JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER

p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 Juni 2026
Hal: 942-950
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i1.1255

Perbandingan Algoritma *Machine Learning* untuk Klasifikasi Hoaks Berbahasa Indonesia pada Dataset Komdigi

Haris Setyo Pratomo[1*], Panny Agustia Rahayuningsih[2], Muhammad Rezki[3]

[1,2,3]Fakultas Teknik & Informatika, Informatika, Universitas Bina Sarana Informatika, Pontianak, Indonesia
Email: [1*]15220713@bsi.ac.id, [2]Panny.par@bsi.ac.id, [3]muhammad.mdk@bsi.ac.id
(*Email Corresponding Author: 15220713@bsi.ac.id)
Received: 18 Juni 2026 | Revision: 20 Juni 2026 | Accepted: 21 Juni 2026

Abstrak

Penyebaran hoaks berbahasa Indonesia terus meningkat seiring berkembangnya platform digital, sehingga dibutuhkan sistem klasifikasi otomatis yang mampu mengelompokkan jenis hoaks secara akurat dan efisien. Penelitian ini membandingkan performa lima algoritma *machine learning*, yaitu Support Vector Machine (SVM), Random Forest, Logistic Regression, Decision Tree, dan Naive Bayes dalam klasifikasi kategori hoaks berbahasa Indonesia menggunakan dataset Komdigi yang terdiri dari 16.308 artikel dengan enam kategori. Representasi fitur dilakukan menggunakan TF-IDF dengan kombinasi n-gram (1,2) yang diperkaya dengan fitur statistik teks, sementara ketidakseimbangan kelas yang sangat ekstrem ditangani menggunakan SMOTE yang diterapkan secara internal di dalam *pipeline* Stratified K-Fold Cross-Validation untuk menghindari data leakage. Hasil evaluasi menunjukkan bahwa SVM (LinearSVC) menghasilkan accuracy tertinggi sebesar 95,9% dan cross-validation 0,960, sedangkan Logistic Regression unggul pada AUC Macro sebesar 0,952 dan F1-Score makro 0,460 yang mencerminkan kemampuan terbaik dalam mengenali seluruh kategori secara seimbang. Decision Tree menunjukkan performa terendah dengan AUC Macro 0,635. Temuan ini mengkonfirmasi bahwa pemilihan algoritma terbaik bergantung pada prioritas metrik evaluasi yang digunakan sesuai kebutuhan. Penelitian ini memberikan kontribusi berupa rekomendasi algoritma yang efektif untuk klasifikasi hoaks berbahasa Indonesia serta kerangka metodologi yang valid dan bebas data leakage.

Kata Kunci: Klasifikasi Hoaks, *Machine Learning*, TF-IDF, SMOTE, Dataset Komdigi

Abstract

The spread of Indonesian-language hoaxes continues to increase along with the development of digital platforms, making it necessary to develop an automatic classification system capable of accurately and efficiently categorizing types of hoaxes. This study compares the performance of five machine learning algorithms, namely Support Vector Machine (SVM), Random Forest, Logistic Regression, Decision Tree, and Naive Bayes, in classifying Indonesian hoax categories using the Komdigi dataset consisting of 16,308 articles across six categories. Feature representation was performed using TF-IDF with n-gram combination (1,2) enriched with text statistical features, while the extreme class imbalance was handled using SMOTE applied internally within the Stratified K-Fold Cross-Validation pipeline to prevent data leakage. Evaluation results show that SVM (LinearSVC) achieved the highest accuracy of 95.9% and cross-validation score of 0.960, while Logistic Regression outperformed others in AUC Macro at 0.952 and macro F1-Score of 0.460, reflecting the best ability to recognize all categories in a balanced manner. Decision Tree showed the lowest performance with an AUC Macro of 0.635. These findings confirm that the selection of the best algorithm depends on the priority of evaluation metrics used according to the needs. This study contributes a recommendation of effective algorithms for Indonesian hoax classification and a valid, data leakage-free methodological framework.

***Keywords**: Hoax Classification, Machine Learning, TF-IDF, SMOTE, Komdigi Dataset*

1. PENDAHULUAN

Perkembangan teknologi informasi dan komunikasi telah mendorong peningkatan penggunaan internet serta media digital dalam kehidupan sehari-hari. Berbagai platform seperti portal berita, media sosial, dan aplikasi pesan instan memungkinkan masyarakat memperoleh serta menyebarkan informasi dengan cepat dan mudah. Kemudahan tersebut memberikan banyak manfaat dalam mendukung aktivitas komunikasi dan pertukaran informasi. Namun, di balik manfaat tersebut, muncul permasalahan berupa penyebaran informasi yang tidak sesuai fakta atau yang dikenal sebagai hoaks[1].

Hoaks merupakan informasi yang mengandung fakta yang dipalsukan, dimanipulasi, atau disebarkan tanpa proses verifikasi yang memadai. Penyebaran hoaks dapat menimbulkan berbagai dampak negatif, seperti terbentuknya opini publik yang keliru, meningkatnya keresahan masyarakat, hingga munculnya konflik sosial. Selain itu, hoaks tidak hanya terbatas pada satu bidang tertentu, tetapi dapat muncul dalam berbagai kategori seperti politik, kesehatan, ekonomi, pendidikan, agama, maupun isu sosial lainnya[2]. Banyaknya informasi yang beredar setiap hari menyebabkan proses identifikasi dan pengelompokan hoaks secara manual menjadi kurang efektif karena membutuhkan waktu dan sumber daya yang besar.

Seiring berkembangnya teknologi kecerdasan buatan, pendekatan *machine learning* banyak dimanfaatkan untuk membantu proses klasifikasi teks secara otomatis. *Machine learning* memungkinkan sistem mempelajari pola dari data yang tersedia sehingga dapat digunakan untuk mengelompokkan informasi ke dalam kategori tertentu dengan tingkat

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 Juni 2026
Hal: 942-950
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i1.1255

akurasi yang baik. Dalam konteks klasifikasi teks, beberapa algoritma yang sering digunakan adalah Support Vector Machine (SVM), Decision Tree, Random Forest, Naive Bayes, dan Logistic Regression. SVM dikenal memiliki kemampuan yang baik dalam menangani data berdimensi tinggi dan sering digunakan pada permasalahan klasifikasi teks [3]. Decision Tree menawarkan proses klasifikasi yang mudah dipahami karena menghasilkan aturan keputusan dalam bentuk pohon [4], sementara Random Forest merupakan pengembangan dari Decision Tree yang memanfaatkan banyak pohon keputusan untuk meningkatkan stabilitas dan performa klasifikasi [5].

Selain pemilihan algoritma, kualitas model klasifikasi juga dipengaruhi oleh distribusi data yang digunakan. Pada dataset klasifikasi teks sering ditemukan kondisi ketidakseimbangan kelas (*class imbalance*), yaitu jumlah data pada masing-masing kategori tidak memiliki proporsi yang sama. Kondisi ini dapat menyebabkan model cenderung menghasilkan performa yang lebih baik pada kelas mayoritas dibandingkan kelas minoritas, sehingga kemampuan model dalam mengenali seluruh kategori data menjadi kurang optimal [6]. Untuk mengatasi permasalahan tersebut, diperlukan teknik penanganan ketidakseimbangan kelas, salah satunya adalah Synthetic Minority *Oversampling* Technique (SMOTE). Metode ini bekerja dengan menghasilkan data sintetis pada kelas minoritas sehingga distribusi data menjadi lebih seimbang dan model dapat belajar secara lebih efektif [7].

Penelitian ini menggunakan dataset hoaks berbahasa Indonesia yang bersumber dari Komdigi (Kementerian Komunikasi dan Digital RI) yang memuat 16.308 artikel hoaks terverifikasi dari tahun 2018 hingga Mei 2026 dengan 6 kategori: Hoaks, Penipuan, Pejabat Publik, Masyarakat Digital, Program/Kebijakan, dan Kesehatan. Kondisi ketidakseimbangan pada dataset ini sangat ekstrem, di mana kategori Hoaks mendominasi sebesar 96% (15.670 data) sementara kategori Kesehatan hanya memiliki 56 data (0,3%). Kondisi ini menjadi tantangan utama yang harus ditangani agar model tidak bias terhadap kelas mayoritas.

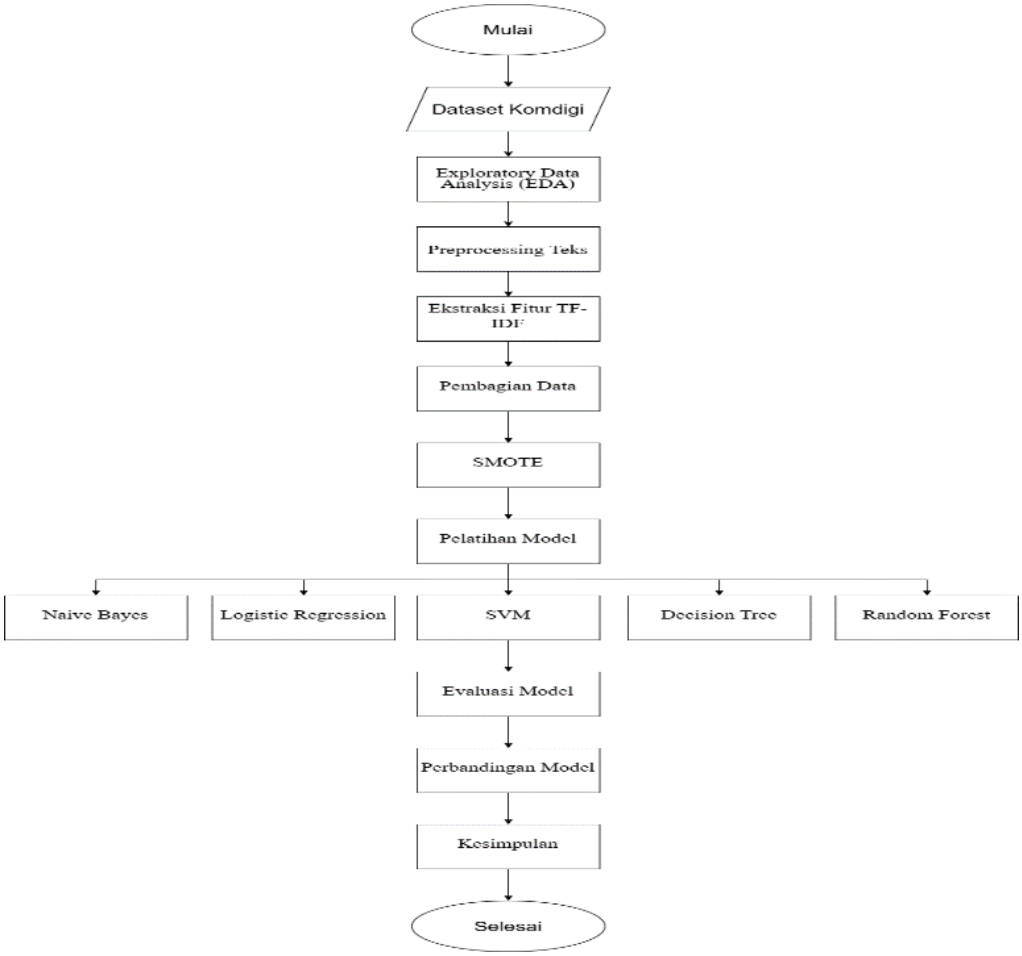
Beberapa penelitian terkait telah mengkaji pendekatan klasifikasi teks dan deteksi hoaks berbahasa Indonesia. Desriansyah et al. [2] menganalisis efektivitas algoritma *machine learning* dalam deteksi hoaks pada berita digital berbahasa Indonesia dan menemukan bahwa pendekatan berbasis teks mampu menghasilkan akurasi yang baik. Haryawan & Ardhana [6] menganalisis perbandingan teknik *oversampling* SMOTE dan membuktikan pengaruhnya terhadap peningkatan performa model pada kelas minoritas. Asep Ripa'i [5] menunjukkan keunggulan Random Forest sebagai metode *ensemble* dalam stabilitas dan akurasi klasifikasi. Tobing et al. [8] membandingkan kinerja IndoBERT dan mBERT untuk deteksi hoaks politik berbahasa Indonesia dan menyoroti pentingnya praproses teks yang tepat. Pulungan et al. [7] membuktikan bahwa penerapan SMOTE secara signifikan meningkatkan performa klasifikasi pada kelas minoritas. Wahid et al. [15] membandingkan performa Logistic Regression dan Random Forest berbasis TF-IDF dalam mendeteksi berita hoaks berbahasa Indonesia dan menemukan bahwa Logistic Regression setelah penerapan *hyperparameter tuning* mencapai akurasi tertinggi sebesar 95,20%, mengungguli Random Forest dalam hal presisi dan F1-score. Berdasarkan kajian tersebut, terdapat *gap* penelitian berupa minimnya studi yang membandingkan lima algoritma sekaligus dengan penerapan SMOTE di dalam *pipeline* Stratified K-Fold *Cross Validation* untuk menghindari *data leakage* pada dataset hoaks Komdigi.

Penelitian ini bertujuan untuk membandingkan performa algoritma SVM, Decision Tree, Random Forest, Naive Bayes, dan Logistic Regression dalam klasifikasi kategori hoaks berbahasa Indonesia menggunakan dataset Komdigi, dengan menerapkan SMOTE di dalam *pipeline* validasi silang serta praproses teks berbahasa Indonesia secara lengkap menggunakan PySastrawi dan representasi fitur TF-IDF yang diperkaya dengan *feature engineering*. Hasil penelitian diharapkan dapat memberikan rekomendasi algoritma yang paling efektif sebagai referensi pengembangan sistem klasifikasi hoaks berbasis *machine learning* di masa mendatang.

2. METODOLOGI PENELITIAN

Penelitian ini menggunakan pendekatan *machine learning* untuk mengklasifikasikan kategori hoaks berbahasa Indonesia secara otomatis menggunakan dataset Indonesian Hoax News Komdigi yang diperoleh melalui platform Kaggle. Proses penelitian dirancang secara sistematis dan terstruktur melalui beberapa tahapan yang saling berkaitan untuk memastikan hasil yang valid dan dapat direproduksi. Tahapan tersebut meliputi eksplorasi data (*Exploratory Data Analysis/EDA*), *preprocessing* teks berbahasa Indonesia, ekstraksi fitur TF-IDF yang diperkaya dengan *feature engineering*, pembagian data latih dan data uji, penanganan ketidakseimbangan kelas menggunakan SMOTE, pelatihan lima algoritma klasifikasi secara paralel, serta evaluasi dan perbandingan performa model menggunakan metrik *accuracy*, *precision*, *recall*, F1-Score, dan AUC Macro. Penerapan SMOTE dilakukan hanya pada data latih setelah pembagian data untuk mencegah *data leakage*, sehingga estimasi performa yang dihasilkan valid dan mencerminkan kemampuan model yang sesungguhnya terhadap data baru. Selain itu, evaluasi konsistensi model dilakukan menggunakan Stratified K-Fold *Cross-Validation* dengan k=5 untuk memastikan performa model tidak hanya baik pada satu subset data tertentu, melainkan konsisten pada seluruh subset data yang digunakan. Tahapan penelitian secara keseluruhan dapat dilihat pada Gambar 1.

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 Juni 2026
Hal: 942-950
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i1.1255



Gambar 1. Tahapan Penelitian

2.1 Dataset Komdigi

Dataset yang digunakan dalam penelitian ini adalah dataset hoaks berbahasa Indonesia yang diperoleh melalui platform Kaggle dengan nama dataset *Indonesian Hoax News Komdigi*. Dataset ini merupakan kumpulan artikel hoaks terverifikasi yang bersumber dari situs resmi Komdigi (Kementerian Komunikasi dan Digital RI) dan memuat 16.308 artikel dari tahun 2018 hingga Mei 2026 dengan enam kategori label, yaitu Hoaks, Penipuan, Pejabat Publik, Masyarakat Digital, Program/Kebijakan, dan Kesehatan. Dataset disimpan dalam format CSV dengan 13 kolom, meliputi kolom id, url, title, slug, published_at, view_count, excerpt, body_html, body_text, main_image_url, category, tags, dan topics. Kolom topics digunakan sebagai label target klasifikasi karena merepresentasikan jenis atau kategori hoaks dari setiap artikel[1].

Dataset ini dipilih karena berasal dari lembaga pemerintah resmi sehingga proses verifikasinya dapat dipertanggungjawabkan, memiliki skala data yang cukup besar, dan representatif untuk konteks hoaks berbahasa Indonesia. Sebelum diproses lebih lanjut, data yang memiliki nilai topics kosong terlebih dahulu dihapus, kemudian label tersebut diubah menjadi bentuk numerik melalui proses *label encoding* agar dapat dikenali oleh algoritma *machine learning.*

2.2 *Exploratory Data Analysis* (EDA)

Exploratory Data Analysis (EDA) adalah proses pemeriksaan awal terhadap dataset untuk memahami karakteristik dan pola yang terdapat di dalamnya sebelum masuk ke tahap pemrosesan lebih lanjut. EDA dilakukan untuk memperoleh gambaran menyeluruh mengenai struktur data, distribusi label, serta karakteristik teks pada setiap kategori hoaks sehingga dapat menjadi dasar penentuan langkah *preprocessing* dan strategi penanganan ketidakseimbangan data yang tepat[10].

Pada penelitian ini, EDA mencakup pemeriksaan informasi umum dataset seperti jumlah baris, kolom, tipe data, dan keberadaan nilai yang hilang. Selain itu, dilakukan analisis distribusi label untuk melihat proporsi setiap kategori, analisis panjang teks dan jumlah kata per kategori, serta identifikasi kata-kata dominan pada masing-masing jenis hoaks. Hasil EDA menunjukkan ketidakseimbangan kelas yang sangat ekstrem, di mana kategori Hoaks mendominasi dengan

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 Juni 2026
Hal: 942-950
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i1.1255

15.670 data (96,1%) sementara kategori Kesehatan hanya memiliki 56 data (0,3%). Kondisi ini menjadi tantangan utama karena model cenderung hanya terfokus pada kelas mayoritas sehingga kelas minoritas berpotensi diklasifikasikan secara tidak tepat[11]

2.3 *Preprocessing Teks*

Preprocessing teks adalah serangkaian tahapan pembersihan dan standarisasi data teks mentah yang dilakukan sebelum data digunakan dalam proses pemodelan. Tujuan utama tahap ini adalah mengurangi *noise* pada data sehingga representasi teks yang dihasilkan lebih berkualitas dan dapat diolah secara optimal oleh algoritma *machine learning*. Kualitas *preprocessing* secara langsung memengaruhi performa model klasifikasi yang dihasilkan karena data teks mentah umumnya masih mengandung karakter tidak relevan, variasi penulisan, dan kata-kata umum yang tidak membawa makna penting dalam proses klasifikasi[12],

Pada penelitian ini, data teks mentah pada kolom judul, ringkasan, dan isi berita diproses melalui beberapa tahapan secara berurutan. Pertama, *case folding* untuk mengubah seluruh teks menjadi huruf kecil guna menghindari duplikasi fitur akibat perbedaan kapitalisasi. Kedua, penghapusan URL, tag HTML, angka, tanda baca, dan karakter non-alfabet lainnya. Ketiga, normalisasi spasi berlebih. Keempat, *stopword removal* untuk menghapus kata-kata umum yang tidak membawa makna penting dalam konteks klasifikasi seperti "yang", "dan", dan "di". Kelima, *stemming* untuk mereduksi kata berimbuhan ke bentuk dasarnya menggunakan library PySastrawi yang mengimplementasikan algoritma Enhanced Confix Stripping (ECS) yang dikembangkan khusus untuk bahasa Indonesia. Hasil akhir tahap ini berupa kolom *clean_text* yang berisi teks bersih dan siap untuk diekstraksi fiturnya pada tahap berikutnya.

2.4 Ekstrasi Fitur TF-IDF

TF-IDF (Term Frequency-Inverse Document Frequency) adalah metode pembobotan kata yang digunakan untuk merepresentasikan dokumen teks dalam bentuk vektor numerik. Metode ini mengukur seberapa penting suatu kata dalam sebuah dokumen relatif terhadap seluruh koleksi dokumen dengan menggabungkan dua komponen utama, yaitu Term Frequency (TF) yang mengukur frekuensi kemunculan kata dalam dokumen dan Inverse Document Frequency (IDF) yang mengukur seberapa jarang kata tersebut muncul di seluruh dokumen dalam korpus [13]. Nilai TF-IDF diformulasikan sebagai berikut:

$$TF - IDF(t,d) = TF(t,d) \times IDF(t) \quad (1)$$

$$\mathrm{IDF(t)} = \log \frac{\mathrm{N}}{\mathrm{df(t)}} \quad (2)$$

Keterangan:

TF(t,d) = frekuensi kemunculan kata t pada dokumen d

IDF(t) = nilai kebalikan frekuensi dokumen

N = jumlah seluruh dokumen

df(t) = jumlah dokumen yang mengandung kata t

Pada penelitian ini, TF-IDF diterapkan menggunakan TfidfVectorizer dengan kombinasi *n-gram* (1,2) untuk menangkap konteks *unigram* dan *bigram*, max_features sebesar 10.000, min_df=2, max_df=0,95, dan *sublinear_tf=True*. Selain fitur TF-IDF, ditambahkan pula fitur statistik teks berupa panjang teks, jumlah kata, panjang judul, jumlah tanda seru, dan rasio huruf kapital. Fitur-fitur statistik tersebut kemudian digabungkan (*concatenate*) dengan matriks TF-IDF untuk membentuk representasi fitur akhir yang digunakan sebagai input model klasifikasi. Pengkombinasian TF-IDF dengan fitur statistik teks terbukti dapat meningkatkan performa klasifikasi dibandingkan menggunakan TF-IDF saja.

2.5 Pembagian Data dan SMOTE

Pembagian data merupakan tahapan penting dalam proses pemodelan untuk memastikan evaluasi model dilakukan secara objektif terhadap data yang belum pernah dilihat selama pelatihan. Dataset dibagi menjadi data latih dan data uji dengan tujuan agar model dapat diukur kemampuan generalisasinya secara valid. Pada penelitian ini, matriks fitur dibagi menjadi data latih sebesar 80% dan data uji sebesar 20% menggunakan train_test_split dengan parameter stratify untuk memastikan proporsi setiap kategori hoaks tetap terjaga secara seimbang pada kedua subset data.

SMOTE (*Synthetic Minority Oversampling Technique*) adalah teknik *oversampling* yang mengatasi masalah ketidakseimbangan kelas dengan membangkitkan sampel sintetis baru pada kelas minoritas melalui interpolasi terhadap

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 Juni 2026
Hal: 942-950
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i1.1255

tetangga terdekat, bukan sekadar menduplikasi sampel yang ada [14]. SMOTE membangkitkan sampel sintetis menggunakan rumus

$$x_{baru} = x_i + \lambda(x_{zi} - x_i) \quad (3)$$

Keterangan:

x_baru = sampel sintetis baru yang dibangkitkan

x_i = sampel minoritas yang dipilih

x_zi = tetangga terdekat dari x_i yang dipilih secara acak

λ = bilangan acak pada interval [0,1]

Pada penelitian ini, SMOTE diterapkan dengan parameter k_neighbors=5 hanya pada data latih setelah pembagian data 80:20, bukan pada keseluruhan dataset. Hal ini penting untuk mencegah *data leakage,* yaitu kondisi di mana informasi dari data uji bocor ke data latih sehingga estimasi performa model menjadi tidak valid. Setelah penerapan SMOTE, jumlah data latih meningkat dari 13.046 menjadi 75.216 sampel dengan setiap kategori memiliki 12.536 sampel secara seimbang.

2.6 Pelatihan dan Evaluasi Model

Pelatihan model dilakukan menggunakan lima algoritma *machine learning* secara paralel, yaitu Naive Bayes (MultinomialNB), Logistic Regression dengan strategi *One-vs-Rest*, Support Vector Machine (LinearSVC yang dikalibrasi menggunakan CalibratedClassifierCV), Decision Tree, dan Random Forest. Kelima algoritma dipilih karena memiliki karakteristik yang beragam, mulai dari pendekatan berbasis probabilitas, regresi linear, margin, hingga pohon keputusan tunggal dan *ensemble*, sehingga perbandingan antar kelimanya dapat memberikan gambaran komprehensif mengenai algoritma yang paling efektif untuk klasifikasi teks hoaks berbahasa Indonesia. Setiap model dilatih menggunakan data latih hasil SMOTE, kemudian digunakan untuk memprediksi data uji yang masih dalam kondisi asli tanpa SMOTE.

Evaluasi performa setiap model dilakukan menggunakan *confusion matrix* dan empat metrik evaluasi utama. Metrik-metrik tersebut diformulasikan sebagai berikut [15]:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

$$Recall = \frac{TP}{TP+FN} \quad (6)$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision+Recall} \quad (7)$$

Keterangan:

TP (True Positive) = data positif diprediksi benar

TN (True Negative) = data negatif diprediksi benar

FP (False Positive) = data negatif diprediksi positif

FN (False Negative) = data positif diprediksi negatif

3. HASIL DAN PEMBAHASAN

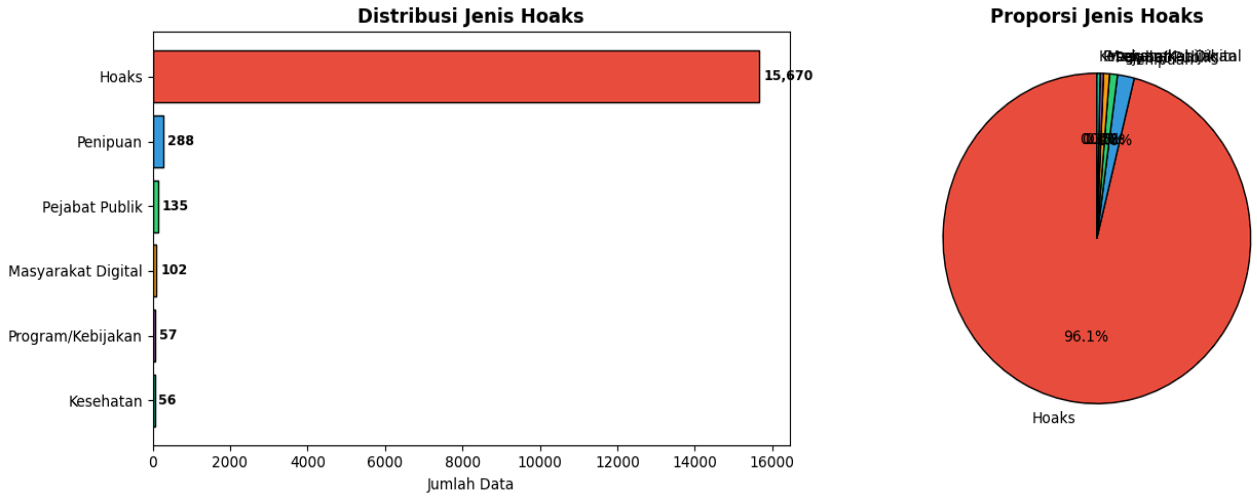
Penelitian ini menghasilkan model klasifikasi kategori hoaks berbahasa Indonesia menggunakan lima algoritma *machine learning* pada dataset Komdigi. Seluruh proses mulai dari eksplorasi data, *preprocessing*, ekstraksi fitur, penanganan ketidakseimbangan kelas, hingga evaluasi model telah dilaksanakan secara sistematis. Berikut diuraikan hasil dan pembahasan dari setiap tahapan penelitian.

3.1 Hasil Eksplorasi Data

Dataset Komdigi terdiri dari 16.308 artikel dengan 13 kolom, di mana kolom topics digunakan sebagai label target klasifikasi. Dataset memuat enam kategori, yaitu Hoaks, Penipuan, Masyarakat Digital, Pejabat Publik, Kesehatan,

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 Juni 2026
Hal: 942-950
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i1.1255

dan Program/Kebijakan. Hasil eksplorasi menunjukkan ketidakseimbangan kelas yang sangat ekstrem, di mana kategori Hoaks mendominasi dengan 15.670 data (96,1%), sementara kategori Kesehatan hanya memiliki 56 data (0,3%), Penipuan 288 data, Pejabat Publik 135 data, Masyarakat Digital 102 data, dan Program/Kebijakan 57 data. Distribusi kategori dataset disajikan pada Gambar 2.

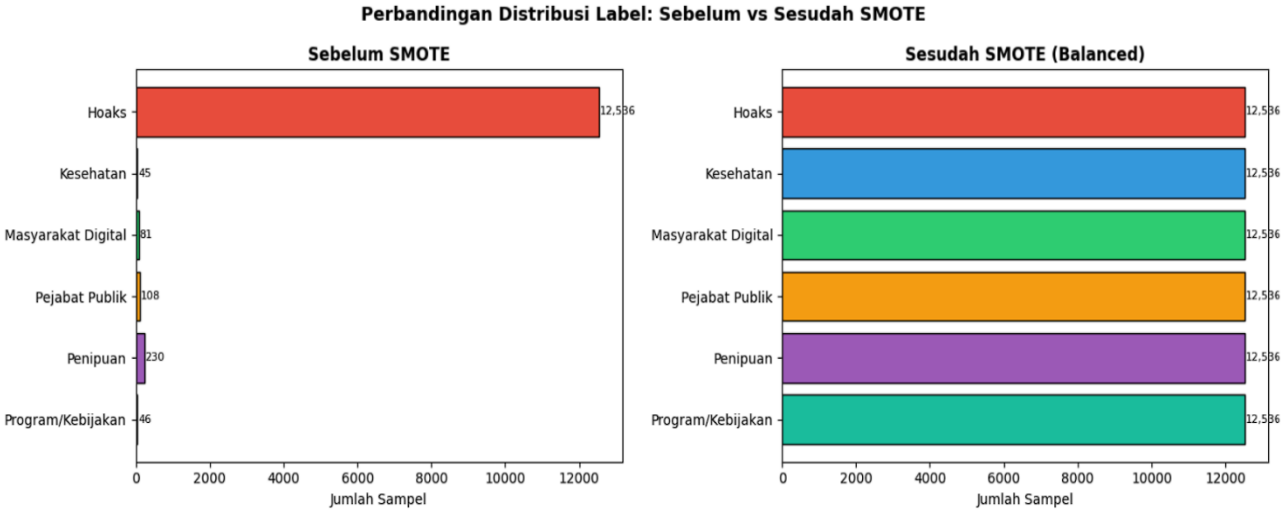


Gambar 2. Distribusi Jenis Hoaks

Analisis distribusi karakteristik teks menunjukkan bahwa sebagian besar artikel memiliki panjang antara 500 hingga 2.000 karakter dengan puncak frekuensi di sekitar 1.000 karakter, dan mayoritas artikel mengandung antara 50 hingga 200 kata. Analisis kata dominan per kategori menunjukkan bahwa setiap kategori memiliki kosakata khas yang membedakannya, seperti kata “beredar” dan “tautan” pada kategori Hoaks, “akun” dan “WhatsApp” pada Penipuan, serta “penyakit” dan “obat” pada Kesehatan. Perbedaan kosakata antar kategori ini menjadi indikasi awal bahwa representasi fitur berbasis TF-IDF memiliki potensi yang baik untuk membedakan setiap jenis hoaks.

3.2 Hasil Penanganan Ketidakseimbangan Kelas

Setelah pembagian data 80:20, data latih yang berjumlah 13.046 sampel ditangani menggunakan SMOTE dengan parameter k_neighbors=5. Setelah penerapan SMOTE, seluruh kategori memiliki jumlah data yang sama yaitu masing-masing 12.536 sampel, sehingga total data latih meningkat menjadi 75.216 sampel. Distribusi data sebelum dan sesudah SMOTE disajikan pada Gambar 3.



Gambar 3. Distribusi data sebelum dan sesudah SMOTE

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 Juni 2026
Hal: 942-950
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i1.1255

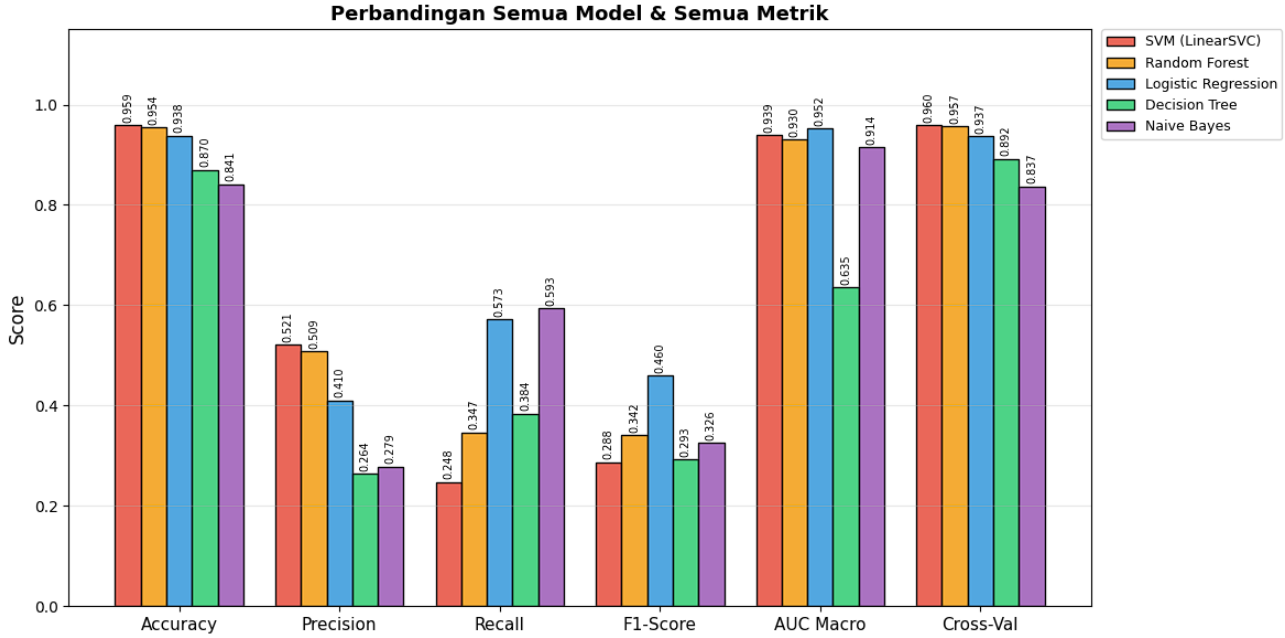
3.3 Has il Evaluasi Model

Evaluasi dilakukan terhadap kelima model menggunakan data uji kondisi asli sebanyak 3.262 sampel. Tabel 1 menyajikan perbandingan performa kelima algoritma pada seluruh metrik evaluasi.

Tabel 1. Perbandingan Peforma Kelima Algoritma

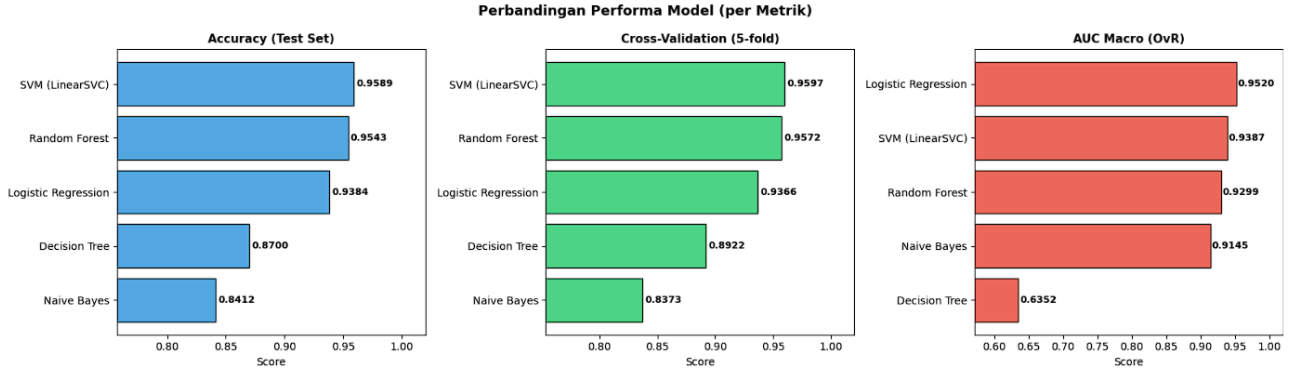
Algoritma	*Accuracy*	*Precision*	*Recall*	F1-Score	AUC Macro	*Cross-Val*
SVM (Linear SVC)	0,959	0,521	0,248	0,288	0,939	0,960
Random Forest	0,954	0,509	0,347	0,342	0,930	0,957
Logistic Regression	0,938	0,410	0,573	0,460	0,952	0,937
Decision Tree	0,870	0,264	0,384	0,293	0,635	0,892
Naïve Bayes	0,841	0,279	0,593	0,326	0,914	0,837

Perbandingan performa seluruh model pada semua metrik secara visual disajikan pada Gambar 4



Gambar 4. Perbandingan Semua Model dan Metrik

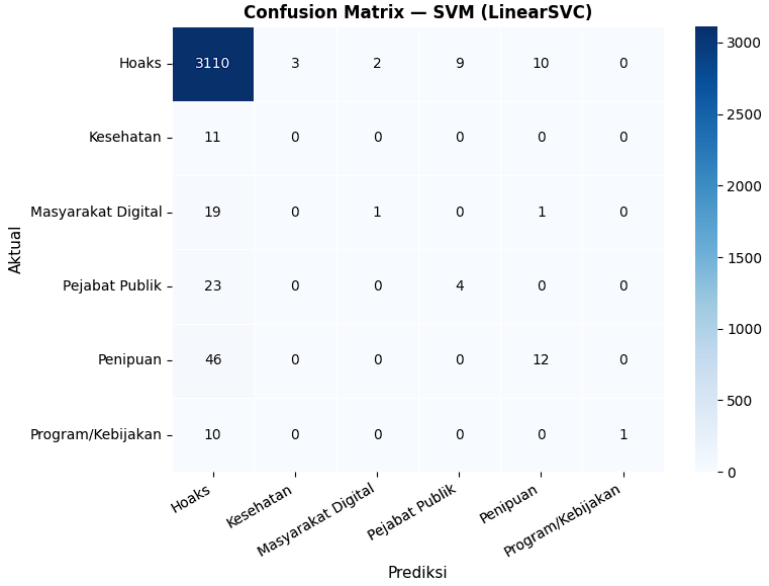
Berdasarkan Tabel 1 dan Gambar 4, terlihat bahwa hasil perbandingan performa berbeda tergantung metrik yang digunakan. Pada metrik *accuracy* dan *cross-validation*, SVM (LinearSVC) menempati posisi teratas dengan nilai masing-masing 0,959 dan 0,960, diikuti Random Forest (0,954 dan 0,957). Namun pada metrik AUC Macro yang lebih mencerminkan kemampuan model mengenali seluruh kelas secara seimbang, Logistic Regression justru unggul dengan nilai 0,952, diikuti SVM (0,939) dan Random Forest (0,930). Perbandingan performa per metrik secara terpisah disajikan pada Gambar 5.



Gambar 5. Perbandingan Peforma Model

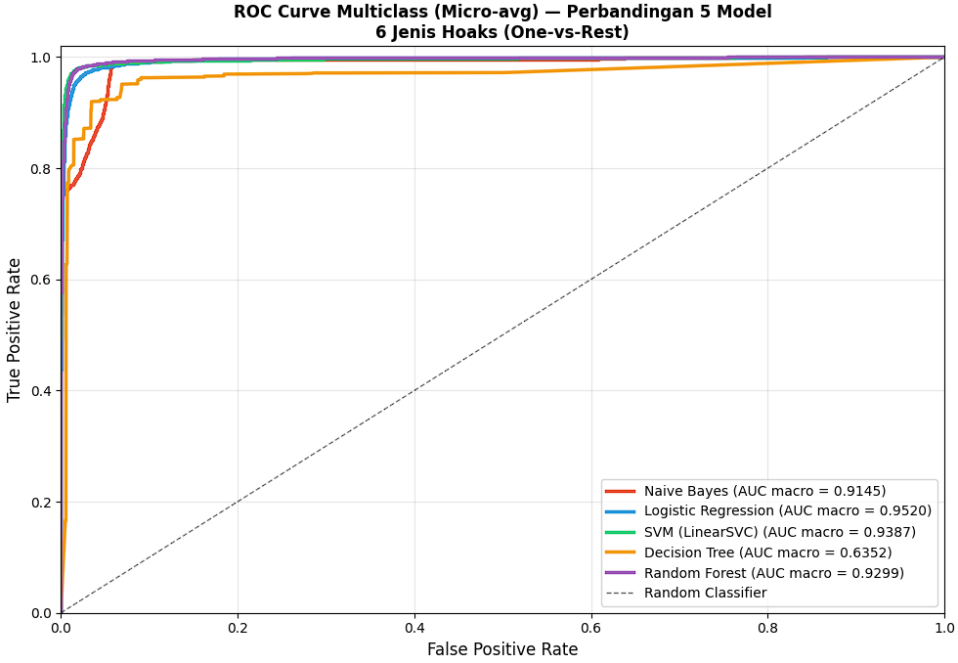
JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 Juni 2026
Hal: 942-950
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i1.1255

Evaluasi detail dilakukan khusus pada model SVM (LinearSVC) sebagai model dengan *accuracy* tertinggi. *Confusion matrix* model SVM disajikan pada Gambar 6.



Gambar 6. *Confusion Matrix* SVM

Berdasarkan *confusion matrix* pada Gambar 6, model SVM berhasil memprediksi 3.110 dari 3.134 data kelas Hoaks dengan benar. Namun hampir seluruh data kelas minoritas salah diklasifikasikan sebagai Hoaks, seperti seluruh 11 data Kesehatan diprediksi sebagai Hoaks. Hal ini dikonfirmasi oleh nilai *recall* makro SVM yang hanya 0,248, meskipun *accuracy* keseluruhannya mencapai 0,959. Kurva ROC kelima model disajikan pada Gambar 7.



Gambar 7. ROC 5 Model

Kurva ROC pada Gambar 8 memperlihatkan bahwa kurva Logistic Regression, SVM, dan Random Forest berada berdekatan di area kiri atas grafik, menandakan kemampuan klasifikasi yang kompetitif di antara ketiganya. Sebaliknya, kurva Decision Tree terlihat jauh lebih rendah dan mendekati garis diagonal acak dengan AUC Macro hanya 0,6352, mengindikasikan kemampuan diskriminasi yang paling lemah di antara kelima algoritma.

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 Juni 2026
Hal: 942-950
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i1.1255

4. KESIMPULAN

Penelitian ini berhasil membandingkan performa lima algoritma machine learning, yaitu SVM (LinearSVC), Random Forest, Logistic Regression, Decision Tree, dan Naive Bayes untuk klasifikasi hoaks berbahasa Indonesia menggunakan dataset Komdigi yang terdiri dari 16.308 artikel dengan enam kategori. Hasil penelitian menunjukkan bahwa penerapan preprocessing teks, kombinasi fitur TF-IDF dan fitur statistik teks, serta penanganan ketidakseimbangan kelas menggunakan SMOTE mampu menghasilkan model klasifikasi dengan performa yang baik. Berdasarkan hasil evaluasi, SVM (LinearSVC) memperoleh performa terbaik pada metrik accuracy dan cross-validation dengan nilai masing-masing sebesar 95,9% dan 0,960, sedangkan Logistic Regression menunjukkan kemampuan paling seimbang dalam mengenali seluruh kategori dengan AUC Macro sebesar 0,952 dan F1-Score makro sebesar 0,460. Sementara itu, Decision Tree menjadi algoritma dengan performa terendah pada sebagian besar metrik evaluasi. Temuan ini menunjukkan bahwa pemilihan algoritma terbaik sangat bergantung pada tujuan penggunaan model, di mana SVM lebih sesuai untuk memaksimalkan akurasi keseluruhan, sedangkan Logistic Regression lebih efektif untuk menjaga keseimbangan klasifikasi pada seluruh kategori hoaks. Dengan demikian, penelitian ini memberikan rekomendasi algoritma yang dapat digunakan sebagai dasar pengembangan sistem klasifikasi hoaks berbahasa Indonesia yang lebih akurat dan andal.

REFERENCES

[1] A. Sarjito, "Hoaks, Disinformasi, dan Ketahanan Nasional: Ancaman Teknologi Informasi dalam Masyarakat Digital Indonesia," *J. Gov. Local Polit.*, vol. 6, no. 2, pp. 175–186, 2024, doi: 10.47650/jglp.v6i2.1547.

[2] M. D. Desriansyah, I. U. Sari, and Z. Zulfahmi, "Analisis Efektivitas Algoritma Machine Learning dalam Deteksi Hoaks: Pada Berita Digital Berbahasa Indonesia," *J. Sist. Inf. Dan Inform.*, vol. 3, no. 2, pp. 63–69, 2025, doi: 10.47233/jiska.v3i1.2024.

[3] N. Arifin, U. Enri, and N. Sulistiyowati, "Penerapan Algoritma Support Vector Machine (SVM) dengan TF-IDF N-Gram untuk Text Classification," *STRING (Satuan Tulisan Ris. dan Inov. Teknol.*, vol. 6, no. 2, p. 129, 2021, doi: 10.30998/string.v6i2.10133.

[4] R. N. Ramadhon, A. Ogi, A. P. Agung, R. Putra, S. S. Febrihartina, and U. Firdaus, "Implementasi Algoritma Decision Tree untuk Klasifikasi Pelanggan Aktif atau Tidak Aktif pada Data Bank," *Karimah Tauhid*, vol. 3, no. 2, pp. 1860–1874, 2024, doi: 10.30997/karimahtauhid.v3i2.11952.

[5] F. L. Asep Ripa'i, Firman Santoso, "Deteksi Berita Hoax dengan Perbandingan Website Menggunakan Pendekatan Deep Learning Algoritma BERT," *G-Tech J. Teknol. Terap.*, vol. 6, no. 2, pp. 295–305, 2022.

[6] C. Haryawan and Y. M. K. Ardhana, "Analisa Perbandingan Teknik Oversampling SMOTE," *JIRE (Jurnal Inform. Rekayasa Elektron.*, vol. 6, no. 1, pp. 73–78, 2023.

[7] M. P. Pulungan, A. Purnomo, and A. Kurniasih, "Penerapan SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Kepribadian MBTI Menggunakan Naive Bayes Classifier," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 11, no. 5, pp. 1033–1042, 2024, doi: 10.25126/jtiik.2024117989.

[8] C. J. L. Tobing, IGN Lanang Wijayakusuma, and Luh Putu Ida Harini, "Perbandingan Kinerja IndoBERT dan MBERT Untuk Deteksi Berita Hoaks Politik dalam Bahasa Indonesia," *JST (Jurnal Sains dan Teknol.*, vol. 14, no. 1, pp. 114–123, 2025, doi: 10.23887/jstundiksha.v14i1.92126.

[9] A. M. Wahid, Turino, K. A. Nugroho, D. Titi Safitri4, and F. S. Utomo, "Optimasi Logistic Regression dan Random Forest untuk Deteksi Berita Hoax Optimasi Logistic Regression dan Random Forest untuk Deteksi Berita Hoax Berbasis Hyperparameter Optimization of Logistic Regression and Random Forest for Hoax News Detection Using T," *J. Pendidik. dan Teknol. Indones.*, vol. 4, no. January, pp. 381–392, 2025.

[10] I. N. Rizki, D. Prayoga, M. L. Puspita, and M. Q. Huda, "Implementasi Exploratory Data Analysis Untuk Analisis Dan Visualisasi Data Penderita Stroke Kalimantan Selatan Menggunakan Platform Tableau," *J. Inform. dan Tek. Elektro Terap.*, vol. 12, no. 1, 2024, doi: 10.23960/jitet.v12i1.3856.

[11] I. A. Rahma and L. H. Suadaa, "Penerapan Text Augmentation untuk Mengatasi Data yang Tidak Seimbang pada Klasifikasi Teks Berbahasa Indonesia," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 10, no. 6, pp. 1329–1340, 2023, doi: 10.25126/jtiik.2023107325.

[12] T. Gori, A. Sunyoto, and H. Al Fatta, "Preprocessing Data dan Klasifikasi untuk Prediksi Kinerja Akademik Siswa," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 11, no. 1, pp. 215–224, 2024, doi: 10.25126/jtiik.20241118074.

[13] A. Arasy and S. Agustian, "Sentiment Classification Using Multilayer Perceptron Algorithm with TF-IDF Features Klasifikasi Sentimen Menggunakan Metode Multilayer Perceptron dengan Fitur TF-IDF," vol. 5, no. July, pp. 908–919, 2025.

[14] M. Sulistiyono *et al.*, "Implementasi Algoritma Synthetic Minority Over - Sampling Technique untuk Menangani Ketidakseimbangan Kelas pada Dataset Klasifikasi," vol. 10, pp. 445–459, 2021.

[15] T. H. Pinem and Z. P. Putra, "Evaluasi Kinerja Algoritma Klasifikasi Deep Learning dalam Prediksi Diabetes," *J. Ilm. FIFO*, vol. 17, no. 1, p. 17, 2025, doi: 10.22441/fifo.2025.v17i1.003.