JUSIFOR: Jurnal Sistem Informasi dan Informatika
Vol. 5, No. 1, Juni 2026, hlmn. 287-294
E-ISSN: 2830-3393 | P-ISSN: 2830-2443
DOI: https://doi.org/10.70609/jusifor.v5i1.9714

Evaluasi Strategi Penanganan Ketidakseimbangan Kelas pada Klasifikasi Sentimen Multikelas Ulasan Aplikasi ChatGPT

Qorry Aina Fitroh[1*], M. Sa'id Abdurrohman Kunta Mardlian[2], Indra Kurniawan[3], Muhammad Kildah Namariq[4]

[1] Teknik Industri, Fakultas Sains dan Teknologi, UIN K.H. Abdurrahman Wahid Pekalongan, Indonesia
[2] Perekam dan Informasi Kesehatan, STIKES Arrahma Mandiri Indonesia, Indonesia
[3] Bisnis Digital, Fakultas Ekonomi dan Bisnis Islam, UIN K.H. Abdurrahman Wahid Pekalongan, Indonesia
[4] Informatika, Fakultas Sains dan Teknologi, UIN K.H. Abdurrahman Wahid Pekalongan, Indonesia

Info Artikel

Riwayat Artikel:
Diterima : **28 April 2026**
Direvisi : **22 Mei 2026**
Diterbitkan : **05 Juni 2026**

Kata Kunci:

Analisis Sentimen,
Ketidakseimbangan Kelas,
SMOTE,
Pembobotan Kelas,
Machine Learning

ABSTRAK

Ketidakseimbangan kelas pada dataset ulasan aplikasi berpotensi menimbulkan bias sistematis pada model klasifikasi sentimen berbasis *machine learning*. Penelitian ini mengevaluasi empat strategi penanganan ketidakseimbangan data, yaitu *baseline*, SMOTE, *class weighting*, dan kombinasi keduanya pada tiga algoritma klasifikasi, yakni Support Vector Machine (SVM), Random Forest, dan Logistic Regression menggunakan 10.000 ulasan dari aplikasi ChatGPT di Google Play Store. Representasi fitur dilakukan dengan TF-IDF dan dievaluasi menggunakan *5-fold cross-validation*. Skenario *baseline* mencapai akurasi tertinggi sebesar 0,90, namun memiliki F1-score makro terendah berkisar antara 0,52 hingga 0,55 yang mengindikasikan dominasi kelas mayoritas. Penerapan *class weighting* secara konsisten menghasilkan F1-score makro tertinggi sebesar 0,58 untuk SVM dan *Logistic Regression*, sementara SMOTE tidak menunjukkan peningkatan yang konsisten. Hasil ini menegaskan adanya trade-off antara akurasi dan keseimbangan kinerja antarkelas, serta efektivitas strategi penanganan ketidakseimbangan bersifat bergantung pada algoritma yang digunakan.

***Keywords*:**

Sentiment Analysis,
Class Imbalance,
SMOTE,
Class Weighting,
Machine Learning

ABSTRACT

Evaluation of Class Imbalance Handling Strategies in Multiclass Sentiment Classification of ChatGPT Application Reviews. Class imbalance in application reviews datasets has the potential to cause systematic bias in machine learning-based sentiment classification models. This study evaluates four imbalance handling strategies, namely baseline, SMOTE, class weighting, and their combination, across three algorithms, Support Vector Machine, Random Forest, and Logistic Regression, using 10,000 ChatGPT reviews from Google Play Store. Feature representation is conducted using TF-IDF with 5-fold cross-validation. The baseline scenario achieves the highest accuracy of 0.90 but produces the lowest macro F1-score ranging from 0.52 to 0.55, indicating dominance of the majority class. Class weighting consistently yields the highest macro F1-score of 0.58 for SVM and Logistic Regression, while SMOTE shows inconsistent improvements. These findings affirm a trade-off between accuracy and balanced class performance, with the effectiveness of imbalance-handling strategies being demonstrably algorithm-dependent.

Penulis Korespondensi:
Qorry Aina Fitroh,
Program Studi Teknik Industri,
Universitas Islam Negeri K.H. Abdurrahman Wahid Pekalongan
Email: qorry.aina.fitroh@uingusdur.ac.id

This is an open access article under the CC BY-SA license

1. PENDAHULUAN

Teknologi yang terdapat dalam kecerdasan buatan generatif saat ini berkembang sangat pesat dengan ChatGPT yang dikembangkan oleh OpenAI sebagai salah satu representasi utamanya [1]. Sebagai pionir dalam penggunaan AI untuk publik berskala luas, aplikasi ini menunjukkan tingkat keterlibatan pengguna yang luar biasa dengan pencapaian rekor 100 juta pengguna aktif hanya dalam kurun waktu dua bulan [2]. Pada platform Google Play Store, aplikasi ChatGPT berhasil menduduki posisi sebagai salah satu dari tiga aplikasi utama dalam kategori produktivitas di awal tahun 2026. Dominasi penggunaan teknologi ini terlihat melalui adopsi skala besar, mulai dari puluhan juta ulasan hingga ratusan juta unduhan di berbagai wilayah global. Akumulasi ulasan pengguna yang sangat masif ini merupakan sumber data empiris yang kaya untuk mengevaluasi pengalaman serta persepsi publik secara mendalam yang mencakup kualitas respons, kendala teknis, dan tingkat kepuasan pengguna secara riil [3].

Pemahaman yang lebih mendalam tentang respons pengguna dan penilaian terhadap fitur serta pengalaman penggunaan aplikasi seperti ChatGPT, memerlukan pendekatan sistematis berbasis analisis opini. Analisis sentimen menjadi salah satu metode yang diterapkan untuk mengidentifikasi dan mengelompokkan opini dalam teks yang berasal dari beragam media sosial melalui pendekatan pembelajaran mesin [4]. Pendekatan ini efektif dalam mengkaji pola respons pengguna terhadap kebijakan dan pembaruan fitur aplikasi [5]. Selain itu, volume ulasan pengguna yang tinggi merupakan sumber informasi yang krusial karena mengandung data berbasis pendapat yang dapat digunakan sebagai bahan evaluasi dan dasar untuk mengoptimalkan serta meningkatkan kualitas fitur aplikasi.

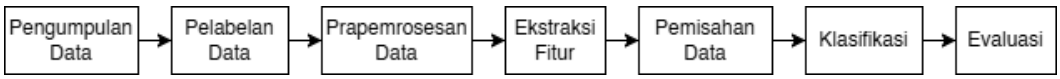
Sejumlah penelitian terdahulu telah membuktikan efektivitas berbagai algoritma *machine learning* dalam melakukan klasifikasi sentimen otomatis. Implementasi algoritma Naïve Bayes pada ulasan ChatGPT menghasilkan akurasi sebesar 94% [6], sedangkan Support Vector Machine (SVM) pada 300 ulasan ChatGPT menghasilkan akurasi 88,74% [7]. Pada konteks yang berbeda, penerapan SVM pada 5.000 ulasan Claude menghasilkan akurasi 99,8% [8]. Komparasi algoritma SVM, Random Forest, dan Naïve Bayes turut menunjukkan kapabilitas yang unggul dalam mengolah umpan balik pengguna dengan pencapaian f1-score tertinggi 90% pada SVM dan Random Forest [9]. Di sisi lain, implementasi algoritma SVM, Decision Tree, dan Logistic Regresion dalam klasifikasi sentimen ulasan aplikasi digital mencatatkan performa yang kompetitif dengan rata-rata akurasi SVM 88,18% dan disusul oleh Logistic Regresion sebesar 87,74% [10]. Keberhasilan berbagai metode komputasional ini dalam mengekstraksi wawasan dari ratusan hingga ribuan ulasan membuktikan bahwa pendekatan *machine learning* merupakan instrumen yang andal untuk mengevaluasi persepsi pengguna dalam skala masif.

Implementasi algoritma *machine learning* menunjukkan hasil yang menjanjikan, meskipun terdapat tantangan mendasar dalam data realitas yaitu fenomena ketidakseimbangan kelas yang terjadi pada *dataset* ulasan [11]. Permasalahan ini muncul ketika salah satu kategori sentimen mendominasi distribusi data secara signifikan yang dapat memicu bias pada model komputasional [2], [12]. Ketidakseimbangan distribusi ini menjadi isu krusial karena algoritma cenderung menggeneralisasi prediksi berdasarkan kelas mayoritas, sehingga sering kali gagal mengenali karakteristik unik dari kelas minoritas [11]. Kondisi tersebut berpotensi menurunkan kualitas evaluasi model karena nilai akurasi yang tinggi tidak sepenuhnya mencerminkan performa model secara menyeluruh, terutama dalam mengidentifikasi dan mengklasifikasikan opini dengan jumlah data yang tidak seimbang [13]. Oleh karena itu, diperlukan strategi penanganan khusus melalui metode penyeimbangan data, seperti metode pembobotan kelas atau *Synthetic Minority Oversampling Technique* (SMOTE), untuk memastikan model memiliki kemampuan generalisasi yang baik [9], [11], [12]. Integrasi teknik-teknik tersebut berkontribusi pada peningkatan kualitas klasifikasi, sehingga model mampu memetakan persepsi publik secara lebih objektif dan reliabel [14], [15].

Penelitian ini bertujuan untuk memberikan kontribusi melalui evaluasi komprehensif terhadap efektivitas strategi penanganan ketidakseimbangan kelas pada klasifikasi sentimen multikelas ulasan aplikasi ChatGPT. Proses evaluasi dilakukan menggunakan tiga algoritma *machine learning*, yaitu SVM, *Random Forest,* dan *Logistic Regression* untuk menguji konsistensi kinerja metode *Synthetic Minority Oversampling Technique* (SMOTE) dan pendekatan pembobotan kelas sebagai upaya meminimalkan bias pada model klasifikasi. Data yang digunakan dalam penelitian ini berjumlah 10.000 ulasan pengguna aplikasi ChatGPT yang dikumpulkan dari platform Google Play Store dengan karakteristik utama berupa distribusi kelas yang tidak seimbang.

2. METODE PENELITIAN

Penelitian dilaksanakan melalui beberapa tahapan, meliputi pengumpulan data, pelabelan, prapemrosesan, ekstraksi fitur, pemisahan data, klasifikasi, dan evaluasi sebagaimana diilustrasikan pada Gambar 1.



Gambar 1. Tahapan Penelitian

2.1. Pengumpulan Data

Penelitian ini mengumpulkan data ulasan pengguna ChatGPT dari Google Play Store secara otomatis melalui metode *scraping* yang dijalankan pada infrastruktur Google Colab. Hasil dari proses ini berupa *file* berformat .CSV untuk selanjutnya diolah menggunakan pustaka Python.

2.2. Pelabelan Data

Data yang diperoleh selanjutnya dilabeli berdasarkan skor *rating* yang diberikan pengguna pada Google Play Store. *Rating* 4 sampai 5 dikategorikan sebagai sentimen positif, *rating* 3 dikategorikan sebagai netral, dan *rating* 1-2 dikategorikan sebagai negatif. Pelabelan ini bertujuan menghasilkan variabel target sebagai acuan dalam proses klasifikasi sentimen.

2.3. Prapemrosesan Data

Tahapan Prapemrosesan data dilakukan untuk mempersiapkan data sebelum digunakan oleh model *machine learning*. Prameprosesan mencakup beberapa tahapan. Pertama, melakukan proses pembersihan data dengan cara menghapus karakter khusus dalam data seperti emoji, angka, simbol, dan elemen lain yang tidak relevan. Dilanjutkan dengan proses *case folding* untuk mengonversi seluruh ulasan menjadi *lowercase* agar analsis data dapat dilakukan dengan konsisten. Langkah ketiga, normalisasi bertujuan untuk menyeragamkan bentuk kata dengan cara penghapusan pengulangan karakter berlebihan, sebagai contoh 'bagusss' menjadi 'bagus'. Tahap berikutnya adalah *stopword removal* untuk mengeliminasi kata-kata umum yang tidak memiliki makna signifikan, dilanjutkan dengan *stemming* menggunakan pustaka Sastrawi untuk mengonversi kata menjadi bentuk dasarnya. Terakhir dilakukan penghapusan baris kosong. Seluruh tahapan telah dilakukan sehingga data siap masuk ke proses berikutnya.

2.4. Ekstraksi Fitur

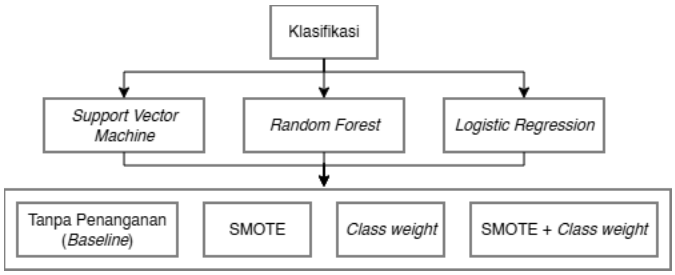
Term Frequency-Inverse Document Frequency (TF-IDF) merupakan salah satu metode pembobotan fitur dengan merepresentasikan teks dalam bentuk numerik berdasarkan frekuensi kemunculan istilah dalam dokumen (*Term Frequency/TF*) dan tingkat keunikannya dalam keseluruhan korpus (*Inverse Document Frequency/IDF*). Metode ini memberikan bobot yang lebih tinggi pada istilah yang informatif dan relevan, sehingga mampu meningkatkan kualitas representasi fitur [4]. Pada penelitian ini, TF-IDF diimplementasikan menggunakan 'TfidfVectorizer()'.

2.5. Pemisahan Data

Pemisahan data (*split data*) dilakukan dengan membagi data yang diperoleh menjadi dua bagian, yaitu 80% untuk data latih dan 20% untuk data uji. Data latih digunakan untuk membangun dan melatih model klasifikasi sentimen, sementara data pengujian dimanfaatkan untuk mengevaluasi kinerja model terhadap data yang tidak pernah dilihat sebelumnya.

2.6. Klasifikasi

Pemodelan klasifikasi yang ditunjukkan dalam Gambar 2 dilakukan menggunakan tiga model algoritma yang meliputi *Support Vector Machine* (SVM), *Random Forest*, dan *Logistic Regression*. Setiap algoritma yang dipilih selanjutnya memiliki tiga empat skema dalam penanganan data tidak seimbang yang mencakup skema tanpa penanganan, skema dengan teknik *oversampling* populer yaitu SMOTE, skema dengan *class weight* (pembobotan kelas), dan penggabungan SMOTE serta *class weight*.



Gambar 2. Tahap Klasifikasi

2.7. Evaluasi

Pengukuran kinerja ketiga model dengan empat skenario penanganan ketidakseimbangan data dievaluasi menggunakan *confusion matrix* pada data uji. Kriteria evaluasi meliputi akurasi, presisi, *recall* (sensitivitas), dan *f1-score* yang dihitung berurutan menggunakan (1), (2), (3), dan (4). Berdasarkan temuan [16], [17], *F1-score* merupakan metrik yang lebih sesuai untuk klasifikasi pada data tidak seimbang. Nilai TP, TN, FP, dan FN masing-masing merepresentasikan *True Positive*, *True Negative*, *False Positive*, dan *False Negative*. Selain itu, untuk memastikan kinerja model dapat menggeneralisasi data baru, proses validasi dilakukan menggunakan *k-fold cross-validation* dengan k = 5 yang membagi data pelatihan ke dalam lima bagian.

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$Presisi = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$F1 - Score = \frac{2\times TP}{2\times(TP+FP+FN)} \quad (4)$$

3. HASIL DAN ANALISIS

3.1. Pengumpulan Data

Pengumpulan data dilakukan melalui teknik *scraping* menggunakan bahasa pemrograman Python dengan pustaka 'google-play-scraper' untuk memperoleh data secara sistematis dari Google Play Store. *Dataset* yang diperoleh dari proses ini sebanyak 10.000 ulasan pengguna aplikasi ChatGPT. Setiap ulasan memuat informasi yang diperlukan dalam penelitian ini, seperti teks ulasan dan skor *rating*. Distribusi data didominasi oleh skor 5 yang mencerminkan tingkat kepuasan tertinggi pengguna aplikasi yaitu sebanyak 7.623 ulasan. Diikuti dengan 958 ulasan untuk skor 1 dan 796 ulasan skor 4. Sementara itu, skor 2 dan 3 secara berturut-turut mendapatkan sebanyak 221 dan 402 ulasan.

3.2. Prapemrosesan Data

Data yang telah diperoleh selanjutnya diberi label sentimen. Untuk skor 4 dan 5 merepresentasikan sentimen positif, skor 3 mencerminkan sentimen netral, serta penggabungan skor 1 dan 2 menunjukkan sentimen negatif. Dengan demikian, total ulasan untuk sentimen positif, negatif, dan netral secara berturut-turut adalah 8.419, 1.179, dan 402 ulasan. Setelah melalui rangkaian tahapan dalam prapemrosesan, total data menjadi 9.779 ulasan, berkurang 221 ulasan dari data awal. Berdasarkan Tabel 1, label sentimen positif mendominasi secara signifikan yaitu sekitar 84%, diikuti dengan label sentimen negatif sekitar 12%, dan data terkecil berasal dari label sentimen netral yaitu sekitar 4%. Distribusi sentimen ini menunjukkan ketidakseimbangan jumlah data yang cukup ekstrim pada ketiga kelas tersebut.

Tabel 1. Distribusi Data Sebelum dan Setelah Prapemrosesan

Label	Sebelum	Setelah
Positif	8.419	8.206
Negatif	1.179	1.176
Netral	402	397
Total	10.000	9.779

3.3. Ekstraksi Fitur dan Pemisahan Data

Ekstraksi fitur menggunakan TF-IDF digunakan untuk menentukan bobot kata dari dataset yang telah dibersihkan sebelumnya. Hasil perhitungan TF-IDF diperoleh matriks berukuran 9779 × 990. Kemudian dilakukan pemisahan data dengan komposisi 80:20 sehingga dihasilkan 7.823 data pelatihan dan 1.956 data pengujian dengan distribusi data seperti yang ditunjukkan Tabel 2.

Tabel 2. Distribusi Data Latih dan Data Uji

Label	Data Latih	Data Uji
Positif	6.559	1.647
Negatif	937	239
Netral	327	70
Total	7.823	1.956

3.4. Performa Model Klasifikasi

3.4.1. Support Vector Machine (SVM)

Model SVM diimplementasikan pada keempat skema pengujian dengan mengintegrasikan kernel *Radial Basis Function* (RBF), nilai regularisasi sebesar C = 1, dan parameter *gamma* bertipe *scale*. Untuk mengukur performa, digunakan teknik *k-fold cross-validation* (*k* = 5) pada data latih. Berdasarkan performa *5-fold-cross-validation* yang disajikan pada Tabel 3, rata-rata akurasi yang dihasilkan menunjukkan konsistensi yang baik pada seluruh skema, dengan capaian dalam rentang 0,85 hingga 0,89. Namun demikian, *f1-score macro* menunjukkan perbedaan yang cukup signifikan yang mencerminkan kemampuan model dalam menangani ketidakseimbangan kelas. Model tanpa penanganan memiliki skor terendah yaitu 0,52, sementara skor tertinggi diperoleh saat menggunakan skema *class weight* sebesar 0,58. Untuk strategi SMOTE serta penggabungan antara SMOTE dan *class weight* menghasilkan nilai serupa, baik akurasi maupun *f1-score*. Dengan demikian, strategi *class weight* pada data latih merupakan pendekatan yang paling optimal.

Tabel 3. Performa 5-*fold-cross-validation* pada SVM

Metrik	***Baseline***	SMOTE	CW	SMOTE+CW
Rata-rata Akurasi	0,89	0,85	**0,87**	0,85
Rata-rata F1-Score	0,52	0,53	**0,58**	0,53

Tabel 4. Hasil Metriks Evaluasi pada SVM

Metrik	***Baseline***	SMOTE	CW	SMOTE+CW
Akurasi	0,90	0,86	**0,87**	0,86
Presisi	0,56	0,57	**0,57**	0,57
Recall	0,52	0,53	**0,60**	0,53
F1-Score	0,53	0,55	**0,58**	0,55

Sementara itu, evaluasi pada data uji menunjukkan performa akurasi yang relatif meningkat pada seluruh skenario dibandingkan dengan evaluasi pada data latih. Model tanpa penanganan keseimbangan data menghasilkan akurasi tertinggi sebesar 0,90, tetapi memiliki *f1-score* terendah yaitu 0,53 yang menunjukkan performa tidak merata antar kelas. Penerapan metode SMOTE serta kombinasi SMOTE dan *class weight* masih serupa dengan hasil sebelumnya, yaitu memberikan hasil evaluasi dengan nilai yang identik. Di sisi lain, skenario *class weight* tidak memberikan peningkatan akurasi, tetapi memiliki performa tertinggi dalam metrik *f1-score* yaitu 0,58 dibandingkan dengan skenanrio lainnya, terutama skenario tanpa penanganan ketidakseimbangan kelas (*baseline*).

3.4.2. Random Forest (RF)

Penggunaan model *Random Forest* dalam penelitian ini menerapkan parameter *n_estimators* sebanyak 100. Evaluasi menggunakan *5-fold-cross-validation* pada data latih menghasilkan nilai rata-rata akurasi tertinggi pada model tanpa penanganan keseimbangan data yaitu sebesar 0,88. Akan tetapi, skenario ini menunjukkan rata-rata *f1-score* terendah, yaitu sebesar 0,52 seperti yang ditunjukkan pada Tabel 5. Hal ini mengindikasikan kecenderungan model untuk bias terhadap kelas mayoritas. Sementara itu, penerapan *class*

weight menghasilkan akurasi sebesar 0,87 lebih tinggi dibanding SMOTE maupun kombinasi SMOTE dan *class weight*. Sementara itu, nilai *f1-score* terbaik saat menerapkan skenario SMOTE dan penggabungan SMOTE serta *class weight*.

Evaluasi data uji menunjukkan bahwa model tanpa penanganan ketidakseimbangan mencapai akurasi tertinggi sebesar 0.90, tetapi dengan *f1-score* terendah sebesar 0.53 yang menunjukkan dominasi prediksi kelas mayoritas. Berdasarkan Tabel 6, penerapan SMOTE dan kombinasi SMOTE dengan *class weight* menurunkan akurasi menjadi 0.85, tetapi berhasil meningkatkan *f1-score* menjadi 0.58, hal tersebut mencerminkan peningkatan sensitivitas terhadap kelas minoritas. Selain itu, penerapan *class weight* memberikan nilai presisi tertinggi sebesar 0.57, sementara *recall* terbaik dicapai oleh SMOTE dan kombinasi SMOTE dengan bobot kelas sebesar 0.61.

Tabel 5. Performa 5-fold-cross-validation pada Random Forest

Metrik	*Baseline*	SMOTE	CW	SMOTE+CW
Rata-rata Akurasi	0,88	0,83	0,87	0,83
Rata-rata F1-Score	0,52	0,55	0,54	0,55

Tabel 6. Hasil Metriks Evaluasi pada Random Forest

Metrik	*Baseline*	SMOTE	CW	SMOTE+CW
Akurasi	0,90	0,85	**0,89**	0,85
Presisi	0,54	0,55	**0,57**	0,55
Recall	0,52	**0,61**	0,53	**0,61**
F1-Score	0,53	**0,58**	0,55	**0,58**

3.4.3. Logistic Regression (LR)

Hasil *5-fold cross-validation* pada data latih menunjukkan bahwa *Logistic Regression* tanpa perlakuan khusus pada ketidakseimbangan kelas mampu mencapai akurasi rata-rata yang cukup tinggi, yaitu 0,89, namun dengan *F1-score* yang relatif rendah di angka 0,52 seperti yang tertulis pada Tabel 7. Penerapan *class weight* berhasil meningkatkan keseimbangan performa model dengan menaikkan F1-score menjadi 0,58, meskipun diiringi penurunan akurasi menjadi 0,82. Di sisi lain, penggunaan SMOTE maupun gabungan SMOTE dengan *class weight* menurunkan akurasi lebih jauh menjadi 0,80, tetapi tetap memberikan peningkatan F1-score jika dibandingkan dengan model tanpa penanganan *imbalance*.

Pada tahap pengujian, hasil metriks evaluasi ditunjukkan pada Tabel 8. Model tanpa penanganan ketidakseimbangan mempertahankan akurasi tertinggi sebesar 0,90, namun performa F1-score hanya mencapai 0,55, yang mengindikasikan dominasi prediksi pada kelas mayoritas. Model dengan skema *class weight* menampilkan keseimbangan terbaik antara akurasi dan kemampuan deteksi kelas minoritas, dengan akurasi 0,82 dan F1-score terbaik 0,57, didukung oleh recall tertinggi sebesar 0,63. Sementara itu, penerapan SMOTE dan kombinasi SMOTE dengan *class weight* memperlihatkan penurunan akurasi ke angka 0,80 dengan F1-score stabil yaitu sebesar 0,55.

Tabel 7. Performa 5-fold-cross-validation pada Logistic Regression

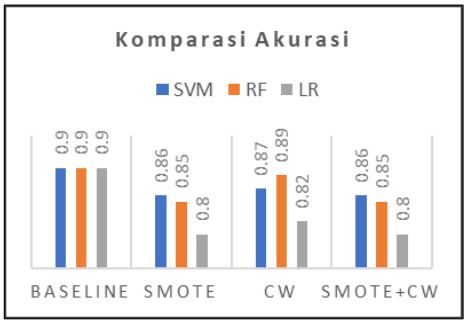
Metrik	*Baseline*	SMOTE	CW	SMOTE+CW
Rata-rata Akurasi	0,89	0,80	0,82	0,80
Rata-rata F1-Score	0,52	0,55	0,58	0,55

Tabel 8. Hasil Metriks Evaluasi pada Logistic Regression

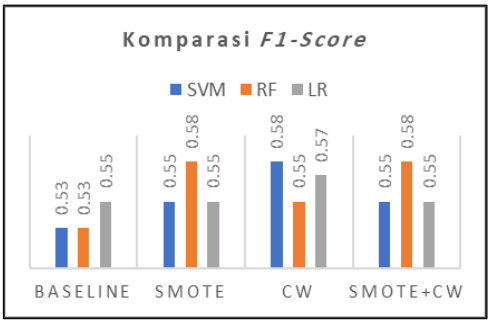
Metrik	*Baseline*	SMOTE	CW	SMOTE+CW
Akurasi	0,90	0,80	**0,82**	0,80
Presisi	0,67	0,54	**0,55**	0,54
Recall	0,53	0,62	**0,63**	0,62
F1-Score	0,55	0,55	**0,57**	0,55

3.5. Analisis Efektivitas Strategi Penanganan Ketidakseimbangan Data

Hasil pengujian menunjukkan adanya perbedaan performa antara model tanpa penanganan ketidakseimbangan data dan model dengan strategi penyeimbangan kelas. Gambar 3 memperlihatkan bahwa akurasi tertinggi pada ketiga algoritma diperoleh pada skema tanpa penanganan (*baseline*), yaitu sebesar 0,90. Namun, Gambar 4 menunjukkan bahwa nilai *f1-score* pada skema tersebut relatif lebih rendah dibandingkan dengan skema yang menerapkan teknik penyeimbangan. Penerapan SMOTE, *class weight*, maupun kombinasi keduanya cenderung menurunkan akurasi, tetapi diikuti dengan peningkatan *f1-score*. Hal ini mengindikasikan adanya *trade-off* antara akurasi keseluruhan dan keseimbangan performa antar kelas sehingga evaluasi kinerja pada data tidak seimbang tidak dapat hanya mengandalkan akurasi, melainkan perlu mempertimbangkan metrik yang lebih presentatif.



Gambar 3 Komparasi Akurasi



Gambar 4 Komparasi F1-Score

Perbandingan metode penanganan ketidakseimbangan menunjukkan bahwa pendekatan *class weight* memiliki konsistensi kinerja yang lebih baik dibandingkan SMOTE maupun kombinasi keduanya. Pada algoritma SVM dan Logistic Regression, *class weight* menghasilkan nilai *macro F1-score* tertinggi, sedangkan pada Random Forest kinerjanya tetap kompetitif meskipun tidak dominan. Sebaliknya, SMOTE tidak memberikan peningkatan performa secara konsisten dan pada beberapa kasus justru menurunkan akurasi secara signifikan. Kombinasi SMOTE dan *class weight* juga tidak selalu memberikan keuntungan tambahan dibandingkan penggunaan metode tunggal.

Perbedaan kinerja antar metode tersebut berkaitan dengan karakteristik masing-masing algoritma, yang menunjukkan bahwa efektivitas strategi penanganan ketidakseimbangan bersifat *model-dependent*. SVM memperoleh peningkatan *macro F1-score* paling optimal melalui *class weight*, sedangkan Random Forest menunjukkan adaptabilitas yang lebih baik terhadap SMOTE. Di sisi lain, Logistic Regression lebih sensitif terhadap perubahan distribusi data, yang tercermin dari penurunan akurasi yang lebih besar meskipun *F1-score* meningkat. Secara metodologis, temuan ini menegaskan bahwa pemilihan strategi penanganan ketidakseimbangan harus disesuaikan dengan karakteristik algoritma dan representasi fitur, serta tidak dapat digeneralisasi untuk seluruh model.

4. KESIMPULAN

Berdasarkan hasil penelitian, penerapan strategi penanganan ketidakseimbangan kelas memberikan dampak terhadap kualitas klasifikasi sentimen multikelas pada ulasan aplikasi ChatGPT, meskipun tidak selalu diikuti oleh peningkatan akurasi. Ketiga algoritma yang diuji, yaitu SVM, Random Forest, dan Logistic Regression, secara konsisten menghasilkan akurasi tertinggi pada skema tanpa penanganan ketidakseimbangan dengan nilai 0,90, namun memiliki nilai macro F1-score yang relatif rendah mengindikasikan adanya kecenderungan bias terhadap kelas mayoritas. Penerapan class weight terbukti menjadi strategi yang paling konsisten dalam meningkatkan macro F1-score, khususnya pada SVM dan Logistic Regression, sehingga pada konteks dataset dan representasi fitur yang digunakan, metrik macro F1-score dinilai lebih merepresentasikan keseimbangan performa antar kelas dibandingkan akurasi semata. Sebaliknya, SMOTE tidak menunjukkan peningkatan performa yang stabil dan pada beberapa skenario justru menurunkan akurasi, sementara kombinasi SMOTE dan class weight tidak selalu memberikan keunggulan tambahan dibandingkan metode tunggal. Hal ini menegaskan bahwa efektivitas strategi penanganan ketidakseimbangan bersifat model-dependent dan perlu disesuaikan dengan karakteristik algoritma serta representasi fitur, sehingga penelitian selanjutnya disarankan untuk mengeksplorasi teknik penyeimbangan alternatif, memanfaatkan representasi teks berbasis contextual embedding, serta memperluas cakupan dataset untuk meningkatkan generalisasi dan ketahanan model pada kondisi data dunia nyata yang lebih kompleks.

REFERENSI

[1] B. Purbayanto dan T. N. Suharsono, "Analisis Sentimen Pengguna X terhadap Chatgpt dengan Algoritme Naive Bayes," *J. Telemat.*, vol. 18, no. 2, hal. 63–71, 2024.

[2] N. Widaad, D. Anggraini, E. Faculty, dan U. Gunadarma, "SENTIMENT ANALYSIS OF CHATGPT APP USER REVIEWS USING SVM AND CNN," *J. Tek. Inform.*, vol. 5, no. 6, hal. 1687–1700, 2024.

[3] F. N. Sukmana, A. Primajaya, dan A. Suharso, "Sentiment Analysis of ChatGPT Application Reviews Using the BERT Algorithm," *J. Info Sains Inform. dan Sains*, vol. 15, no. 02, hal. 436–442, 2025, doi: 10.54209/infosains.v15i02.

[4] I. P. Ariyani, K. D. Tania, A. Wedhasmara, dan A. Meiriza, "Ekstraksi Pengetahuan dari Ulasan Aplikasi CapCut Menggunakan Metode Aspect-Based Sentiment Analysis dan Klasifikasi," *J. Buana Inform.*, vol. 16, no. 1, hal. 80–92, 2025.

[5] Z. Xu, F. Qingguo, Y. Huang, dan M. Xie, "The public attitude towards ChatGPT on reddit : A study based on unsupervised learning from sentiment analysis and topic modeling," *PLoS One*, vol. 19, no. 5, hal. 1–26, 2024, doi: 10.1371/journal.pone.0302502.

[6] F. A. Tohir, B. Irawan, dan A. Bahtiar, "Analisis Sentimen Aplikasi ChatGPT Mobile Menggunakan Agoritma Naïve Bayes," *ICIT J.*, vol. 10, no. 2, hal. 179–192, 2024.

[7] A. Ardiansyah, E. A. Pratama, dan N. I. Fadlilah, "Analisis Sentimen Pengguna Terhadap Aplikasi ChatGPT Di Google Play Store: Penerapan Algoritma Support Vector Machine," *J. Tek. Inform. dan Sist. Inf.*, vol. 11, no. 2, hal. 247–254, 2024.

[8] M. Z. Mubarok, S. Rizqi, R. M. Rizal, A. B. N. Syalim, dan N. Iksan, "Klasifikasi Sentimen Ulasan Aplikasi Claude Menggunakan Naive Bayes dan Support Vector Machine," *JUSIFOR J. Sist. Inf. Dan Inform.*, vol. 4, no. 1, hal. 110–119, 2025, doi: https://doi.org/10.70609/jusifor.v4i1.7054.

[9] G. J. Sagala dan Y. T. Samuel, "Sentiment Analysis on ChatGPT App Reviews on Google Play Store Using Random Forest Algorithm , Support Vector Machine and Naïve Bayes," *Int. J. Eng. Bus. Soc. Sci.*, vol. 2, no. 04, hal. 1194–1204, 2024.

[10] N. C. Ramadani, I. Tahyudin, dan A. S. Barkah, "Perbandingan Algoritma Support Vector Machine , Decision Tree , dan Logistic Regresion Pada Analisis Sentimen Ulasan Aplikasi Netflix," *J. Nas. Teknol. dan Sist. Inf.*, vol. 10, no. 02, hal. 110–117, 2024.

[11] E. M. Pusung dan I. N. Dewi, "Optimasi RoBERTa dengan Hyperparameter Tuning untuk Deteksi Emosi berbasis Teks," *J. Nas. Teknol. dan Sist. Inf.*, vol. 10, no. 03, hal. 240–248, 2024.

[12] A. A. Qolbu, N. Fitriyati, dan N. Inayah, "Performa Naïve Bayes , SVM , dan IndoBERT pada Analisis Sentimen Twitter IndiHome dengan Strategi Penanganan Data Tidak Seimbang," *J. Fourier*, vol. 814, no. 1, hal. 29–44, 2025, doi: 10.14421/fourier.2025.141.29-44.

[13] H. Kelvin, Erlin, Y. Desnelita, dan D. Oktarina, "Analisis Sentimen Terhadap Ulasan Aplikasi IKD di Play Store Menggunakan Random Forest," *J. Nas. Tek. Elektro dan Teknol. Inf.*, vol. 14, no. 3, hal. 171–180, 2025.

[14] A. S. Rizkia, Wufron, dan F. F. Roji, "Analisis Sentimen Coretax : Perbandingan Pelabelan Data Manual, Transformers-Based, dan Lexicon-Based pada Performa IndoBERT," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 5, no. July, hal. 1037–1048, 2025.

[15] A. F. Zain, H. Al Azies, dan I. K. Ananda, "Analisis Sentimen Ulasan Pengguna iPhone dengan Pendekatan Hibrida RoBERTa dan XGBoost," *J. Algoritm.*, vol. 22, no. 1, hal. 1039–1049, 2025, doi: 10.33364/algoritma/v.22-1.2277.

[16] T. T. Wong, "Linear Approximation of F-Measure for the Performance Evaluation of Classification Algorithms on Imbalanced Data Sets," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 2, hal. 753–763, Feb 2022, doi: 10.1109/TKDE.2020.2986749.

[17] K. M. Sujon, R. Hassan, K. Choi, dan A. Samad, "empirical evidence from advanced statistics , ML , and XAI for evaluating business predictive models," *J. Big Data*, vol. 12, no. 268, hal. 1–45, 2025, doi: https://doi.org/10.1186/s40537-025-01313-4.