Eksplorasi Teknik Balancing pada Klasifikasi Sentimen SVM dengan K-Fold Cross Validation

Septi Kirana Damai Lestari[1], Anita Qoiriah[2]

[1,2] Teknik Informatika, Universitas Negeri Surabaya
[1]Septi.22050@universitas.ac.id
[2]Anitaqoiriah@unesa.ac.id

***Abstrak*—Media social merupakan ruang public digital yang memungkinkan Masyarakat bebas dalam menyampaikan opini tentang berbagai hal, salah satunya kinerja pemerintah daerah. Komentar media social menjadi sumber informasi penting dalam memahami opini public dan kecenderungan sentiment publik. Namun, terdapat beberapa tantangan dalam analisis komentar media social, seperti karakteristik teks yang pendek dan informal serta distribusi kelas yang cenderung tidak seimbang sehingga model cenderung bias ke kelas mayoritas. Penelitian ini malakukan eksplorasi Teknik balancing pada model SVM dengan representasi fitur TF-IDF. Penelitian mengeksplorasi lima strategi penanganan ketidakseimbangan kelas, yaitu tanpa balancing (baseline), class weighting, random undersampling, random oversampling, dan SMOTE Seluruh eksperimen dievaluasi menggunakan Stratified K-Fold Cross-Validation (K=5) untuk memperoleh estimasi performa yang lebih stabil dan representatif. Dataset terdiri atas 3.957 komentar dari platform TikTok dan X/Twitter terkait kinerja Pemerintah Kota Malang periode Maret–Desember 2025 dengan distribusi 86,92% komentar negatif dan 13,08% komentar positif. Teknik resampling diterapkan hanya pada data training di setiap fold dengan tujuan menghindari data leakage. Hasil eksplorasi menunjukkan bahwa penerapan class weighting menghasilkan Mean Macro F1 tertinggi sebesar 0,6399 ± 0,0232, melampaui baseline (0,6350), random undersampling (0,5690), oversampling (0,6327), dan SMOTE (0,5956). Analisis confusion matrix agregat menunjukkan bahwa random undersampling menghasilkan False Positive Rate sebesar 30,9%, hal ini mengindikasikan bahwa hampir sepertiga kritik masyarakat salah diklasifikasikan sebagai apresiasi. Berdasarkan eksplorasi dapat disimpulkan bahwa class weighting merupakan strategi yang paling sesuai untuk dataset komentar media sosial berbahasa Indonesia karena kemampuan sensitivitas model terhadap kelas minoritas tanpa mengubah distribusi data dan risiko overfitting akibat duplikasi atau sampel sintetik.**

***Kata Kunci*— analisis sentimen, class imbalance, K-Fold cross-validation, media sosial, SVM, strategi balancing.**

I. PENDAHULUAN

Media sosial telah menjadi salah satu sarana bagi masyarakat Indonesia dalam menyampaikan opini dan pandangannya terkait berbagai isu, termasuk kinerja pemerintahan. Seiring dengan berkembangnya teknologi, terdapat peningkatan dalam penggunaan internet, Dimana tercatat penggunaan internet mencapai 79,5% dari total populasi. Platform seperti TikTok dan X/Twitter semakin banyak dimanfaatkan sebagai ruang untuk mengekspresikan penilaian, kritik, maupun apresiasi terhadap kebijakan dan pelayanan pemerintah [1]. Analisis otomatis terhadap komentar di media sosial dapat memberikan kemudahan dalam menganalisis informasi yang bermanfaat bagi pemerintah daerah untuk memahami persepsi masyarakat serta mengevaluasi kinerja dan implementasi kebijakan.

Pemilihan Kota Malang sebagai objek penelitian dikarenakan beberapa hal, seperti tingginya populasi pada kota ini yang mencapai 885.270 jiwa dengan 65,52% berusia produktif, keberadaan 51 perguruan tinggi yang membentuk masyarakat kritis, serta pergantian kepemimpinan pada Februari 2025 yang memicu peningkatan diskusi publik di media sosial terkait kinerja pemimpin baru [2].

Selama lima tahun terakhir, penelitian analisis sentimen media sosial berbahasa Indonesia terhadap kinerja pemerintah telah banyak dilakukan. Penelitian oleh Sebastian et al. Pada tahun 2024 Menggunakan algoritma SVM untuk menganalisis opini publik terhadap hak angket DPR pada platform X/Twitter [3]. Penelitian Setyaningsih dan Sipayung tahun 2026 menerapkan SVM pada data media sosial X untuk menganalisis sentimen terhadap kinerja Kementerian Keuangan Indonesia, memperoleh akurasi 86,27% dengan F1-Score 0,86 [4]. Penelitian lainnya oleh Rahman et al. Tahun 2025 yang berfokus pada opini publik terhadap lembaga pemerintah Indonesia melakukan eksplorasi terhadap beberapa model klasifikasi seperti SVM, Naïve Bayes, Rocchio, dan KNN dan mencatat performa terbaik pada model SVM dengan nilai *weighted average* F1 sebesar 0,68 [5]. Berdasarkan beberapa penelitian sebelumnya, dapat disimpulkan bahwa model SVM+TF-IDF merupakan pendekatan yang seringkali digunakan karena terbukti efektif dan kompetitif untuk klasifikasi sentimen teks berbahasa Indonesia yang pendek dan informal.

Analisis sentimen pada komentar media sosial berbahasa Indonesia masih menghadapi sejumlah tantangan, seperti karakteristik teks yang pendek dan informal, adanya fenomena code-mixing antara Bahasa Indonesia dan Bahasa Jawa yang menyulitkan proses NLP konvensional, serta distribusi kelas yang sangat tidak seimbang. Komentar media sosial terhadap kinerja pemerintah secara alami didominasi oleh sentimen negatif, hal ini sejalan dengan penelitian oleh Setyaningsih dan Sipayung (2026) yang menganalisis komentar dengan distribusi data negatif sebesar 81,1% dan positif 18,9% [4]. Model yang dilatih pada data imbalanced cenderung bias pada kelas mayoritas dan menghasilkan recall yang sangat rendah pada kelas minoritas meskipun akurasi keseluruhan tampak tinggi, fenomena ini dikenal sebagai fenomena accuracy paradox [6]. Dalam konteks monitoring kinerja pemerintah, kegagalan model dalam mendeteksi sentimen minoritas (positif) memiliki efek serius karena apresiasi masyarakat yang tidak terdeteksi dapat menghilangkan penilaian terhadap program yang berjalan baik.

Berbagai strategi penanganan ketidakseimbangan kelas telah diusulkan dalam literatur, mencakup pendekatan berbasis manipulasi data (data-level) seperti random undersampling, random oversampling, dan SMOTE (Synthetic Minority Oversampling Technique), serta pendekatan berbasis algoritma (algorithm-level) seperti class weighting. Penelitian Fajar et al. (2023) menunjukkan bahwa pemilihan strategi balancing yang tidak tepat dapat menurunkan performa model secara keseluruhan dan justru memperburuk kemampuan deteksi kelas minoritas [6] . Studi komparatif terbaru oleh Ritonga et al. tentang sentimen pada platform BPJS Kesehatan menegaskan bahwa strategi balancing perlu dipilih berdasarkan karakteristik domain data, bukan hanya berdasarkan nilai metrik agregat [7].

Beberapa penelitian sebelumnya masih terdapat beberapa keterbatasan metodologis yang dapat dikembangkan. Sebagian besar penelitian hanya mengevaluasi satu atau dua strategi balancing tanpa melakukan perbandingan yang komprehensif. Selain itu, beberapa penelitian masih menggunakan skema *hold-out* tunggal, padahal pendekatan tersebut lebih rentan terhadap bias pembagian data apabila dibandingkan *Stratified K-Fold Cross-Validation*. Di sisi lain, aspek evaluasi juga umumnya masih berfokus pada metrik performa utama tanpa melakukan analisis pada pola kesalahan klasifikasi, seperti *False Positive Rate* (FPR) dan *False Negative Rate* (FNR), yang sebenarnya memiliki implikasi penting dalam konteks pemantauan kinerja pemerintah. Oleh karena itu, penelitian ini mengisi keterbatasan tersebut melalui eksplorasi terhadap lima strategi penanganan ketidakseimbangan kelas pada model SVM berbasis TF-IDF menggunakan *Stratified K-Fold Cross-Validation* (K=5), dengan memastikan seluruh proses preprocessing, termasuk pembentukan representasi TF-IDF dan penerapan teknik resampling, dilakukan hanya pada data latih di setiap fold untuk mencegah *data leakage*. Selain membandingkan kinerja model berdasarkan nilai *Macro F1*, penelitian ini juga menganalisis pola kesalahan klasifikasi melalui FPR dan FNR dari *confusion matrix* agregat dengan tujuan memberikan pemahaman yang lebih kontekstual mengenai efektivitas analisis sentimen sebagai sarana monitoring kinerja pemerintah.

II. Tinjauan Pustaka

A. Klasifikasi Sentimen

Klasifikasi sentimen merupakan salah satu bidang dalam *Natural Language Processing* (NLP) yang bertujuan untuk mengidentifikasi dan mengelompokkan polaritas opini yang terkandung dalam teks, seperti sentimen positif, negatif, maupun netral. Dalam penelitian ini, klasifikasi sentimen digunakan untuk mengidentifikasi polaritas komentar masyarakat di media sosial terhadap kinerja Pemerintah Kota Malang ke dalam dua kelas, yaitu sentimen positif dan sentimen negatif. Proses klasifikasi dilakukan melalui tahapan pra-pemrosesan teks, meliputi cleaning, case folding, normalisasi bahasa, dan stopword removal, yang kemudian diikuti dengan representasi teks dan proses pemodelan menggunakan algoritma klasifikasi.

B. Representasi Teks

Representasi teks merupakan tahapan mengubah data teks menjadi bentuk numerik agar dapat diproses oleh model komputasi. TF-IDF. TF-IDF merupakan salah satu metode representasi fitur yang mengubah kata menjadi nilai numerik berdasarkan frekuensi kemunculannya pada suatu dokumen serta tingkat kelangkaan kata tersebut pada keseluruhan korpus [8]. Karena prosesnya relatif sederhana, efisien, dan mampu menangani data berdimensi tinggi, metode ini banyak digunakan pada berbagai penelitian klasifikasi teks. Dalam penelitian ini, TF-IDF dimanfaatkan untuk merepresentasikan fitur teks sebelum diproses oleh model SVM.

C. Penanganan Ketidakseimbangan Kelas (Imbalanced Data Handling)

Class imbalance merupakan kondisi Ketika distribusi jumlah data antar kelas tidak proporsional, sehingga model cenderung mempelajari dan mengenali pola dari kelas mayoritas dan mengabaikan kelas minoritas [9]. Pada penelitian ini, distribusi sentimen sangat tidak seimbang, yaitu sekitar 86,92% komentar negatif dan 13,08% komentar positif. Oleh karena itu, dilakukan beberapa eksplorasi terkait penanganan ketidakseimbangan data dengan penerapan pendekatan resampling data dan *cost-sensitive* lalu melakukan penilaian pendekatan mana yang lebih sesuai untuk dataset penelitian. Berikut merupakan beberapa pendekatan penanganan ketidakseimbangan kelas.

1. Class Weighting

Class weighting merupakan pendekatan cost-sensitive learning yang memberikan bobot penalti lebih besar pada kelas minoritas tanpa mengubah distribusi data asli [10]. Bobot kelas dihitung dengan rumus:

$$w_c = \frac{N}{|C| \times N_c}$$

dengan:

N = jumlah total sampel

$|C|$ = jumlah kelas

N_c = jumlah sampel pada kelas *c*

Dengan memberikan bobot yang lebih besar pada kelas minoritas, kesalahan prediksi pada kelas tersebut akan dihitung lebih berat sehingga model tidak hanya fokus pada kelas mayoritas. Class weighting memungkinkan model belajar secara lebih adil tanpa perlu menambah data sintetis atau melakukan oversampling yang berisiko mengganggu representasi semantik, sehingga efektif diterapkan pada model berbasis neural network termasuk transformer [11].

2. Random Undersampling

Random undersampling merupakan teknik resampling yang dilakukan dengan mengurangi jumlah sampel pada kelas mayoritas secara acak hingga distribusi kelas menjadi lebih seimbang [12]. Teknik ini dapat mengurangi kompleksitas komputasi, namun berpotensi menyebabkan hilangnya informasi penting akibat pembuangan sebagian data mayoritas.

3. Random Oversampling

Random oversampling merupakan teknik resampling data dengan tujuan menyeimbangkan distribusi kelas dengan menduplikasi sampel pada kelas minoritas secara acak [12]. Pendekatan ini tidak mengurangi data dari kelas mayoritas, namun berpotensi menyebabkan model mengalami overfitting karena adanya duplikasi data pada kelas minoritas.

4. SMOTE (Synthetic Minority Oversampling Technique)

SMOTE merupakan teknik oversampling yang menghasilkan sampel sintetik pada kelas minoritas melalui proses interpolasi antara suatu sampel dan beberapa tetangga terdekatnya [13]. Sampel sintetik dibentuk menggunakan persamaan:

$$x_{new} = x_i + \lambda\,(x_{nn} - x_i) \qquad 0 \le \lambda \le 1$$

x_i merupakan sampel kelas minoritas, x_{nn} adalah salah satu tetangga terdekat dari x_i, dan λ merupakan bilangan acak pada rentang 0 hingga 1. Berbeda dengan random oversampling yang hanya menduplikasi sampel yang sudah ada, SMOTE menghasilkan sampel baru yang tidak identik dengan data asli sehingga dapat meningkatkan variasi representasi kelas minoritas.

D. Model Support Machine Learning (SVM)

Support Vector Machine (SVM) merupakan algoritma machine learning yang bekerja dengan mencari hyperplane optimal yang dapat memisahkan data dari dua kelas secara maksimal [14]. SVM LinearSVC terbukti efektif falam menangani data berdimensi tinggi dan memberikan performa baik ketika dikombinasikan dengan representasi TF-IDF, oleh karena itu, metode ini seringkali digunakan pada klasifikasi teks [15].

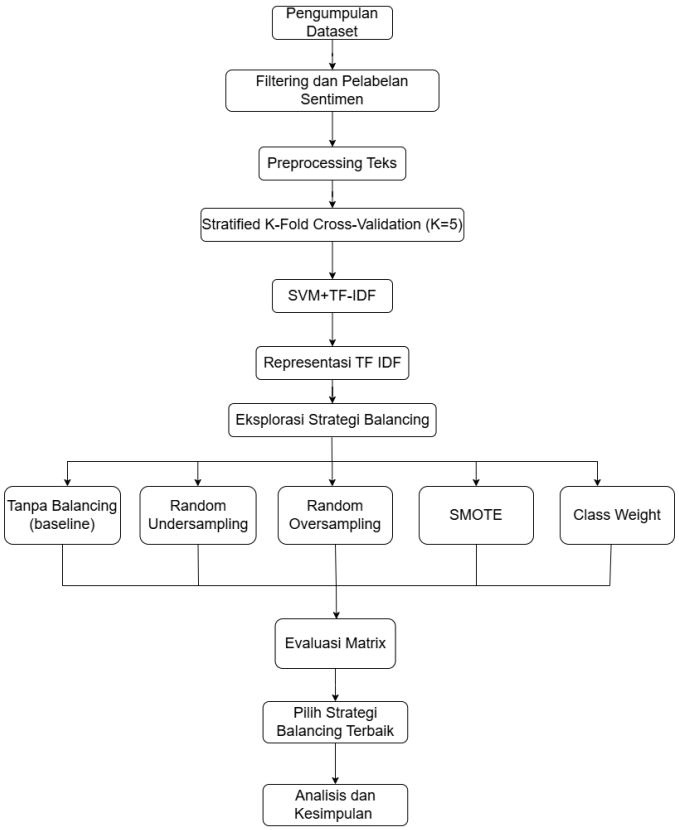
E. Stratified K-Fold Cross-Validation dan Data Leakage

Stratified K-Fold Cross-Validation merupakan metode evaluasi yang membagi dataset ke dalam K bagian (fold) dengan mempertahankan proporsi setiap kelas pada masing-masing fold [16]. Proses pelatihan dan evaluasi dilakukan sebanyak K kali, di mana setiap fold memperoleh kesempatan menjadi data validasi satu kali. Pendekatan ini menghasilkan estimasi performa yang lebih stabil dan mengurangi ketergantungan terhadap satu pembagian data (single split).

Pada penerapan teknik resampling, proses balancing harus dilakukan secara eksklusif pada data training di dalam setiap fold, bukan sebelum pembagian fold. Penerapan yang tidak tepat dapat menyebabkan data leakage, yaitu kondisi dimana sampel duplikat atau sampel sintetik tersebar ke data validasi sehingga berpotensi menghasilkan estimasi performa yang terlalu optimistis dan tidak merepresentasikan kemampuan generalisasi model pada data yang belum pernah dilihat sebelumnya [16].

III. METODOLOGI

Penelitian ini menerapkan pendekatan kuantitatif berbasis Machine Learning untuk melakukan analisis sentimen komentar masyarakat terkait kinerja Pemerintah Kota Malang pada platform TikTok dan X/Twitter. Tahapan penelitian meliputi pengumpulan dataset dan pelabelan sentimen, filtering data non opini dan quality control, preprocessing teks, eksplorasi strategi penanganan ketidakseimbangan kelas, pemodelan klasifikasi, dan evaluasi performa. Model yang digunakan adalah SVM dengan representasi TF-IDF dan fine-tuning IndoBERT berbasis contextual embedding. Eksplorasi dilakukan menggunakan Stratified K-Fold Cross-Validation untuk memperoleh konfigurasi model dengan performa terbaik. Seluruh tahapan penelitian ditunjukkan pada Gambar 1.



Gbr. 1 Flowchart Alur Penelitian

A. *Pengumpulan Dataset dan Pelabelan Sentimen*

Dataset dikumpulkan dari dua platform yaitu TikTok (70,71%) menggunakan Apify, dan X/Twitter (29,29%) menggunakan tweet-harvest. Komentar berkaitan dengan kinerja Pemerintah Kota Malang pada periode Maret–Desember 2025. Setelah preprocessing dan filtering, diperoleh 3.957 komentar yang digunakan sebagai dataset final.

Pelabelan dilakukan secara otomatis menggunakan InSet (Indonesian Sentiment Lexicon) dengan mekanisme skor = Σ(kata positif) − Σ(kata negatif); skor ≤ 0 dikategorikan negatif. Distribusi kelas disajikan pada Tabel I.

Setelah proses pelabelan otomatis, dilakukan quality check terhadap 70% data untuk mengevaluasi hasil pelabelan serta memverifikasi kesesuaian label secara manual.

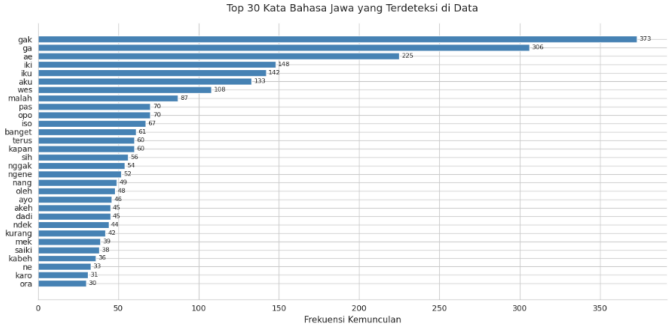
TABEL I

DISTRIBUSI KELAS SENTIMEN PADA DATASET

Kelas	Jumlah	Proporsi
Negatif	3.440	86,92%
Positif	517	13,08%
Total	3.957	100%

B. *Preprocessing Data*

Tahap preprocessing terdiri atas penerapan *case folding*, penghapusan *noise*, normalisasi Bahasa Jawa, normalisasi *slang*, *stemming*, dan *stopword removal* untuk menghasilkan representasi fitur yang lebih konsisten. Kamus normalisasi Bahasa Jawa dibangun berdasarkan analisis frekuensi token pada dataset dan mencakup kata-kata dengan frekuensi tinggi, seperti *gak* (1.847 kali), *ra* (934 kali), dan *wes* (812 kali), guna mengurangi variasi kosakata akibat fenomena *code-mixing* Bahasa Indonesia–Jawa.



Gbr. 2 Kata Bahasa Jawa yang Terdeteksi di Dataset

C. *Stratified K-Fold Validation*

Pada penelitian ini menggunakan metode Stratified K-Fold Cross Validation untuk proses evaluasi model dengan jumlah fold (k) sebanyak 5. Pemilihan nilai k = 5 bertujuan untuk memperoleh hasil evaluasi yang lebih stabil dan representatif, dengan tetap mempertahankan proporsi distribusi kelas pada setiap fold. Selanjutnya, model akan dilatih dan diuji sebanyak lima kali, dengan pembagian pada setiap iterasi satu fold digunakan sebagai data pengujian, sedangkan empat fold lainnya digunakan sebagai data pelatihan. Nilai evaluasi akhir diperoleh dari rata-rata hasil pengujian pada kelima fold tersebut.

D. *Desain Eksperimen SVM dan Strategi Penanganan Ketidakseimbangan Kelas*

Evaluasi model SVM+TF-IDF dilakukan menggunakan Stratified K-Fold Cross-Validation dengan k = 5 dan random_state=42. Pendekatan stratified dipilih untuk mempertahankan proporsi kelas sentimen pada setiap fold sehingga distribusi data latih dan validasi tetap representatif. Evaluasi dilakukan secara konsisten pada lima skenario penanganan ketidakseimbangan kelas, yaitu *baseline* tanpa balancing, *class weighting*, *random undersampling*, *random oversampling*, dan *Synthetic Minority Oversampling Technique* (SMOTE).

Untuk mencegah terjadinya *data leakage*, proses pembentukan fitur TF-IDF dan teknik balancing diterapkan secara eksklusif pada data training di setiap fold. TfidfVectorizer di-*fit* hanya menggunakan data training pada fold yang bersangkutan, sedangkan data validasi hanya ditransformasi menggunakan *vocabulary* dan nilai IDF yang diperoleh dari data training. Demikian pula, seluruh teknik resampling diterapkan hanya pada data training, sementara data validasi dipertahankan pada distribusi aslinya yang tidak seimbang. Prosedur ini memastikan bahwa estimasi performa yang diperoleh lebih merepresentasikan kemampuan generalisasi model pada data yang belum pernah dilihat sebelumnya.

Model klasifikasi yang digunakan adalah LinearSVC dari *scikit-learn* dengan parameter max_iter=2000 dan random_state=42. Pada skenario *class weighting*, bobot kelas dihitung secara otomatis menggunakan fungsi compute_class_weight('balanced'), sedangkan implementasi *random undersampling*, *random oversampling*, dan SMOTE dilakukan menggunakan library *imbalanced-learn*.

E. *Metric Evaluasi*

Metrik utama yang digunakan adalah Macro F1-Score. Metric ini dipilih karena memberikan bobot yang setara untuk setiap kelas, tidak terpengaruh dominasi kelas mayoritas seperti metric accuracy. Karena menggunakan evaluasi K-Fold dilaporkan mean ± standar deviasi dari 5 fold. Selain itu dilakukan juga penilaian terhadap confusion matrix agregat (kumulatif dari seluruh fold), serta False Positive Rate (FPR) dan False Negative Rate (FNR) untuk menganalisis implikasi pola kesalahan dalam konteks monitoring kinerja pemerintah.

IV. HASIL DAN PEMBAHASAN

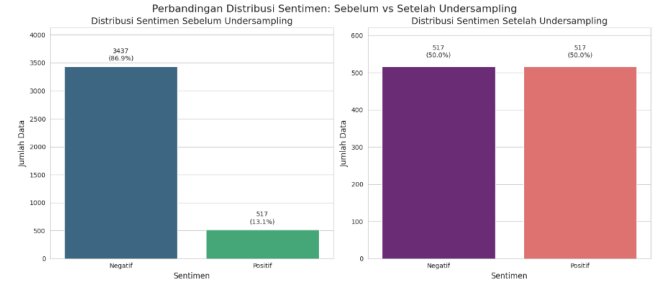
A. *Karakteristik Dataset*

Dataset terdiri dari 3.957 komentar terkait kinerja Pemerintah Kota Malang yang diperoleh dari platform TikTok dan X/Twitter. Hasil pelabelan menunjukkan distribusi kelas yang sangat tidak seimbang, yaitu 3.440 komentar negatif (86,92%) dan 517 komentar positif (13,08%). Adanya imbalance kelas pada dataset media social dikarenakan dominasi Masyarakat yang lebih condong mengungkapkan kritik dari pada apresiasi. Berdasarkan kondisi tersebut mengindikasikan dominasi komentar terhadap kelas mayoritas (negatif) sehingga model berpotensi bias pada kelas mayoritas apabila tidak dilakukan penanganan ketidakseimbangan kelas. Oleh karena itu, diperlukan penerapan balancing untuk mengatasi ketidakseimbangan data. Pada penelitian ini dilakukan eksplorasi beberapa pendekatan balancing mulai dari pendekatan balancing berbasis resampling data dan pendekatan berbasis cost-sensitive learning.

B. *Pengaruh Strategi Penanganan Imbalance pada Distribusi Data dan Pembobotan Kelas*

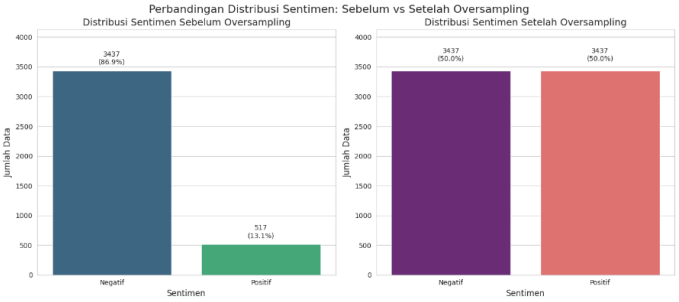
Pada penelitian dilakukan eskplorasi terhadap beberapa pendekatan penanganan katidakseimbangan kelas yang meliputi pendekatan berbasis resampling data dan pendekatan berbasis cost-sensitive. Pendekatan berbasis resampling dilakukan dengan mengubah distribusi data pada dataset,

sedangkan pendekatan berbasis cost-sensitive menangani ketidakseimbangan kelas dengan memberikan bobot atau biaya yang lebih besar pada kelas minoritas selama proses pelatihan model. Berikut merupakan pengaruh penerapan strategi balancing pada distribusi dataset dan bobot kelas.



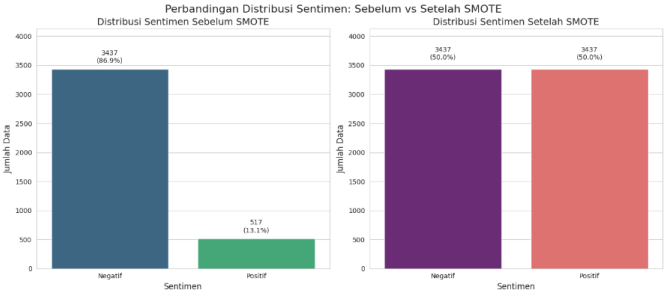
Gbr. 3 Distribusi Dataset per Kelas Sebelum dan Setelah Undersampling

Berdasarkan Gambar 3, penerapan Random Undersampling mengurangi jumlah data pada kelas mayoritas (negatif) dari 3.437 komentar menjadi 517 komentar sehingga jumlahnya setara dengan kelas minoritas. Setelah undersampling, distribusi kedua kelas menjadi seimbang, masing-masing sebesar 50%. Namun, proses ini menyebabkan sebagian besar data pada kelas negatif dihapus sehingga berpotensi menghilangkan informasi penting yang diperlukan model dalam mempelajari karakteristik sentimen negatif.



Gbr. 4 Perbandingan Distribusi Data Sebelum dan Sesudah Oversampling

Berdasarkan grafik pada Gambar 4, penerapan Random Oversampling menambah data pada kelas minoritas (positif) dari 517 menjadi 3.437 komentar sehingga jumlahnya sama dengan kelas minoritas (negatif) melalui proses duplikasi sample secara acak. Setelah penerapan oversampling, distribusi setiap kelas menjadi seimbang, masing-masing sebesar 50%.

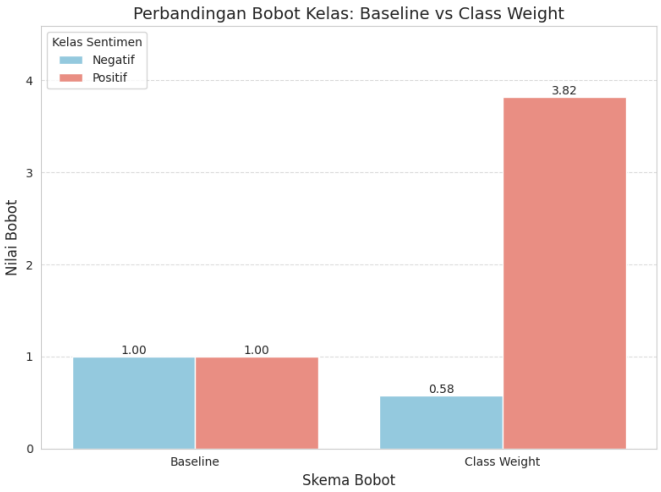


Gbr. 5 Perbandingan Distribusi Data Sebelum dan Sesudah SMOTE

Berdasarkan Gambar 5, penerapan SMOTE meningkatkan jumlah data pada kelas minoritas (positif) dari 517 komentar menjadi 3.437 komentar sehingga distribusi kedua kelas menjadi seimbang, masing-masing sebesar 50%. Berbeda dengan Random Oversampling yang menambahkan data melalui duplikasi sampel, SMOTE menambahkan data pada kelas minoritas dengan menghasilkan sampel sintetik baru sehingga jumlah data meningkat tanpa mengulang data yang sama secara identik.

Sample sintetik yang dibentuk oleh SMOTE didapatkan melalui proses interpolasi linear antara suatu sampel kelas minoritas dan salah satu tetangga terdekatnya (nearest neighbor) di ruang fitur TF-IDF. Dengan demikian, data baru yang dihasilkan memiliki karakteristik yang berbeda diantara sampel-sampel kelas minoritas yang sudah ada. Pendekatan ini diharapkan mampu memperkaya representasi kelas minoritas serta mengurangi risiko overfitting akibat penggunaan sampel duplikat.

Sedangkan pada pendekatan *cost-sensitive* yaitu *Class Weight* tidak merubah distribusi dataset pada setiap kelas, namun memberikan bobot yang lebih tinggi pada kelas minoritas sehingga kesalahan klasifikasi pada kelas tersebut memperoleh penalti yang lebih besar selama proses pelatihan model. Berikut merupakan distribusi pengaruh penerapan class weighting pada nilai pembobotan per kelas.



Gbr. 6 Perbandingan Bobot Kelas Setelah Penerapan Class Weight

Berdasarkan grafik distribusi pembobotan kelas sebelum dan setelah penerapan Class Weight, dapat dilihat bahwa penalti per kelas yang awalnya seimbang berubah setelah penerapan clas weight menjadi 0,58 untuk kelas negatif dan 3,82 untuk kelas positif, dimana setelah perhitungan diberikan bobot yang lebih tinggi pada kelas minoritas (kelas positif) dengan tujuan model tidak lagi terlalu bias pada kelas mayoritas serta tetap mempertimbangkan pembelajaran dan penilaian terhadap kelas minoritas.

C. *Pengaruh Strategi Penanganan Imbalance pada Performa Model SVM+TF-IDF*

TABEL II
HASIL EKSPLORASI BALANCING PADA SVM+TF-IDF

Skema Balancing	Mean Macro F1	±Std	Akurasi	FPR	FNR
Baseline (tanpa balancing)	0,6350	0,0183	0,87	3,1%	75,2%
SVM+TF-IDF+Class Weight	0,6399	0,0232	0,82	11,4%	58,6%
SVM+TF-IDF+Oversampling	0,6327	0,0195	0,82	10,9%	61,3%
SVM+TF-IDF+SMOTE	0,5956	0,0211	0,77	17,3%	58,0%
SVM+TF-IDF+Undersampling	0,5690	0,0145	0,68	30,9%	35,8%

Tabel II menunjukkan bahwa penerapan class weighting menghasilkan Mean Macro F1 tertinggi sebesar 0,6399, sedikit lebih tinggi dibandingkan baseline tanpa balancing (0,6350). Meskipun peningkatan tersebut relatif kecil, analisis distribusi kesalahan menunjukkan perubahan yang lebih signifikan dibandingkan yang tercermin dari nilai Macro F1 saja.

Pada skenario baseline, model menghasilkan False Negative Rate (FNR) sebesar 75,2%, yang mengindikasikan bahwa terdapat tiga dari empat komentar positif gagal dikenali. Sebaliknya, False Positive Rate (FPR) hanya sebesar 3,1%, menunjukkan bahwa model cenderung memprediksi hampir semua data sebagai kelas mayoritas (negatif). Kondisi ini menyebabkan akurasi keseluruhan terlihat tinggi, namun kemampuan model dalam mengenali kelas minoritas menjadi sangat terbatas.

Penerapan class weighting mampu menurunkan nilai FNR menjadi 58,6%, yang menunjukkan peningkatan kemampuan model dalam mengenali komentar positif. Di sisi lain, nilai FPR meningkat menjadi 11,4%. Namun, peningkatan tersebut masih tergolong wajar jika dibandingkan dengan penurunan FNR yang diperoleh. Hasil ini menunjukkan bahwa keunggulan class weighting tidak hanya pada nilai F1 Macro nya yang lebih unggul namun juga keunggulannya pada distribusi kesalahan yang lebih seimbang antara kedua kelas.

Pola yang sama juga ditunjukkan oleh pendekatan random oversampling, yang memperoleh nilai Mean Macro F1 sebesar 0,6327 dengan distribusi FPR dan FNR yang mendekati class weighting. Kemiripan ini terjadi karena keduanya sama-sama meningkatkan perhatian model terhadap kelas minoritas. Perbedaannya, oversampling melakukannya dengan menggandakan sampel minoritas, sedangkan class weighting memberikan penalti yang lebih besar terhadap kesalahan pada kelas minoritas tanpa mengubah distribusi data asli.

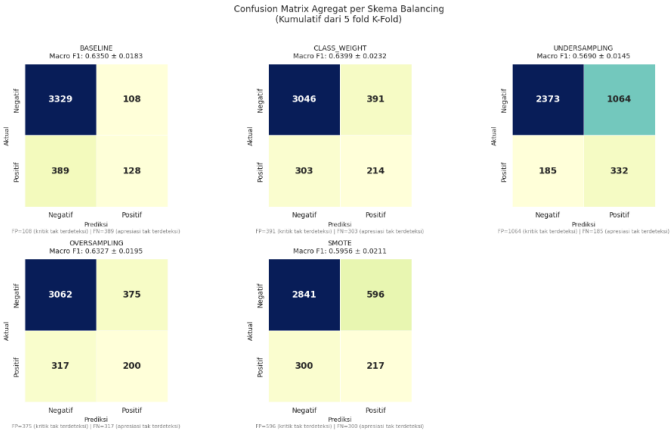
Berbeda dengan dua pendekatan tersebut, eksplorasi pada SMOTE menghasilkan Mean Macro F1 sebesar 0,5956 dan FPR sebesar 17,3%. Pada data komentar media sosial yang memiliki karakteristik cenderung pendek, informal, serta mengandung code-mixing, proses interpolasi SMOTE berpotensi menghasilkan sampel sintetik yang kurang merepresentasikan pola bahasa yang sebenarnya. Hal tersebut diduga menyebabkan batas keputusan model menjadi kurang optimal sehingga meningkatkan kesalahan klasifikasi.

Performa terendah diperoleh pada eksplorasi random undersampling, dengan Mean Macro F1 sebesar 0,5690 dan FPR tertinggi, yaitu 30,9%. Hal ini menunjukkan bahwa sekitar 1.064 komentar negatif atau kritik masyarakat justru diprediksi sebagai komentar positif. Kondisi tersebut terjadi karena undersampling mengurangi sebagian besar dari sampel negatif yang mendominasi 86,92% data, sehingga model kehilangan banyak variasi pola kritik yang diperlukan untuk mempelajari karakteristik kelas mayoritas secara memadai.

Pada konteks monitoring opini publik terhadap kinerja pemerintah, adanya kesalahan ini memiliki efek yang serius karena mengindikasikan bahwa hampir sepertiga kritik masyarakat tidak berhasil dikenali dan justru tercatat sebagai apresiasi. Oleh karena itu, meskipun perbedaan nilai F1-Macro antar strategi balancing relatif kecil, class weight dipilih sebagai strategi balancing terbaik tidak hanya berdasarkan nilai F1-Macro, namun juga mempertimbangkan distribusi kesalahan yang lebih seimbang serta risiko kesalahan interpretasi opini publik yang lebih rendah.

D. *Analisis Classification Report per Kelas*

Untuk memperoleh pemahaman yang lebih mendalam terkait pengaruh dari setiap strategi balancing terhadap masing-masing kelas, dilakukan analisis menggunakan classification report yang akan disajikan pada Tabel III. Berbeda dengan analisis pada subbab sebelumnya yang berfokus pada penilaina terhadap performa model secara keseluruhan, pembahasan pada bagian ini bertujuan mengevaluasi nilai precision, recall, dan F1-score dari setiap kelas pada setiap pengujian balancing sehingga perubahan distribusi kesalahan dapat diamati dan dinilai dengan lebih rinci.



Gbr. 7 Confusion Matrix Eksplorasi Balancing (SVM+TF IDF)

TABEL III
PERBANDINGAN CLASSIFICATION REPORT PER KELAS

Model	Kelas	Precision	Recall	F1 Score Negatif
SVM+TF IDF (Baseline)	Negatif	0,89	0,96	0,933
	Positif	0,54	0,24	0,34
SVM+TF IDF+Class Weight	Negatif	0,91	0,88	0,89
	Positif	0,35	0,41	0,38

SVM+TF IDF+ Undersampling	Negatif	0,92	0,69	0,79
	Positif	0,23	0,64	0,34
SVM+TF IDF+ Oversampling	Negatif	0,90	0,89	0,89
	Positif	0,34	0,38	0,36
SVM+TF IDF+ SMOTE	Negatif	0,90	0,82	0,86
	Positif	0,26	0,42	0,32

Berdasarkan tabel III, pada skenario baseline model menunjukkan performa yang sangat baik dalam mengenali kelas negative dengan nilai F1-Score sebesar 0,933. Namun performa pada kelas positif jauh lebih rendah dengan nilai F1-Score hanya sebesar 0,34. Hal ini mengindikasikan bahwa dengan rendahnya nilai recall positif yang hanya mencapai 0,24, menunjukkan bahwa Sebagian besar komentar positif gagal dikenali. Kondisi ini konsisten dengan tingginya nilai False Negative Rate pada Subbab C dan mengindikasikan bahwa model cenderung memprediksi komentar ke kelas mayoritas.

Penerapan class weighting meningkatkan kemampuan model dalam mengenali kelas positif. Hal ini terlihat dari kenaikan nilai recall kelas positif dari 0,24 menjadi 0,41, yang diikuti peningkatan F1-score kelas positif menjadi 0,38. Meskipun F1-score kelas negatif sedikit menurun dari 0,933 menjadi 0,89, performa kelas mayoritas masih tetap tinggi sehingga distribusi performa antar kelas menjadi lebih seimbang dibandingkan baseline dan menunjukkan bahwa model tidak lagi terlalu bias pada kelas mayoritas.

Random oversampling menunjukkan pola yang mirip dengan class weighting. Pada eksplorasi ini menunjukkan nilai F1-score kelas positif meningkat menjadi 0,36 dengan recall sebesar 0,38, sedangkan F1-score kelas negatif tetap berada pada kisaran 0,89. Hasil ini menunjukkan bahwa penambahan sampel kelas minoritas terbukti mampu meningkatkan kemampuan model dalam mengenali komentar positif tanpa menurunkan performa kelas negatif secara signifikan.

Pada SMOTE menghasilkan recall kelas positif sebesar 0,42, dan precision yang hanya mencapai 0,26. Hal ini menyebabkan nilai F1-score kelas positif tetap rendah, yaitu 0,32. Hasil ini menunjukkan bahwa peningkatan jumlah prediksi positif tidak sepenuhnya diikuti oleh peningkatan ketepatan prediksi, sehingga masih banyak komentar negatif yang salah diklasifikasikan sebagai komentar positif.

Sedangkan pada eksplorasi dengan pendekatan Random undersampling menghasilkan recall kelas positif tertinggi, yang mencapai nilai 0,64. Hal ini menunjukkan bahwa model mampu mengenali lebih banyak komentar positif apabila dibandingkan strategi lainnya. Namun, precision kelas positif hanya sebesar 0,23, sehingga sebagian besar prediksi positif merupakan kesalahan klasifikasi. Di sisi lain, recall kelas negatif turun menjadi 0,69, yang menunjukkan berkurangnya kemampuan model dalam mengenali komentar negatif akibat hilangnya sebagian besar sampel kelas mayoritas selama proses pelatihan.

Berdasarkan eksplorasi keseluruhan teknik balancing, analisis classification report menunjukkan bahwa setiap strategi balancing menghasilkan trade-off yang berbeda antara kemampuan mengenali kelas mayoritas dan kelas minoritas. Class weighting memberikan keseimbangan terbaik dengan meningkatkan kemampuan deteksi kelas positif tanpa menurunkan performa kelas negatif secara drastis. Sebaliknya, random undersampling memang menghasilkan recall kelas positif tertinggi, namun diikuti penurunan precision yang signifikan sehingga menghasilkan banyak kesalahan prediksi pada kelas negatif.

E. *Implikasi Hasil terhadap Monitoring Kinerja Pemerintah*

Dalam konteks pemanfaatan analisis sentimen sebagai sarana monitoring kinerja pemerintah, temuan penelitian ini memberikan efek yang penting. Pemilihan strategi *balancing* tidak hanya memengaruhi nilai evaluasi model, seperti *Macro F1*, tetapi juga memengaruhi jenis kesalahan yang dihasilkan selama proses klasifikasi. Pada model *baseline* tanpa *balancing*, nilai *False Negative Rate* (FNR) mencapai 75,2%. Hal ini menunjukkan bahwa sekitar tiga dari empat komentar positif yang berisi apresiasi masyarakat tidak berhasil dikenali oleh model, sehingga model cenderung lebih berpihak pada kelas mayoritas, yaitu sentimen negatif. Di sisi lain, penerapan *random undersampling* menghasilkan *False Positive Rate* (FPR) sebesar 30,9%. Artinya, hampir sepertiga komentar negatif yang berisi kritik justru diklasifikasikan sebagai sentimen positif, sehingga berpotensi terlewat dalam proses evaluasi kinerja pemerintah.

Kedua jenis kesalahan tersebut tentu memiliki konsekuensi yang berbeda dalam penerapan sistem monitoring. Namun, dalam konteks evaluasi kinerja pemerintah, FPR yang tinggi perlu mendapat perhatian lebih karena dapat memberikan gambaran yang terlalu positif terhadap kondisi sebenarnya. Akibatnya, berbagai kritik atau permasalahan yang disampaikan masyarakat berisiko tidak teridentifikasi dan tidak menjadi bahan evaluasi. Sementara itu, penerapan *class weighting* menghasilkan FPR yang jauh lebih rendah, yaitu 11,4%, serta memperoleh nilai *Macro F1* tertinggi dibandingkan metode lainnya. Hasil ini menunjukkan bahwa *class weighting* mampu memberikan keseimbangan yang lebih baik antara kemampuan mendeteksi apresiasi dan kritik masyarakat, sehingga lebih sesuai untuk mendukung proses monitoring kinerja pemerintah.

V. KESIMPULAN

Penelitian ini membandingkan lima strategi penanganan ketidakseimbangan kelas pada klasifikasi sentimen komentar media sosial berbahasa Indonesia menggunakan Stratified K-Fold (K=5). Hasil menunjukkan bahwa class weighting terbukti secara empiris sebagai strategi terbaik dengan Mean Macro F1 sebesar 0,6399 ± 0,0232, melampaui random undersampling (0,5690), random oversampling (0,6327), dan SMOTE (0,5956). Analisis *confusion matrix* secara agregat menunjukkan bahwa *random undersampling* menghasilkan *False Positive Rate* (FPR) sebesar 30,9%, sehingga hampir sepertiga komentar negatif yang berisi kritik masyarakat salah diklasifikasikan sebagai komentar positif. Dalam konteks monitoring kinerja pemerintah, kondisi ini berpotensi menyebabkan kritik yang seharusnya menjadi bahan evaluasi

justru terlewat. Di sisi lain, SMOTE belum menunjukkan kinerja yang optimal pada representasi TF-IDF karena sampel sintetis yang dihasilkan belum mampu merepresentasikan karakteristik linguistik data asli dengan baik. Oleh karena itu, class weighting direkomendasikan sebagai strategi penanganan ketidakseimbangan kelas yang lebih sesuai untuk sistem monitoring opini publik berbasis media sosial.

Untuk penelitian selanjutnya, terdapat beberapa hal yang dapat dipertimbangkan. Pertama, melakukan perbandingan dengan model berbasis *deep learning*, seperti IndoBERT atau IndoRoBERTa, untuk mengetahui apakah representasi kontekstual yang dimiliki model tersebut mampu mengatasi keterbatasan TF-IDF dalam menangani komentar yang mengandung *code-mixing* Bahasa Indonesia dan Jawa. Kedua, mengeksplorasi teknik *balancing* berbasis augmentasi teks, seperti *paraphrase* atau *back-translation*, sebagai alternatif terhadap SMOTE karena dinilai lebih mampu menghasilkan data sintetis yang tetap valid secara linguistik. Ketiga, mengembangkan klasifikasi sentimen menjadi tiga kelas, yaitu positif, netral, dan negatif, sehingga analisis opini publik dapat dilakukan secara lebih rinci dan memberikan gambaran yang lebih komprehensif.

UCAPAN TERIMA KASIH

Puji syukur dipanjatkan kepada Tuhan yang maha esa atas segala limpahan rahmat sehingga penulis dapat diberikan kelancaran dan kemudahan selama penelitian hingga penulisan artikel ini. Rasa terima kasih penulis tujukan kepada keluarga serta teman-teman atas doa dan segala dukungan moral yang menjadi kekuatan serta penyemangat penulis dalam menyelesaikan penelitian ini.

Penulis juga mengucapkan terima kasih kepada Ibu Anita Qoiriah, S.Kom., M.Kom selaku dosen pembimbing atas segala arahan, bimbingan, doa, dan kesabarannya dalam membimbing penulis selama ini. Tidak lupa ucapan terima kasih kepada Bapak Ervin Yohannes, S.Kom., M.Kom., M.Sc., Ph.D dan Ibu Dr. Yuni Yamasari, S.Kom., M.Kom selaku dosen penguji yang telah banyak memberikan masukan serta saran yang membangun bagi penelitian ini.

REFERENSI

[1] Asosiasi Penyelenggara Jasa Internet Indonesia (APJII), "Laporan Survei Pengguna Internet Indonesia 2024," Jakarta, 2024.

[2] Badan Pusat Statistik Kota Malang, "Kota Malang Dalam Angka 2024," Malang, 2024.

[3] D. F. Sebastian, H. Sulistiani, and A. R. Isnain, "SENTIMENT ANALYSIS OF PUBLIC OPINION ON THE RIGHT OF INQUIRY IN INDONESIA IN 2024 USING THE SUPPORT VECTOR MACHINE (SVM) METHOD ANALISIS SENTIMEN OPINI MASYARAKAT MENGENAI HAK ANGKET DI INDONESIA TAHUN 2024 MENGGUNAKAN METODE SUPPORT VECTOR MACHINE (SVM)," vol. 5, no. 4, pp. 1025–1034, 2024.

[4] A. Tyas, W. Setyaningsih, and Y. R. Sipayung, "Public Sentiment Analysis on the Performance of the Ministry of Finance for the 2025-2029 Period Using the Support Vector Machine (SVM) Method on Social Media Data X," vol. 10, no. 2, pp. 1338–1346, 2026.

[5] D. H. Rahman, Z. Dhiya, A. Alistin, and S. Pramana, "Analisis Sentimen dan Pemodelan Topik Opini Publik Terkait Data Badan Pusat Statistik Tahun 2024," vol. 2024, pp. 227–236, 2024.

[6] M. Fajar, D. Kurnianingtyas, P. Studi, T. Informatika, F. I. Komputer, and U. Brawijaya, "Analisis Kinerja Naive Bayes pada Klasifikasi Sentimen Twitter dengan Teknik Oversampling ADASYN," vol. 1, no. 1, pp. 1–8, 2017.

[7] I. S. Ritonga and D. Hartama, "Sentiment Classification in Imbalanced Data : Trade-Offs Between," vol. 18, no. 2, pp. 303–315, 2025.

[8] K. Sentimen, A. Davina, M. Putri, N. Sulistianingsih, and R. Rismayati, "JTIM : Jurnal Teknologi Informasi dan Multimedia Pengaruh Teknik Representasi Teks Bag-of-Words dan TF-IDF," vol. 7, no. 4, pp. 675–688, 2025.

[9] A. B. Siva and L. Hoki, "Comparison of IndoBERT and SVM Performance in Sentiment Analysis of Digital Education Platforms," vol. 10, no. 1, pp. 64–74, 2026.

[10] W. Chen, K. Yang, Z. Yu, Y. Shi, and C. L. P. Chen, *A survey on imbalanced learning : latest research , applications and future directions*, vol. C. 2024. doi: 10.1007/s10462-024-10759-6.

[11] N. F. Adhim and N. Cahyono, "Optimization of IndoBERT for Sentiment Analysis of FOMO on Social Media Through Fine-Tuning and Hybrid Labeling," *J. Appl. Informatics Comput.*, vol. 9, no. 6, pp. 3786–3797, 2025, doi: 10.30871/jaic.v9i6.11686.

[12] T. Wongvorachan and S. He, "A Comparison of Undersampling , Oversampling , and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining," 2023.

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

[14] H. I. Pratama and P. T. Prasetyaningrum, "Penerapan Support Vector Machine untuk Analisis Sentimen pada Google Review Hotel," vol. 6, no. 2, pp. 1244–1252, 2025, doi: 10.47065/josh.v6i2.6645.

[15] F. Putri, Z. Awliya, and U. Budiyanto, "ANALISIS SENTIMEN PUBLIK TWITTER DENGAN TF-IDF DAN ANALYSIS OF PUBLIC SENTIMENT ON TWITTER USING TF-IDF AND," vol. 4, no. September, pp. 346–354, 2025.

[16] A. Demircioğlu, "Applying oversampling before cross - validation will lead to high bias in radiomics," *Sci. Rep.*, pp. 1–12, 2024, doi: 10.1038/s41598-024-62585-z.