JURNAL INFORMATIKA TEKNOLOGI DAN SAINS (JINTEKS) Hal. 1045-1054
E-ISSN : 2686-3359 Vol. 8; No. 2
DOI PREFIX : 10.51401 Mei 2026

PENANGANAN EXTREME CLASS IMBALANCE PADA ANALISIS SENTIMEN PROVIDER INTERNET MENGGUNAKAN SMOTE RANDOM FOREST DAN SUPPORT VECTOR MACHINE

Risca Lusiana Pratiwi[1], Ginabila[2*], Zulia Imami Alfianti[3]
[1]Sistem Informasi, Universitas Nusa Mandiri
[2,3]Sistem Informasi, Universitas Bina Sarana Informatika
email: gina.gnb@bsi.ac.id[2*]

Abstrak: Ketimpangan distribusi kelas yang ekstrem pada data ulasan pelanggan menjadi tantangan utama yang memicu bias prediksi pada model pembelajaran mesin. Penelitian ini bertujuan untuk menguji efektivitas penanganan *extreme class imbalance* pada ulasan penyedia jasa internet menggunakan pendekatan *Synthetic Minority Over-sampling Technique* (SMOTE) yang dikombinasikan dengan algoritma *Random Forest* dan *Support Vector Machine* (SVM). Dataset berjumlah 699 rekaman ulasan diolah melalui rangkaian pra-pemrosesan teks, ekstraksi fitur *Term Frequency-Inverse Document Frequency* (TF-IDF), serta diuji menggunakan skema *Cross Validation*. Hasil evaluasi memperlihatkan bahwa kombinasi SMOTE dan SVM mencatatkan akurasi sebesar 91,56%, nilai *Area Under Curve* (AUC) 0,769, presisi kelas positif 100,00%, dan *recall* kelas positif sebesar 4,84%. Di sisi lain, skema SMOTE dan *Random Forest* memperoleh akurasi 91,13%, nilai AUC 0,740, serta presisi dan *recall* kelas positif sebesar 0,00%. Penelitian ini menyimpulkan bahwa batas pemisah linier pada SVM lebih responsif memanfaatkan sampel sintetis SMOTE dalam ruang fitur berdimensi tinggi, serta mengonfirmasi bahwa kedua algoritma memiliki kecenderungan prediktif yang dominan dalam mengenali sampel kelas mayoritas.

Kata Kunci : Analisis Sentimen, *Extreme Class Imbalance*, *Random Forest*, SMOTE, *Support Vector Machine*

PENDAHULUAN

Tingginya aktivitas masyarakat dalam mengakses platform konten dan aplikasi pada ponsel cerdas mendorong kenaikan lalu lintas data internet secara berkelanjutan di Indonesia. Berdasarkan data terbaru, tingkat penetrasi internet telah menjangkau 79,5% populasi atau setara dengan 221,6 juta penduduk. Peningkatan performa jaringan juga terlihat dari kecepatan unduh *mobile broadband* di beberapa provinsi yang telah menembus batas 25 Mbps, bersamaan dengan perluasan akses jaringan 5G yang kini telah beroperasi di 48 kota [1]. Peningkatan kualitas pelayanan dari pihak ISP secara langsung berbanding lurus dengan persepsi positif pelanggan, di mana keandalan dan kerapian layanan menjadi kunci utama keberhasilan menjaga kepuasan pengguna [2]. Opini langsung pengguna yang dituangkan dalam bentuk ulasan tertulis menjadi sumber data penting yang merekam realitas lapangan, mulai dari keandalan koneksi, keterjangkauan tarif, sampai kesigapan petugas dalam menangani gangguan [3]. Kondisi ini menyebabkan setiap keluhan maupun pujian yang diunggah pengguna menjadi cerminan langsung dari kualitas operasional penyedia jasa di lapangan yang tidak boleh diabaikan.

Proporsi ulasan publik terhadap penyedia layanan internet cenderung didominasi oleh tanggapan negatif yang mengindikasikan bahwa media digital umumnya dimanfaatkan pengguna sebagai wadah penyampaian keluhan teknis maupun administratif [4]. Ketidaknyamanan saat terjadi gangguan jaringan berfungsi sebagai reaksi emosional, sehingga pengguna cenderung jauh lebih responsif menuliskan keluhan daripada saat jaringan berjalan lancar [5]. Perilaku pengguna yang reaktif ini secara tidak langsung membentuk kumpulan data teks yang tidak seimbang, di mana keluhan teknis jaringan, gangguan modem, dan kelemahan customer service jauh lebih mendominasi ketimbang ekspresi kepuasan. Kondisi ketidakseimbangan kelas yang ekstrem (*extreme class imbalance*) tersebut menyebabkan model klasifikasi konvensional mengalami kecenderungan untuk memprediksi seluruh data ke dalam kelas mayoritas. Bias tersebut memicu munculnya akurasi semu, di mana performa model terlihat tinggi secara keseluruhan padahal gagal mengidentifikasi pola sentimen positif dari kelas minoritas [6].

Pendekatan *Synthetic Minority Over-sampling Technique* (SMOTE) merupakan salah satu variasi teknik *over-sampling* untuk menyelesaikan isu *extreme imbalanced class* dan mampu mendongkrak tingkat sensitivitas model [7]. Penanganan *class imbalance* menggunakan *Synthetic Minority Over-sampling Technique* (SMOTE) efektif menyeimbangkan proporsi dataset ulasan pengguna, seperti yang ditunjukkan dalam penelitian Ginabila dan Fauzi pada tahun 2023[8], serta Pratiwi pada tahun 2025 [9]. Penggunaan SMOTE memastikan model klasifikasi seperti SVM dan Naive Bayes tidak terbias pada kelas mayoritas dan menghasilkan evaluasi performa yang lebih presisi. Melalui pembentukan data sintetis pada kelas minoritas ini, model dilatih agar memiliki kepekaan yang seimbang dalam mengenali pola kalimat positif maupun negatif secara lebih adil.

Pemanfaatan *Random Forest* memberikan keunggulan komputasi saat berhadapan dengan data bertipe *high-dimensional*. Kombinasi dari variasi pohon keputusan yang dibangun secara acak membantu algoritma ini dalam mengekstrak fitur-fitur krusial sekaligus mencegah penyimpangan hasil akibat *overfitting* [10]. Sementara itu, algoritma *Support Vector Machine* (SVM) memiliki keandalan tinggi dalam memetakan *hyperplane* pemisah yang presisi, khususnya ketika dihadapkan pada struktur data teks berdimensi tinggi yang kompleks [11]. Meski perbandingan kedua

JURNAL INFORMATIKA TEKNOLOGI DAN SAINS (JINTEKS)
E-ISSN : 2686-3359
DOI PREFIX : 10.51401
Hal. 1045-1054
Vol. 8; No. 2
Mei 2026

algoritma tersebut cukup sering dilakukan pada kajian analisis sentimen, pengujian khusus yang mengombinasikan teknik SMOTE untuk menangani ketimpangan kelas ekstrem pada data ulasan penyedia internet masih terhitung minim. Dengan menandingkan skema *ensemble* milik *Random Forest* dan pemisah bidang vektor dari SVM secara langsung, kita bisa memetakan keunggulan tiap model secara utuh dalam mengeksekusi data ulasan yang berdistribusi tidak seimbang.

Penelitian ini menerapkan tahapan *text preprocessing* secara sistematis yang mencakup *Transform Cases*, *Tokenize*, *Stemming*, penghapusan *stopwords*, serta pembuatan *n-Grams*. Informasi dari teks kemudian diekstrak melalui skema pembobotan TF-IDF, sebelum akhirnya data latih diselaraskan menggunakan metode SMOTE. Evaluasi komparatif kemudian dijalankan dengan mengimplementasikan dua algoritma utama, yakni *Random Forest* dan *Support Vector Machine* (SVM), untuk melihat efektivitas masing-masing model.

Penelitian ini dibuat untuk memberikan kerangka kerja klasifikasi yang objektif dalam mengolah ulasan berdistribusi tidak seimbang secara akurat tanpa terpengaruh oleh bias kelas mayoritas. Hasil komparasi dalam penelitian ini diharapkan dapat menjadi rujukan akademis sekaligus memberikan kontribusi praktis bagi pihak penyedia jasa internet dalam memahami keluhan dan persepsi pelanggan secara tepat. Dengan demikian, penelitian ini tidak hanya menyelesaikan permasalahan teknis berupa ketidakseimbangan data pada pembelajaran mesin, tetapi juga memberikan solusi nyata bagi industri telekomunikasi dalam memetakan kualitas pelayanan secara otomatis.

TINJAUAN PUSTAKA

Berbagai kombinasi pemodelan machine learning telah banyak diterapkan untuk menjawab tantangan analisis sentimen dan ketidakseimbangan kelas. Sebagai gambaran komprehensif, tabel di bawah memperlihatkan perbandingan sejumlah penelitian terdahulu yang mendasari sekaligus menjadi pembanding dalam kajian ini:

Tabel 1. Penelitian Terdahulu

No.	Peneliti & Tahun	Judul	Metode	Hasil	Keterbatasan
1	Muhammad Hafizd Mustaqim & Angga Bayu Santoso [12]	*Penerapan Penyeimbangan Data Pada Analisis Sentimen Ulasan Game Magic Chess Go Go di Play Store dengan Naive Bayes*	Naive Bayes, SMOTE, ADASYN, Random Oversampling (ROS), Random Undersampling (RUS), TF-IDF	SMOTE + Naive Bayes (skema 80:20) memberikan akurasi tertinggi sebesar **84,2%**, diikuti ADASYN (83,8%), ROS (82,9%), dan RUS (77,9%). Tanpa penyeimbangan akurasinya 71,0%.	Penelitian terbatas pada algoritma probabilistik tunggal (Naive Bayes) dan belum menguji algoritma Support Vector Machine atau Random Forest.
2	Anis Nikmatul Kasanah, Muladi, & Utomo Pujianto [13]	*Penerapan Teknik SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Objektivitas Berita Online Menggunakan Algoritma KNN*	K-Nearest Neighbor (KNN), SMOTE, Case Folding, Stopwords, Stemming	Kombinasi SMOTE + KNN (k=1) menghasilkan akurasi **87,50%**, Recall **0,98**, dan F1-Score **0,88**, serta terbukti meningkatkan sensitivitas terhadap kelas minoritas.	Menggunakan KNN yang komputasinya sangat lambat jika dataset berukuran besar serta sensitif terhadap *noise* atribut.
3	Dini Andriyani, Ahmad Faqih, & Sandy Eka Permana [14]	*The Effect of SMOTE Application on Support Vector Machine Performance in Sentiment Classification on Imbalanced Datasets*	Support Vector Machine (SVM), SMOTE, TF-IDF, BERT (Labeling)	Penerapan SMOTE berhasil meningkatkan akurasi model SVM sebesar **12,48%** (dari **71, 41%** tanpa SMOTE menjadi **83,89%** dengan SMOTE). F1-Score naik dari 0,68 menjadi 0,84.	Hanya mengevaluasi algoritma SVM tunggal tanpa membandingkan dengan algoritma *ensemble tree* seperti Random Forest.
4	Ivan Muliansyah, Nuralamsah Zulkarnaim, & Andi Amirul Asnan Cirua [15]	*Perbandingan Algoritma Support Vector Machine dan Random Forest untuk Analisis Sentimen Terkait Danantara*	Support Vector Machine (SVM), Random Forest (RF), TF-IDF, Lexicon InSet	**SVM** lebih unggul dengan akurasi **75%** (F1-score positif 85%), sedangkan **Random Forest** memperoleh akurasi **72%** (F1-score positif 72%).	Menggunakan pelabelan otomatis Lexicon yang menghasilkan banyak data imbalanced tanpa menerapkan teknik penyeimbangan data sintetis (SMOTE).
5	Cherfly Kaope & Yoga Pristyanto [16]	*The Effect of Class Imbalance Handling on Datasets Toward Classification Algorithm Performance*	ADASYN, SMOTE, SMOTE-ENN, AdaBoost, KNN, Random Forest (RF)	Penggabungan metode **ADASYN + Random Forest** memberikan peningkatan performa klasifikasi 5% hingga 10% lebih baik dibandingkan skema model lainnya.	Evaluasi dilakukan pada dataset tabular umum (KEEL repository) dan belum diuji pada dataset teks ulasan Bahasa Indonesia (*NLP/Text Mining*).

JURNAL INFORMATIKA TEKNOLOGI DAN SAINS (JINTEKS) Hal. 1045-1054
E-ISSN : 2686-3359 Vol. 8; No. 2
DOI PREFIX : 10.51401 Mei 2026

Hasil tinjauan pada tabel di atas mengonfirmasi bahwa teknik SMOTE secara konsisten berhasil memperbaiki ketimpangan data dan menaikkan performa klasifikasi pada berbagai model *machine learning*. Namun, sebagian besar studi terdahulu masih berfokus pada satu algoritma saja atau menguji dataset umum di luar domain opini teks Bahasa Indonesia. Oleh sebab itu, penelitian ini dirancang untuk menutup celah tersebut melalui pengujian komparatif antara *Random Forest* dan SVM dalam menangani data ulasan provider internet yang mengalami ketidakseimbangan kelas secara ekstrim.

Setelah melihat gambaran dari penelitian terdahulu, bagian tinjauan pustaka akan menjelaskan teori-teori dasar yang digunakan dalam penelitian. Penjelasan ini mencakup konsep pengolahan teks ulasan, ekstraksi fitur, penyeimbangan data, sampai algoritma klasifikasi yang digunakan. Ringkasan teori dan metodenya dapat dilihat pada uraian berikut:

1. *Text Mining*

   Text mining merupakan teknik pengolahan data berbasis komputasi yang berfungsi mengekstraksi dan menemukan pola informasi bermakna dari kumpulan dokumen teks tidak terstruktur. Metode ini bekerja dengan menguraikan sekumpulan teks mentah untuk mengidentifikasi pengetahuan serta informasi berharga secara otomatis [17].

2. Analisis Sentimen

   Analisis sentimen merupakan bagian dari penambangan teks (*text mining*) yang memanfaatkan teknik Pemrosesan Bahasa Alami (*Natural Language Processing*) untuk mengidentifikasi dan mengelompokkan polaritas dokumen. Metode ini bekerja dengan menguraikan ekspresi opini, sikap emosional, serta evaluasi tertulis pengguna terhadap suatu isu atau layanan tertentu ke dalam kategori sentimen yang spesifik [18].

3. Pra-Pemrosesan Teks (*Text Preprocessing*)

   Data preprocessing merupakan tahapan krusial dalam pemrosesan data mentah yang bertujuan untuk mengeliminasi gangguan, mengubah format, serta menyelaraskan struktur data agar dapat diolah oleh algoritma pembelajaran mesin secara lebih akurat [19].

4. SMOTE (*Synthetic Minority Over-sampling Technique*)

   SMOTE *(Synthetic Minority Over-sampling Technique)* berfungsi menyeimbangkan distribusi data yang timpang dengan cara membangkitkan sampel sintetis baru pada kelas minoritas. Pendekatan *oversampling* ini bertujuan untuk menyetarakan proporsi antarkelas serta meningkatkan kepekaan dan performa akurasi model klasifikasi [20].

5. *Random Forest*

   Random Forest adalah metode *machine learning* berbasis *ensembel* yang mengombinasikan prediksi dari banyak pohon keputusan *(decision trees)* secara mandiri. Keputusan akhir diambil berdasarkan mekanisme *voting* terbanyak *(majority voting)*, yang terbukti tangguh dalam mengurangi masalah overfitting dan menangani fitur data berdimensi tinggi [21].

6. *Support Vector Machine*

   Support Vector Machine (SVM) adalah metode *machine learning* berbasis statistik yang bekerja dengan memetakan data ke dalam ruang fitur dan menentukan bidang pemisah *(hyperplane)* optimal. Dengan memanfaatkan *support vectors* untuk memaksimalkan marjin pemisah, SVM terbukti sangat andal dalam menangani klasifikasi data teks yang kompleks [22].

7. Confusion Matrix

   Confusion Matrix berfungsi sebagai alat ukur kinerja model klasifikasi dengan membandingkan hasil prediksi sistem terhadap label sampel yang sebenarnya. Melalui pemetaan empat elemen utama (TP, TN, FP, dan FN), metrik evaluasi seperti akurasi, presisi, *recall*, hingga *F1-score* dapat dihitung untuk menilai keandalan model secara obyektif [23].

8. AUC (Area Under Curve)

   Area Under Curve (AUC) merepresentasikan luas wilayah di bawah kurva *Receiver Operating Characteristic* (ROC) yang mengukur kinerja pemisahan kelas oleh model klasifikasi. Nilai AUC yang mendekati angka 1,0 mengindikasikan bahwa model memiliki diferensiasi dan sensitivitas prediksi yang sangat baik pada berbagai tingkat ambang batas *(threshold)* [24].

METODE

Metode penelitian ini disusun secara sistematis agar penelitian yang dilakukan dapat direproduksi secara mandiri (*reproducible*). Tahapan penelitian dirancang menggunakan perangkat lunak AI Studio untuk mengolah data ulasan *provider* internet yang berkondisi *extreme class imbalance*. Alur penelitian dibagi menjadi lima fase utama yaitu pengintegrasian data mentah, pra-pemrosesan teks, ekstraksi fitur TF-IDF, penyeimbangan data sintetis dengan SMOTE, serta pemodelan komparatif menggunakan algoritma *Random Forest* dan *Support Vector Machine* (SVM) melalui skema *Cross Validation*. Berikut adalah gambaran dari kerangka alur penelitian:

JURNAL INFORMATIKA TEKNOLOGI DAN SAINS (JINTEKS)
E-ISSN : 2686-3359
DOI PREFIX : 10.51401
Hal. 1045-1054
Vol. 8; No. 2
Mei 2026

Gambar 1. Kerangka Alur Penelitian

1. Pengintegrasian Data Mentah
   Proses diawali dengan mengumpulkan dan membaca berkas dataset ulasan pelanggan penyedia jasa internet (*Dataset_Oke.xlsx*) sejumlah 699 rekaman data yang diambil dari review aplikasi MyRepublic, serta mengonfigurasi kolom *review* sebagai atribut teks utama dan *sentiment* sebagai variabel label target.
2. Pra-Pemrosesan Teks (*Text Preprocessing*)
   Pembersihan dokumen dilakukan melalui lima tahapan yaitu *Transform Cases* untuk menyelaraskan kapitalisasi huruf, *Tokenize* untuk memotong kalimat menjadi kata dasar, *Stemming* berbasis kamus untuk mengembalikan kata dasar, *Filter Stopwords* untuk membuang kata-kata umum, serta *Generate N-Grams* untuk membentuk kombinasi frasa kontekstual.
3. Ekstraksi Fitur TF-IDF
   Dokumen yang telah dibersihkan selanjutnya diubah menjadi representasi matriks berbobot numerik mengacu pada metode *Term Frequency-Inverse Document Frequency* (TF-IDF) untuk mengukur tingkat kepentingan kata. Pembobotan TF-IDF untuk kata t pada dokumen ulasan d dihitung menggunakan rumus pada persamaan:

   $$W_{t,d} = TF(t,d) \times \log\left(\frac{N}{DF(t)}\right) \quad (1)$$

   Keterangan:
   a. $W_{t,d}$: Bobot fitur kata t pada dokumen ulasan d.
   b. $\mathrm{TF}(t,d)$: Frekuensi kemunculan kata t dalam dokumen d.
   c. N: Total jumlah seluruh dokumen ulasan dalam dataset.
   d. $\mathrm{DF}(t)$: Jumlah dokumen yang memuat kata t.

4. Skema Cross Validation
   Pemisahan dataset menjadi data latih (*training fold*) dan data uji (*testing fold*) mengimplementasikan teknik *k-fold cross validation* untuk menjamin objektifitas serta keandalan validasi model.
5. Penyeimbangan Data Sintetis (SMOTE)
   Synthetic Minority Over-sampling Technique (SMOTE) diterapkan khusus pada data latih untuk meningkatkan sampel sintetis kelas minoritas (positif), sehingga proporsi distribusi antarkelas menjadi seimbang sebelum memasuki pelatihan. Persamaan matematika pembentukan sampel sintetis baru ($\mathbf{x}_{\text{baru}}$) dirumuskan pada persamaan (2):

   $$xbaru = xi + \lambda \cdot (xzi - xi) \quad (2)$$

   Keterangan:
   a. x_{baru} : Sampel data sintetis baru yang dihasilkan dari proses *oversampling*.
   b. x_i : Sampel data kelas minoritas aktual yang sedang diproses.
   c. x_{zi} : Salah satu sampel data tetangga terdekat (*k-nearest neighbors*) yang dipilih secara acak dari x_i.
   d. δ : Bilangan acak kontinu dalam rentang 0 hingga 1 ($\delta \in [0,1]$) yang menentukan titik posisi sampel sintetis baru.
6. Pemodelan Klasifikasi Komparatif
   Tahap pelatihan dijalankan secara komparatif dengan menguji dua algoritma utama yaitu *Random Forest* (berbasis *ensemble decision tree*) dan *Support Vector Machine* (SVM linier) untuk memetakan keunggulan masing-masing model. SVM bekerja dengan mencari bidang pemisah (*hyperplane*) linier dengan marjin maksimal yang memisahkan vektor fitur kelas positif dan negatif. Persamaan bidang pemisah linier SVM dirumuskan pada persamaan:

JURNAL INFORMATIKA TEKNOLOGI DAN SAINS (JINTEKS) Hal. 1045-1054
E-ISSN : 2686-3359 Vol. 8; No. 2
DOI PREFIX : 10.51401 Mei 2026

$$wT \cdot x + b = 0 \quad (3)$$

Keterangan:
a. w : Vektor bobot (*weight vector*) yang menentukan arah dan orientasi bidang pemisah (*hyperplane*).
b. w^T : Transpos dari vektor bobot w.
c. x : Vektor fitur masukan dokumen ulasan (hasil pembobotan numerik TF-IDF).
d. b : Nilai bias (*bias term*) yang menentukan pergeseran letak bidang pemisah terhadap titik pusat (*origin*).
e. 0 : Nilai ambang batas keputusan (*decision boundary*) untuk memisahkan kelas positif dan negatif.

Random Forest mengombinasikan prediksi dari kumpulan pohon keputusan (*ensemble of decision trees*) yang dibangun secara independen dari sampel *bootstrap* acak. Pemisahan simpul (*node splitting*) pada setiap pohon keputusan dihitung berdasarkan tingkat kemurnian *Gini Impurity*, yang dirumuskan pada persamaan:

$$Gini(D) = 1 - \sum_{i=1}^{m} p_i^2 \quad (4)$$

Keterangan:
a. *Gini (D)* : Nilai *Gini Index* untuk himpunan data D, yang mengukur tingkat ketidakmurnian (*impurity*) cabang data.
b. m : Jumlah total kategori atau kelas target yang ada pada dataset (dalam penelitian ini $m = 2$, yaitu kelas positif dan negatif).
c. p_i : Proporsi atau probabilitas kemunculan sampel data yang tergolong dalam kelas ke-i terhadap keseluruhan himpunan data D.

7. Evaluasi Performa Model

Pengujian kinerja model diukur mengacu pada pemetaan *Confusion Matrix* yang menghasilkan indikator objektif berupa Akurasi, Presisi, *Recall*, dan *F1-Score* serta *AUC Curve*. Metrik evaluasi kinerja dihitung berdasarkan persamaan :

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (5)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (6)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (7)$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (8)$$

Keterangan Variabel:
a. TP (*True Positive*): Jumlah data positif yang diprediksi dengan bagai kelas positif.
b. TN (*True Negative*): Jumlah data negatif yang diprediksi deng sebagai kelas negatif.
c. FP (*False Positive*): Jumlah data negatif yang salah diprediksi sebagai kelas positif.
d. FN (*False Negative*): Jumlah data positif yang salah diprediksi sebagai kelas negatif.

Keterangan Metrik Evaluasi:
a. Accuracy (Akurasi) : Mengukur seberapa besar persentase total prediksi yang benar (positif dan negatif) dari seluruh data yang dievaluasi.
b. Precision (Presisi) : Mengukur tingkat ketepatan model, yaitu sejauh mana data yang diprediksi positif benar-benar bernilai positif.
c. Recall (Sensitivitas) : Mengukur kemampuan model dalam menemukan kembali seluruh data yang sebenarnya bernilai positif.
d. F1-Score : Rata-rata harmonis antara *Precision* dan *Recall*, yang sangat berguna untuk mengukur performa model saat menghadapi data dengan kelas tidak seimbang (*imbalance*).

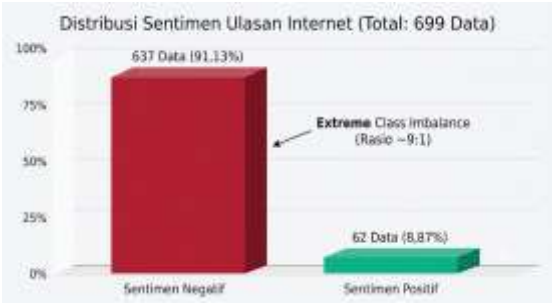
HASIL DAN PEMBAHASAN

1. Deskripsi Dataset

Dataset yang diuji pada penelitian ini bersumber dari berkas ulasan pelanggan provider internet sebanyak 699 data. Struktur data terbagi dalam dua variabel utama, yaitu *review* yang memuat teks tanggapan atau keluhan pengguna, serta

JURNAL INFORMATIKA TEKNOLOGI DAN SAINS (JINTEKS)
E-ISSN : 2686-3359
DOI PREFIX : 10.51401
Hal. 1045-1054
Vol. 8; No. 2
Mei 2026

sentiment yang bertindak sebagai label target. Teks ulasan didominasi oleh kosakata bahasa Indonesia yang sarat akan variasi ungkapan sehari-hari, hingga istilah teknis seputar gangguan modem, kelambatan jaringan, dan respons *customer service*. Dibawah ini terdapat grafik dari distribusi dataset:



Gambar 2. Distribusi Dataset

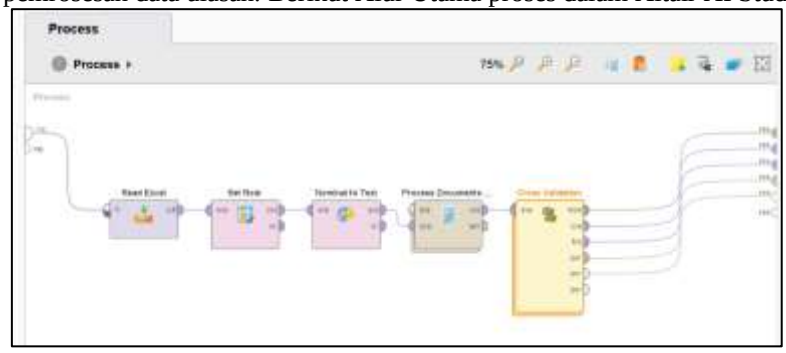
Secara kuantitatif, distribusi kelas menunjukkan ketimpangan yang sangat tajam, di mana sentimen negatif mendominasi sebanyak 637 data (91,13%), sementara sentimen positif hanya sebesar 62 data (8,87%). Rasio ketidakseimbangan yang menyentuh angka 9:1 ini menjadi tantangan teknis utama dalam pemodelan *machine learning*, sebab model rawan mengalami bias prediksi dan cenderung mengabaikan pola pada kelas minoritas. Berikut adalah contoh review pengguna dari dataset:

Tabel 2. Contoh Data Review

No	Review	Sentimen
1	tiap bulan pasti ada aja koneksinya jelek	Negatif
2	layanan internetnya bagus, kalau ada kendala di jaringan misalnya tidak lancar maka langsung ada pemberitahuan lewat email ,bgitu juga kalau sdh pulih ada pemberitahuan.., pakai my rep,OK.	Positif
3	setiap bulan ada aja gangguannya,gak ada internet lah ,gangguan dllnya,pokoknya ada aja gangguannya,bayar juga udh	Negatif
...	…	…
699	sangat jelek dalam penanganan gangguan dan sangat lambat sekali responnya	Negatif

2. Alur Utama Proses (*Main Process Pipeline)*

Alur eksperimen pada *Main Process* Altair AI Studio mengintegrasikan rangkaian operator terstruktur untuk menjamin transparansi pemrosesan data ulasan. Berikut Alur Utama proses dalam Altair AI Studio:



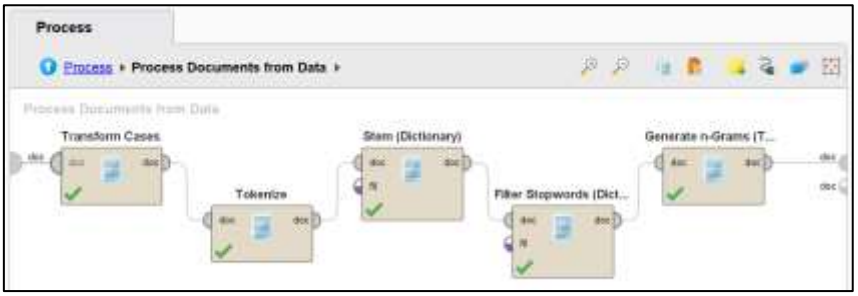
Gambar 3. Rangkaian *Main Process Pipeline*

Pada rangkaian *Main Process* Pipeline tahap awal diimplementasikan melalui operator Read Excel yang berfungsi untuk mengimpor berkas data mentah ke dalam lembar kerja sistem. Penataan struktur atribut kemudian dilakukan lewat operator *Set Role*, yang mengonfigurasi kolom sentiment sebagai variabel dependen (*label*) serta kolom review sebagai variabel independen. Untuk memastikan kompatibilitas format pada tahapan *text mining*, operator *Nominal to Text* melakukan transformasi tipe data ulasan menjadi *text*. Teks ulasan ini selanjutnya dialirkan menuju sub-proses *Process Documents from Data* untuk menjalani serangkaian pembersihan kata dan ekstraksi fitur berbobot. Pemrosesan pada alur utama diakhiri oleh operator *Cross Validation* yang membagi data ke dalam lipatan (*folds*) validasi guna menguji konsistensi kinerja model klasifikasi.

3. *Preprocessing Data*

Tahap pra-pemrosesan teks pada penelitian ini dirancang secara berurutan menggunakan lima operator dalam aplikasi Altair AI Studio (*Process Documents from Data*) untuk membersihkan dan menstandarkan data ulasan mentah. Berikut adalah gambaran tahap pra-pemrosesan data pada aplikasi:

JURNAL INFORMATIKA TEKNOLOGI DAN SAINS (JINTEKS)
E-ISSN : 2686-3359
DOI PREFIX : 10.51401
Hal. 1045-1054
Vol. 8; No. 2
Mei 2026



Gambar 4. Rangkaian *Preprocessing Data*

Proses diawali dengan penerapan operator *Transform Cases* untuk merata-kirikan format kapitalisasi menjadi huruf kecil seluruhnya (*case folding*), sehingga perbedaan penulisan huruf besar dan kecil tidak dianggap sebagai kata yang berbeda. Selanjutnya, operator Tokenize bekerja memotong rentetan kalimat menjadi unit kata terpisah (*token*). Kata-kata hasil pemotongan ini kemudian dilanjutkan ke tahap Stem (*Dictionary*), yang menggunakan referensi kamus untuk mengembalikan kata-kata berimbuhan menjadi bentuk kata dasarnya. Setelah kata dasar terbentuk, operator *Filter Stopwords* (*Dictionary*) diterapkan untuk menyaring dan membuang kata-kata umum yang kurang memiliki bobot informasi dalam analisis sentimen, seperti kata hubung atau kata depan. Rangkaian pemrosesan ini diakhiri dengan operator *Generate n-Grams (Terms)* guna membentuk susunan frasa majemuk berurutan, sehingga konteks dan makna gabungan kata dalam ulasan pelanggan tetap terjaga dengan baik sebelum diproses ke tahap pembobotan fitur.

4. *Data Balancing and Modeling Sub-process* (Sub-proses Penyeimbangan Data dan Pemodelan)

Di dalam sub-proses *Cross Validation*, alur kerja terbagi secara independen menjadi dua blok utama, yaitu blok pelatihan (*Training*) dan blok pengujian (*Testing*). Berikut adalah rangkaian *Data Balancing and Modeling Sub-process* dari kedua algoritma yang digunakan pada penelitian*:*

Gambar 5. Rangkaian *Data Balancing and Modeling Sub-process*

Pada blok pelatihan, data latih pertama-tama diproses oleh operator *SMOTE Upsampling* untuk menyeimbangkan proporsi kelas minoritas melalui pembentukan sampel sintetis. Data latih yang telah seimbang ini kemudian digunakan oleh operator *SVM* atau *Random Forest* untuk mempelajari pola dan membangun bidang pemisah (*hyperplane*). Sementara itu, pada blok pengujian, kedua model algoritma yang terbentuk diaplikasikan pada data uji murni menggunakan operator *Apply Model*. Hasil prediksi dari data uji tersebut selanjutnya dievaluasi secara akurat oleh operator *Performance* untuk menghasilkan indikator kinerja berupa nilai akurasi, presisi, *recall*, hingga *F1-score*.

5. Hasil Evaluasi Performa Model

Pada tahapan ini, performa algoritma *Support Vector Machine* (SVM) dan *Random Forest* dievaluasi berdasarkan pengujian *Cross Validation* menggunakan dataset ulasan pelanggan. Indikator kinerja yang diukur meliputi *Accuracy*, *Class Precision*, *Class Recall*, dan *Area Under Curve* (AUC). Berikut adalah hasil perbandingan akurasi dari kedua algoritma:

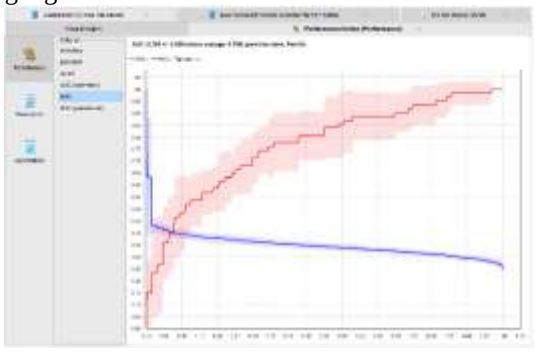
Gambar 6. Hasil Akurasi SVM

JURNAL INFORMATIKA TEKNOLOGI DAN SAINS (JINTEKS)
E-ISSN : 2686-3359
DOI PREFIX : 10.51401
Hal. 1045-1054
Vol. 8; No. 2
Mei 2026

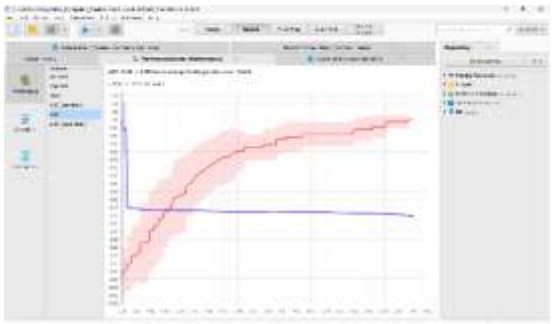
Gambar 7. Hasil Akurasi Random Forest

Berdasarkan matriks konfusi (*confusion matrix*) yang diperoleh, dataset memiliki ketimpangan kelas (*class imbalance*) yang sangat dominan, di mana mayoritas ulasan berada pada kelas sentimen negatif. Hasil pengujian menunjukkan bahwa algoritma SVM menghasilkan performa keseluruhan yang sedikit lebih unggul dibandingkan *Random Forest*. Akurasi yang dicapai oleh SVM berada pada angka 91,56%, sedangkan *Random Forest* mencatatkan akurasi sebesar 91,13%.

Pada metrik AUC, SVM juga menunjukkan kemampuan pemisahan kelas yang lebih baik dengan nilai 0,769 (kategori *Fair Classification*), dibanding *Random Forest* yang memperoleh nilai AUC sebesar 0,740. Berikut adalah hasil dari grafik AUC dari masing-masing algoritma:



Gambar 8. AUC SVM



Gambar 9. AUC Random Forest

Hasil pengujian menunjukkan bahwa penggabungan SMOTE + SVM menghasilkan performa keseluruhan yang lebih baik dibandingkan SMOTE + Random Forest. Kombinasi SMOTE + SVM mencatatkan nilai akurasi sebesar 91,56% dengan AUC sebesar 0,769 (kategori *Fair Classification*). Sementara itu, kombinasi SMOTE + Random Forest memperoleh akurasi sebesar 91,13% dengan nilai AUC sebesar 0,740. Meskipun teknik SMOTE telah diterapkan pada data latih (*training data*) untuk membalanskan proporsi kelas, analisis terhadap matriks konfusi mengungkapkan fakta bahwa model masih mengalami kesulitan dalam mengenali data uji pada kelas positif.

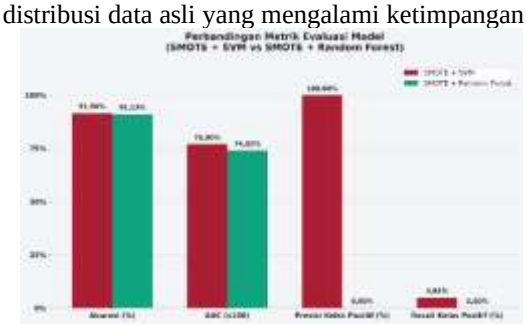
Pengujian terhadap kedua model menunjukkan bahwa penerapan teknik penyeimbangan SMOTE pada data latih memberikan efektivitas yang berbeda secara signifikan antara algoritma *Support Vector Machine* (SVM) dan *Random Forest* dalam mengenali data uji pada kelas minoritas. Pada skenario SMOTE + SVM, integrasi ini berhasil membantu model memprediksi 3 sampel positif secara tepat dari total 62 data uji aktual positif dengan tingkat presisi kelas positif mencapai 100,00%. Kendati demikian, tingkat *recall* kelas positif pada SVM masih tergolong rendah yaitu 4,84%, mengingat mayoritas sampel positif (59 data) masih teridentifikasi sebagai kelas negatif. Di sisi lain, skenario SMOTE + Random Forest menunjukkan bahwa penambahan sampel sintetis pada data latih belum mampu mengatasi kecenderungan model terhadap kelas mayoritas (*majority class bias*). Hal ini terbukti dari seluruh data uji yang diprediksi sebagai kelas negatif oleh *Random Forest*, sehingga menghasilkan nilai *recall* maupun presisi sebesar 0,00% untuk kelas positif.

JURNAL INFORMATIKA TEKNOLOGI DAN SAINS (JINTEKS)
E-ISSN : 2686-3359
DOI PREFIX : 10.51401
Hal. 1045-1054
Vol. 8; No. 2
Mei 2026

Tabel 3. Hasil Evaluasi Performa Algoritma SVM dan Random Forest dengan SMOTE

Metrik Evaluasi	SMOTE + Support Vector Machine	SMOTE + Random Forest
Akurasi (*Accuracy*)	91,56%	91,13%
Akurasi Area (*AUC*)	0,769	0,740
Presisi Kelas Negatif	91,52%	91,13%
Presisi Kelas Positif	100,00%	0,00%
Recall Kelas Negatif	100,00%	100,00%
Recall Kelas Positif	4,84%	0,00%

Data pada tabel tersebut mengindikasikan bahwa meskipun penyeimbangan sampel melalui teknik SMOTE telah diterapkan pada tahap pelatihan, batas pemisah (*decision boundary*) yang terbentuk pada fitur teks ulasan tetap terpengaruh secara signifikan oleh distribusi data asli yang mengalami ketimpangan ekstrem (*extreme imbalance*).



Gambar 10. Perbandingan Metrik Evaluasi Model

Berdasarkan grafik hasil evaluasi performa model yang disajikan, terlihat perbandingan capaian antara skema SMOTE + Support Vector Machine (SVM) dan SMOTE + Random Forest secara mendalam. Kedua model memang mencatatkan tingkat akurasi yang tergolong tinggi dan hampir setara, yaitu sebesar 91,56% untuk SVM dan 91,13% untuk Random Forest. Begitu pula pada indikator Area Under Curve (AUC), SVM sedikit lebih unggul dengan nilai 76,90% (0,769) dibanding Random Forest yang meraih 74,00% (0,740). Namun, perbedaan mendasar antara kedua pendekatan baru terlihat jelas saat meninjau metrik evaluasi kelas minoritas (positif). Kombinasi SMOTE + SVM terbukti mampu memetakan batas pemisah (hyperplane) dengan tingkat Presisi Kelas Positif mencapai 100,00% walau nilai Recall-nya masih terbatas pada angka 4,84%. Sebaliknya, skema SMOTE + Random Forest mengalami kegagalan total (*complete bias*) dalam mengidentifikasi sampel positif di data uji, yang ditunjukkan oleh nilai Presisi maupun Recall Kelas Positif sebesar 0,00%.

Kekuatan utama dari penelitian ini terletak pada ketelitian skenario pemodelan yang menguji efektivitas interaksi antara penyeimbangan sampel sintetis (SMOTE) dan struktur ruang fitur berdimensi tinggi (*high-dimensional text features*) hasil pembobotan TF-IDF. Penelitian ini secara obyektif membuktikan bahwa penambahan sampel sintetis pada data latih tidak serta-merta menjamin algoritma *ensemble tree* seperti Random Forest dapat mengenali kelas minoritas pada data uji yang mengalami ketimpangan ekstrem (*extreme class imbalance* dengan rasio 9:1). Di sisi lain, *novelty* atau kebaruan dari studi ini terletak pada pengungkapan karakteristik kerja SVM yang terbukti jauh lebih adaptif dan responsif dalam memanfaatkan sampel buatan SMOTE pada data teks ulasan penyedia jasa internet Bahasa Indonesia. Melalui temuan ini, penelitian memberikan kontribusi ilmiah penting bahwa untuk domain analisis sentimen teks yang timpang ekstrem, pendekatan pemisah bidang linier (SVM) memiliki ketahanan batas keputusan yang jauh lebih handal ketimbang mekanisme pemisahan berbasis pohon (*decision trees*) yang cenderung didominasi oleh kelas mayoritas.

KESIMPULAN DAN SARAN

Penelitian ini menunjukkan bahwa penerapan *oversampling* SMOTE yang dikombinasikan dengan Support Vector Machine (SVM) dan Random Forest menghasilkan karakteristik performa yang berbeda dalam mengklasifikasikan data sentimen berketimpangan kelas (91,13% negatif : 8,87% positif). Model SVM berhasil mencatatkan tingkat akurasi sebesar 91,56%, nilai AUC 0,769, serta presisi sempurna sebesar 100,00% pada kelas positif. Sementara itu, model Random Forest memperoleh akurasi sebesar 91,13% dan nilai AUC 0,723. Hasil pengujian secara terperinci mencatat tingkat *recall* kelas positif sebesar 4,84% untuk SVM dan 0,00% untuk Random Forest, yang mengonfirmasi bahwa kedua model memiliki kecenderungan prediktif yang jauh lebih dominan dalam mengenali sampel pada kelas mayoritas (negatif). Temuan ini memberikan kontribusi ilmiah penting yang membuktikan secara empiris bahwa pemisahan ruang vektor berbasis hyperplane pada SVM jauh lebih responsif memanfaatkan sampel sintetis SMOTE dalam ruang fitur teks berdimensi tinggi dibandingkan pendekatan pemisahan simpul decision tree pada Random Forest yang rentan mengalami keterpukulan bias total terhadap kelas mayoritas pada kondisi ketimpangan ekstrem.

Berdasarkan keterbatasan dan temuan dalam studi ini, peneliti selanjutnya disarankan untuk mengeksplorasi teknik penanganan ketimpangan data alternatif selain SMOTE baku, seperti ADASYN atau pendekatan *undersampling* (misalnya *Tomek Links*), guna mengoptimalkan nilai *Recall* kelas positif yang pada model SVM masih relatif rendah. Selain itu, pengembangan penelitian ke depan dapat mempertimbangkan penggunaan teknik ekstraksi fitur berbasis

pembagian kata berkonteks (*word embeddings*) seperti IndoBERT, serta memperluas pengujian pada arsitektur algoritma *deep learning* untuk mengatasi kompleksitas struktur bahasa non-formal pada teks ulasan.

DAFTAR PUSTAKA

[1] KOMDIGI, "Infrastruktur Digital Fondasi Akselerasi Inovasi dan Pertumbuhan Ekonomi."
[2] G. P. Nurzhavira, S. Iriani, and J. Manajemen, "PENGARUH KUALITAS LAYANAN DAN KEPUASAN PELANGGAN TERHADAP LOYALITAS PELANGGAN INDIHOME," *Jurnal Ilmiah Mahasiswa Akuntansi) Universitas Pendidikan Ganesha*, vol. 13, no. 2, pp. 692–704, 2022.
[3] M. A. N. Dzaki, H. Hindarto, A. Eviyanti, and N. L. Azizah, "Analisis Sentimen Layanan Pelanggan Provider Internet dengan Algoritma Support Vector Machine dan Naïve Bayes," *SemanTIK : Jurnal Teknik Informasi*, vol. 10, pp. 84–93, 2025.
[4] D. F. Sari, D. Kurniawati, E. Wahyuningsih, and T. R. Wibowo, "Analisis Sentimen Keluhan Pelanggan ISP menggunakan Support Vector Machine (SVM) dan TF-IDF," *Bulletin of Computer Science Research*, vol. 5, no. 6, pp. 1387–1394, 2025, doi: 10.47065/bulletincsr.v5i6.822.
[5] P. Y. Ng and J. K. M. Sia, "Managers' perspectives on restaurant food waste separation intention: The roles of institutional pressures and internal forces," *Int. J. Hosp. Manag.*, vol. 108, Jan. 2023, doi: 10.1016/j.ijhm.2022.103362.
[6] Y. Guo *et al.*, "Bias in Large Language Models: Origin, Evaluation, and Mitigation," *Electronics (Switzerland)*, vol. 15, no. 9, May 2026, doi: 10.3390/electronics15091824.
[7] F. Dwi Astuti and F. Nova Lenti, "Implementasi SMOTE untuk mengatasi Imbalance Class pada Klasifikasi Car Evolution menggunakan K-NN," *Jurnal JUPITER*, vol. 13, pp. 89–98, 2021.
[8] Ginabila and A. Fauzi, "Analisis Sentimen Terhadap Pemutar Musik Online Spotify Dengan Algoritma Naive Bayes dan Support Vector Machine," *Jurnal Ilmiah ILKOMINFO - Jurnal Ilmu Komputer dan Informatika*, vol. 6, pp. 111–122, 2023.
[9] R. Lusiana Pratiwi, Z. Imami Alfianti, and A. Fauzi, "Penerapan Algoritma Naive Bayes dan SVM untuk Analisis Sentimen Terhadap Penggunaan True Wireless Stereo (TWS)," *SKANIKA: Sistem Komputer dan Teknik Informatika*, vol. 8, no. 2, pp. 257–268, 2025.
[10] S. Emilita, P. S. Rahayu, S. R. Larasati, G. Taufiq, and J. T. Kumalasari, "Penerapan Algoritma Random Forest Dalam Memprediksi Harga Properti di Jakarta Selatan Berdasarkan Karakteristik Fisik dan Lokasi," *Jurnal Minfo Polgan*, vol. 14, no. 2, pp. 3415–3421, Jan. 2026, doi: 10.33395/jmp.v14i2.15735.
[11] R. T. Nugraha, D. Hidayat, and F. Mahardika, "Analisis Perbandingan Algoritma Random Forest, SVM, dan Neural Network untuk Klasifikasi Risiko Keamanan E-Learning," *Jurnal Teknologi dan Informasi*, vol. 16, no. 1, pp. 1–10, Mar. 2026, doi: 10.34010/jati.v16i1.17191.
[12] M. H. Mustaqim and A. B. Santoso, "Penerapan Penyeimbangan Data Pada Analisis Sentimen Ulasan Game Magic Chess Go Go di Play Store dengan Naive Bayes," *Building of Informatics, Technology and Science (BITS)*, vol. 7, no. 2, pp. 1078–1089, Sep. 2025, doi: 10.47065/bits.v7i2.7845.
[13] N. A. Kasanah, Muliadi, and U. Pujianto, "Penerapan Teknik SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Objektivitas Berita Online Menggunakan Algoritma KNN," *Jurnal RESTI: Rekayasa Sistem dan Teknologi Informasi*, vol. 1, pp. 196–201, 2021.
[14] D. Andriyani, A. Faqih, and S. E. Permana, "The Effect of SMOTE Application on Support Vector Machine Performance in Sentiment Classification on Imbalanced Datasets," *Journal of Artificial Intelligence and Engineering Applications*, vol. 4, no. 2, pp. 2808–4519, 2025, [Online]. Available: https://ioinformatic.org/
[15] I. Muliansyah, N. Zulkarnaim, A. Amirul, and A. Cirua, "Perbandingan Algoritma Support Vector Machine dan Random Forest untuk Analisis Sentimen Terkait Danantara," *Jurnal Sistem Komputer dan Kecerdasan Buatan*, vol. 9, pp. 263–270, 2026.
[16] C. Kaope and Y. Pristyanto, "The Effect of Class Imbalance Handling on Datasets Toward Classification Algorithm Performance," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 22, no. 2, pp. 227–238, Mar. 2023, doi: 10.30812/matrik.v22i2.2515.
[17] Y. A. P. G. S, G. L. Ginting, and M. Panjaitan, "Optimasi Penerimaan Siswa Baru dengan Penerapan Algortima Text Mining dan TF-IDF," *Journal of Computing and Informatics Research*, vol. 2, no. 3, pp. 110–117, Jul. 2023, doi: 10.47065/comforch.v2i3.941.
[18] I. S. K. Idris, Y. A. Mustofa, and I. A. Salihi, "Analisis Sentimen terhadap Penggunaan Aplikasi Shopee Menggunakan Algoritma Support Vector Machine (SVM)," *Jambura Journal of Electrical and Electronics Engineering*, vol. 5, pp. 32–35, Jan. 2023, doi: 10.1177/0165551510388123.
[19] A. A. A. Daniswara and I. K. D. Nuryana, "Data Preprocessing Pola Pada Penilaian Mahasiswa Program Profesi Guru," *Journal of Informatics and Computer Science*, vol. 05, 2023.
[20] F. Handayani and R. M. Taufiq, "Komparasi Algoritma Menggunakan Teknik Smote Dalam Melakukan Klasifikasi Penyakit Stroke Otak," *Jurnal CoSciTech (Computer Science and Information Technology)*, vol. 5, no. 2, pp. 367–372, Aug. 2024, doi: 10.37859/coscitech.v5i2.7439.
[21] M. F. R. Aditya, N. L. Azizah, and U. Indahyati, "Prediksi Penyakit Hipertensi Menggunakan Metode Decison Tree dan Random Forest," *Jurnal Ilmiah Komputasi*, vol. 23, no. 1, Mar. 2024, doi: 10.32409/jikstik.23.1.3503.
[22] N. Hadi and D. Sugiarto, "Analisis Sentimen Pembangunan IKN pada Media Sosial X Menggunakan Algoritma SVM, Logistic Regression dan Naïve Bayes," *Jurnal Informatika: Jurnal Pengembangan IT*, vol. 10, no. 1, pp. 37–49, Jan. 2025, doi: 10.30591/jpit.v10i1.7106.
[23] G. Rininda, I. H. Santi, and S. Kirom, "PENERAPAN SVM DALAM ANALISIS SENTIMEN PADA EDLINK MENGGUNAKAN PENGUJIAN CONFUSION MATRIX," *Jurnal Mahasiswa Teknik Informatika*, vol. 7, no. 5, 2023.
[24] M. Iqbal, S. Miskiyah, S. L. Sham, S. Anwar, and M. H. Fuad, "PERBANDINGAN METODE DECISION TREE DAN NAIVE BAYES PADA TINGKAT PENJUALAN MINUMAN KOPI DI KOPI PAWON NUSANTARA," *Jurnal INSAN (Journal of Information Systems Management Innovation*, vol. 4, no. 1, pp. 27–34, 2024.