

Analisis *Trade-off* antara Performa dan Stabilitas pada Klasifikasi Sentimen Wisata Berbahasa Indonesia dengan Teknik *Data Balancing*

Nurul Chamidah*, Intan Hesti Indriana, Bambang Triwahyono, Didit Widiyanto, Indra Permana Solihin, Rio Wirawan

Fakultas Ilmu Komputer, Universitas Pembangunan Nasional "Veteran" Jakarta, Indonesia
Email: nurul.chamidah@upnvj.ac.id, hesti@upnvj.ac.id, bambang.triwahyono@upnvj.ac.id,
didit.widiyanto@upnvj.ac.id, indrapermana@upnvj.ac.id, rio.wirawan@upnvj.ac.id

Abstrak

Analisis sentimen pada dataset berukuran kecil sering menghadapi masalah ketidakseimbangan kelas yang dapat menurunkan performa klasifikasi. Berbagai teknik penanganan seperti *Synthetic Minority Oversampling Technique* (SMOTE) dan *Random Under Sampling* (RUS) telah digunakan untuk meningkatkan performa, tapi aspek stabilitas model masih relatif jarang diperhatikan. Penelitian ini bertujuan untuk menganalisis performa dan stabilitas model klasifikasi sentimen pada dataset ulasan wisata berbahasa Indonesia yang berukuran kecil dan tidak seimbang. Selain itu, penelitian ini mengintegrasikan metrik stabilitas, yaitu *Stability Sensitivity Index* (SSI) dan *Relative Robustness Score* (RRS), yang masih jarang dieksplorasi dalam klasifikasi sentimen berbahasa Indonesia. Model yang digunakan adalah *Logistic Regression* dengan representasi fitur TF-IDF. Eksperimen dilakukan menggunakan tiga skenario, yaitu tanpa *balancing*, SMOTE, dan RUS, serta dievaluasi menggunakan *5-fold cross-validation*. Hasil menunjukkan bahwa SMOTE menghasilkan F1-score tertinggi sebesar 0.926, sedangkan metode RUS memiliki stabilitas terbaik dengan standar deviasi terendah dan nilai RRS tertinggi. Temuan ini menunjukkan adanya *trade-off* antara performa dan stabilitas, di mana model dengan performa terbaik tidak selalu paling stabil. Penelitian ini memberikan kerangka evaluasi yang lebih komprehensif untuk klasifikasi sentimen pada dataset kecil dan tidak seimbang.

Kata kunci: dataset kecil, ketidakseimbangan data, klasifikasi sentimen, SMOTE, stabilitas model

Abstract

Sentiment analysis on small datasets often faces class imbalance issues that can degrade classification performance. Various techniques, such as *Synthetic Minority Oversampling Technique* (SMOTE) and *Random Under Sampling* (RUS), have been widely used to improve performance; however, model stability remains relatively underexplored. This study aims to analyze the performance and stability of sentiment classification models on a small and imbalanced Indonesian tourism review dataset. It also integrates stability metrics, namely *Stability Sensitivity Index* (SSI) and *Relative Robustness Score* (RRS), which are still rarely explored in sentiment classification for Indonesian-language datasets. *Logistic Regression* with TF-IDF feature representation is employed as the classification model. Experiments are conducted using three scenarios: no balancing, SMOTE, and RUS, evaluated through *5-fold stratified cross-validation*. The results show that SMOTE achieves the highest F1-score of 0.926, while the RUS approach provides the best stability with the lowest standard deviation and highest RRS. These findings reveal a *trade-off* between performance and stability, where the best-performing model is not necessarily the most stable. This study provides a more comprehensive evaluation framework for sentiment classification on small and imbalanced datasets.

Keywords: imbalanced data, model stability, sentiment classification, small dataset, SMOTE

This work is an open access article and licensed under a Creative Commons Attribution-NonCommercial ShareAlike 4.0 International (CC BY-NC-SA 4.0)



1. PENDAHULUAN

Perkembangan teknologi informasi dan media sosial telah menghasilkan sejumlah besar data tekstual berupa ulasan pengguna yang dapat dimanfaatkan untuk memahami opini publik. Salah satu pendekatan yang umum digunakan untuk menganalisis opini tersebut adalah analisis sentimen, yaitu proses mengidentifikasi polaritas emosi dalam teks, seperti sentimen positif atau negatif [1], [2]. Dalam

konteks pariwisata, analisis sentimen terhadap ulasan pengunjung memiliki peran penting dalam membantu pengelola destinasi memahami tingkat kepuasan serta aspek-aspek yang perlu ditingkatkan.

Namun, penerapan analisis sentimen pada dataset berukuran kecil menghadapi sejumlah tantangan, salah satunya adalah ketidakseimbangan distribusi kelas (*class imbalance*). Dataset ulasan umumnya didominasi oleh sentimen positif, sehingga model cenderung bias terhadap kelas mayoritas dan mengabaikan kelas minoritas [3], [4]. Kondisi ini dapat menyebabkan penurunan kualitas model, terutama dalam mendeteksi sentimen negatif yang justru penting sebagai bahan evaluasi layanan. Permasalahan ketidakseimbangan data ini telah menjadi perhatian luas dalam penelitian machine learning, karena dapat mempengaruhi kemampuan model dalam melakukan generalisasi [5], [6], [7].

Berbagai teknik telah dikembangkan untuk mengatasi masalah ketidakseimbangan data, di antaranya adalah metode *oversampling* seperti *Synthetic Minority Oversampling Technique* (SMOTE), serta *undersampling* seperti *Random Under Sampling* (RUS). Teknik-teknik tersebut telah terbukti mampu meningkatkan performa model klasifikasi, khususnya pada metrik *precision*, *recall*, dan *F1-score* [8], [9], [10]. Selain itu, variasi metode SMOTE yang lebih modern juga dikembangkan untuk meningkatkan kualitas data sintesis dan performa model [11]. Kombinasi teknik *balancing* tertentu juga dilaporkan dapat memberikan hasil yang lebih optimal dibandingkan penggunaan metode tunggal [12].

Meskipun demikian, banyak penelitian sebelumnya masih berfokus pada peningkatan performa model berdasarkan metrik evaluasi konvensional, seperti *precision*, *recall*, dan *F1-score*, sementara aspek stabilitas dan robustness model terhadap variasi data relatif kurang mendapat perhatian, khususnya pada klasifikasi sentimen dengan dataset berukuran kecil dan tidak seimbang. Padahal, dalam konteks penerapan nyata, stabilitas model merupakan faktor penting untuk memastikan konsistensi prediksi dan kemampuan generalisasi [13]. Variasi hasil antar percobaan juga menunjukkan bahwa evaluasi model berbasis satu kali pembagian data dapat menghasilkan kesimpulan yang kurang reliabel [14], terutama ketika variasi antar eksperimen tidak diperhitungkan secara memadai [13].

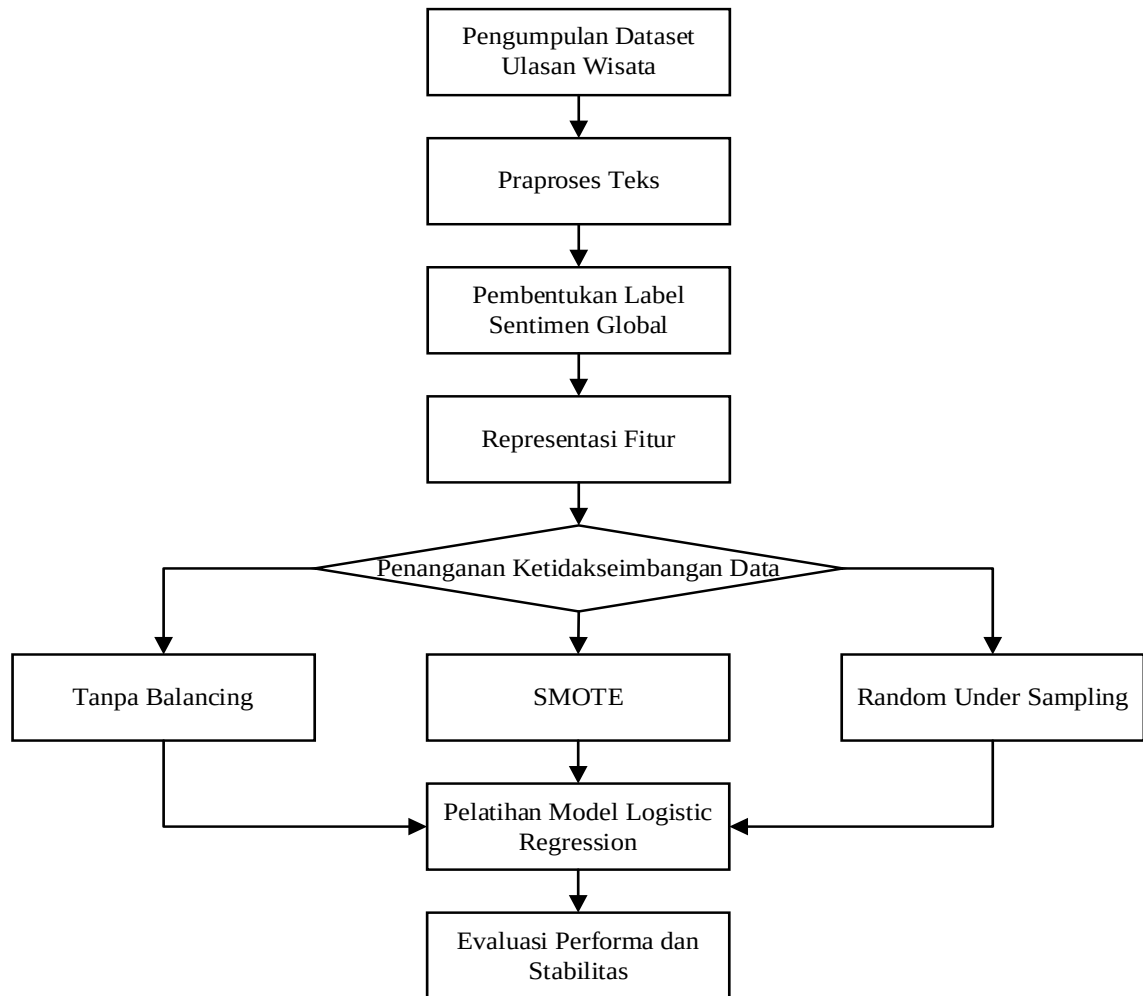
Di sisi lain, penelitian sebelumnya juga banyak menggunakan pendekatan *aspect-based sentiment analysis* (ABSA) untuk memperoleh informasi yang lebih rinci dari ulasan pengguna [15], [16]. Meskipun pendekatan ini memberikan informasi yang lebih granular, penerapannya pada dataset kecil berpotensi menyebabkan fragmentasi data dan memperburuk masalah ketidakseimbangan kelas. Hal ini dapat berdampak pada penurunan performa dan stabilitas model klasifikasi.

Berdasarkan permasalahan tersebut, penelitian ini bertujuan untuk menganalisis performa dan stabilitas model klasifikasi sentimen pada dataset berukuran kecil dengan membandingkan beberapa teknik penanganan data tidak seimbang, yaitu tanpa *balancing*, SMOTE, dan *Random Under Sampling*. Model yang digunakan adalah *Logistic Regression* dengan representasi fitur TF-IDF. Model ini dipilih karena telah terbukti efektif dan efisien dalam tugas klasifikasi teks, khususnya pada representasi berbasis TF-IDF dan dataset berukuran kecil, serta memiliki kemampuan generalisasi yang baik dibandingkan model yang lebih kompleks [17], [18]. Selain itu, kesederhanaan model ini memungkinkan interpretasi yang lebih mudah terhadap hasil klasifikasi. Selain menggunakan metrik evaluasi umum seperti *precision*, *recall*, dan *F1-score*, penelitian ini juga mengintegrasikan metrik stabilitas berupa *Stability Sensitivity Index* (SSI) dan *Relative Robustness Score* (RRS) untuk memberikan perspektif evaluasi yang lebih komprehensif.

Kontribusi utama dari penelitian ini adalah: (1) evaluasi komparatif teknik data *balancing* pada klasifikasi sentimen dataset kecil, (2) integrasi analisis stabilitas dalam evaluasi model klasifikasi sentimen, serta (3) identifikasi *trade-off* antara performa dan stabilitas model. Hasil penelitian ini diharapkan dapat memberikan panduan dalam pemilihan metode yang tepat untuk klasifikasi sentimen pada kondisi data terbatas.

2. METODE PENELITIAN

Penelitian ini bertujuan untuk menganalisis performa dan stabilitas model klasifikasi sentimen pada dataset berukuran kecil dengan kondisi ketidakseimbangan kelas. Tahapan penelitian dapat dilihat pada Gambar 1, yang meliputi pengumpulan data, praproses teks, pembentukan label sentimen global, representasi fitur menggunakan TF-IDF, pelatihan model *Logistic Regression* dengan berbagai skema *balancing*, serta evaluasi performa dan stabilitas menggunakan *cross-validation*.



Gambar 1. Tahapan Penelitian

2.1. Dataset

Dataset yang digunakan dalam penelitian ini merupakan dataset ulasan pengguna pada objek wisata di Kabupaten Indramayu yang diambil dari platform *Google Reviews*. Dataset ini mengacu pada penelitian sebelumnya [16] yang telah melakukan pengumpulan dan anotasi data berbasis aspek.

Secara keseluruhan, dataset terdiri dari 1.677 ulasan yang telah dianotasi ke dalam beberapa aspek, yaitu kebersihan, fasilitas, aksesibilitas, daya tarik, citra, harga, dan sumber daya manusia. Setiap aspek memiliki label sentimen yang diklasifikasikan ke dalam dua kategori, yaitu positif dan negatif.

2.2. Preprocessing

Tahap *preprocessing* dilakukan untuk membersihkan dan menormalisasi teks sebelum digunakan dalam proses pelatihan model. Tahapan ini bertujuan untuk mengurangi *noise* serta meningkatkan kualitas representasi fitur pada data teks [19]. Proses *preprocessing* merupakan langkah penting dalam *pipeline text mining* karena kualitas data awal sangat mempengaruhi performa model klasifikasi [1]. Proses *preprocessing* yang dilakukan meliputi:

- *Case folding*, yaitu mengubah seluruh huruf menjadi huruf kecil untuk menghindari duplikasi fitur akibat perbedaan kapitalisasi.
- Penghapusan tanda baca (*punctuation removal*) untuk menghilangkan karakter non-alfabet yang tidak memiliki kontribusi terhadap makna teks.
- Normalisasi, yaitu mengubah kata tidak baku atau informal menjadi bentuk baku menggunakan kamus normalisasi. Proses ini bertujuan untuk menyeragamkan variasi kata yang memiliki makna yang sama sehingga meningkatkan konsistensi representasi teks [20].

- *Stopword removal*, yaitu menghapus kata-kata umum yang tidak memiliki makna signifikan dalam proses klasifikasi, sehingga dapat meningkatkan fokus model terhadap kata-kata yang lebih informatif [21].
- *Stemming*, yaitu mengubah kata ke bentuk dasar untuk menyatukan variasi kata dengan makna yang sama [22]. Proses ini membantu mengurangi dimensi fitur dan meningkatkan konsistensi representasi kata dalam dokumen. Dalam penelitian ini, proses *stemming* dilakukan menggunakan library Sastrawi, yang merupakan *stemmer* berbasis aturan untuk Bahasa Indonesia [23].

Tabel 1. Contoh Tahapan *Preprocessing* Teks

Tahapan	Hasil Teks
Teks Asli	Akses lokasi hanya bisa motor, banyak pohon jd adem, ada beberapa spot foto yg bagus jg. Tiket masuk murah.
Case Folding	akses lokasi hanya bisa motor, banyak pohon jd adem, ada beberapa spot foto yg bagus jg. tiket masuk murah.
Penghapusan Tanda Baca	akses lokasi hanya bisa motor banyak pohon jd adem ada beberapa spot foto yg bagus jg tiket masuk murah
Normalisasi	akses lokasi hanya bisa motor banyak pohon jadi adem ada beberapa tempat foto yang bagus juga tiket masuk murah
Stopword Removal	akses lokasi motor banyak pohon jadi adem beberapa tempat foto bagus tiket masuk murah
Stemming	akses lokasi motor banyak pohon jadi adem beberapa tempat foto bagus tiket masuk murah

Tabel 1 menunjukkan contoh hasil dari setiap tahapan preprocessing pada satu kalimat ulasan. Proses ini dilakukan untuk mengurangi noise, menyeragamkan variasi kata tidak baku melalui normalisasi, serta menghasilkan representasi teks yang lebih konsisten sebelum tahap ekstraksi fitur.

2.3. Pembentukan Label Sentimen Global

Pada penelitian ini, anotasi berbasis aspek (*aspect-based sentiment*) dikonversi menjadi label sentimen global untuk setiap ulasan. Proses konversi ini dilakukan melalui pendekatan agregasi terhadap seluruh label sentimen pada aspek yang tersedia dalam satu ulasan. Pendekatan ini sejalan dengan konsep klasifikasi sentimen tingkat dokumen (*document-level sentiment*), di mana informasi sentimen pada berbagai aspek dapat diintegrasikan untuk merepresentasikan sentimen keseluruhan suatu teks [24], [25].

Setiap label sentimen pada aspek terlebih dahulu direpresentasikan dalam bentuk numerik, yaitu nilai 1 untuk sentimen positif dan 0 untuk sentimen negatif. Selanjutnya, label sentimen global ditentukan berdasarkan aturan agregasi berbasis *polarity dominance*, yaitu apabila terdapat minimal satu aspek dengan sentimen negatif, maka ulasan diklasifikasikan sebagai sentimen negatif. Sebaliknya, apabila seluruh aspek menunjukkan sentimen positif, maka ulasan diklasifikasikan sebagai sentimen positif. Pendekatan ini didasarkan pada prinsip bahwa sentimen negatif memiliki pengaruh yang lebih dominan terhadap persepsi keseluruhan dibandingkan sentimen positif (*negativity bias*) [26]. Oleh karena itu, keberadaan satu aspek dengan sentimen negatif dianggap cukup untuk merepresentasikan sentimen global suatu ulasan.

Pendekatan ini dipilih karena sentimen negatif pada salah satu aspek umumnya memiliki pengaruh yang lebih signifikan terhadap persepsi keseluruhan pengguna, sehingga lebih representatif dalam mencerminkan opini global. Selain itu, strategi ini juga berfungsi untuk menyederhanakan kompleksitas anotasi berbasis aspek menjadi klasifikasi biner yang lebih sesuai untuk dataset berukuran kecil. Data yang tidak memiliki informasi sentimen pada aspek tertentu atau tidak lengkap dieliminasi dari dataset untuk menjaga konsistensi dan kualitas data dalam proses pelatihan model.

Data yang tidak memiliki informasi sentimen pada seluruh aspek dieliminasi dari dataset untuk menjaga konsistensi dan kualitas data dalam proses pelatihan model. Setelah proses pembentukan label sentimen global dan pembersihan data, jumlah dataset menjadi 1.639 ulasan, dengan distribusi 1.195 data positif dan 444 data negatif. Distribusi ini menunjukkan bahwa dataset tetap didominasi oleh

sentimen positif. Meskipun pendekatan agregasi yang digunakan bersifat sensitif terhadap sentimen negatif, dominasi ulasan positif menyebabkan distribusi kelas tetap tidak seimbang.

2.4. Representasi Fitur

Teks hasil preprocessing direpresentasikan menggunakan metode *Term Frequency-Inverse Document Frequency* (TF-IDF), yang merupakan salah satu teknik pembobotan kata yang digunakan dalam *text mining* [27]. Meskipun merupakan metode klasik, TF-IDF masih digunakan dalam berbagai tugas klasifikasi teks karena kesederhanaan dan efektivitasnya, terutama pada dataset berukuran kecil [1].

Metode ini mengubah teks menjadi vektor numerik berdasarkan frekuensi kemunculan kata dalam dokumen serta tingkat kepentingannya dalam keseluruhan korpus. Secara matematis, bobot TF-IDF suatu term t pada dokumen d didefinisikan pada Persamaan (1).

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

Pada persamaan (1), $TF - IDF(t, d)$ merupakan bobot suatu term t dalam dokumen d , yang diperoleh dari perkalian antara nilai *Term Frequency* (TF) dan *Inverse Document Frequency* (IDF). Nilai *Term Frequency* (TF) dihitung menggunakan Persamaan (2).

$$TF(t, d) = f(t, d) / \sum_k f(k, d) \quad (2)$$

Pada persamaan (2), $f(t, d)$ merupakan jumlah kemunculan term t dalam dokumen d , sedangkan $\sum_k f(k, d)$ adalah total jumlah seluruh term dalam dokumen tersebut. Selanjutnya, nilai *Inverse Document Frequency* (IDF) dihitung menggunakan Persamaan (3).

$$IDF(t) = \log(N / df(t)) \quad (3)$$

Pada persamaan (3), N adalah jumlah total dokumen dalam korpus, dan $df(t)$ adalah jumlah dokumen yang mengandung term t . Nilai IDF digunakan untuk mengurangi bobot kata yang sering muncul dan meningkatkan bobot kata yang lebih spesifik. Pendekatan ini memungkinkan representasi fitur yang lebih diskriminatif dalam proses klasifikasi teks.

2.5. Model Klasifikasi

Model yang digunakan dalam penelitian ini adalah *Logistic Regression*, yaitu model klasifikasi linier yang digunakan dalam analisis teks karena kesederhanaan dan efektivitasnya [28]. Model ini dipilih karena memiliki performa yang baik pada dataset berukuran kecil serta efisien dalam proses pelatihan [1], [29]. *Logistic Regression* digunakan untuk memprediksi probabilitas suatu ulasan termasuk ke dalam kelas sentimen positif atau negatif.

Persamaan (4) merupakan fungsi sigmoid yang digunakan dalam model *Logistic Regression* untuk menghitung probabilitas suatu data termasuk ke dalam kelas tertentu.

$$\sigma(z) = 1 / (1 + e^{(-z)}) \quad (4)$$

Pada persamaan (4), $\sigma(z)$ merupakan fungsi sigmoid yang menghasilkan nilai probabilitas antara 0 dan 1. Variabel z merupakan kombinasi linear dari fitur input. Nilai z dihitung menggunakan persamaan (5) berikut:

$$z = w^T x + b \quad (5)$$

Pada persamaan (5), x merupakan vektor fitur hasil representasi TF-IDF, w adalah vektor bobot (*weight*), dan b adalah bias.

2.6. Penanganan Ketidakseimbangan Data

Masalah ketidakseimbangan kelas ditangani menggunakan tiga pendekatan, yaitu tanpa *balancing* sebagai *baseline*, *Synthetic Minority Oversampling Technique* (SMOTE) untuk menambah data pada kelas minoritas, dan *Random Under Sampling* (RUS) untuk mengurangi jumlah data pada kelas mayoritas. Permasalahan ketidakseimbangan data merupakan tantangan umum dalam *machine learning* karena dapat menyebabkan bias model terhadap kelas mayoritas dan menurunkan kemampuan generalisasi [30], [31].

SMOTE merupakan metode *oversampling* yang menghasilkan data sintesis pada kelas minoritas dengan memanfaatkan kedekatan antar sampel [32]. Selain metode klasik tersebut, berbagai pengembangan SMOTE telah diusulkan untuk meningkatkan kualitas data sintesis [11]. Metode ini telah terbukti efektif dalam meningkatkan performa klasifikasi pada data tidak seimbang [8], [10].

2.7. Skema Evaluasi

Evaluasi model dilakukan menggunakan *Stratified K-Fold Cross Validation* dengan $k = 5$, yang memastikan distribusi kelas tetap seimbang pada setiap *fold* [33] serta mengurangi bias akibat pembagian data tunggal. Selain itu, evaluasi berbasis beberapa *fold* juga penting untuk menangkap variasi performa model yang dapat terjadi akibat perbedaan subset data pelatihan [34]. Proses penanganan ketidakseimbangan data menggunakan SMOTE dan Random Under Sampling (RUS) diterapkan di dalam setiap *fold* pelatihan (*training split*) saja. Dengan demikian, data uji pada setiap *fold* tetap bersih dan tidak terpengaruh oleh proses *balancing*, sehingga menghindari terjadinya *data leakage* dan menghasilkan evaluasi yang lebih valid.

Metrik evaluasi yang digunakan meliputi *precision*, *recall*, dan *F1-score* [35]. Persamaan (6) merupakan rumus untuk menghitung nilai *precision* pada model klasifikasi. Pada persamaan (6), TP (*True Positive*) adalah jumlah data yang diklasifikasikan dengan benar sebagai kelas positif, sedangkan FP (*False Positive*) adalah jumlah data yang salah diklasifikasikan sebagai positif.

$$Precision = TP / (TP + FP) \quad (6)$$

Persamaan (7) digunakan untuk menghitung nilai *recall*. FN (*False Negative*) adalah jumlah data positif yang salah diklasifikasikan sebagai negatif.

$$Recall = TP / (TP + FN) \quad (7)$$

Selanjutnya, *F1-score* dihitung menggunakan persamaan (8). *F1-score* merupakan rata-rata harmonik dari *precision* dan *recall*.

$$F1 = 2 * (Precision * Recall) / (Precision + Recall) \quad (8)$$

Selain itu, stabilitas model diukur menggunakan *Stability Sensitivity Index* (SSI) dan *Relative Robustness Score* (RRS), yang merepresentasikan variasi dan konsistensi performa model terhadap perubahan data. Pendekatan ini digunakan untuk mengukur variasi dan konsistensi performa model antar *fold*. Rata-rata performa dihitung menggunakan persamaan (9).

$$\mu(S) = (1/k) \times \sum_{i=1}^k S_i$$

Pada persamaan (9), S_i merupakan nilai *F1-score* pada fold ke- i dan k adalah jumlah *fold*. Selanjutnya, nilai SSI dihitung menggunakan persamaan (10).

$$SSI = \sqrt{((1/k) \times \sum_{i=1}^k (S_i - \mu(S))^2)} \quad (10)$$

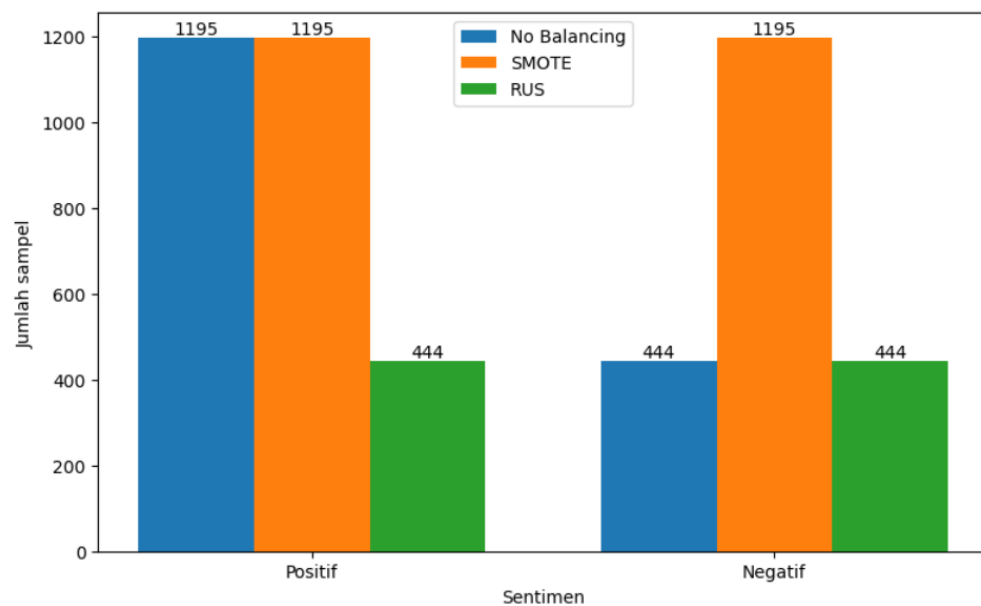
Pada persamaan (10), SSI merepresentasikan standar deviasi dari nilai *F1-score* yang menunjukkan tingkat variasi performa model. Nilai RRS dihitung menggunakan persamaan (11).

$$RRS = \mu(S) / SSI \quad (11)$$

Pada persamaan (11), $\mu(S)$ adalah rata-rata *F1-score* dan SSI adalah standar deviasi. Nilai RRS yang lebih tinggi menunjukkan model yang lebih stabil dan konsisten.

3. HASIL DAN PEMBAHASAN

Eksperimen dilakukan menggunakan tiga skenario penanganan ketidakseimbangan data, yaitu tanpa *balancing*, *Synthetic Minority Oversampling Technique* (SMOTE), dan *Random Under Sampling* (RUS). Gambar 2 menunjukkan distribusi data sebelum dan sesudah penerapan teknik *balancing*. Metode SMOTE meningkatkan jumlah data pada kelas minoritas, sedangkan metode RUS mengurangi jumlah data pada kelas mayoritas sehingga menghasilkan distribusi yang lebih seimbang.



Gambar 2. Distribusi Data Sebelum dan Sesudah *Balancing*

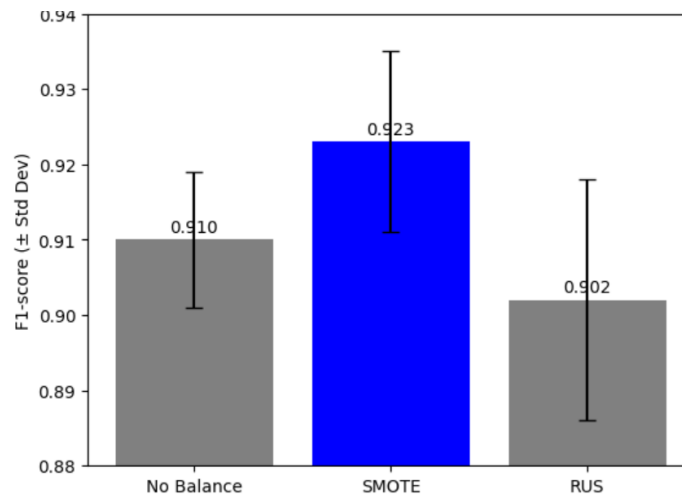
Evaluasi dilakukan menggunakan metode *5-fold cross-validation* dengan metrik *precision*, *recall*, *F1-score*, serta metrik stabilitas berupa *Stability Sensitivity Index* (SSI) atau standar deviasi, dan *Relative Robustness Score* (RRS). Hasil eksperimen dapat dilihat pada Tabel 1.

Tabel 1. Hasil Eksperimen *Balancing*

Metode	<i>F1-score</i>	Std (SSI)	<i>Precision</i>	<i>Recall</i>	RRS
No Balance	0.919	0.008	0.874	0.968	116.272
SMOTE	0.926	0.010	0.926	0.926	92.217
RUS	0.914	0.007	0.949	0.882	136.840

Berdasarkan Tabel 1, metode SMOTE menghasilkan nilai *F1-score* tertinggi sebesar 0.926, menunjukkan bahwa teknik *oversampling* efektif dalam meningkatkan performa model pada dataset yang tidak seimbang. Sementara itu, metode RUS menunjukkan nilai standar deviasi (SSI) terendah sebesar 0.007 serta nilai RRS tertinggi sebesar 136.840, yang mengindikasikan tingkat stabilitas model yang lebih baik dibandingkan metode lainnya.

Gambar 2 memperjelas perbandingan antara performa dan stabilitas model, yang menunjukkan bahwa metode SMOTE memiliki nilai *F1-score* yang lebih tinggi, namun dengan variasi performa yang lebih besar dibandingkan metode tanpa *balancing*.



Gambar 2. Perbandingan performa ($F1$ -score) dan stabilitas (standar deviasi) antar metode

3.1. Analisis Performa Model

Berdasarkan hasil eksperimen, metode SMOTE menghasilkan nilai $F1$ -score tertinggi sebesar 0.926, meningkat sebesar 0.7% dibandingkan metode tanpa balancing (0.919) dan 1.2% dibandingkan metode RUS (0.914). Hal ini menunjukkan bahwa teknik oversampling mampu memperbaiki representasi kelas minoritas, sehingga model dapat membangun batas keputusan (*decision boundary*) yang lebih seimbang antara kelas positif dan negatif.

Selain itu, metode SMOTE menunjukkan keseimbangan yang sangat baik antara precision (0.926) dan recall (0.926). Keseimbangan ini mengindikasikan bahwa model tidak hanya mampu mengidentifikasi data positif dengan akurat, tetapi juga mampu menekan kesalahan klasifikasi secara proporsional, baik *false positive* maupun *false negative*. Hal ini menjadi indikator penting dalam evaluasi model klasifikasi pada data tidak seimbang.

Sebaliknya, metode tanpa balancing menghasilkan nilai recall yang tinggi (0.968), namun precision yang lebih rendah (0.874). Kondisi ini menunjukkan bahwa model cenderung mengklasifikasikan lebih banyak data sebagai kelas positif, sehingga meningkatkan jumlah *false positive*. Fenomena ini umum terjadi pada dataset tidak seimbang, di mana model cenderung mengikuti distribusi kelas mayoritas.

Sementara itu, metode RUS menghasilkan nilai precision tertinggi (0.949), namun recall yang lebih rendah (0.882). Hal ini menunjukkan bahwa model menjadi lebih selektif dalam memprediksi kelas positif, sehingga mampu mengurangi *false positive*, tetapi pada saat yang sama meningkatkan jumlah *false negative*.

Tabel 2. Uji Signifikansi McNemar

Perbandingan	McNemar (p-value)
SMOTE vs No Balance	0.013
SMOTE vs RUS	0.034
No Balance vs RUS	0.681

Untuk mengevaluasi signifikansi perbedaan antar metode, dilakukan uji McNemar berdasarkan hasil prediksi model. Hasil pengujian menunjukkan bahwa perbedaan antara metode SMOTE dan metode tanpa balancing signifikan secara statistik ($p = 0.013$), demikian pula antara SMOTE dan RUS ($p = 0.034$). Sebaliknya, perbedaan antara metode tanpa balancing dan RUS tidak signifikan ($p = 0.681$), yang mengindikasikan bahwa kedua metode tersebut menghasilkan pola prediksi yang relatif serupa.

Secara keseluruhan, metode SMOTE menunjukkan performa terbaik dalam hal keseimbangan antara precision dan recall, serta menghasilkan perbedaan yang signifikan dibandingkan metode lainnya pada tingkat prediksi individual. Namun, perbedaan nilai $F1$ -score antar metode relatif kecil, sehingga menunjukkan bahwa peningkatan performa tidak terlalu besar pada tingkat agregat.

3.2. Analisis Stabilitas Model

Selain performa, stabilitas model merupakan aspek penting dalam evaluasi klasifikasi, terutama pada dataset berukuran kecil yang cenderung sensitif terhadap variasi data pelatihan. Dalam penelitian ini, stabilitas diukur menggunakan standar deviasi hasil *F1-score* antar *fold*, yang direpresentasikan sebagai *Stability Sensitivity Index* (SSI), serta *Relative Robustness Score* (RRS), yang mencerminkan rasio antara performa rata-rata dan variasinya.

Berdasarkan hasil eksperimen, metode *Random Under Sampling* (RUS) menunjukkan nilai standar deviasi (SSI) terendah sebesar 0.007, yang mengindikasikan bahwa performa model relatif konsisten pada setiap *fold*. Nilai ini lebih rendah dibandingkan metode tanpa *balancing* (0.008) dan *SMOTE* (0.010), yang menunjukkan bahwa RUS menghasilkan variasi performa yang paling kecil terhadap perubahan subset data pelatihan.

Jika dianalisis lebih lanjut, metode *SMOTE* memiliki nilai standar deviasi yang lebih tinggi dibandingkan metode lainnya, yang menunjukkan bahwa teknik *oversampling* cenderung menghasilkan variasi performa yang lebih besar antar *fold*. Hal ini dapat disebabkan oleh proses pembangkitan data sintetis yang memperkenalkan variasi tambahan dalam distribusi data, sehingga model menjadi lebih adaptif terhadap perubahan data, namun kurang konsisten dalam menghasilkan performa yang stabil.

Di sisi lain, metode tanpa *balancing* menunjukkan tingkat stabilitas yang relatif baik, meskipun tidak sebaik metode *RUS*. Hal ini menunjukkan bahwa distribusi data asli masih mampu menghasilkan performa yang cukup konsisten tanpa adanya intervensi pada komposisi data pelatihan.

Stabilitas model juga dapat dianalisis melalui nilai *Relative Robustness Score* (RRS). Metode *RUS* menghasilkan nilai RRS tertinggi sebesar 136.840, yang menunjukkan rasio performa terhadap variasi yang paling optimal. Sementara itu, metode tanpa *balancing* dan *SMOTE* memiliki nilai RRS yang lebih rendah, masing-masing sebesar 116.272 dan 92.217.

Temuan ini menunjukkan bahwa model dengan performa tertinggi tidak selalu memiliki stabilitas terbaik. Dalam konteks dataset berukuran kecil, stabilitas menjadi faktor krusial karena model yang tidak stabil cenderung menghasilkan prediksi yang tidak konsisten ketika dihadapkan pada variasi data baru. Dengan demikian, metode *RUS* memberikan tingkat stabilitas terbaik, meskipun tidak menghasilkan performa tertinggi, sehingga menunjukkan adanya *trade-off* yang signifikan antara performa dan stabilitas dalam klasifikasi sentimen pada dataset kecil.

3.3. Trade-off antara Performa dan Stabilitas

Hasil eksperimen menunjukkan adanya *trade-off* yang jelas antara performa dan stabilitas model pada klasifikasi sentimen dengan dataset berukuran kecil dan tidak seimbang. Metode *SMOTE* menghasilkan nilai *F1-score* tertinggi, namun menunjukkan tingkat stabilitas yang lebih rendah dibandingkan metode lainnya. Sebaliknya, metode *Random Under Sampling* (RUS) menghasilkan performa yang sedikit lebih rendah, tetapi memiliki stabilitas yang paling tinggi, yang ditunjukkan oleh nilai standar deviasi (SSI) paling kecil serta nilai *Relative Robustness Score* (RRS) paling besar.

Meskipun perbedaan nilai *F1-score* antar metode relatif kecil, hasil uji McNemar menunjukkan bahwa metode *SMOTE* menghasilkan perbedaan yang signifikan pada tingkat prediksi individual dibandingkan metode lainnya ($p < 0,05$). Hal ini mengindikasikan bahwa *SMOTE* mampu mempengaruhi pola klasifikasi secara lebih sensitif terhadap data minoritas, meskipun peningkatan performa agregat tidak terlalu besar.

Di sisi lain, metode tanpa *balancing* menunjukkan performa dan stabilitas yang berada pada tingkat menengah dibandingkan metode lainnya. Sementara itu, metode *RUS* menunjukkan bahwa pengurangan data pada kelas mayoritas tidak selalu berdampak negatif terhadap stabilitas model, melainkan dapat menghasilkan performa yang lebih konsisten antar *fold*. Hal ini mengindikasikan bahwa dalam kondisi tertentu, pengurangan jumlah data dapat membantu mengurangi variasi model terhadap perubahan data pelatihan.

Implikasi dari temuan ini sangat penting dalam konteks aplikasi nyata. Pada sistem analisis sentimen untuk evaluasi pariwisata, kestabilan model menjadi krusial karena hasil analisis digunakan sebagai dasar pengambilan keputusan. Model dengan performa tinggi namun tidak stabil berpotensi menghasilkan rekomendasi yang inkonsisten ketika diterapkan pada data baru.

Dengan demikian, pemilihan metode terbaik tidak hanya bergantung pada nilai performa, tetapi juga harus mempertimbangkan stabilitas model. Pada kondisi dataset berukuran kecil, metode *RUS*

dapat menjadi pilihan yang lebih tepat ketika konsistensi hasil menjadi prioritas utama, sedangkan SMOTE lebih sesuai digunakan ketika fokus utama adalah peningkatan performa klasifikasi. Temuan ini menegaskan bahwa terdapat *trade-off* inheren antara performa dan stabilitas, sehingga evaluasi model klasifikasi sentimen perlu mempertimbangkan kedua aspek tersebut secara simultan untuk memperoleh model yang optimal dan andal.

4. KESIMPULAN

Penelitian ini bertujuan untuk menganalisis performa dan stabilitas model klasifikasi sentimen pada dataset berukuran kecil dengan kondisi ketidakseimbangan kelas. Hasil eksperimen menunjukkan bahwa metode SMOTE menghasilkan nilai *F1-score* tertinggi, yang mengindikasikan kemampuannya dalam meningkatkan keseimbangan performa antara kelas mayoritas dan minoritas. Namun, metode ini menunjukkan tingkat stabilitas yang lebih rendah dibandingkan metode lainnya.

Sebaliknya, metode *Random Under Sampling* (RUS) menghasilkan tingkat stabilitas terbaik, yang ditunjukkan oleh nilai standar deviasi (SSI) paling kecil serta nilai *Relative Robustness Score* (RRS) paling tinggi, meskipun performanya sedikit lebih rendah dibandingkan SMOTE. Sementara itu, metode tanpa *balancing* menunjukkan performa dan stabilitas yang berada pada tingkat menengah.

Temuan utama penelitian ini adalah adanya *trade-off* yang jelas antara performa dan stabilitas model. Peningkatan performa melalui teknik *balancing* tidak selalu diikuti dengan peningkatan stabilitas, sehingga model dengan performa tertinggi belum tentu menjadi model yang paling andal pada kondisi data yang bervariasi.

Lebih lanjut, hasil analisis menunjukkan bahwa perbedaan kinerja model juga bergantung pada perspektif evaluasi yang digunakan. Meskipun perbedaan nilai *F1-score* antar metode relatif kecil, uji berbasis prediksi menunjukkan adanya perbedaan signifikan pada tingkat individual. Hal ini mengindikasikan bahwa perubahan pola prediksi tidak selalu tercermin secara jelas dalam metrik agregat, sehingga evaluasi model klasifikasi perlu mempertimbangkan kedua perspektif tersebut secara komplementer.

Secara keseluruhan, penelitian ini menunjukkan bahwa pemilihan metode klasifikasi sentimen pada dataset berukuran kecil tidak dapat hanya didasarkan pada performa, tetapi juga harus mempertimbangkan stabilitas dan tujuan penggunaan model. Metode SMOTE lebih sesuai digunakan ketika prioritas utama adalah peningkatan performa, sedangkan RUS lebih tepat digunakan ketika konsistensi hasil menjadi faktor utama.

Untuk penelitian selanjutnya, disarankan untuk mengeksplorasi model klasifikasi yang lebih kompleks, pendekatan *balancing* yang lebih adaptif, serta menguji stabilitas model pada dataset yang lebih besar dan beragam. Selain itu, penelitian lanjutan dapat mengkaji integrasi metode evaluasi berbasis stabilitas pada berbagai domain dan jenis data untuk memperoleh generalisasi yang lebih luas.

DAFTAR PUSTAKA

- [1] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 8, no. 4, 2018, doi: 10.1002/widm.1253.
- [2] M. K. Chandan and S. Mandal, "A comprehensive survey on sentiment analysis: Framework, techniques, and applications," *Comput. Sci. Rev.*, vol. 58, p. 100777, Nov. 2025, doi: 10.1016/J.COSREV.2025.100777.
- [3] M. Rezapour, "Sentiment classification of skewed shoppers' reviews using machine learning techniques," *Engineering Reports*, 2020.
- [4] M. Lango, "Tackling the Problem of Class Imbalance in Multi-class Sentiment Classification," *Foundations of Computing and Decision Sciences*, 2019.
- [5] W. Chen, K. Yang, Z. Yu, Y. Shi, and C. L. P. Chen, "A survey on imbalanced learning: latest research, applications and future directions," *Artif. Intell. Rev.*, vol. 57, no. 6, p. 137, 2024, doi: 10.1007/s10462-024-10759-6.
- [6] M. M. Khan and M. Alkhathami, "Anomaly detection in IoT-based healthcare: machine learning for enhanced security," *Sci. Rep.*, vol. 14, no. 1, p. 5872, 2024, doi: 10.1038/s41598-024-56126-x.

- [7] J. Park, S. Kwon, and S.-P. Jeong, "A study on improving turnover intention forecasting by solving imbalanced data problems: focusing on SMOTE and generative adversarial networks," *J. Big Data*, vol. 10, no. 1, p. 36, 2023, doi: 10.1186/s40537-023-00715-6.
- [8] D. Andriyani, A. Faqih, and S. E. Permana, "The Effect of SMOTE Application on SVM Performance in Sentiment Classification," *Journal of Artificial Intelligence and Engineering Applications*, 2025.
- [9] N. S. Sediarmoko, Y. Nataliani, and I. Suryady, "Sentiment Analysis of Customer Review Using Classification Algorithms and SMOTE," *Indonesian Journal of Information Systems*, 2024.
- [10] M. Al-Khazaleh, M. Alian, and M. Jaradat, "Sentiment analysis of imbalanced Arabic data using sampling techniques," *Bulletin of Electrical Engineering and Informatics*, 2024.
- [11] G. Douzas and F. Bacao, "Geometric SMOTE: a geometrically enhanced drop-in replacement for SMOTE," *Inf. Sci. (N. Y.)*, vol. 501, pp. 118–135, 2019, doi: 10.1016/j.ins.2019.06.007.
- [12] A. Saekhu, B. Berilana, and D. Saputra, "Comparative Analysis of Data Balancing Techniques for Machine Learning Classification," *Jurnal Teknik Informatika*, 2025.
- [13] X. Bouthillier, C. Laurent, and P. Vincent, "Accounting for variance in machine learning benchmarks," *Proceedings of Machine Learning and Systems*, vol. 3, pp. 747–772, 2021, doi: 10.48550/arXiv.2106.04838.
- [14] C. Nadeau and Y. Bengio, "Inference for the generalization error," *Mach. Learn.*, vol. 52, no. 3, pp. 239–281, 2003, doi: 10.1023/A:1024068626366.
- [15] H. Do, P. W. C. Prasad, A. Maag, and A. Alsadoon, "Deep learning for aspect-based sentiment analysis: A comparative review," *Expert Syst. Appl.*, vol. 118, pp. 272–299, 2019, doi: 10.1016/j.eswa.2018.10.003.
- [16] N. Chamidah, D. Widiyanto, H. B. Seta, and A. A. Aziz, "The impact of oversampling and undersampling on aspect-based sentiment analysis of Indramayu tourism using logistic regression," *Revue d'Intelligence Artificielle*, vol. 38, no. 3, pp. 795–804, 2024, doi: 10.18280/ria.380306.
- [17] Ni Made Gita Satviki Nirmala and Ngurah Agus Sanjaya ER, "Analisis Sentimen dengan Logistic Regression untuk Deteksi Kata pada Livin' by Mandiri," *Jurnal Nasional Teknologi Informasi dan Aplikasinya*, vol. 2, no. 4, pp. 869–878, 2024, doi: 10.24843/JNATIA.2024.v02.i04.p25.
- [18] H. P. Singh, N. Singh, A. Mishra, S. K. Sen, M. Swarnkar, and D. Pandey, "Logistic Regression based Sentiment Analysis System: Rectify," in *2024 IEEE International Conference on Big Data & Machine Learning (ICBDML)*, 2024, pp. 186–191. doi: 10.1109/ICBDML60909.2024.10577296.
- [19] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008.
- [20] Md. S. Islam *et al.*, "Challenges and future in deep learning for sentiment analysis: a comprehensive review and a proposed novel hybrid approach," *Artif. Intell. Rev.*, vol. 57, no. 3, p. 62, 2024, doi: 10.1007/s10462-023-10651-9.
- [21] S. Dixit, W. Mao, K. K. McDade, M. Schäferhoff, O. Ogbuaji, and G. Yamey, "Tracking financing for global common goods for health: A machine learning approach using natural language processing techniques," *Front. Public Health*, vol. Volume 10-2022, 2022, [Online]. Available: <https://www.frontiersin.org/journals/public-health/articles/10.3389/fpubh.2022.1031147>
- [22] M. F. Porter, "An algorithm for suffix stripping," *Program*, 1980.
- [23] A. P. Pamungkas, A. Mahendra, I. Wahid, and F. Aziz, "Comparison of Machine Learning Models for Classifying Consumer Sentiment of Coffee Shops on Social Media X," *Jurnal Teknik Pertanian Lampung (Journal of Agricultural Engineering)*, vol. 14, no. 5, pp. 1905–1912, Oct. 2025, doi: 10.23960/JTEPL.V14I5.1905-1912.
- [24] K. Schouten and F. Frasincar, "Aspect-based sentiment analysis: A survey," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 3, pp. 813–830, 2016, doi: 10.1109/TKDE.2015.2485209.
- [25] X. Huang, J. Li, J. Wu, J. Chang, and D. Liu, "Transfer Learning With Document-Level Data Augmentation for Aspect-Level Sentiment Classification," *IEEE Trans. Big Data*, vol. 9, no. 6, pp. 1643–1657, 2023, doi: 10.1109/TBDATA.2023.3310267.

- [26] R. F. et al. Baumeister, "Bad is stronger than good," *Review of General Psychology*, vol. 5, no. 4, pp. 323–370, 2001, doi: 10.1037/1089-2680.5.4.323.
- [27] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Inf. Process. Manag.*, 1988.
- [28] D. Hosmer and S. Lemeshow, *Applied Logistic Regression*. Wiley, 2013.
- [29] A. Genkin, D. Lewis, and D. Madigan, "Large-scale Bayesian logistic regression for text categorization," in *Technometrics*, 2007.
- [30] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," *Progress in Artificial Intelligence*, vol. 5, no. 4, pp. 221–232, 2016, doi: 10.1007/s13748-016-0094-0.
- [31] H. He and Y. Ma, "Learning from class-imbalanced data: Review of methods and applications," *Expert Syst. Appl.*, vol. 73, pp. 220–239, 2017, doi: 10.1016/j.eswa.2016.12.035.
- [32] N. V. et al. Chawla, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, 2002.
- [33] A. O. Mostafa and T. M. Ahmed, "Enhanced Emotion Analysis Model using Machine Learning in Saudi Dialect: COVID-19 Vaccination Case Study," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 1, 2024, doi: 10.14569/IJACSA.2024.0150134.
- [34] N. Sathiparan, S. H. Wijekoon, P. Jeyananthan, and D. N. Subramaniam, "Prediction of characteristics of pervious concrete by machine learning technique using mix parameters and non-destructive test measurements," *Nondestructive Testing and Evaluation*, vol. 41, no. 1, pp. 314–363, Jan. 2026, doi: 10.1080/10589759.2025.2468273.
- [35] D. Powers, "Evaluation: From precision, recall and F-measure to ROC," *Journal of Machine Learning Technologies*, 2011.