Perbandingan Kinerja Algoritma Machine Learning pada Sentimen #KaburAjaDulu dengan Penanganan Ketidakseimbangan Data Menggunakan SMOTE

Alisha Sumahesa[1], Cahyono Budy Santoso[2*]
[1,2*]Sistem Informasi / Fakultas Teknologi dan Desain, Universitas Pembangunan Jaya
*email: cahyono.budy@upj.ac.id

DOI: https://doi.org/10.31603/komtika.v10.i1.16851

ABSTRACT

The phenomenon of labor migration abroad has become a widely discussed social issue on social media, particularly through the hashtag #KaburAjaDulu on the X platform. This study aims to analyze public sentiment toward the phenomenon using machine learning classification methods with the implementation of Synthetic Minority Over-sampling Technique (SMOTE) to address data imbalance. The research was conducted using 1,750 data collected from the X platform through several stages, including data collection, text preprocessing, sentiment labeling, TF-IDF weighting, SMOTE implementation, and classification using Support Vector Machine (SVM), Naive Bayes, Random Forest, and Logistic Regression algorithms. Model evaluation was carried out using accuracy, precision, recall, and f1-score metrics. The results show that the implementation of SMOTE significantly improved classification performance. Logistic Regression achieved the best performance with an accuracy of 91.36%, followed by Random Forest at 90.30%, Support Vector Machine at 88.89%, and Naive Bayes at 82.89%. These findings indicate that Logistic Regression has the best capability in recognizing sentiment patterns within unstructured social media data. This study proves that data balancing using SMOTE plays an important role in improving sentiment classification performance and in understanding public opinion regarding social phenomena developing in digital media.

Keywords: *Sentiment Analysis, Machine Learning, SMOTE, Logistic Regression, Support Vector Machine.*

ABSTRAK

Fenomena migrasi tenaga kerja ke luar negeri menjadi salah satu isu sosial yang banyak diperbincangkan di media sosial, khususnya melalui tagar #KaburAjaDulu pada platform X. Penelitian ini bertujuan untuk menganalisis sentimen publik terhadap fenomena tersebut menggunakan metode klasifikasi machine learning dengan penerapan Synthetic Minority Over-sampling Technique (SMOTE) untuk mengatasi ketidakseimbangan data. Penelitian dilakukan menggunakan 1.750 data yang diperoleh dari media sosial X melalui tahapan pengumpulan data, preprocessing teks, pelabelan sentimen, pembobotan TF-IDF, penerapan SMOTE, serta klasifikasi menggunakan algoritma Support Vector Machine (SVM), Naive Bayes, Random Forest, dan Logistic Regression. Evaluasi model dilakukan menggunakan metrik accuracy, precision, recall, dan f1-score. Hasil penelitian menunjukkan bahwa penerapan SMOTE mampu meningkatkan performa klasifikasi secara signifikan. Algoritma Logistic Regression memperoleh performa terbaik dengan accuracy sebesar 91,36%, diikuti Random Forest sebesar 90,30%, Support Vector Machine sebesar 88,89%, dan Naive Bayes sebesar 82,89%. Hasil tersebut menunjukkan bahwa Logistic Regression memiliki kemampuan terbaik dalam mengenali pola sentimen pada data media sosial yang tidak terstruktur. Penelitian ini membuktikan bahwa penyeimbangan data menggunakan SMOTE berperan penting dalam meningkatkan kualitas klasifikasi sentimen serta membantu memahami opini publik terhadap fenomena sosial yang berkembang di media digital..

Kata Kunci: Analisis Sentimen, Machine Learning, SMOTE, Logistic Regression, Support Vector Machine.

PENDAHULUAN

Fenomena migrasi tenaga kerja ke luar negeri semakin menjadi perhatian masyarakat Indonesia dalam beberapa tahun terakhir. Berbagai faktor sosial dan ekonomi, seperti keterbatasan lapangan pekerjaan, ketidaksesuaian antara tingkat pendapatan dan kebutuhan hidup, serta keinginan memperoleh peluang karier yang lebih baik mendorong masyarakat mempertimbangkan bekerja di luar negeri sebagai alternatif peningkatan kesejahteraan [1]. Perkembangan media sosial turut memperluas ruang diskusi publik mengenai fenomena tersebut. Salah satu yang menarik perhatian adalah munculnya tagar #KaburAjaDulu pada platform X yang digunakan oleh masyarakat untuk menyampaikan pandangan, aspirasi, maupun kritik terhadap kondisi sosial dan ekonomi di dalam negeri. Tingginya aktivitas pengguna pada tagar tersebut menunjukkan bahwa media sosial dapat dimanfaatkan sebagai sumber data untuk memahami opini publik terkait fenomena migrasi tenaga kerja.

Analisis sentimen merupakan salah satu pendekatan dalam Natural Language Processing (NLP) yang digunakan untuk mengidentifikasi kecenderungan opini masyarakat terhadap suatu isu berdasarkan data teks [2]. Pemanfaatan analisis sentimen pada media sosial telah banyak dilakukan karena mampu menghasilkan informasi yang berguna untuk memahami persepsi publik secara cepat dan dalam skala besar [3]. Namun, data media sosial memiliki karakteristik yang kompleks karena mengandung bahasa tidak baku, singkatan, slang, emotikon, serta berbagai bentuk noise yang dapat memengaruhi kualitas hasil klasifikasi. Oleh karena itu, diperlukan tahapan preprocessing seperti cleansing, case folding, tokenizing, stopword removal, dan stemming untuk menghasilkan data yang lebih terstruktur dan siap digunakan dalam proses klasifikasi [4].

Selain permasalahan kualitas data, analisis sentimen pada media sosial juga sering menghadapi ketidakseimbangan distribusi kelas (imbalanced dataset), yaitu kondisi ketika jumlah data pada suatu kelas sentimen jauh lebih dominan dibandingkan kelas lainnya. Ketidakseimbangan data dapat menyebabkan model klasifikasi cenderung menghasilkan prediksi yang bias terhadap kelas mayoritas sehingga kemampuan model dalam mengenali kelas minoritas menjadi menurun. Untuk mengatasi permasalahan tersebut, beberapa penelitian telah menerapkan Synthetic Minority Oversampling Technique (SMOTE) yang bekerja dengan menghasilkan sampel sintetis pada kelas minoritas sehingga distribusi data menjadi lebih seimbang [5]. Penerapan SMOTE dilaporkan mampu meningkatkan performa model klasifikasi, terutama pada nilai precision, recall, dan F1-score pada berbagai kasus analisis sentimen [6].

Seiring dengan perkembangan machine learning, berbagai algoritma telah digunakan untuk melakukan klasifikasi sentimen, di antaranya Naïve Bayes, Logistic Regression, Random Forest, dan Support Vector Machine (SVM). SVM merupakan salah satu algoritma yang banyak digunakan pada klasifikasi teks karena mampu menangani data berdimensi tinggi serta memiliki kemampuan generalisasi yang baik dalam menemukan hyperplane optimal untuk memisahkan kelas data [7], [8]. Naïve Bayes banyak digunakan sebagai baseline karena memiliki proses komputasi yang sederhana dan efisien untuk data teks berukuran besar [9]. Logistic Regression dipilih karena merupakan algoritma klasifikasi linear yang memiliki kemampuan interpretasi yang baik dan sering menunjukkan performa yang kompetitif pada tugas analisis sentimen [10]. Sementara itu, Random Forest merepresentasikan pendekatan ensemble learning yang mampu mengurangi risiko overfitting

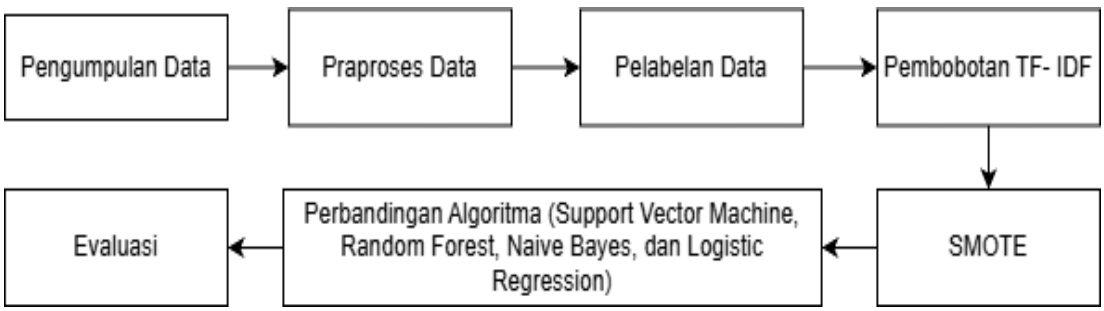
dan menangani kompleksitas data melalui kombinasi banyak decision tree [11]. Perbedaan karakteristik tersebut menyebabkan performa masing-masing algoritma dapat berbeda ketika diterapkan pada data sentimen media sosial.

Meskipun penelitian analisis sentimen menggunakan SVM, Naïve Bayes, Logistic Regression, dan Random Forest telah banyak dilakukan, sebagian besar penelitian masih mengevaluasi algoritma secara terpisah atau tanpa mempertimbangkan permasalahan ketidakseimbangan distribusi kelas yang umum terjadi pada data media sosial. Di sisi lain, penelitian yang menerapkan Synthetic Minority Oversampling Technique (SMOTE) umumnya hanya berfokus pada peningkatan performa satu algoritma tertentu sehingga efektivitas SMOTE terhadap berbagai pendekatan klasifikasi, seperti probabilistik, linear classifier, dan ensemble learning, masih belum banyak dikaji secara komprehensif. Akibatnya, pengaruh penerapan SMOTE terhadap kinerja masing-masing algoritma dalam analisis sentimen media sosial masih belum diketahui secara jelas.

Berdasarkan celah penelitian tersebut, penelitian ini bertujuan untuk menganalisis sentimen publik terhadap fenomena #KaburAjaDulu pada platform X dengan menerapkan SMOTE untuk mengatasi ketidakseimbangan distribusi kelas pada data sentimen. Selanjutnya, penelitian ini membandingkan kinerja algoritma Support Vector Machine (SVM), Naïve Bayes, Logistic Regression, dan Random Forest pada dataset yang telah diseimbangkan untuk mengidentifikasi algoritma yang paling efektif dalam klasifikasi sentimen. Kontribusi penelitian ini terletak pada penyajian analisis komparatif beberapa algoritma machine learning dengan karakteristik yang berbeda pada data sentimen berbahasa Indonesia yang telah melalui proses penyeimbangan kelas menggunakan SMOTE.

METODE

Penelitian ini menggunakan algoritma machine learning untuk menganalisis sentimen publik terhadap fenomena #KaburAjaDulu pada platform X. Tahapan penelitian meliputi Frequency-Inverse Document Frequency (TF-IDF), penyeimbangan data menggunakan Synthetic Minority Oversampling Technique (SMOTE), klasifikasi sentimen menggunakan algoritma machine learning, serta evaluasi [12]. Alur penelitian yang digunakan ditunjukkan pada Gambar 1.



Gambar 1. Alur penelitian untuk perancangan model SVM

Proses ini merupakan pengambilan data dari sumber yang digunakan dalam penelitian, yaitu media sosial X. Data dikumpulkan berdasarkan kata kunci yang mengandung tagar

#KaburAjaDulu. Pengumpulan data dilakukan menggunakan *web scraping* dan hasil dari tahap ini berupa data mentah yang masih belum terstruktur [13].

Tahap ini bertujuan untuk meningkatkan kualitas data teks sebelum proses klasifikasi. Tahapan yang diterapkan meliputi cleaning, case folding, normalisasi, tokenisasi, stopword removal, dan stemming. Tahap cleaning digunakan untuk menghapus elemen yang tidak relevan, sedangkan case folding dan normalisasi bertujuan menyeragamkan bentuk kata. Selanjutnya, tokenisasi memecah teks menjadi token kata, stopword removal menghilangkan kata yang kurang informatif, dan stemming mengubah kata menjadi bentuk dasarnya. Tahapan ini menghasilkan data yang lebih terstruktur untuk proses ekstraksi fitur dan klasifikasi.

Data diberi label positif, negatif, dan netral berdasarkan konteks dan makna unggahan. Proses pelabelan ini dilakukan secara otomatis dengan bantuan bahasa pemrograman *python* untuk menjaga konsistensi data [14]. Pelabelan data yang tepat sangat penting karena menjadi dasar pelatihan model klasifikasi.

Tahap pembobotan TF-IDF mengubah data teks menjadi bentuk numerik menggunakan metode TF-IDF (*Term Frequency-Inverse Document Frequency*). Setiap kata diberi bobot berdasarkan tingkat kepentingannya dalam dokumen dan keseluruhan data. TF-IDF dipilih karena efektif dalam analisis teks dan sering digunakan dalam klasifikasi sentimen [12].

Penggunaan SMOTE (*Synthetic Minority Over-sampling Technique*) terhadap data yaitu untuk mengatasi ketidakseimbangan jumlah data pada masing-masing kelas sentimen [5]. Data akan dibagi menjadi data latih dan data uji serta direpresentasikan menggunakan TF-IDF, SMOTE diterapkan pada data latih untuk menghasilkan sampel sintetis pada kelas minoritas. Pembentukan data sintetis dilakukan berdasarkan metode k-nearest neighbors, yaitu dengan menghitung kedekatan antar data pada kelas minoritas kemudian membentuk sampel baru melalui interpolasi nilai fitur antara data dan tetangga terdekatnya. Hasil proses ini menghasilkan distribusi kelas yang lebih seimbang sehingga dapat mengurangi bias model terhadap kelas mayoritas dan meningkatkan kemampuan klasifikasi pada kelas minoritas [6].

Pada penelitian ini dilakukan perbandingan performa empat algoritma machine learning, yaitu Support Vector Machine (SVM), Naïve Bayes, Logistic Regression, dan Random Forest. Keempat algoritma tersebut dipilih karena mewakili pendekatan klasifikasi yang berbeda. SVM dan Logistic Regression mewakili pendekatan linear classifier yang banyak digunakan pada klasifikasi teks berdimensi tinggi [7], [10]. Naïve Bayes mewakili pendekatan probabilistik yang sederhana dan efisien, sedangkan Random Forest mewakili pendekatan ensemble learning yang mampu menangkap pola data yang lebih kompleks. Penggunaan algoritma dengan karakteristik yang berbeda bertujuan untuk memperoleh gambaran yang lebih komprehensif mengenai efektivitas masing-masing algoritma dalam mengklasifikasikan sentimen pada dataset yang telah diseimbangkan menggunakan SMOTE. Seluruh algoritma dilatih dan dievaluasi menggunakan dataset yang sama sehingga hasil perbandingan dapat dilakukan secara objektif dan konsisten.

Tahapan ini dilakukan untuk mengukur kinerja model klasifikasi. Pengukuran dilakukan menggunakan metrik seperti akurasi, *precision*, *recall*, dan *F1-score*. Hasil evaluasi digunakan untuk membandingkan performa model sebelum dan setelah penerapan SMOTE, sehingga dapat diketahui pengaruh metode tersebut terhadap peningkatan kinerja model.

HASIL DAN PEMBAHASAN

Bagian ini menguraikan hasil penelitian berdasarkan tahapan yang telah dirancang dalam alur penelitian, mulai dari pengumpulan data hingga evaluasi model. Setiap tahapan dianalisis untuk memahami kontribusinya terhadap kinerja model klasifikasi sentimen yang dihasilkan.

Data penelitian dikumpulkan menggunakan teknik web scraping melalui pustaka Tweet Harvest yang memanfaatkan API X menggunakan bahasa pemrograman Python [13]. Proses scraping dilakukan dengan memasukkan kata kunci tagar #KaburAjaDulu untuk memperoleh unggahan yang relevan dari platform X. Pengambilan data dilakukan pada periode Februari 2025 yang merupakan puncak tren percakapan terkait tagar tersebut. Hasil scraping menghasilkan lebih 2.104 data mentah yang kemudian melalui proses seleksi dan validasi sehingga diperoleh 1.750 data yang digunakan dalam penelitian.

Praproses Data dilakukan untuk meningkatkan kualitas data teks sebelum digunakan dalam proses klasifikasi sentimen. Mengingat data bersumber dari media sosial X yang cenderung bersifat tidak terstruktur dan informal, diperlukan serangkaian tahapan untuk mengurangi *noise* serta menyeragamkan bentuk teks [9].

1. *Cleaning*

   Pada tahap ini, teks dibershihkan dari elemen-elemen yang tidak relevan, seperti tanda baca berlebih, URL, simbol, angka, *mention*, dan elemen lain yang tidak merepresentasikan opini pengguna, seperti pada kata "#KaburAjaDulu" menjadi "KaburAjaDulu". Proses ini menghasilkan teks yang lebih bersih dan mudah diproses pada tahap selanjutnya.

Tabel 1. Hasil *Data Cleaning*

No	Sebelum *Cleaning*	Setelah *Cleaning*
1	Ngelihat Indonesia gini aku yang berjuang di akar rasanya ingin menyerah juga. Ingin #KaburAjaDulu tapi masa ninggalin Bangsaku dan keluarga? Meski di LN juga tetep bisa terus nasionalis hanya saja.... Nanti siapa yang tersisa?	Ngelihat Indonesia gini aku yang berjuang di akar rasanya ingin menyerah juga Ingin KaburAjaDulu tapi masa ninggalin Bangsaku dan keluarga Meski di LN juga tetep bisa terus nasionalis hanya saja Nanti siapa yang tersisa

2. *Case folding*

   Seluruh teks dikonversi menjadi huruf kecil (*lowercase*) untuk menyeragamkan bentuk kata, contohnya pada Tabel 2 kata "Ngelihat" masih menggunakan huruf kapital diawal kata, lalu diubah menjadi "ngelihat" agar dapat mengurangi redudansi fitur dan meningkatkan konsistensi data.

Tabel 2. Hasil *Case Folding*

No	Sebelum *Case Folding*	Setelah *Case Folding*
1	Ngelihat Indonesia gini aku yang berjuang di akar rasanya ingin menyerah juga. Ingin #Ngelihat Indonesia gini aku yang berjuang di akar rasanya ingin menyerah juga Ingin KaburAjaDulu tapi masa ninggalin Bangsaku dan keluarga Meski di LN juga tetep bisa terus nasionalis hanya saja Nanti siapa yang tersisatapi masa ninggalin Bangsaku dan keluarga? Meski di LN juga tetep bisa terus nasionalis hanya saja.... Nanti siapa yang tersisa?	ngelihat indonesia gini aku yang berjuang di akar rasanya ingin menyerah juga ingin kaburajadulu tapi masa ninggalin bangsaku dan keluarga meski di ln juga tetep bisa terus nasionalis hanya saja nanti siapa yang tersisa

3. Normalisasi

   Tahap normalisasi dilakukan untuk mengubah kata tidak baku, singkatan, atau bahasa gaul menjadi bentuk baku. Pada Tabel 3, kata "ngelihat" bukanlah bentuk baku, maka dari itu kata tersebut diubah ke bentuk bakunya menjadi "melihat". Hasil normalisasi dapat merepresentasikan makna lebih konsisten dan membantu model dalam mengenali hubungan antar kata dengan lebih akurat.

Tabel 3. Hasil Normalisasi

No	Sebelum Normalisasi	Setelah Normalisasi
1	ngelihat indonesia gini aku yang berjuang di akar rasanya ingin menyerah juga ingin kaburajadulu tapi masa ninggalin bangsaku dan keluarga meski di ln juga tetep bisa terus nasionalis hanya saja nanti siapa yang tersisa	melihat indonesia begini aku yang berjuang di akar rasanya ingin menyerah juga ingin kaburajadulu tapi masa meninggalkan bangsaku dan keluarga meski di ln juga tetap bisa terus nasionalis hanya saja nanti siapa yang tersisa

4. Tokenisasi

   Tokenisasi merupakan proses pemecahan teks menjadi unit kata atau token. Hasil dari tahap ini menunjukkan bahwa setiap kalimat diubah menjadi kumpulan kata yang terpisah sehingga memudahkan proses analisis berbasis fitur. Pada Tabel 4 diperlihatkan hasil tokenisasi yang semula berbentuk kalimat panjang lalu dipecah menjadi token seperti "melihat", "indonesia", "begini", dan seterusnya. Dengan adanya tokenisasi, sistem dapat mengolah setiap kata secara individual sehingga dapat menjadi dasar dalam pembentukan representasi numerik seperti TF-IDF.

Tabel 4. Hasil Tokenisasi

No	Sebelum Tokenisasi	Setelah Tokenisasi
1	melihat indonesia begini aku yang berjuang di akar rasanya ingin menyerah juga ingin kaburajadulu tapi masa meninggalkan bangsaku dan keluarga meski di ln juga tetap bisa terus nasionalis hanya saja nanti siapa yang tersisa	['melihat', 'indonesia', 'begini', 'aku', 'yang', 'berjuang', 'di', 'akar', 'rasanya', 'ingin', 'menyerah', 'juga', 'ingin', 'kaburajadulu', 'tapi', 'masa', 'meninggalkan', 'bangsaku', 'dan', 'keluarga', 'meski', 'di', 'ln', 'juga', 'tetap', 'bisa', 'terus', 'nasionalis', 'hanya', 'saja', 'nanti', 'siapa', 'yang', 'tersisa']

5. *Stopword Removal*

Tahap *stopword removal* bertujuan untuk menghapus kata-kata umum yang sering muncul namun tidak memiliki makna signifikan dalam analisis sentimen, seperti kata hubung dan kata depan. Proses ini membantu mengurangi dimensi data sekaligus meningkatkan efisiensi dan akurasi model dalam mengidentifikasi sentimen.

Tabel 5. Hasil *Stopword Removal*

No	Sebelum ***Stopword Removal***	Setelah ***Stopword Removal***
1	['melihat', 'indonesia', 'begini', 'aku', 'yang', 'berjuang', 'di', 'akar', 'rasanya', 'ingin', 'menyerah', 'juga', 'ingin', 'kaburajadulu', 'tapi', 'masa', 'meninggalkan', 'bangsaku', 'dan', 'keluarga', 'meski', 'di', 'ln', 'juga', 'tetap', 'bisa', 'terus', 'nasionalis', 'hanya', 'saja', 'nanti', 'siapa', 'yang', 'tersisa']	['indonesia', 'berjuang', 'akar', 'menyerah', 'kaburajadulu', 'meninggalkan', 'bangsaku', 'keluarga', 'ln', 'nasionalis', 'tersisa']

6. *Stemming*

Tahap *stemming* dilakukan untuk mengubah kata berimbuhan menjadi bentuk dasar, seperti kata “berjuang” diubah ke bentuk dasarnya menjadi “juang”. Hasil dari *stemming* menunjukkan bahwa variasi kata yang memiliki akar kata yang sama dan dapat disederhanakan menjadi satu bentuk, sehingga meningkatkan konsistensi fitur.

Tabel 6. Hasil *Stemming*

No	Sebelum ***Stemming***	Setelah ***Stemming***
1	['indonesia', 'berjuang', 'akar', 'menyerah', 'kaburajadulu', 'meninggalkan', 'bangsaku', 'keluarga', 'ln', 'nasionalis', 'tersisa']	indonesia juang akar serah kaburajadulu tinggal bangsa keluarga ln nasionalis sisa

Pelabelan sentimen dibagi ke dalam tiga kelas, yaitu positif, negatif, dan netral. Berdasarkan Tabel 7, label sentimen positif didominasi oleh kata-kata optimis yang menunjukkan bahwa audiens melihat isu tagar #KaburAjaDulu sebagai langkah strategis untuk pengembangan diri. Sedangkan label sentimen negatif didominasi oleh kata-kata berkonotasi negatif sebagai rasa frustasi terhadap kondisi ekonomi atau kebijakan yang dianggap menghambat kebijakan individu. Label sentimen netral didominasi oleh informasi

atau pemberitaan terkait yang berfungsi sebagai penyampai informasi tanpa tendensi emosional tertentu.

Tabel 7. Hasil Pelabelan Data

No	Teks Hasil Stemming	Label
1	digit range ya luas ya juta digit good luck finding orang gaji segitu timbal sesuai gua sih encourage kaburajadulu indonesia sih hehe	Positif
2	indonesia gue umur ngelamar waiters jaga konter kagak terima anjing mending kaburajadulu persetan makan still better tahan rotten indonesia gue kangen mie ayam seblak ayam hisana	Negatif
3	info kaburajadulu olahragane kejar drone	Netral

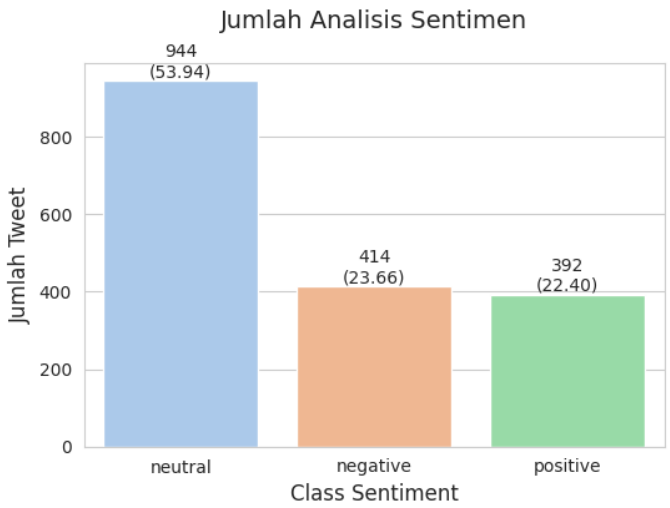
Tahap pembobotan menggunakan metode TF-IDF dilakukan untuk mengubah data teks menjadi representasi numerik berdasarkan tingkat kepentingan setiap kata dalam dokumen. Hasil pembobotan menunjukkan bahwa kata-kata yang relevan terhadap konteks sentimen memiliki nilai bobot yang lebih tinggi dibandingkan kata yang sering muncul secara umum. Dengan demikian, representasi fitur yang dihasilkan mampu mendukung proses klasifikasi secara lebih efektif pada tahap selanjutnya.

```
===== Top 20 Terms by Total TF-IDF Score =====
                term   tfidf_sum
2218    kaburajadulu  168.015404
4958              ya   50.222590
1424             gue   34.463832
3136          negara   32.125040
3471          pengin   30.882795
3344           orang   30.740171
1980       indonesia   28.912995
4166             sih   27.098980
2217           kabur   24.129457
2385           kerja   23.653680
2316           kayak   21.537404
3142          negeri   21.313722
370           banget   19.503780
499             biar   17.130273
1938            ikut   16.793263
2911         mending   16.640187
3011            moga   15.635940
2240             kak   14.941202
1984  indonesiagelap   14.675906
3529           pilih   13.843116
```

Gambar 2. Hasil TF-IDF

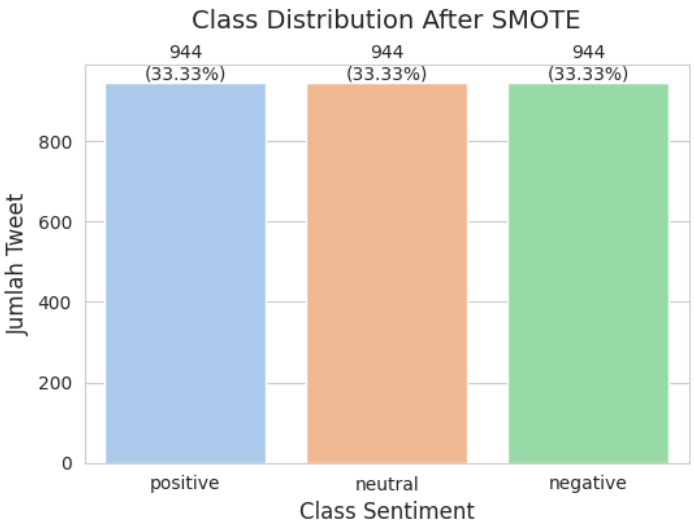
Ketidakseimbangan kelas merupakan tantangan utama dalam klasifikasi sentimen pada media sosial. Sebelum dilakukan penanganan, distribusi data sentimen menunjukkan ketimpangan yang signifikan, di mana kelas netral mendominasi dengan jumlah 944 sampel, sementara kelas negatif (414 sampel) dan positif (392 sampel) berada dalam posisi minoritas.

Dominasi kelas netral ini berisiko menyebabkan model klasifikasi menjadi bias, di mana model akan cenderung mengklasifikasikan hampir semua data sebagai kelas netral demi mendapatkan akurasi yang tinggi secara semu [15].



Gambar 3. Jumlah Analisis Sentimen Sebelum SMOTE

Untuk mengatasi hal tersebut, teknik SMOTE diterapkan untuk menyeimbangkkan distribusi data. Hasil penerapan SMOTE berhasil merekonstruksi data pada kelas minoritas dengan menciptakan sampel sintesis, sehingga menghasilkan distribusi yang setara: positif (944 sampel), negatif (944 sampel), dan netral (944 sampel), dengan total dataset meningkat menjadi 2.832 sampel.

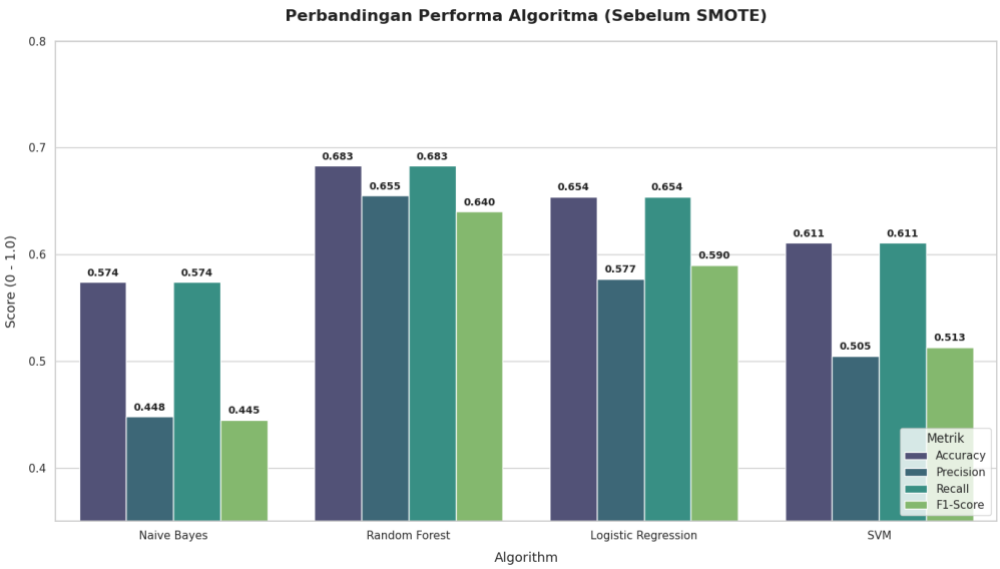


Gambar 4. Jumlah Analisis Sentimen Setelah SMOTE

Pada penelitian ini dilakukan perbandingan performa algoritma Support Vector Machine (SVM) dengan tiga algoritma klasifikasi lainnya, yaitu Naïve Bayes, Random Forest, dan Logistic Regression setelah penerapan SMOTE untuk mengatasi ketidakseimbangan distribusi kelas. Pemilihan ketiga algoritma tersebut didasarkan pada

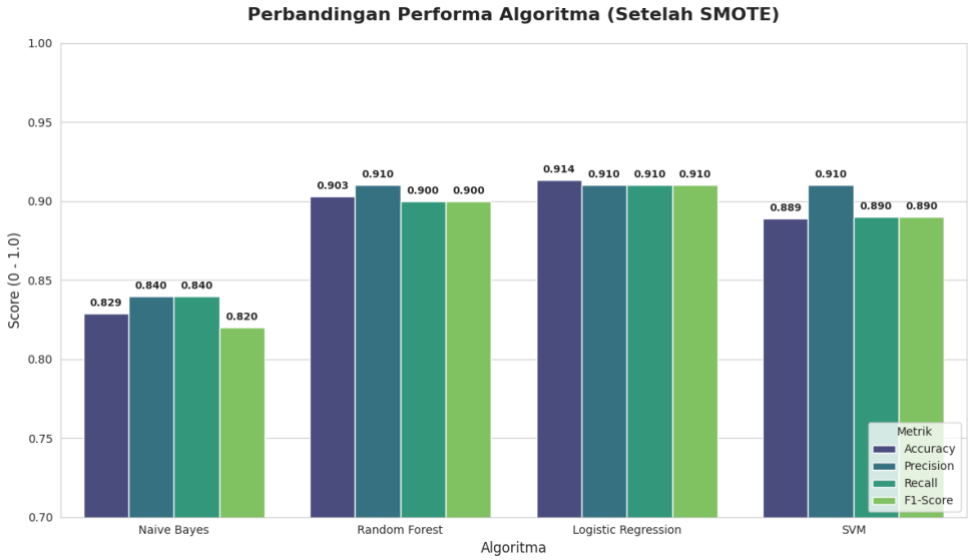
perbedaan pendekatan klasifikasi yang digunakan. Naïve Bayes mewakili pendekatan probabilistik yang sederhana dan banyak digunakan sebagai baseline pada klasifikasi teks karena memiliki kompleksitas komputasi yang rendah [9]. Logistic Regression dipilih karena merupakan algoritma klasifikasi linear yang terbukti efektif pada data teks berdimensi tinggi hasil pembobotan TF-IDF [10], [14]. Sementara itu, Random Forest digunakan untuk merepresentasikan pendekatan ensemble learning yang mampu meningkatkan kemampuan generalisasi model melalui kombinasi beberapa decision tree [11]. Perbandingan ini dilakukan untuk mengevaluasi efektivitas masing-masing pendekatan klasifikasi pada dataset sentimen yang telah diseimbangkan menggunakan SMOTE.

Berdasarkan hasil pengujian pada Gambar 5, terlihat bahwa penerapan SMOTE memberikan peningkatan performa pada seluruh algoritma yang digunakan. Sebelum dilakukan penyeimbangan data, nilai accuracy algoritma berada pada rentang 57,4% hingga 68,3%. Random Forest dan Logistic Regression menunjukkan performa terbaik dengan accuracy masing-masing sebesar 68,3% dan 65,4%, sedangkan SVM dan Naïve Bayes memperoleh accuracy sebesar 61,1% dan 57,4%. Nilai precision, recall, dan F1-score yang relatif rendah menunjukkan bahwa ketidakseimbangan distribusi kelas menyebabkan model cenderung mengalami kesulitan dalam mengenali seluruh kelas sentimen secara optimal [16].



Gambar 5. Hasil Perbandingan Algoritma Setelah SMOTE

Setelah penerapan SMOTE, terjadi peningkatan performa yang signifikan pada seluruh algoritma. Accuracy Naïve Bayes meningkat dari 57,4% menjadi 82,92%, Random Forest meningkat dari 68,3% menjadi 90,30%, Logistic Regression meningkat dari 65,4% menjadi 91,36%, dan SVM meningkat dari 61,1% menjadi 88,89%. Peningkatan serupa juga terlihat pada nilai precision, recall, dan F1-score. Hasil ini menunjukkan bahwa penyeimbangan data menggunakan SMOTE berhasil mengurangi bias model terhadap kelas mayoritas sehingga setiap algoritma dapat mempelajari karakteristik data secara lebih representatif [10]. Dengan demikian, proses klasifikasi menjadi lebih akurat dan konsisten dibandingkan ketika menggunakan data yang belum diseimbangkan.



Gambar 6. Hasil Perbandingan Algoritma Setelah SMOTE

Berdasarkan hasil pengujian pada Gambar 6, algoritma Naïve Bayes memperoleh nilai accuracy sebesar 82,89%, dengan precision sebesar 0,84, recall sebesar 0,84, dan F1-score sebesar 0,82. Nilai tersebut merupakan yang terendah dibandingkan algoritma lainnya. Kondisi ini dapat terjadi karena Naïve Bayes menggunakan asumsi bahwa setiap fitur bersifat independen satu sama lain. Pada data teks hasil pembobotan TF-IDF, hubungan antar kata sering kali tidak sepenuhnya independen sehingga asumsi tersebut dapat mengurangi kemampuan model dalam membedakan sentimen yang memiliki karakteristik kata yang mirip. Akibatnya, model masih mengalami kesalahan klasifikasi terutama pada kelas netral yang memiliki karakteristik linguistik yang cenderung tumpang tindih dengan kelas positif maupun negatif.

Algoritma Random Forest menghasilkan peningkatan performa dengan nilai accuracy sebesar 90,30%, precision sebesar 0,91, recall sebesar 0,90, dan F1-score sebesar 0,90. Hasil ini menunjukkan bahwa pendekatan ensemble learning mampu mempelajari pola yang lebih kompleks dibandingkan Naïve Bayes. Kombinasi banyak decision tree memungkinkan Random Forest menangkap hubungan non-linear antar fitur sehingga proses klasifikasi menjadi lebih stabil [17]. Selain itu, mekanisme agregasi hasil dari banyak pohon keputusan membantu mengurangi risiko overfitting yang umum terjadi pada model berbasis decision tree tunggal.

Logistic Regression memperoleh performa terbaik dengan nilai accuracy sebesar 91,36%, serta nilai precision, recall, dan F1-score masing-masing sebesar 0,91. Hasil ini menunjukkan bahwa Logistic Regression mampu memisahkan kelas sentimen secara efektif pada ruang fitur TF-IDF. Tingginya performa Logistic Regression dapat disebabkan oleh karakteristik data teks yang umumnya memiliki dimensi tinggi dan bersifat sparse, sehingga lebih sesuai dimodelkan menggunakan pendekatan linear. Selain itu, penerapan SMOTE membantu menghasilkan distribusi data yang lebih seimbang sehingga model dapat mempelajari batas keputusan (decision boundary) antar kelas dengan lebih baik [10]. Kondisi tersebut menyebabkan Logistic Regression mampu menghasilkan performa yang sedikit lebih unggul dibandingkan Random Forest maupun SVM pada dataset yang digunakan.

Sementara itu, algoritma SVM memperoleh nilai accuracy sebesar 88,89% dengan precision, recall, dan F1-score masing-masing sebesar 0,91, 0,89, dan 0,89. Meskipun SVM dikenal efektif dalam klasifikasi teks berdimensi tinggi, hasil penelitian menunjukkan bahwa performanya masih berada di bawah Logistic Regression dan Random Forest. Hal ini mengindikasikan bahwa pola distribusi data setelah penerapan SMOTE lebih sesuai dimodelkan menggunakan pendekatan linear Logistic Regression dibandingkan pembentukan hyperplane pada SVM [17]. Namun demikian, perbedaan performa antar kedua algoritma relatif kecil sehingga keduanya tetap menunjukkan kemampuan klasifikasi yang baik pada kasus analisis sentimen #KaburAjaDulu [18].

Secara keseluruhan, hasil penelitian menunjukkan bahwa Logistic Regression merupakan algoritma dengan performa terbaik pada dataset yang digunakan. Temuan ini mengindikasikan bahwa pendekatan klasifikasi linear lebih sesuai untuk data sentimen berbasis TF-IDF yang telah diseimbangkan menggunakan SMOTE [11]. Di sisi lain, Random Forest juga menunjukkan performa yang kompetitif karena kemampuannya menangkap pola yang lebih kompleks, sedangkan Naïve Bayes memiliki keterbatasan akibat asumsi independensi fitur yang kurang sesuai dengan karakteristik data teks pada penelitian ini [10].

KESIMPULAN

Penelitian ini berhasil menganalisis sentimen publik terhadap fenomena #KaburAjaDulu pada media sosial X menggunakan pendekatan machine learning dengan menerapkan SMOTE untuk mengatasi ketidakseimbangan data. Hasil penelitian menunjukkan bahwa penyeimbangan data dan tahapan preprocessing yang sistematis berperan penting dalam meningkatkan kemampuan model untuk mengklasifikasikan sentimen secara lebih akurat. Berdasarkan hasil perbandingan algoritma, Logistic Regression memperoleh performa terbaik dengan nilai akurasi 91,36%, diikuti oleh Random Forest sebesar 90,30%, Support Vector Machine sebesar 88,89%, dan Naïve Bayes sebesar 82,89%.

Kontribusi penelitian ini terletak pada pembuktian bahwa kualitas *preprocessing* data dan penanganan ketidakseimbangan kelas memiliki peran yang sangat penting dalam meningkatkan performa analisis sentimen pada data media sosial yang tidak terstruktur [19]. Penelitian ini juga memberikan referensi empiris mengenai efektivitas penggunaan beberapa algoritma *machine learning* dalam klasifikasi sentimen berbahasa Indonesia.

Untuk penelitian selanjutnya, disarankan untuk menggunakan dataset dengan jumlah yang lebih besar serta mengembangkan metode klasifikasi menggunakan pendekatan *deep learning* seperti BERT, LSTM, atau metode *hybrid* agar mampu memahami konteks bahasa, sarkasme, dan penggunaan bahasa informal pada media sosial secara lebih mendalam [20]. Selain itu, penelitian berikutnya dapat menambahkan analisis temporal atau topik untuk memperoleh gambaran yang lebih komprehensif mengenai dinamika opini publik terhadap isu sosial di ruang digital.

DAFTAR PUSTAKA

[1] Ravenstein, “Analisis Faktor-Faktor yang Mempengaruhi Tingkat Migrasi Tenaga Kerja Indonesia (TKI) (Studi Kasus Pada 6 Provinsi Tahun 2008-2017),” *FEB UIN Syarif Hidayatullah*, 2019, [Online]. Available: http://repository.uinjkt.ac.id/dspace/handle/123456789/46978

[2] P. Nandwani and R. Verma, "A review on sentiment analysis and emotion detection from text," *Soc. Netw. Anal. Min.*, vol. 11, no. 1, pp. 1–19, 2021, doi: 10.1007/s13278-021-00776-6.

[3] R. C. Rivaldi, T. D. Wismarini, J. T. Lomba, and J. Semarang, "Analisis Sentimen Pada Ulasan Produk Dengan Metode Natural Language Processing (NLP) (Studi Kasus Zalika Store 88 Shopee)," *Elkom*, vol. 17, no. 1, pp. 120–128, 2024.

[4] E. R. Kaburuan and N. R. Setiawan, "Sentimen Analisis Review Aplikasi Digital Korlantas Pada Google Play Store Menggunakan Metode SVM," *J. Sisfokom (Sistem Inf. dan Komputer)*, vol. 12, no. 1, pp. 105–116, 2023, doi: 10.32736/sisfokom.v12i1.1614.

[5] P. P. Putra, M. K. Anam, A. S. Chan, A. Hadi, N. Hendri, and A. Masnur, "Optimizing Sentiment Analysis on Imbalanced Hotel Review Data Using SMOTE and Ensemble Machine Learning Techniques," *J. Appl. Data Sci.*, vol. 6, no. 2, pp. 936–951, 2025, doi: 10.47738/jads.v6i2.618.

[6] U. Hasanah, A. M. Soleh, and K. Sadik, "Effect of Random Under sampling, Oversampling, and SMOTE on the Performance of Cardiovascular Disease Prediction Models," *J. Mat. Stat. dan Komputasi*, vol. 21, no. 1, pp. 88–102, 2024, doi: 10.20956/j.v21i1.35552.

[7] D. Valkenborg, A. J. Rousseau, M. Geubbelmans, and T. Burzykowski, "Support vector machines," *Am. J. Orthod. Dentofac. Orthop.*, vol. 164, no. 5, pp. 754–757, 2023, doi: 10.1016/j.ajodo.2023.08.003.

[8] H. C. Husada and A. S. Paramita, "Analisis Sentimen Pada Maskapai Penerbangan di Platform Twitter Menggunakan Algoritma Support Vector Machine (SVM)," *Teknika*, vol. 10, no. 1, pp. 18–26, 2021, doi: 10.34148/teknika.v10i1.311.

[9] M. F. As Shidiq and D. Alita, "Analisis Sentimen Masyarakat Terhadap Kasus Judi Online Menggunakan Data Dari Media Sosial X Pendekatan Naive Bayes Dan Svm," *J. Sist. Inf. dan Inform.*, vol. 8, no. 1, pp. 24–35, 2025, doi: 10.47080/simika.v8i1.3624.

[10] B. Ramadhani and R. R. Suryono, "Komparasi Algoritma Naïve Bayes dan Logistic Regression Untuk Analisis Sentimen Metaverse," *J. Media Inform. Budidarma*, vol. 8, no. 2, p. 714, 2024, doi: 10.30865/mib.v8i2.7458.

[11] E. Erlin, Y. Desnelita, N. Nasution, L. Suryati, and F. Zoromi, "Dampak SMOTE terhadap Kinerja Random Forest Classifier berdasarkan Data Tidak seimbang," *MATRIK J. Manajemen, Tek. Inform. dan Rekayasa Komput.*, vol. 21, no. 3, pp. 677–690, 2022, doi: 10.30812/matrik.v21i3.1726.

[12] O. I. Gifari, M. Adha, F. Freddy, and F. F. S. Durrand, "Analisis Sentimen Review Film Menggunakan TF-IDF dan Support Vector Machine," *J. Inf. Technol.*, vol. 2, no. 1, pp. 36–40, 2022, doi: 10.46229/jifotech.v2i1.330.

[13] A. Firdaus and W. I. Firdaus, "Text Mining Dan Pola Algoritma Dalam Penyelesaian Masalah Informasi : (Sebuah Ulasan)," *J. JUPITER*, vol. 13, no. 1, p. 66, 2021.

[14] Muhammad Fernanda Naufal Fathoni, Eva Yulia Puspaningrum, and Andreas Nugroho Sihananto, "Perbandingan Performa Labeling Lexicon InSet dan VADER pada Analisa Sentimen Rohingya di Aplikasi X dengan SVM," *Modem J. Inform. dan Sains Teknol.*, vol. 2, no. 3, pp. 62–76, 2024, doi: 10.62951/modem.v2i3.112.

[15] H. Hairani, T. Widiyaningtyas, and D. D. Prasetya, "Addressing Class Imbalance of Health Data: a Systematic Literature Review on Modified Synthetic Minority Oversampling Technique (SMOTE) Strategies," *Int. J. Informatics Vis.*, vol. 8, no. 3, pp. 1310–1318, 2024, doi: 10.62527/joiv.8.3.2283.

[16] R. Blagus and L. Lusa, "SMOTE for high-dimensional class-imbalanced data," *BMC Bioinformatics*, vol. 14, 2013, doi: 10.1186/1471-2105-14-106.

[17] A. N. Laily, M. A. Barata, and D. Nurdiansyah, "Analisis Klasifikasi Hepatitis

Menggunakan Synthetic Minority Oversampling Technique , Support Vector Machine , dan Random Forest," vol. 6, no. 1, pp. 223–236, 2026.

[18] A. A. Qolbu, N. Fitriyati, and 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, pp. 29–44, 2025, doi: 10.14421/fourier.2025.141.29-44.

[19] W. Nugraha and M. Syarif, "Teknik Weighting untuk Mengatasi Ketidakseimbangan Kelas Pada Prediksi Churn Menggunakan XGBoost, LightGBM, dan CatBoost," *Techno.Com*, vol. 22, no. 1, pp. 97–108, 2023, doi: 10.33633/tc.v22i1.7191.

[20] Olih Solihin, Dwi Firmansyah, Ahmad Zakki Abdullah, and Ahmad Prawira Dhahiyat, "Pemanfaatan AI dalam Analisis Isi Digital : Studi Kasus Komentar Media Sosial," *J. Pendidik. Dan Ilmu Sos.*, vol. 3, no. 2, pp. 117–129, 2025, doi: 10.54066/jupendis.v3i2.3141.