Jurnal Ilmiah Sistem Informasi dan Ilmu Komputer
Volume 6, Nomor 3, November 2026
e-ISSN: 2827-7953; p-ISSN: 2827-8135, Hal. 75-89
DOI: https://doi.org/10.55606/juisik.v6i3.2726
Tersedia: http://journal.sinov.id/index.php/juisik

Analisis Perbandingan Kinerja Algoritma *Random Forest* dan *Logistic Regression* untuk Deteksi Berita Hoaks

Retno Anjani[1*], Deni Arifianto[2], Ulya Anisatur Rosyidah[3]
[1-3]Program Studi Sistem Informasi, Universitas Muhammadiyah Jember, Indonesia

**Penulis Korespondensi: anjaniretno1125@gmail.com*

Abstract. *The widespread circulation of hoax news across digital media in Indonesia presents serious challenges to public trust and social stability. This study compares the classification performance of Random Forest and Logistic Regression in detecting Indonesian-language hoax news while addressing class imbalance using Customized Instance Random Undersampling (CIRUS). A dataset of 17,185 news articles from the official Kominfo repository (December 2018 to March 2026) was utilized, focusing combination on two primary text attributes: tittle (headline) and body_text (news content). Following data cleaning, 15,411 clean articles were obtained (10,274 hoaxes and 5,137 non-hoaxes). Preprocessing on both attributes included case folding, cleaning, tokenization, stopword removal, and Sastrawi stemming, followed by TF-IDF n-gram feature extraction (10,000 features; 5,000 from tittle and 5,000 from body_text). Models were evaluated using K-Fold Cross Validation (K=5) on imbalanced and CIRUS-balanced datasets (10,274 articles, 1:1 ratio). On the balanced dataset, Random Forest outperformed Logistic Regression across all metrics, achieving an accuracy of 95.16%, precision of 95.27%, recall of 95.16%, and F1-score of 95.16%, compared to 92.21% accuracy and 92.19% F1-score for Logistic Regression. These results show that CIRUS effectively mitigates the accuracy paradox by retaining representative boundary samples, while Random Forest provides superior predictive stability for Indonesian text classification.*

Keywords: *CIRUS; Hoax Detection; Imbalanced Data; Logistic Regression; Random Forest.*

Abstrak. Peredaran berita hoaks di media digital Indonesia menjadi persoalan nyata yang dapat mengganggu kepercayaan publik dan ketertiban sosial. Penelitian ini bertujuan untuk membandingkan performa algoritma *Random Forest* dan *Logistic Regression* dalam klasifikasi berita hoaks berbahasa Indonesia dengan penanganan ketidakseimbangan kelas menggunakan metode *Customized Instance Random Undersampling* (CIRUS). Data penelitian bersumber dari portal resmi Kominfo periode Desember 2018 hingga Maret 2026 sebanyak 17.185 artikel berita dengan fokus pada gabungan dua atribut teks utama, yaitu *title* (judul berita) dan *body_text* (isi berita). Tahap pembersihan menghasilkan 15.411 data bersih (10.274 hoaks dan 5.137 tidak hoaks). Tahap *preprocessing* data pada kedua atribut meliputi *case folding, cleaning, tokenization, stopword removal,* dan *stemming* Sastrawi, dilanjutkan dengan pembobotan fitur TF-IDF *unigram* dan *bigram* dengan batasan 10.000 fitur (5.000 fitur *tittle* dan 5.000 fitur *body_text*). Pengujian dilakukan menggunakan *K-Fold Cross Validation* (K=5) pada data tidak seimbang dan seimbang pasca-CIRUS (10.274 data, rasio 1:1). Hasil pengujian pada data seimbang menunjukkan bahwa *Random Forest* mengungguli *Logistic Regression* pada seluruh metrik, dengan memperoleh Akurasi 95,16%, Presisi 95,27%, *Recall* 95,16%, dan *F1-Score* 95,16%, dibandingkan *Logistic Regression* yang memperoleh Akurasi 92,21% dan *F1-Score* 92,19%. Hasil ini membuktikan bahwa metode CIRUS mampu mengatasi masalah *accuracy paradox* tanpa menghilangkan informasi penting, serta *Random Forest* menghasilkan klasifikasi yang lebih akurat dan stabil pada teks berita berbahasa Indonesia.

Kata kunci: CIRUS; Deteksi Hoaks; *Imbalanced Data*; *Logistic Regression*; *Random Forest*.

1. LATAR BELAKANG

Kemajuan teknologi informasi dan internet telah memungkinkan masyarakat mengakses serta menyebarluaskan informasi secara cepat melalui berbagai platform digital, di mana jumlah pengguna internet di Indonesia tercatat mencapai 221,5 juta jiwa atau 79,5% dari total populasi pada tahun 2024 (Adebesin et al., 2023). Kemudahan komunikasi daring melalui media sosial dan situs berita sayangnya memicu eskalasi penyebaran berita palsu atau hoaks yang sulit diverifikasi secara manual oleh publik. Disinformasi di media digital tidak hanya

Naskah Masuk: 16 Juli 2026; Revisi: 20 Agustus 2026; Diterima: 16 September 2026; Tersedia: 26 September

menurunkan kepercayaan masyarakat terhadap informasi resmi (Zhou & Zafarani, 2020), tetapi juga menimbulkan dampak psikososial nyata seperti kepanikan massal dan distorsi opini publik (Iizuka et al., 2022). Menghadapi volume data berita digital yang masif, diperlukan otomasi sistem deteksi hoaks yang andal dan terstruktur. Salah satu pendekatan utama yang efektif adalah *text classification* berbasis *Natural Language Processing* (NLP) dan representasi fitur teks seperti TF-IDF untuk memetakan pola kebahasaan berita ke dalam kategori hoaks atau fakta secara otomatis (Jahan et al., 2024).

Berbagai penelitian terdahulu telah mengkaji implementasi algoritma pembelajaran mesin untuk mendeteksi berita palsu melalui klasifikasi teks. Berdasarkan penelitian Hassan et al. (2024) membuktikan efektivitas algoritma *Random Forest* berkombinasi pembobotan fitur TF-IDF dalam menangani teks berdimensi tinggi dan mereduksi risiko *overfitting* melalui mekanisme *ensemble* banyak pohon keputusan, meski menuntut komputasi yang lebih kompleks. Dengan demikian, performa algoritma tersebut pada literatur sering kali menunjukkan ketidakkonsistensian akibat variasi prapemrosesan, perbedaan dimensionalitas korpus berita, serta keterbatasan eksperimen terdahulu yang mayoritas hanya menguji model pada kondisi data yang telah seimbang.

Pada skenario dunia nyata, data berita digital selalu dihadapkan pada permasalahan ketimpangan distribusi kelas (*class imbalance*). Ketidakseimbangan data merupakan hambatan kritis dalam pembelajaran mesin karena algoritma klasifikasi cenderung mengalami bias terhadap kelas mayoritas dan mengabaikan kelas minoritas (Swana & Doorsamy, 2022). Kondisi ini melahirkan fenomena *accuracy paradox*, di mana model klasifikasi menghasilkan skor akurasi keseluruhan yang tampak sangat tinggi semata-mata karena memprediksi kelas mayoritas secara dominan, padahal performa deteksi terhadap kelas minoritas sangat rendah (Annie & Jung, 2023). Dalam domain deteksi berita hoaks, kegagalan mengenali kelas minoritas menjadi fatal karena menyamarkan disinformasi berbahaya yang seharusnya dicegah sedini mungkin.

Dataset yang digunakan dalam penelitian ini juga memiliki distribusi kelas yang tidak seimbang, di mana jumlah data pada kategori hoaks jauh lebih banyak dibandingkan kategori tidak hoaks. Kondisi tersebut berpotensi mempengaruhi kinerja model klasifikasi, di mana adanya *overlapping* data antarkelas dapat semakin menurunkan performa model karena batas pemisahan antarkelas menjadi tidak jelas. Oleh karena itu, diperlukan teknik penanganan seperti *undersampling* untuk menyeimbangkan distribusi data sebelum dilakukan proses klasifikasi dengan memangkas kuantitas data pada kelompok mayoritas agar proporsinya

mendekati kelompok minoritas (Elsoud et al., 2024). Namun, penelitian yang dilakukan oleh Hoyos-osorio et al. (2021) menggarisbawahi bahwa pemotongan sampel secara acak manual seperti *Random Undersampling* (RUS) berisiko tinggi membuang instans data krusial yang berada di sekitar batas keputusan (*decision boundary*), sehingga mereduksi kapasitas generalisasi model. Menjawab keterbatasan tersebut, penelitian Tusell-rey et al. (2022) memperkenalkan metode *Customized Instance Random Undersampling* (CIRUS) yang membentuk himpunan kompak (*compact sets*) guna merepresentasikan struktur distribusi kelas mayoritas sebelum melakukan seleksi terarah. Hingga saat ini, integrasi metode CIRUS pada kumpulan berita berbahasa Indonesia berskala besar dengan komparasi algoritma *Random Forest* dan *Logistic Regression* belum pernah dieksplorasi secara komprehensif, sehingga menjadi celah riset (*research gap)* penting yang memiliki urgensi empiris tinggi.

Berdasarkan permasalahan dan celah penelitian tersebut, penelitian ini bertujuan untuk menganalisis dan membandingkan performa algoritma *Random Forest* dan *Logistic Regression* dalam mendeteksi berita hoaks berbahasa Indonesia, baik sebelum maupun sesudah penanganan ketidakseimbangan data berbasis *Customized Instance Random Undersampling* (CIRUS). Pengujian model dievaluasi secara ketat menggunakan skema *K-Fold Cross Validation* dengan ekstraksi fitur kombinasi unigram dan bigram TF-IDF berdimensi 10.000 fitur pada dataset resmi Kominfo. Hasil penelitian ini diharapkan tidak hanya mampu mengidentifikasi algoritma yang paling unggul dalam kondisi data yang realistis, tetapi juga memberikan bukti empiris terkait efektivitas CIRUS sebagai solusi penanganan *imbalanced* data, serta memberikan kontribusi nyata pada pendekatan *text classification* deteksi hoaks berbahasa Indonesia yang lebih akurat.

2. KAJIAN TEORITIS

Pengkategorian teks secara komputasi mengacu pada prosedur pemetaan dokumen tak berstruktur ke dalam himpunan kelas target yang telah didefinisikan sebelumnya menurut karakteristik linguistik yang dikandungnya (Kowsari et al., 2019). Dalam tata kelola korpus bahasa Indonesia, data mentah memerlukan transformasi terstruktur melalui prapemrosesan (*text preprocessing*) agar derau leksikal tereliminasi dan representasi masukan siap diinterpretasi oleh algoritma pembelajar. Rangkaian prapemrosesan standar mengintegrasikan penyeragaman kapitalisasi huruf ke format kecil (*case folding*), penapisan tanda baca, angka, dan tautan pengganggu (*cleaning*), segmentasi deretan kalimat menjadi token leksikal mandiri (*tokenizing*), reduksi kata-kata fungsional yang minim muatan informasi pembeda (*stopword*

removal), serta pemulihan leksem berimbuhan menuju bentuk lema dasarnya (*stemming*) mengandalkan modul morfologi komputasi seperti pustaka Sastrawi (Septiani & Isabela, 2022).

Transformasi teks menuju representasi numerik pada studi ini direalisasikan lewat algoritma *Term Frequency-Inverse Document Frequency* (TF-IDF) yang diperluas dengan skema n-gram gabungan unigram dan bigram. Nilai pembobotan TF-IDF mencerminkan tingkat kekhasan suatu istilah melalui perpaduan frekuensi kemunculan kata pada dokumen spesifik (*Term Frequency*) yang dibobotkan secara terbalik terhadap prevalensi kemunculan kata tersebut di seluruh himpunan korpus (*Inverse Document Frequency*). Pengadopsian rentang *n-gram* (1, 2) berfungsi memelihara keutuhan relasi semantik antarkata yang berdekatan, seperti idiom atau frasa khusus yang lazim mewarnai gaya penulisan pers, sehingga nuansa kontekstual tidak tereduksi saat dokumen dikonversi menjadi ruang vektor berdimensi tinggi (Hashim et al., 2023).

Pendekatan *Random Forest* bertumpu pada paradigma pembelajaran ensambel (*ensemble learning*) yang mengagregasi puluhan hingga ratusan pohon keputusan (*decision trees*) (Breiman, 2001). Selama proses pembentukan model, masing-masing pohon dibangun menggunakan subset sampel teracak lewat teknik *bootstrap* (*bagging*) dan mengevaluasi subset prediktor secara acak pada tiap simpul pembagian (*split*). Prediksi final ditentukan melalui mekanisme pemungutan suara mayoritas (*majority voting*) dari segenap pohon penilai. Desain desentralistik ini menjadikan *Random Forest* sangat imun terhadap perangkap *overfitting* yang sering melanda pohon keputusan tunggal, seraya mempertahankan akurasi dan ketahanan komputasi yang kokoh pada matriks teks berdimensi masif (Paramita et al., 2024).

Sebagai pembanding linier, *Logistic Regression* merepresentasikan metode klasifikasi terbimbing yang mengestimasi peluang terjadinya label biner dengan menerapkan transformasi fungsi logistik atau kurva *sigmoid* terhadap kombinasi linier variabel fitur (Lavalley, 2008). Nilai luaran model terpetakan secara teratur ke dalam skala probabilitas kontinu antara interval 0 hingga 1, di mana batas ambang standar 0,5 dijadikan acuan penentu kategori: probabilitas melampaui 0,5 dimasukkan sebagai kelas hoaks, sedangkan nilai di bawahnya dikelompokkan sebagai tidak hoaks. Model ini menawarkan keunggulan berupa latensi komputasi yang minimal, interpretabilitas tinggi, serta efisiensi alokasi memori yang luar biasa saat mengolah matriks data teks yang renggang (*sparse matrix*) (Putra & Yulianingsih, 2023).

Untuk mengatasi distribusi kelas tersebut, penelitian ini mengimplementasikan *Customized Instance Random Undersampling* (CIRUS) yang dirancang untuk mereduksi kelemahan undersampling acak standar. Metode CIRUS mengeksekusi penyeimbangan data melalui empat fase algoritmis yang tertata rapi. Pertama, dilakukan kalkulasi rasio ketidakseimbangan (*imbalance ratio*) guna menetapkan batas reduksi matematis yang diperlukan untuk menyetarakan kelas dominan dengan kelas minoritas. Kedua, pembentukan kluster terkompresi (*compact sets*), yakni pemetaan sampel kelas mayoritas ke dalam beberapa kelompok fitur berbasis algoritma *K-Means* dengan metrik jarak *Euclidean* terhadap titik pusat (*centroid*) kluster (Gillala et al., 2024). Ketiga, penarikan instansi representatif (*representative instance selection*), di mana subsampel diambil secara terukur dan proporsional dari tiap compact set agar keanekaragaman pola semantik dan sebaran informasi penting kelas mayoritas tetap terlindungi. Keempat, penyesuaian seleksi acak akhir secara terfokus guna mengunci kuota terpilih kelas mayoritas tepat pada proporsi 1:1 terhadap kelas minoritas. Melalui mekanisme bertahap ini, CIRUS menyeimbangkan partisi data secara optimal tanpa memicu degradasi pola atau mengabaikan sampel batas krusial (Tusell-rey et al., 2022).

Protokol pengujian reliabilitas klasifikasi diatur mengacu pada metode *K-Fold Cross Validation* demi menjamin stabilitas metrik dan meniadakan subjektivitas yang timbul dari pembagian data tunggal (*train-test* split acak). Di bawah skema ini, keseluruhan korpus dibagi secara adil ke dalam K (*fold*) berukuran seimbang. Pengujian dieksekusi dalam K putaran berulang (diatur pada parameter K=5), di mana pada setiap putaran, satu partisi disisihkan sebagai data uji independen sementara K-1 partisi sisanya dimanfaatkan untuk fase pelatihan model (Lumumba et al., 2024). Pola rotasi menyeluruh ini memastikan bahwa seluruh sampel data pernah diuji tepat satu kali, menghasilkan agregasi skor performa dan simpangan baku yang akurat serta mencerminkan daya generalisasi model pada data yang belum pernah dipelajari sebelumnya. Tingkat keberhasilan klasifikasi dikuantifikasi menggunakan instrumen matriks konfusi (*confusion matrix*) yang mengorganisasikan perbandingan luaran prediksi sistem terhadap kebenaran faktual (*ground truth*) ke dalam empat sel luaran: *True Positive* (TP) untuk hoaks yang berhasil diidentifikasi tepat, *True Negative* (TN) untuk berita fakta yang dikonfirmasi benar, *False Positive* (FP) untuk berita fakta yang salah dituduh sebagai hoaks, serta *False Negative* (FN) untuk berita hoaks yang gagal terdeteksi dan lolos sebagai fakta (Gillala et al., 2024). Dari komposisi matriks ini diturunkan empat parameter kuantitatif: rasio akurasi (*Accuracy*) untuk mengukur rasio ketepatan menyeluruh, presisi (*Precision*) untuk menilai keandalan luaran positif, kepekaan (*Recall*) guna mengukur rasio penemuan seluruh

entitas hoaks, serta *F1-Score* sebagai rata-rata harmonis presisi dan *recall* yang menjadi tolok ukur utama kestabilan performa model di bawah sebaran data teks yang bervariasi (Othman et al., 2024), yang dirumuskan sebagai berikut:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \, x \, 100\% \quad \ldots (i)$$

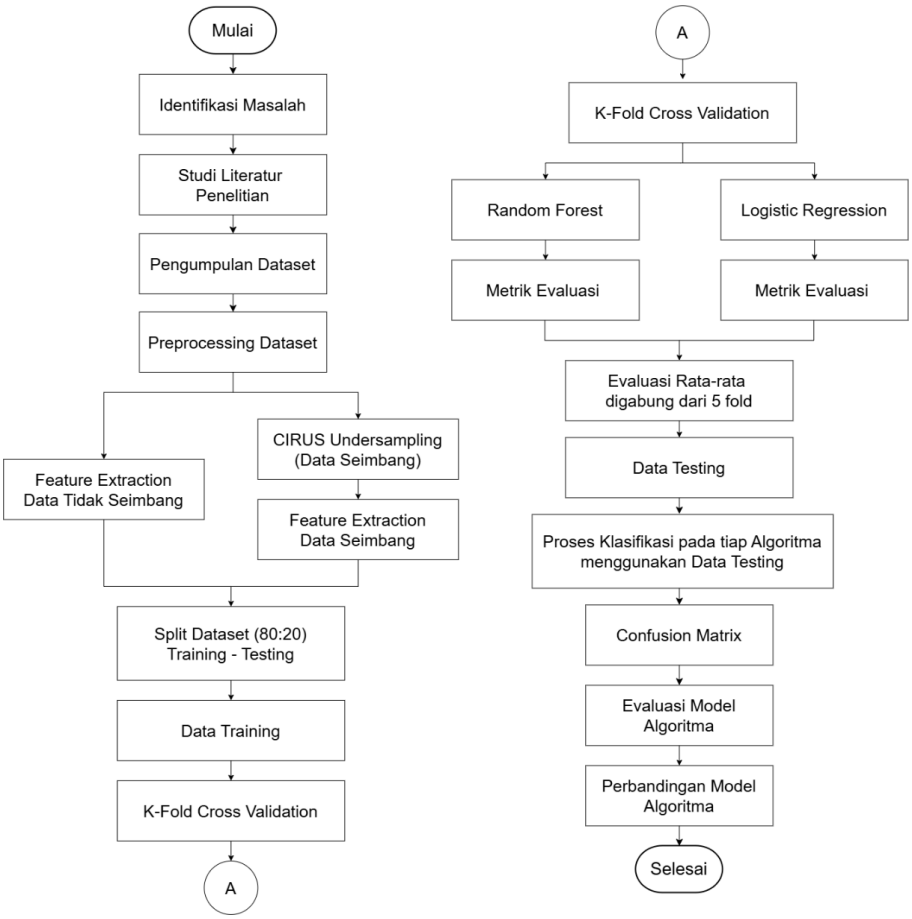
$$Precision = \frac{TP}{TP + FP} \, x \, 100\% \quad \ldots (ii)$$

$$Recall = \frac{TP}{TP + FN} \, x \, 100\% \quad \ldots (iii)$$

$$F1 - Score = \frac{2 \times Recall \times Precision}{Recall + Precision} \quad \ldots (iv)$$

3. METODE PENELITIAN

Penelitian ini menggunakan pendekatan eksperimen kuantitatif untuk membandingkan performa algoritma *Random Forest* dan *Logistic Regression* dalam klasifikasi teks berita hoaks. Berikut tahapan penelitian secara keseluruhan dapat ditunjukkan pada Gambar 1.



Gambar 1. Tahapan Penelitian.

Dataset yang digunakan diperoleh dari repositori publik Kaggle berjudul *Indonesian Hoax News Dataset*, yang memuat data klarifikasi berita dari situs resmi Kementerian Komunikasi dan Informatika Republik Indonesia (*komdigi.go.id).* Dataset awal terdiri atas 17.185 artikel berita yang dipublikasikan pada periode 06 Desember 2018 sampai 04 Maret 2026. Dari struktur dataset yang tersedia, penelitian ini secara khusus berfokus pada dua atribut teks utama sebagai fitur klasifikasi, yaitu *tittle* (judul berita) dan *body_text* (isi berita), yang dipadukan dengan atribut label kelas biner (Hoaks dan Tidak Hoaks).

Inspeksi awal dijalankan guna menguji validitas korpus melalui pendeteksian rekaman kosong *(missing values)* serta catatan duplikasi. Dari audit tersebut, teridentifikasi 1.557 data tanpa isi pada ruas teks dan 217 rekaman redundan, sehingga eliminasi keduanya menyisakan 15.411 artikel bersih yang siap diolah lebih lanjut. Tahap prapemrosesan kemudian diterapkan serentak pada kolom *title* dan *body_text* melalui rangkaian berurutan: pembersihan leksikal (*cleaning*) dari entitas numerik, tanda baca, pranala web, emotikon, dan spasi ganda; transformasi kapitalisasi (*case folding*) ke huruf kecil; fragmentasi teks menjadi token kata terpisah (*tokenization*); eliminasi leksikon non-diskriminatif berbasis senarai stopword bahasa Indonesia; serta pengembalian morfem kata ke bentuk lema asalnya (*stemming*) memanfaatkan algoritma morfologi Sastrawi.

Eksperimen dirancang ke dalam dua skenario komparatif untuk menilai dampak penyeimbangan data secara objektif. Pada skenario pertama memproses korpus awal yang masih asimetris sebanyak 15.411 sampel (10.274 dokumen hoaks berbanding 5.137 non-hoaks). Sebaliknya, skenario kedua mengaplikasikan metode CIRUS pada kelompok hoaks melalui pengelompokan *compact sets* dan penarikan terarah 50% data secara representatif, menyusutkan volume hoaks menjadi 5.137 sampel sehingga menghasilkan korpus seimbang sempurna (rasio 1:1 dengan total 10.274 rekaman). Vektorisasi teks diterapkan lewat metode TF-IDF gabungan unigram dan bigram yang dibatasi ketat pada kuota 10.000 fitur unggulan (terbagi simetris: 5.000 atribut dari judul dan 5.000 atribut dari teks isi) guna memastikan efisiensi memori dan komputasi. Skema evaluasi dijalankan menggunakan validasi silang *K-Fold Cross Validation* (K=5), di mana data dipecah merata ke dalam lima partisi; pada setiap iterasi, empat partisi difungsikan sebagai data latih dan satu partisi dialokasikan sebagai data uji. Pengukuran komparatif mencakup parameter Akurasi, Presisi, *Recall*, dan *F1-Score*, lengkap dengan simpangan baku (standar deviasi) dari seluruh *fold* pengujian.

4. HASIL DAN PEMBAHASAN

Karakteristik Dataset dan Hasil *Preprocessing* Data

Pemeriksaan data awal terhadap 17.185 artikel berita menunjukkan adanya 1.557 data dengan nilai kosong pada kedua atribut teks utama, yaitu *tittle* (judul berita) dan *body_text* (isi berita), serta 217 data artikel yang terduplikasi. Setelah kedua kelompok data tersebut dilakukan *cleaning*, diperoleh kumpulan dataset akhir sebanyak 15.411 artikel berita bersih yang terdiri dari 10.274 berita hoaks (66,67%) dan 5.137 berita tidak hoaks (33,33%). Sebelum dilakukan pembobotan fitur dan pemodelan klasifikasi, data teks mentah dari gabungan kedua atribut utama yaitu *tittle* (judul berita) dan *body_text* (isi berita) diproses melalui tahapan *preprocessing* data. Berikut hasil tahapan *preprocessing* data pada atribut *tittle* (judul berita) dan *body_text* (Isi berita) yang disajikan pada Tabel 1.

Tabel 1. Hasil *Preprocessing* Data pada Gabungan Atribut *Tittle* (Judul Berita) dan *Body_text* (Isi berita)

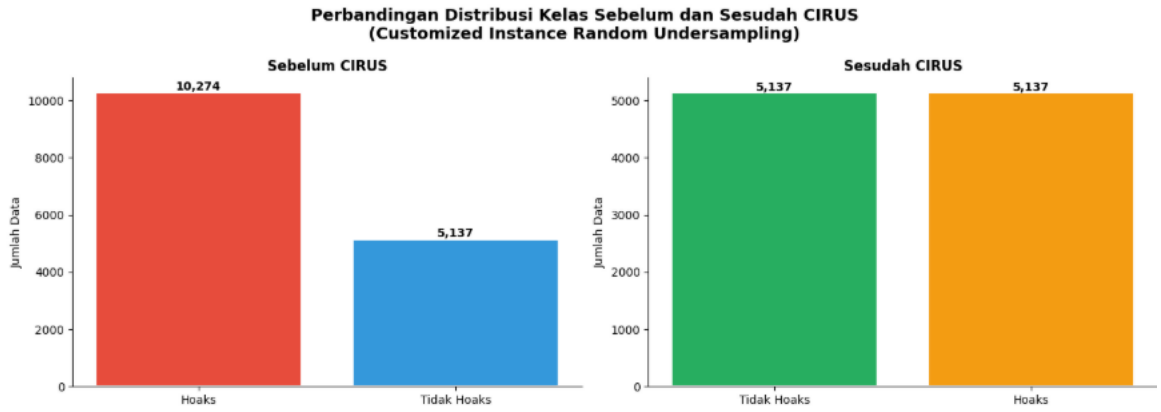
Tahapan *Preprocessing* Data	Hasil
Data Awal	Penjelasan: Beredar narasi di media sosial Facebook berisi klaim bahwa minum air garam hangat dapat membunuh virus dan menyembuhkan penyakit. Faktanya, Kementerian Kesehatan menegaskan informasi tersebut tidak berdasar medis!!! Bukti lengkap di https://komdigi.go.id/klarifikasi #cekfakta
Cleaning Data	Penjelasan Beredar narasi di media sosial Facebook berisi klaim bahwa minum air garam hangat dapat membunuh virus dan menyembuhkan penyakit Faktanya Kementerian Kesehatan menegaskan informasi tersebut tidak berdasar medis Bukti lengkap di cekfakta
Case Folding	penjelasan beredar narasi di media sosial facebook berisi klaim bahwa minum air garam hangat dapat membunuh virus dan menyembuhkan penyakit faktanya kementerian kesehatan menegaskan informasi tersebut tidak berdasar medis bukti lengkap di cekfakta
Tokenization	["penjelasan", "beredar", "narasi", "di", "media", "sosial", "facebook", "berisi", "klaim", "bahwa", "minum", "air", "garam", "hangat", "dapat", "membunuh", "virus", "dan", "menyembuhkan", "penyakit", "faktanya", "kementerian", "kesehatan", "menegaskan", "informasi", "tersebut", "tidak", "berdasar", "medis", "bukti", "lengkap", "di", "cekfakta"]
Stopword Removal	["penjelasan", "beredar", "narasi", "media", "sosial", "facebook", "klaim", "minum", "air", "garam", "hangat", "membunuh", "virus", "menyembuhkan", "penyakit", "kementerian", "kesehatan", "menegaskan", "informasi", "berdasar", "medis", "bukti", "lengkap", "cekfakta"]
Stemming	["jelas", "edar", "narasi", "media", "sosial", "facebook", "klaim", "minum", "air", "garam", "hangat", "bunuh", "virus", "sembuh", "sakit", "menteri", "sehat", "tegas", "informasi", "dasar", "medis", "bukti", "lengkap", "cekfakta"]

Berdasarkan pada Tabel 1, menunjukkan bahwa hasil proses *preprocessing* berhasil mengikis derau non-konseptual seperti tanda baca berulang, tautan situs web, simbol acak, dan tagar pada tahap *cleaning*, yang dilanjutkan dengan standardisasi karakter menjadi huruf kecil melalui *case folding*. Rangkaian tahapan berikutnya mulai dari tokenisasi, penyaringan kata henti, hingga *stemming* Sastrawi diterapkan secara identik pada atribut judul dan isi naskah,

yang pada akhirnya memadatkan panjang rata-rata tubuh berita menjadi 73,1 leksem per dokumen.

Implementasi Penyeimbangan Data Menggunakan Metode CIRUS

Kondisi sebaran label pada 15.411 data bersih memperlihatkan polarisasi kelas yang mencolok sebelum tahap ekstraksi fitur dilakukan. Kategori hoaks mendominasi ruang sampel sebesar 66,67% (10.274 artikel), sementara porsi berita tidak hoaks sebesar 33,33% (5.137 artikel), merefleksikan rasio ketimpangan 2:1. Melalui intervensi *Customized Instance Random Undersampling* (CIRUS), kelas mayoritas direduksi secara terukur via pembentukan *compact sets* berbasis *K-Means* serta sampling representatif. Mekanisme ini berhasil memangkas volume hoaks tepat menjadi 5.137 dokumen (50,00%), selaras dengan kuantitas kelas tidak hoaks yang berada pada angka 5.137 dokumen (50,00%). Keseluruhan dataset pasca-perlakuan CIRUS menghasilkan 10.274 sampel dengan perbandingan seimbang 1:1. Pergeseran proporsi kelas sebelum dan sesudah penanganan CIRUS dapat disajikan pada Gambar 2.



Gambar 2. Hasil Perbandingan Distribusi Kelas Sebelum dan Sesudah *Penerapan* CIRUS.

Ekstraksi Fitur Berbasis TF-IDF pada Atribut *Title* dan *Body_text*

Pasca-tahap prapemrosesan, korpus ditransformasikan ke dalam representasi numerik melalui TF-IDF dengan jangkauan *n-gram unigram* dan *bigram.* Mengingat disparitas panjang yang kontras antara judul (rata-rata 8,8 kata) dan badan warta (rata-rata 73,1 kata), kuota ekstraksi dibagi secara seimbang: 5.000 dimensi dari elemen judul serta 5.000 dimensi dari badan teks. Pengalokasian simetris ini krusial agar kunci pada judul tidak tereliminasi oleh dominasi volume kata pada isi naskah. Penapisan parameter *min_df=2* dan *max_df=0,95* diintegrasikan bersama normalisasi logaritmik guna mengeliminasi istilah langka maupun kata yang terlampau umum. Berikut ringkasan karakteristik matriks fitur TF-IDF pada kedua skenario pengujian dapat disajikan pada pada Tabel 2.

Tabel 2. Spesifikasi Matriks Ekstraksi Fitur TF-IDF pada Elemen Judul dan Isi Berita.

Komponen Fitur	Skenario 1 (Data *Imbalanced*)	Skenario 2 (Data *Balanced* Pasca-CIRUS)
Jumlah Dokumen Berita	15.411 artikel	10.274 artikel
Total Dimensi Fitur Gabungan	10.000 fitur (kolom)	10.000 fitur (kolom)
Filter Frekuensi Minimum (*min_df*)	2 (hapus kata *typo*/unik)	2 (hapus kata *typo*/unik)
Filter Frekuensi Maksimum (*max_df*)	0.95 (hapus kata terlalu umum)	0.95 (hapus kata terlalu umum)
Penskalaan Frekuensi (*sublinear_tf*)	*True (1 + log(tf))*	*True (1 + log(tf))*
Dimensi Matriks Akhir (Baris × Kolom)	15.411 × 10.000 = 54.110.000	10.274 × 10.000 = 102.740.000
Kepadatan Matriks	0,63%	0,82%

Berdasarkan pada Tabel 2, menunjukkan bahwa kombinasi *unigram-bigram* sukses mengonstruksi 10.000 atribut prediktor terpadu. Konfigurasi min_df = 2 dan max_df = 0,95 secara efektif menyingkirkan salah eja leksikal dan istilah nirfaedah, sementara opsi sublinear_tf meredam lonjakan bobot pada kata yang berulang. Matriks skenario pertama berdimensi 15.411 baris × 10.000 kolom dengan tingkat kepadatan 0,63%, sedangkan matriks skenario kedua (pasca-CIRUS) mencakup 10.274 baris × 10.000 kolom dengan densitas meningkat menjadi 0,82%. Kepadatan yang lebih tinggi pada skenario kedua menegaskan bahwa eliminasi instansi acak terarah oleh CIRUS berhasil menciptakan ruang fitur yang lebih padat, ringkas, dan sarat informasi diskriminatif. berikut hasil nilai bobot TF-IDF gabungan dari fitur judul dan isi berita dapat disajikan pada Tabel 3.

Tabel 3. Hasil Nilai Bobot Fitur TF-IDF Gabungan Atribut *Tittle* dan *Body_text.*

ID	Kategori Aktual	*Title*_vaksin	*Title*_covid	*Body*_warga	*Body*_korupsi
63285	Hoaks	0,214	0,198	0,052	0,000
47310	Hoaks	0,176	0,000	0,000	0,085
51206	Tidak Hoaks	0,000	0,153	0,072	0,000
39824	Tidak Hoaks	0,000	0,000	0,049	0,037

Berdasarkan Tabel 3, menunjukkan bahwa hasil pembobotan di mana tiap kata dari gabungan judul (*tittle*) dan isi berita (*body_text*) telah berhasil diubah menjadi nilai angka bobot TF-IDF. Setiap baris mewakili satu artikel berita, sedangkan kolomnya memuat kata-kata kunci penting yang diperoleh dari judul (seperti *Title*_vaksin dan *Title*_covid) serta dari isi berita (seperti *Body*_warga dan *Body*_korupsi). Terdapat nilai 0,000 menandakan bahwa kata tersebut tidak muncul sama sekali di dalam artikel berita yang bersangkutan. Sebaliknya, angka yang lebih besar dari nol menunjukkan bahwa kata tersebut muncul dan memiliki bobot pengaruh pada berita hoaks dengan ID 63285, di mana kata vaksin (0,214) dan covid (0,198) pada judul memperoleh nilai bobot yang cukup tinggi. Hal ini membuktikan bahwa penggabungan kedua fitur tersebut berhasil menyatukan informasi penting dari kedua bagian teks.

Hasil Evaluasi Pengujian Model *Random Forest* dan *Logistic Regression* Berdasarkan *K-Fold Cross Validation* (K=5)

Pengujian performa model dilakukan secara ketat menggunakan metode *K-Fold Cross Validation* dengan nilai K=5 untuk memastikan bahwa evaluasi bebas dari bias pembagian data acak tunggal. Pada setiap iterasi *fold*, 80% data untuk pelatihan model dan 20% data untuk data uji independen. Pada skenario pertama menggunakan data *imbalanced* sebanyak 15.411 sampel, setiap *fold* terdiri atas 12.328 data latih dan 3.083 data uji. Berikut hasil metrik evaluasi pada setiap iterasi *fold* untuk skenario pertama dapat disajikan pada Tabel 4.

Tabel 4. Hasil Evaluasi Pengujian *K-Fold Cross Validation* pada Data *Imbalanced.*

Fold	*RF: Accuracy*	*RF: Precision*	*RF: Recall*	*RF: F1-Score*	*LR: Accuracy*	*LR: Precision*	*LR: Recall*	*LR: F1-Score*
Fold 1	99,74%	99,74%	99,74%	99,74%	99,48%	99,48%	99,48%	99,48%
Fold 2	99,84%	99,84%	99,84%	99,84%	99,54%	99,55%	99,54%	99,54%
Fold 3	99,80%	99,81%	99,80%	99,80%	99,58%	99,58%	99,58%	99,58%
Fold 4	99,74%	99,74%	99,74%	99,74%	99,38%	99,39%	99,38%	99,38%
Fold 5	99,90%	99,90%	99,90%	99,90%	99,54%	99,55%	99,54%	99,54%
Rata-rata	99,81%	99,81%	99,81%	99,81%	99,51%	99,51%	99,51%	99,51%
Standar Deviasi	±0,08%	±0,08%	±0,08%	±0,08%	±0,07%	±0,07%	±0,07%	±0,07%

Berdasarkan pada Tabel 4, menunjukkan bahwa pada data yang timpang, *Random Forest* dan *Logistic Regression* mencatatkan rata-rata akurasi semu yang nyaris sempurna, yakni 99,81% (simpangan baku ±0,08%). Angka tinggi yang seragam ini mengindikasikan kehadiran *accuracy paradox* secara nyata, di mana model memperoleh skor agregat fantastis murni karena kecenderungan algoritma memprediksi kelas dominan. Untuk mengatasi distorsi evaluasi tersebut, pengujian dilanjutkan pada skenario kedua memanfaatkan korpus berimbang pasca-CIRUS (10.274 artikel), dengan pembagian 8.219 data latih (80%) dan 2.055 data uji (20%) per *fold* pada rasio 1:1. Hasil pengujian per *fold* pada data *balanced* pasca-CIRUS dapat disajikan pada Tabel 5.

Tabel 5. Hasil Evaluasi Pengujian *K-Fold Cross Validation* pada Data *Balanced.*

Fold	*RF: Accuracy*	*RF: Precision*	*RF: Recall*	*RF: F1-Score*	*LR: Accuracy*	*LR: Precision*	*LR: Recall*	*LR: F1-Score*
Fold 1	94,65%	94,88%	94,65%	94,64%	90,80%	91,64%	90,80%	90,76%
Fold 2	95,38%	95,46%	95,38%	95,38%	92,26%	92,70%	92,26%	92,25%
Fold 3	95,77%	95,81%	95,77%	95,77%	93,14%	93,14%	93,14%	93,14%
Fold 4	94,94%	95,01%	94,94%	94,94%	91,82%	92,16%	91,82%	91,81%
Fold 5	95,08%	95,21%	95,08%	95,07%	93,04%	93,37%	93,04%	93,01%
Rata-rata	95,16%	95,27%	95,16%	95,16%	92,21%	92,60%	92,21%	92,19%
Standar Deviasi	±0,43%	±0,37%	±0,43%	±0,43%	±0,96%	±0,77%	±0,96%	±0,97%

Berdasarkan Tabel 5, menunjukkan hasil evaluasi pengujian tersebut terhadap data *balanced*, di mana *Random Forest* mengungguli *Logistic Regression* dengan rata-rata *accuracy* 95,16% dengan standar deviasi ±0,43% dibanding *Logistic Regression* 92,21% dengan standar deviasi ±0,96%. Keseimbangan nilai *precision, recall,* dan *f1-score* pada kedua model juga menegaskan bahwa penyeimbangan data berhasil mengeliminasi bias prediksi tanpa gejala *accuracy paradox.*

Confusion Matrix dan Metrik Evaluasi pada Model *Random Forest* dan *Logistic Regression*

Evaluasi mendalam terhadap efektivitas klasifikasi dilakukan menggunakan confusion matrix berbasis akumulasi data uji dari seluruh lipatan (*fold*) terdiri atas *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN) untuk memberikan gambaran rinci mengenai ketepatan prediksi model. Berikut hasil confusion matrix untuk kedua algoritma pada masing-masing skenario disajikan pada Tabel 6.

Tabel 6. Hasil *Confusion Matrix* Model *Random Forest* dan *Logistic Regression*.

Skenario (Total Data)	Algoritma Model	TP	FN	FP	TN	Prediksi Benar	Prediksi Salah
Imbalanced (15.411)	Random Forest	10.070	3	27	5.311	15.381 (99,81%)	30 (0,19%)
Imbalanced (15.411)	Logistic Regression	10.072	1	75	5.232	15.304 (99,31%)	76 (0,49%)
Balanced (10.274)	Random Forest	5.010	127	370	4.767	9.777 (95,16%)	497 (4,84%)
Balanced (10.274)	Logistic Regression	4.994	143	657	4.480	9.474 (92,21%)	800 (7,79%)

Berdasarkan pada Tabel 6, menunjukkan bahwa evaluasi model performa sebelum dan sesudah intervensi CIRUS. Pada Skenario 1 (15.411 data *imbalanced*), kedua algoritma menunjukkan jumlah prediksi benar yang sangat besar *Random Forest* hanya mencatat 30 galat (99,81%) dan *Logistic Regression* 76 galat (99,31%) yang sepenuhnya dipicu oleh dominasi berita hoaks di dataset awal. Pada skenario kedua (10.274 data seimbang), *Random Forest* membuktikan keunggulannya dengan membukukan 9.777 prediksi tepat (95,16%) dan menekan kesalahan hingga 497 kasus (4,84%), yang mencakup 127 berita hoaks tak terdeteksi (FN) serta 370 berita fakta yang terlabeli keliru (FP). Sebaliknya, *Logistic Regression* menghasilkan 9.474 prediksi tepat (92,21%) dengan tingkat galat mencapai 800 data (7,79%), di mana kelemahan utamanya terletak pada salah klasifikasi 657 berita fakta menjadi hoaks. Berikut hasil performa mode pada masing-masing kelas melalui perbandingan *classification report* dapat disajikan pada Tabel 7.

Tabel 7. Hasil *Classification Report* Model Random Forest dan Logistic Regression.

Skenario Data	Algoritma Model	Kategori Kelas	*Accuracy (Mean ± Std)*	*Precision*	*Recall*	*F1-Score*	*Support*
Imbalanced	Random Forest	Hoaks	99,81% (±0,08%)	99,73%	99,97%	99,85%	10.274
Imbalanced	Random Forest	Tidak Hoaks	99,81% (±0,08%)	99,94%	99,49%	99,72%	5.137
Imbalanced	Logistic Regression	Hoaks	99,51% (±0,07%)	99,26%	99,99%	99,62%	10.274
Imbalanced	Logistic Regression	Tidak Hoaks	99,51% (±0,07%)	99,98%	98,59%	99,28%	5.137
Balanced (CIRUS)	Random Forest	Hoaks	95,16% (±0,43%)	93,12%	97,53%	95,27%	5.137
Balanced (CIRUS)	Random Forest	Tidak Hoaks	95,16% (±0,43%)	97,41%	92,80%	95,05%	5.137
Balanced (CIRUS)	Logistic Regression	Hoaks	92,21% (±0,96%)	88,37%	97,22%	92,58%	5.137
Balanced (CIRUS)	Logistic Regression	Tidak Hoaks	92,21% (±0,96%)	96,91%	87,21%	91,81%	5.137

Berdasarkan Tabel 7, menunjukkan bahwa algoritma *Random Forest* terbukti lebih unggul pada data *balanced* dengan akurasi sebesar 95,16% serta standar deviasi ±0,43% dan nilai F1-Score di atas 95% pada kedua kelas berita. Model ini sangat efektif mendeteksi berita hoaks dengan *Recall* 97,53% sekaligus menjaga kebenaran berita fakta (*Precision* 97,41%). Sementara itu, *Logistic Regression* dengan akurasi 92,21% serta standar deviasi ±0,96% dengan penurunan *Recall* pada kelas tidak hoaks (87,21%).

5. KESIMPULAN DAN SARAN

Berdasarkan hasil penelitian dan pengujian yang telah dilakukan, dapat disimpulkan bahwa algoritma *Random Forest* memiliki performa yang lebih unggul dan stabil dibandingkan *Logistic Regression* dalam mengklasifikasikan berita hoaks berbahasa Indonesia. Pada data yang telah diseimbangkan dengan metode CIRUS (10.274 data), *Random Forest* mencapai Akurasi 95,16% dengan standar deviasi ±0,43%, Presisi 95,27%, *Recall* 95,16%, dan F1-Score 95,16%, sedangkan *Logistic Regression* memperoleh Akurasi 92,21% dengan standar deviasi ±0,96%, Presisi 92,66%, Recall 92,21%, dan F1-Score 92,19%. Keunggulan *Random Forest* ini menunjukkan bahwa pendekatan *ensemble* lebih efektif dalam menangani fitur teks berdimensi tinggi dari TF-IDF. Selain itu, penerapan metode CIRUS terbukti mampu mengatasi masalah *accuracy paradox* pada data yang tidak seimbang, sehingga evaluasi model menjadi lebih objektif dan dapat diandalkan.

Penggunaan teknik penanganan ketidakseimbangan data (*imbalanced data*) lain, seperti *Synthetic Minority Over-sampling Technique* (SMOTE), *Adaptive Synthetic Sampling*

(ADASYN), maupun pendekatan *hybrid*, dapat dipertimbangkan pada penelitian berikutnya. Perbandingan antara metode-metode tersebut dan *Customized Instance Random Undersampling* (CIRUS) memungkinkan evaluasi yang lebih mendalam terkait efektivitas relatifnya terhadap klasifikasi teks. Dengan demikian, hasil analisis diharapkan dapat memberikan gambaran yang lebih menyeluruh mengenai efektivitas penyeimbangan data pada deteksi berita hoaks. Selain itu, eksplorasi ekstraksi fitur berbasis *word embedding* kontekstual, seperti FastText dan Word2Vec perlu diterapkan. Dengan demikian, pengembangan teknik penyeimbangan data dan representasi fitur tersebut diharapkan dapat meningkatkan akurasi performa algoritma pendekatan *text classification* serta presisi deteksi berita hoaks secara lebih optimal.

DAFTAR REFERENSI

Adebesin, F., Smuts, H., & Mawela, T. (2023). The Role of Social Media in Health Misinformation and Disinformation During the COVID-19 Pandemic: Bibliometric Analysis Corresponding Author: 3. https://doi.org/10.2196/48620

Annie, K., & Jung, I. (2023). Optimal selection of resampling methods for imbalanced data with high complexity. 1–18. https://doi.org/10.1371/journal.pone.0288540

Breiman, L. E. O. (2001). Random Forests. 5–32.

Elsoud, E. A., Hassan, M., Alidmat, O., Henawi, E. Al, Alshdaifat, N., Iqtait, M., Ghaben, A., Katrawi, A., & Dmour, M. (2024). Under Sampling Techniques for Handling Unbalanced Data with Various Imbalance Rates: A Comparative Study. IJACSA International Journal of Advanced Computer Science and Applications, 15(8), 1274–1284. https://doi.org/10.14569/IJACSA.2024.01508124

Gillala, R., Reddy, V. K., & Kumar, A. (2024). CIRUS: Critical Instances Removal based Under-Sampling - A solution for class imbalance problem. https://doi.org/10.3233/HIS-200279

Hashim, A., Almarashy, J., & Salehpour, P. (2023). Enhancing Fake News Detection by Multi-Feature Classification. IEEE Access, PP, 1. https://doi.org/10.1109/ACCESS.2023.3339621

Hassan, K. F., Abdullah, M. N., & Al-hchimy, Z. S. (2024). The Application of Random Forest to the Classification of Fake News. 00049, 1–12. https://doi.org/10.1051/bioconf/20249700049

Hoyos-osorio, J., Alvarez-meza, A., Daza-santacoloma, G., & Orozco-gutierrez, A. (2021). Relevant information undersampling to support imbalanced data classification. Neurocomputing, 436, 136–146. https://doi.org/10.1016/j.neucom.2021.01.033

Iizuka, R., Toriumi, F., Nishiguchi, M., Takano, M., Yoshida, M., & Guidi, B. (2022). Impact of correcting misinformation on social disruption. PLOS ONE, 1–21. https://doi.org/10.1371/journal.pone.0265734

Jahan, I., Hasan, N., Islam, S. N., Akter, L., & Onik, K. R. (2024). Advanced machine learning techniques for fake news detection: A comprehensive analysis.

Kowsari, K., Meimandi, K. J., Heidarysafa, M., & Mendu, S. (2019). Text Classification Algorithms: A Survey. 1–68. https://doi.org/10.3390/info10040150

Lavalley, M. P. (2008). Statistical Primer for Cardiovascular Research Logistic Regression. 2395–2399. https://doi.org/10.1161/CIRCULATIONAHA.106.682658

Lumumba, V. W., Kiprotich, D., Makena, N. G., Kavita, M. D., & Mpaine, M. L. (2024). Comparative Analysis of Cross-Validation Techniques: LOOCV, K-folds Cross-Validation, and Repeated K-folds Cross-Validation in Machine Learning Models. 13(5), 127–137. https://www.sciencepublishinggroup.com/article/10.11648/j.ajtas.20241305.13

Othman, N. A., Elzanfaly, D. S., & Elhawary, M. M. M. (2024). Arabic Fake News Detection Using Deep Learning. IEEE Access, 12(September), 122363–122376. https://doi.org/10.1109/ACCESS.2024.3451128

Paramita, N., Adiati, R., Febriyan, D., & Indarjani, S. (2024). Comparative study of predictive models for hoax and disinformation detection in indonesian news. International Journal of Advances in Intelligent Informatics, 10(3), 504–516.

Putra, B. T., & Yulianingsih, E. (2023). Analisis Tingkat Akurasi Prediksi Gejala COVID-19 Dengan Menggunakan Metode Logistic Regression dan Support Vector Machine. 13(2), 269–278. https://ejurnal.umri.ac.id/index.php/JIK/article/view/5629

Septiani, D., & Isabela, I. (2022). Analisis Term Frequency Inverse Document Frequency (Tf-Idf) Dalam Temu Kembali Informasi Pada Dokumen Teks. SINTESIA: Jurnal Sistem Dan Teknologi Informasi Indonesia, 1, 81–88.

Swana, E. F., & Doorsamy, W. (2022). Tomek Link and SMOTE Approaches for Machine Fault Classification with an Imbalanced Dataset.

Tusell-rey, C. C., Camacho-nieto, O., & Cornelio, Y. (2022). Customized Instance Random Undersampling to Increase Knowledge Management for Multiclass Imbalanced Data Classification. Sustainability. https://doi.org/10.3390/su142114398

Zhou, X., & Zafarani, R. (2020). A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities.