split data* untuk membagi dataset opini publik terkait keberlanjutan Ibu Kota Nusantara (IKN) menjadi data latih (*training data*) dan data uji (*testing data*).

Model dibangun menggunakan data latih dan dievaluasi dengan data uji untuk menilai kinerja serta kemampuan generalisasi (Oktafiani, Hermawan, and Avianto 2023; Rahmadani, Rahim, and Rudiman 2024). Dalam penelitian ini digunakan skema *split data* sebesar 60:40 sesuai acuan Adhi Putra, (2021).

2.8 Validasi Silang (*Cross-Validation*)

K-Fold Cross-Validation merupakan teknik validasi yang digunakan untuk menilai kinerja generalisasi model dan mengurangi risiko *overfitting* dengan cara melatih serta menguji model secara berulang pada subset data yang berbeda [22]. Pada metode ini, seluruh data dibagi ke dalam beberapa lipatan (*fold*), di mana setiap lipatan secara bergantian digunakan sebagai data uji sementara lipatan lainnya berfungsi sebagai data latih, sehingga setiap lipatan memperoleh kesempatan yang sama sebagai data uji. Dengan demikian, *K-Fold Cross-Validation* mampu menghasilkan estimasi kinerja yang lebih stabil dan andal.

Dalam penelitian ini, metode tersebut tidak hanya digunakan untuk menentukan nilai K yang optimal pada algoritma *K-Nearest Neighbors* (*KNN*), tetapi juga dijadikan dasar evaluasi nilai *fitness* dalam proses seleksi fitur menggunakan *Genetic Algorithm* (*GA*). Jumlah lipatan yang digunakan ditetapkan sebanyak 10, merujuk pada konfigurasi parameter terbaik yang diusulkan oleh Adhi Putra, (2021).

2.9 Klasifikasi *K-Nearest Neighbor* (KNN)

Metode *K-Nearest Neighbor* (*KNN*) digunakan dalam penelitian ini karena kesederhanaan dan kemudahan implementasinya, di mana data baru diklasifikasikan berdasarkan kedekatannya dengan data latih menggunakan pendekatan pembobotan sederhana namun efektif [23]. Proses klasifikasi KNN meliputi beberapa tahap, yaitu penentuan parameter K (jumlah tetangga terdekat), perhitungan jarak menggunakan rumus *Euclidean Distance*, pemilihan K tetangga terdekat berdasarkan jarak terkecil, dan penentuan kelas prediksi melalui proses *voting* mayoritas.

$$d(x, x_i) = \sqrt{\sum_{j-1}^{m} (xj - xij)^2}$$

Dalam penelitian ini, nilai K = 9 digunakan sebagai acuan berdasarkan penelitian Adhi Putra, (2021) untuk memperoleh hasil klasifikasi yang optimal.

2.10 Evaluasi

Evaluasi dilakukan untuk menilai kinerja model klasifikasi yang digunakan dalam penelitian ini. Metode evaluasi yang diterapkan adalah *Confusion Matrix*, yaitu matriks prediksi yang dibandingkan dengan kategori asli dari data masukan [24]. Pendekatan *One-vs-All* (*OvA*) digunakan untuk mengurai masalah klasifikasi multi-kelas menjadi beberapa klasifikasi biner, sehingga memungkinkan penerapan metrik evaluasi biner seperti *True Positive* (TP), *False Positive* (FP), dan *False Negative* (FN) pada setiap kelas secara spesifik [25]. Dengan demikian, *Confusion Matrix* berfungsi sebagai alat visualisasi yang efektif untuk menganalisis seberapa baik model dapat mengenali data dari kelas yang berbeda [26].

Empat metrik utama digunakan untuk mengukur performa model, yaitu akurasi, presisi, *recall*, dan *F1-score*. Akurasi mengukur persentase prediksi yang benar terhadap total data, presisi menunjukkan ketepatan model dalam memprediksi kelas positif, *recall* menilai kemampuan model dalam mendeteksi data positif yang sebenarnya, sedangkan *F1-score* merupakan rata-rata harmonis dari presisi dan *recall* yang berguna saat distribusi data antar kelas tidak seimbang. Evaluasi dilakukan pada setiap model, baik sebelum maupun sesudah optimasi menggunakan *Genetic Algorithm* (GA), untuk menilai peningkatan performa dalam mengklasifikasikan opini publik berdasarkan fitur hasil ekstraksi *BERT*, sehingga diperoleh model terbaik yang digunakan dalam penelitian.

3. HASIL DAN PEMBAHASAN

3.1 Hasil Pengumpulan Data

Data penelitian ini diperoleh dari 1.274 komentar pengguna pada salah satu video *YouTube* yang membahas mengenai Ibu Kota Negara (IKN). Pengambilan data dilakukan secara otomatis menggunakan teknik *web scraping* dengan bahasa pemrograman. Teknik ini memungkinkan proses pengumpulan data dilakukan secara efisien tanpa perlu mengakses komentar secara manual, sehingga mempercepat dan mempermudah proses akuisisi data. Hasil pengambilan data disajikan pada Tabel 1.

Tabel 1 Hasil *web scrapping youtube* (sampel 5 data)

Nama	Komentar
@MjoleSANJAYASanjaya	Mantap bas ikn penting ibukota ikn top markotop❤👍💪🙏
@budilesnoto4598	Korupsi di pertamina kan lebih buat nerusin pembangunan ikn..
@Tamyiz	Ada apa Bpk ini ambisi sekali mindahkan IKN ya?
@JayadiKopana	Monyet² ikn lagi bikin singgasana buat para monyet² politik
@EhaJulaeha-p4p	Smoga Bos IKN dan SATGAS IKN Sehat Selalu.....

Tabel 1 menampilkan lima sampel data komentar mentah (*raw data*) yang diperoleh melalui proses *web scraping*. Data tersebut terdiri atas nama pengguna serta isi komentar yang mencerminkan opini publik terhadap isu pembangunan IKN. Sampel ini hanya merupakan sebagian kecil dari keseluruhan data yang digunakan dalam penelitian

3.2 Hasil Labeling Data

Proses pelabelan data dilakukan dengan bantuan seorang pakar yang memiliki kompetensi di bidang analisis sentimen. Proses ini bertujuan untuk memastikan akurasi klasifikasi awal sehingga dapat digunakan sebagai dasar pembelajaran model dalam tahap pemrosesan selanjutnya. Hasil pelabelan sampel data ditampilkan pada Tabel 2.

Tabel 2 Hasil pelabelan data

Komentar	Sentimen
Mantap bas ikn penting ibukota ikn top markotof❤👍💪🙏	Positif
Korupsi di pertamina kan lebih buat nerusin pembangunan ikn..	Negatif
Ada apa Bpk ini ambisi sekali mindahkan IKN ya?	Netral
Monyet² ikn lagi bikin singgasana buat para monyet² politik	Negatif
Smoga Bos IKN dan SATGAS IKN Sehat Selalu.....	Positif

Tabel 2 menyajikan hasil pelabelan data komentar yang diperoleh dari *YouTube*. Data diklasifikasikan ke dalam tiga kelas sentimen, yakni positif sebanyak 294 komentar, negatif sebanyak 831 komentar, dan netral sebanyak 149 komentar. Proses pelabelan ini memberikan struktur yang jelas terhadap data sehingga memudahkan model klasifikasi dalam mempelajari pola sentimen serta berkontribusi dalam meningkatkan kinerja model secara keseluruhan.

3.3 Hasil *Pre-Processing*

Pada tahap ini, data melalui proses *preprocessing* untuk dibersihkan dan dipersiapkan sebelum digunakan dalam pelatihan model klasifikasi. Proses ini bertujuan menghapus elemen yang tidak relevan serta menyederhanakan struktur data agar lebih efisien secara komputasional. Beberapa tahapan yang dilakukan meliputi normalisasi, penghapusan simbol atau karakter khusus, serta penyederhanaan bentuk kata sehingga konsisten dan mudah dipahami sistem, khususnya dalam analisis berbasis *machine learning*. Setelah tahap *preprocessing* selesai, hasil akhirnya ditampilkan pada Tabel 3.

Tabel 3 Hasil akhir *pre-processing*

Komentar	Hasil Akhir
Mantap bas ikn penting ibukota ikn top markotof❤👍💪🙏	mantap banget ikn penting ibu ikn top markotop
Korupsi di pertamina kan lebih buat nerusin pembangunan ikn..	korupsi pertamina mungkin lebih untuk menerus bangun ikn
Ada apa Bpk ini ambisi sekali mindahkan IKN ya?	apa bapak ambisi sekali memindah ikn
Monyet² ikn lagi bikin singgasana buat para monyet² politik	monyet ikn membuat singgasana untuk monyet politik
Smoga Bos IKN dan SATGAS IKN Sehat Selalu.....	semoga bos ikn satgas ikn sehat selalu

Berdasarkan Tabel 3 penerapan tahap *preprocessing* membuat data komentar lebih bersih, terstruktur, dan siap digunakan untuk proses *feature extraction* serta pelatihan model klasifikasi. Tahapan ini berperan penting dalam memastikan kualitas data bebas dari *noise* sehingga dapat meningkatkan efektivitas model dan menghasilkan akurasi prediksi sentimen yang lebih optimal.

3.4 Hasil Ekstraksi Fitur (BERT)

Pada tahap ini, data yang telah melalui proses pra-pemrosesan diekstraksi fiturnya menggunakan *IndoBERT*. Proses ini bertujuan untuk mengubah data teks menjadi representasi vektor numerik yang dapat dipahami oleh algoritma pembelajaran mesin. *IndoBERT* merupakan model *pre-trained* berbasis *BERT* yang dilatih secara khusus untuk bahasa Indonesia sehingga mampu menangkap makna semantik dalam teks dengan lebih baik. Hasil ekstraksi fitur ditampilkan pada Tabel 4.

Tabel 4 Hasil ekstraksi fitur dari vektor

Data	Fitur_1, Fitur_2, Fitur_3, ... , Fitur_768
Kalimat 1	([-0.0387, 0.0768, -0.0101, ... -0.2672,])
Kalimat 2	([-0.3155, 0.1337, -0.2458, ... 0.0216,])
Kalimat 3	([-0.2751, 0.1042, -0.4319, ... 0.0305,])
Kalimat 4	([-0.3582, 0.0331, -0.1884, ... , -0.0142])
Kalimat 5	([-0.2523, 0.0817, -0.1251, ... , -0.0163])

Tabel 4 menyajikan hasil representasi vektor dari setiap kalimat dengan dimensi 768 fitur numerik. Representasi tersebut merepresentasikan makna semantik dari teks dan menjadi masukan bagi proses klasifikasi sentimen. Dengan demikian, setiap data dapat dipetakan ke dalam kategori Positif, Negatif, atau Netral berdasarkan hasil pemrosesan vektor.

3.5 Hasil Normalisasi (*Min-Max Scaling*)

Sebelum dilakukan proses seleksi fitur, seluruh vektor hasil ekstraksi dari *Bidirectional Encoder Representations from Transformers* (BERT) dinormalisasi menggunakan metode *Min-Max Scaling*. Normalisasi ini bertujuan untuk mengubah skala nilai fitur ke dalam rentang 0 hingga 1 agar setiap fitur memiliki bobot yang sebanding.

Hasil perhitungan sederhana ditampilkan pada Tabel 5.

Tabel 5 Hasil normalisasi

Data ke-	Fitur BERT ke-			Normalisasi fitur BERT ke-		
	1	2	3	1	2	3
1	0,31	3,08	-0,33	0,47	0,82	0,32
2	-0,68	0,86	0,56	0,29	0,2	0,7
3	0,96	1,63	0,59	0,74	0,41	0,71
4	-0,22	1,50	-0,61	0,141	0,38	0,2
5	0,12	1,72	-0,26	0,51	0,44	0,35

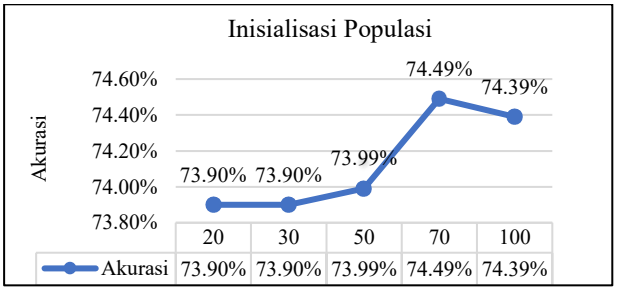
Tabel 5 memperlihatkan hasil normalisasi pada tiga fitur pertama dari total 768 fitur yang dihasilkan oleh BERT. Setiap nilai telah dipetakan ke dalam rentang 0–1 sesuai prinsip *Min-Max Scaling*, sehingga distribusi fitur menjadi lebih seragam. Setelah proses ini, tahap selanjutnya adalah seleksi fitur untuk menentukan subset fitur yang paling relevan dan berkontribusi optimal terhadap kinerja klasifikasi.

3.6 Seleksi Fitur *Genetic Algorithm*

Hasil ekstraksi fitur yang telah dinormalisasi kemudian diseleksi menggunakan *Genetic Algorithm* (GA) untuk mendapatkan subset fitur terbaik.

1. Inisialisasi Populasi

Pengujian dilakukan dengan variasi ukuran populasi (20, 30, 50, 70, 100). Hasil menunjukkan populasi 70 dengan 10 generasi memberikan akurasi tertinggi sebesar 74,49%. Pengujian dapat dilihat pada Gambar 2.

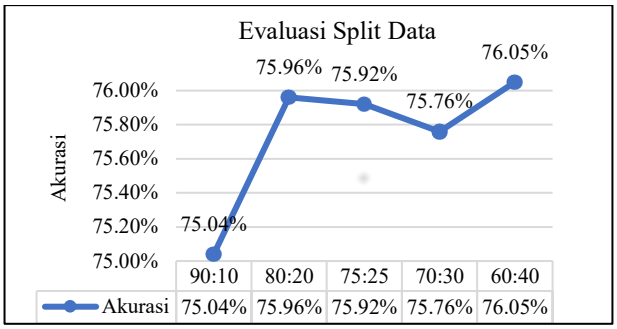


Gambar 2 Grafik pengujian Ukuran Populasi

2. Evaluasi *Fitness*

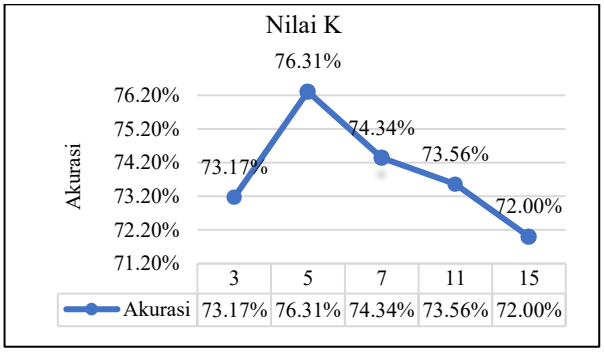
Setiap individu (kromosom) dievaluasi menggunakan algoritma *K-Nearest Neighbors* (KNN). Pengujian parameter meliputi rasio *split* data, nilai K, dan jumlah *K-Fold*. Hasil menunjukkan:

- Rasio terbaik: 60:40 dengan akurasi 76,05% (Gambar 3).



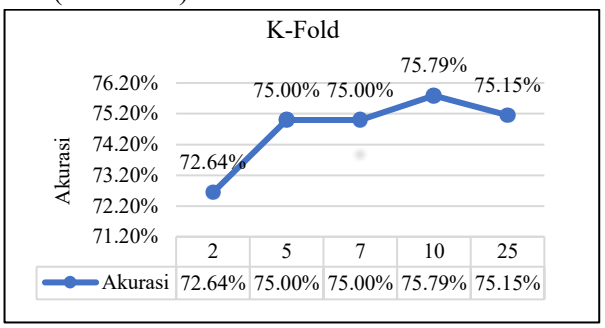
Gambar 3 Grafik pengujian *Split* data

- Nilai K optimal: K=5 dengan akurasi 76,31% (Gambar 4).



Gambar 4 Grafik pengujian Nilai K

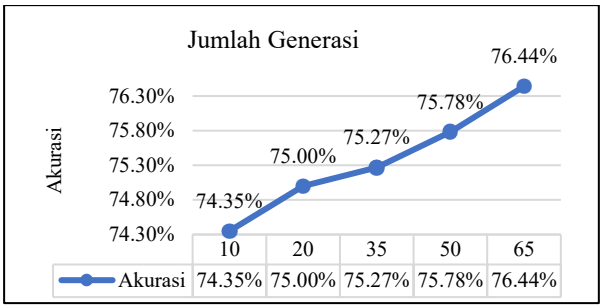
- Teknik validasi silang terbaik: 10-Fold (Gambar 5).



Gambar 5 Grafik pengujian K-Fold

3. Iterasi (Generasi)

Tahap akhir *Genetic Algorithm* (GA) dilakukan melalui proses evolusi yang mencakup seleksi, *crossover*, dan mutasi hingga jumlah generasi yang ditentukan tercapai. Pengujian dengan variasi jumlah generasi (10, 20, 35, 50, dan 65) menunjukkan bahwa konfigurasi terbaik diperoleh pada 65 generasi dengan ukuran populasi 70, menghasilkan akurasi sebesar 76,44%. Hasil pengujian variasi jumlah generasi ditunjukkan pada Gambar 6.



Gambar 6 Grafik pengujian jumlah generasi

Berdasarkan hasil tersebut, diperoleh konfigurasi parameter optimal GA dan KNN yang ditunjukkan pada Tabel 6.

Tabel 6 Konfigurasi parameter terbaik

Parameter	Nilai
Ukuran Populasi	70
Jumlah Generasi	65
Tingkat Mutasi	1%
Tingkat *Crossover*	80%
Split data	60:40
Nilai *K*	5
K-Fold	10

Dengan konfigurasi tersebut, GA dijalankan secara penuh untuk mengevaluasi performa individu terbaik. Hasil pengujian ditunjukkan pada Tabel 7.

Tabel 7 Hasil pengujian parameter terbaik GA

Individu	Akurasi (*Fitnes*)	Jumlah Fitur Terpilih
Kromosom_7	76.05%	371
Kromosom_65	75.92%	376
Kromosom_19	75.92%	369
Kromosom_67	75.79%	366
Kromosom_46	75.79%	371

Berdasarkan Tabel 7 individu terbaik adalah *Kromosom_7* dengan akurasi 76,05% dan 371 fitur terpilih dari 768 fitur awal. Subset fitur tersebut ditetapkan sebagai kombinasi paling optimal untuk melatih dan menguji model klasifikasi final.

3.7 Hasil Pengujian Tanpa Seleksi Fitur

Setelah diperoleh konfigurasi parameter terbaik, dilakukan pengujian terhadap model klasifikasi *K-Nearest Neighbors* (KNN) dengan menggunakan seluruh 768 fitur hasil ekstraksi IndoBERT tanpa seleksi fitur. Tahap ini bertujuan memperoleh kinerja dasar (*baseline*) yang digunakan sebagai pembanding terhadap model yang dioptimasi melalui seleksi fitur. Hasil klasifikasi ditunjukkan pada Tabel 8.

Tabel 8 *Confusion Matrix* Pengujian Tanpa Seleksi Fitur

	Pred: Negatif	Pred: Netral	Pred: Positif
Actual: Negatif	82	0	1
Actual: Netral	15	0	0
Actual: Positif	13	1	16

Ringkasan metrik evaluasi ditunjukkan pada Tabel 9.

Tabel 9 Laporan Evaluasi Kinerja Model Pengujian Tanpa Seleksi Fitur

	Precision	Recall	F1-Score
Negatif	0.745	0.988	0.849
Netral	0.000	0.000	0.000
Positif	0.941	0.533	0.680
Akurasi			**0.7656**

Berdasarkan hasil evaluasi, model menunjukkan kinerja baik pada kelas Negatif dan Positif, namun gagal mendeteksi kelas Netral. Hal ini diduga disebabkan oleh ketidakseimbangan distribusi data atau keterbatasan fitur dalam membedakan kelas tersebut.

3.8 Hasil Pengujian Dengan Seleksi Fitur

Pada tahap ini, pengujian dilakukan terhadap model klasifikasi *K-Nearest Neighbors* (KNN) dengan menggunakan fitur yang telah diseleksi melalui proses optimasi *Genetic Algorithm* (GA). Dari total 768 fitur hasil ekstraksi IndoBERT, sebanyak 371 fitur terpilih sebagai subset paling relevan berdasarkan evaluasi *fitness*. Proses klasifikasi dilakukan dengan menghitung jarak antar data menggunakan rumus *Euclidean*, sebagaimana pada pengujian tanpa seleksi fitur. Hasil klasifikasi ditunjukkan pada Tabel 10.

Tabel 10 *Confusion Matrix* Pengujian Dengan Seleksi Fitur

	Pred: Negatif	Pred: Netral	Pred: Positif
Actual: Negatif	81	0	2
Actual: Netral	15	0	0
Actual: Positif	15	0	15

Selanjutnya, perhitungan metrik evaluasi berdasarkan confusion matrix ditunjukkan pada Tabel 11.

Tabel 11 Laporan Evaluasi Kinerja Model Tanpa Seleksi Fitur

	Precision	Recall	F1-Score
Negatif	0.730	0.976	0.835
Netral	0.000	0.000	0.000
Positif	0.882	0.500	0.638
Akurasi			**0.750**

Berdasarkan hasil evaluasi, model menunjukkan kinerja cukup baik dalam mengklasifikasikan kelas Negatif dan Positif. Namun, serupa dengan pengujian tanpa seleksi fitur, model masih gagal mendeteksi kelas Netral. Hal ini diduga dipengaruhi oleh ketidakseimbangan distribusi data antar kelas atau keterbatasan fitur dalam membedakan kelas tersebut secara efektif.

3.9 Perbandingan dan Pembahasan

Bagian ini membahas perbandingan kinerja model *K-Nearest Neighbors* (KNN) pada dua skenario, yaitu tanpa seleksi fitur dan dengan seleksi fitur menggunakan *Genetic Algorithm* (GA). Analisis difokuskan pada dampak seleksi fitur terhadap efektivitas dan efisiensi model klasifikasi.

1. Perbandingan Hasil Evaluasi Model

Perbandingan dilakukan terhadap metrik akurasi serta nilai *F1-Score* pada setiap kelas. Hasil perbandingan ditunjukkan pada Tabel 12.

Tabel 12 Perbandingan kinerja model KNN tanpa dan dengan seleksi fitur

Metrik Evaluasi	Tanpa Seleksi Fitur	Dengan Seleksi Fitur	Perubahan
Jumlah Fitur	768	371	Berkurang 397 fitur
Akurasi	76%	75%	Berkurang 1%
F1-Score (Negatif)	0.849	0.835	Berkurang 0.014
F1-Score (Positif)	0.680	0.638	Berkurang 0.042
F1-Score (Netral)	0.000	0.000	Tidak ada

Tabel 3.34 menunjukkan bahwa model tanpa seleksi fitur memiliki akurasi dan *F1-Score* yang sedikit lebih tinggi. Penurunan akurasi sebesar 1% serta *F1-Score* pada kelas Negatif dan Positif menunjukkan bahwa dampak seleksi fitur terhadap performa model relatif kecil. Kedua model juga gagal dalam mengklasifikasikan kelas Netral, yang tercermin dari nilai *F1-Score* sebesar 0.

2. Analisis Dampak Seleksi Fitur pada Kinerja KNN

Penerapan GA berhasil mereduksi jumlah fitur dari 768 menjadi 371 (pengurangan sebesar 51,7%). Reduksi ini meningkatkan efisiensi model dengan mempercepat proses pelatihan dan prediksi, serta mengurangi risiko *overfitting* akibat keberadaan fitur yang tidak relevan.

Meskipun akurasi model dengan seleksi fitur mengalami penurunan sebesar 1,56% dibandingkan dengan model tanpa seleksi fitur, penurunan ini tergolong kecil dan masih dapat diterima. Kedua model tetap gagal dalam mengklasifikasikan kelas Netral, yang diduga disebabkan oleh ketidakseimbangan jumlah data antar kelas dan keterbatasan fitur dalam membedakan karakteristik kelas tersebut. Dengan demikian, seleksi fitur memberikan manfaat utama dalam efisiensi tanpa mengorbankan performa model secara signifikan.

4. KESIMPULAN

Berdasarkan analisis dan pembahasan hasil penelitian, dapat ditarik dua kesimpulan utama yang menjawab rumusan masalah yaitu (i) Optimasi parameter untuk model klasifikasi *K-Nearest Neighbors* (KNN) menunjukkan bahwa konfigurasi terbaik untuk kasus klasifikasi opini publik terkait IKN adalah penggunaan rasio pembagian data 90:10, metode validasi silang 10-Fold *Cross-Validation*, dan nilai K=3. (ii) Perbandingan kinerja model menunjukkan bahwa KNN tanpa seleksi fitur (menggunakan 768 fitur) mencapai akurasi 76%, sementara model dengan seleksi fitur *Genetic Algorithm* (menggunakan 371 fitur) mencapai akurasi 75%. Meskipun terjadi sedikit penurunan akurasi, seleksi fitur berhasil mereduksi dimensi fitur hingga 51,7%. Hal ini membuktikan bahwa metode seleksi fitur mampu meningkatkan efisiensi model secara signifikan tanpa mengorbankan kinerja secara drastis.

5. SARAN

Untuk penelitian di masa mendatang, beberapa pengembangan dapat dilakukan, antara lain: (i) Mengatasi permasalahan ketidakseimbangan kelas data, terutama pada kelas Netral, dengan menerapkan teknik seperti *oversampling* (misalnya metode SMOTE) atau *undersampling*. (ii) Mengeksplorasi dan membandingkan algoritma seleksi fitur lain seperti *Particle Swarm Optimization* (PSO). (iii) Menggunakan dan membandingkan algoritma klasifikasi lain di luar KNN, seperti *Support Vector Machine*, *Decision Tree*, atau *Random Forest*. (iv) Meningkatkan generalisasi model dengan menambah variasi sumber data, tidak hanya dari satu video *YouTube*, tetapi juga dari platform media sosial lain atau portal berita daring. (v) Melakukan *fine-tuning* model IndoBERT pada korpus data yang lebih spesifik dan relevan dengan konteks kebijakan untuk meningkatkan kualitas representasi fitur.

DAFTAR PUSTAKA

[1] T. R. Ayuningtyas, N. Laila, I. Septiana, A. Sallwa, and B. Rizki, "Analisa Hukum Terhadap Pengaturan Hak Guna Usaha di Ibu Kota Negara," vol. 6, no. 4, pp. 11766–11776, 2024.

[2] C. Huda and M. Betty Yel, "Analisa Sentimen Tentang Ibu Kota Nusantara (IKN) Dengan Menggunakan Algoritma K-Nearest Neighbors (KNN) dan Naïve Bayes," *J. Ilmu Komput. dan Sist. Inf.*, vol. 7, no. 1, pp. 126–130, 2024, doi: 10.55338/jikomsi.v7i1.2846.

[3] S. Prasetyo and T. Dewayanto, "Penerapan Machine Learning, Deep Learning, Dan Data Mining Dalam Deteksi Kecurangan Laporan Keuangan-a Systematic Literature Review," *Diponegoro J. Account.*, vol. 13, no. 3, pp. 1–12, 2024, [Online]. Available: http://ejournal-s1.undip.ac.id/index.php/accounting

[4] M. Bagas Dikal Putra and E. Setiawan, "Metode Lexicon Based Untuk Analisis Sentimen Pengguna Twitter Terhadap Kinerja Isp," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 6, pp. 12079–12087, 2024, doi: 10.36040/jati.v8i6.11765.

[5] I. Tupan Tri Muryono, Ahmad Taufik, "PERBANDINGAN ALGORITMA K-NEAREST NEIGHBOR, DECISION TREE, DAN NAIVE BAYES UNTUK MENENTUKAN KELAYAKAN PEMBERIAN KREDIT," *INFOTECH J. Technol. Inf.*, vol. 7, no. 1, pp. 55–62, 2021.

[6] A. Budiyantara, I. Irwansyah, E. Prengki, P. A. Pratama, and N. Wiliani, "Komparasi Algoritma Decision Tree, Naive Bayes Dan K-Nearest Neighbor Untuk Memprediksi Mahasiswa Lulus Tepat Waktu," *JITK (Jurnal Ilmu Pengetah. dan Teknol. Komputer)*, vol. 5, no. 2, pp. 265–270, 2020, doi: 10.33480/jitk.v5i2.1214.

[7] M. S. Louis and B. S. Dian, "Perbandingan Algoritma Knn, Decision Tree,*Dan Random*Forest Pada Data Imbalanced Class Untuk Klasifikasi Promosi Karyawan," *J. INSTEK (Informatika Sains dan Teknol.*, vol. 7, no. 2, pp. 192–200, 2022, doi: 10.24252/instek.v7i2.31385.

[8] A. Mulyani, S. Khoerunisa, and D. Kurniadi, "Comparison of KNN and SVM Algorithms Performance Using SMOTE to Classify Diabetes," vol. 14, no. 1, pp. 25–34, 2025.

[9] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep Learning Based Text Classification: A Comprehensive Review," vol. 1, no. 1, pp. 1–43, 2020, [Online]. Available: http://arxiv.org/abs/2004.03705

[10] L. Efrizoni, S. Defit, M. Tajuddin, and A. Anggrawan, "Komparasi Ekstraksi Fitur dalam Klasifikasi Teks Multilabel Menggunakan Algoritma Machine Learning," *MATRIK J. Manajemen, Tek. Inform. dan Rekayasa Komput.*, vol. 21, no. 3, pp. 653–666, 2022, doi: 10.30812/matrik.v21i3.1851.

[11] N. E. Azarin, R. A. Saputra, and Subardin, "Prediksi Popularitas Novel Berbasis Fitur-Fitur Teks Menggunakan Metode Random Forest," vol. 9, no. 1, pp. 57–62, 2024.

[12] G. Fadillah Grandis and Y. Arumsari, "Seleksi Fitur Gain Ratio pada Analisis Sentimen Kebijakan Pemerintah Mengenai Pembelajaran Jarak Jauh dengan K-Nearest Neighbor," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 8, pp. 3507–3514, 2021, [Online]. Available: http://j-ptiik.ub.ac.id

[13] F. Septianingrum and A. S. Y. Irawan, "Metode Seleksi Fitur Untuk Klasifikasi Sentimen Menggunakan Algoritma Naive Bayes: Sebuah Literature Review," *J. Media Inform. Budidarma*, vol. 5, no. 3, p. 799, 2021, doi: 10.30865/mib.v5i3.2983.

[14] G. Abdullah and Z. A. Haddi Hassan, "A Comparison between Genetic Algorithm and Practical Swarm to Investigate the Reliability Allocation of Complex Network," *J. Phys. Conf. Ser.*, vol. 1818, no. 1, 2021, doi: 10.1088/1742-6596/1818/1/012163.

[15] F. M. N. Akbar, "Metode KNN (K-Nearest Neighbor) untuk Menentukan Kualitas Air," *J. Tekno Kompak*, vol. 18, no. 1, p. 28, 2024, doi: 10.33365/jtk.v18i1.3241.

[16] F. Putra, H. F. Tahiyat, R. M. Ihsan, R. Rahmaddeni, and L. Efrizoni, "Penerapan Algoritma K-Nearest Neighbor Menggunakan Wrapper Sebagai Preprocessing untuk Penentuan Keterangan Berat Badan Manusia," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 4, no. 1, pp. 273–281, 2024, doi: 10.57152/malcom.v4i1.1085.

[17] Y. A. Hafiz and E. Sudarmilah, "Implementasi Web Scraping Pada Portal Berita Online," *Inisiasi*, pp. 55–60, 2023, doi: 10.59344/inisiasi.v12i1.120.

[18] O. Sihombing, E. Sitanggang, E. Luis, and K. W. Winata, "72 Analisis Spare Part Harbour Tag Pada Divisi Workshop Menggunakan Algoritma Knn Min-Max Scaling," *J. TEKINKOM*, vol. 6, no. 1, pp. 72–80, 2023, doi: 10.37600/tekinkom.v6i1.742.

[19] R. Oktafiani, A. Hermawan, and D. Avianto, "Pengaruh Komposisi Split data Terhadap Performa Klasifikasi Penyakit Kanker Payudara Menggunakan Algoritma Machine Learning," *J. Sains dan Inform.*, vol. 9, no. April, pp. 19–28, 2023, doi: 10.34128/jsi.v9i1.622.

[20] R. Rahmadani, A. Rahim, and R. Rudiman, "Analisis Sentimen Ulasan 'Ojol the Game' Di Google Play Store Menggunakan Algoritma Naive Bayes Dan Model Ekstraksi Fitur Tf-Idf Untuk Meningkatkan Kualitas Game," *J. Inform. dan Tek. Elektro Terap.*, vol. 12, no. 3, 2024, doi: 10.23960/jitet.v12i3.4988.

[21] A. D. Adhi Putra, "Analisis Sentimen pada Ulasan pengguna Aplikasi Bibit Dan Bareksa dengan Algoritma KNN," *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, vol. 8, no. 2, pp. 636–646, 2021, doi: 10.35957/jatisi.v8i2.962.

[22] M. Kanwal, M. M. Ur Rehman, M. U. Farooq, and D. K. Chae, "Mask-Transformer-Based Networks for Teeth Segmentation in Panoramic Radiographs," *Bioengineering*, vol. 10, no. 7, 2023, doi: 10.3390/bioengineering10070843.

[23] M. R. Elfansyah, Rudiman, and F. Yulianto, "Perbandingan Metode K–Nearest Neighbor (Knn) Dan Naive Bayes Terhadap Analisis Sentimen Pada Pengguna E-Wallet Aplikasi Dana Menggunakan Fitur Ekstraksi Tf-Idf," *J. Teknol. Inf. J. Keilmuan dan Apl. Bid. Tek. Inform.*, vol. 8, no. 2, pp. 139–159, 2024.

[24] Ainurrohma, "Akurasi Algoritma Klasifikasi pada Software Rapidminer dan Weka," *Prism. Pros. Semin. Nas. Mat.*, vol. 4, pp. 493–499, 2021.

[25] A. Ben-Yishai and O. Ordentlich, "Constructing Multiclass Classifiers using Binary Classifiers under Log-Loss," *IEEE Int. Symp. Inf. Theory - Proc.*, vol. 2021-July, no. 1791, pp. 2435–2440, 2021, doi: 10.1109/ISIT45174.2021.9518152.

[26] P. W. Sudarmadji, N. Fallo, and Y. S. Peli, "Komparasi Algoritma Klasifikasi Untuk Memprediksi Kelulusan Mahasiswa Program Studi Teknik Komputer Jaringan," *J. Ilm. Flash*, vol. 8, no. 2, p. 109, 2023, doi: 10.32511/flash.v8i2.998.