JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 3, Juni 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i3.8548
Hal 161-171
http://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Identifikasi Polaritas Sikap Pengguna Aplikasi X terhadap Coretax di Indonesia Menggunakan Algoritma Naïve Bayes

Dina Rahma Prasilda, Wenty Dwi Yuniarti, Maya Rini Handayani, Khothibul Umam*

Fakultas Sains dan Teknologi, Teknologi Informasi, UIN Walisongo, Semarang, Indonesia
Email: [1]2208096028@student.walisongo.ac.id, [2]wenty@walisongo.ac.id, [3]maya@walisongo.ac.id,
[4,*]khothibul_umam@walisongo.ac.id
Email Penulis Korespondensi: khothibul_umam@walisongo.ac.id
Submitted **22-04-2025;** Accepted **23-05-2025;** Published **30-06-2025**

Abstrak

Core Tax Administration System (Coretax) diluncurkan oleh Direktorat Jenderal Pajak (DJP) pada bulan Januari 2025 sebagai sistem perpajakan terintegrasi berbasis teknologi. Tujuan awalnya adalah untuk meningkatkan efisiensi dan kepatuhan pajak, Coretax menghadapi tantangan teknis, termasuk kesalahan sistem, kecepatan pemrosesan yang lambat, dan kritik dari masyarakat. Platform utama yang digunakan untuk mengatasi tantangan-tantangan ini adalah aplikasi X (sebelumnya dikenal sebagai Twitter). Penelitian ini bertujuan untuk memahami pandangan dan tanggapan masyarakat terhadap layanan Coretax dengan menganalisis pola sentimen pengguna yang terlihat di media sosial. Penelitian ini mengidentifikasi polaritas sikap pengguna dengan memanfaatkan pemrosesan bahasa alami (NLP) dan algoritma *Naïve Bayes*, yang diterapkan pada dataset 1.628 *tweet* yang dikumpulkan antara Januari hingga Maret 2025. Data yang dianalisis mencerminkan beragam reaksi publik yang mencakup opini positif dan negatif terhadap implementasi Coretax, baik dari aspek fungsionalitas maupun kemudahan penggunaannya. Hasil menunjukkan bahwa model memiliki tingkat akurasi sebesar 93.07%, nilai *precision* sebesar 95%, nilai serta *recall* sebesar 96%, dan nilai F1-Score 96%. Hasil dari penelitian ini diharapkan mampu memberikan pemetaan yang tepat terkait perubahan opini publik terhadap Coretax, sehingga dapat menjadi sumber informasi yang bernilai bagi pengembang aplikasi, perumus kebijakan di bidang perpajakan, serta analisis di sektor teknologi dalam menanggapi kebutuhan serta ekspektasi masyarakat di era digital.

Kata Kunci: Coretax; Polaritas Sikap Publik; Pemrosesan Bahasa Alami; Leksikon Bahasa Indonesia; *Naïve Bayes*

Abstrak

The Core Tax Administration System (Coretax) was launched by the Directorate General of Taxes (DGT) in January 2025 as a technology-based integrated tax system. While its initial goal was to improve tax efficiency and compliance, Coretax faced technical challenges, including system errors, slow processing speed, and criticism from the public. The main platform used to address these challenges is the X app (formerly known as Twitter). This research aims to understand the public's views and responses to Coretax's services by analyzing user sentiment patterns seen on social media. The research identifies the polarity of user attitudes by utilizing natural language processing (NLP) and *Naïve Bayes* algorithms, applied to a dataset of 1,628 tweets collected between January and March 2025. The analyzed data reflects a wide range of public reactions that include both positive and negative opinions towards the Coretax implementation, both in terms of functionality and ease of use. The results show that the model has an accuracy rate of 93.07%, a precision value of 95%, a recall value of 96%, and an F1-Score value of 96%. The results of this study are expected to be able to provide precise mapping related to changes in public opinion towards Coretax, so that it can be a valuable source of information for application developers, policy makers in the field of taxation, and analysis in the technology sector in responding to the needs and expectations of society in the digital era.

Keywords: Coretax; Public Sentiment Analysis; Natural Language Processing; Indonesian *Lexicon*; *Naïve Bayes*

1. PENDAHULUAN

Dalam konteks eksplorasi data besar *(big data)* dari platform media sosial, penilaian terhadap kecenderungan opini pengguna menjadi pendekatan penting. Dengan pesatnya pertumbuhan platform sosial, seperti X (sebelumnya dikenal sebagai Twitter), terdapat lautan informasi yang mencerminkan pandangan dan opini pengguna mengenai berbagai isu, termasuk aplikasi layanan seperti Coretax [1]. Coretax menawarkan kemudahan bagi pengguna dalam mengelola kewajiban perpajakan dengan platform yang inovatif dan user-friendly. Identifikasi polaritas sikap ini tidak hanya berguna untuk memahami penerimaan dan kepuasan pengguna, tetapi juga untuk memberikan umpan balik yang konstruktif bagi pengembang aplikasi dalam rangka perbaikan dan peningkatan fitur.

Namun demikian, permasalahan yang muncul adalah belum tersedianya sistem yang secara otomatis mampu mengklasifikasikan opini publik terhadap aplikasi Coretax secara cepat dan akurat. Sebagian besar evaluasi terhadap kepuasan pengguna masih dilakukan secara manual melalui survei atau ulasan langsung, yang cenderung memakan waktu, tidak efisien, dan kurang mampu menangkap opini yang tersebar secara masif di media sosial. Selain itu, tingginya volume data dan keberagaman cara pengguna dalam menyampaikan pendapat menjadi tantangan tersendiri dalam proses analisis sentimen, khususnya dalam bahasa Indonesia yang memiliki kekhasan struktur dan kosakata.

Keberhasilan pemodelan polaritas sikap sangat dipengaruhi oleh mutu data yang diperoleh serta metode yang diterapkan dalam tahap pemrosesan dan klasifikasi. Algoritma machine learning seperti *Naïve Bayes* dikenal efektif untuk klasifikasi teks berkat kemudahannya dalam penerapan serta kecepatan dalam proses pengolahan [2]. Namun demikian, penerapannya pada topik spesifik seperti CoreTax - sistem pajak digital pemerintah - masih belum banyak dieksplorasi. Beberapa penelitian sebelumnya yang dilakukan Dewan dan Said mencatat bahwa diperoleh hasil akurasi sebesar 89.97% dalam analisis terkait topik perpajakan di Twitter[3]. Berbeda dengan penelitian Hani Brilianti dkk berfokus pada tanggapan publik terhadap kebijakan kenaikan PPN 12% di media sosial X[4]. Namun, kedua penelitian tersebut berfokus

JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 3, Juni 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i3.8548
Hal 161-171
http://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

pada analisis sentimen umum di bidang perpajakan, bukan secara spesifik pada aplikasi Coretax sebagai objek studi. Demikian pula penelitian Ramadhani dan Wahyudin yang membandingkan kinerja dua algoritma klasifikasi, yaitu *Naïve Bayes* (90,71%) dan K-NN (74,78%) lebih berfokus pada sentimen vaksinasi[5]. Sementara itu, penelitian-penelitian tentang Coretax yang ada masih didominasi oleh pendekatan kualitatif melalui survei mahasiswa atau wawancara konsultan pajak, sehingga belum mampu menangkap persepsi publik secara real-time dan komprehensif.

Berdasarkan Peraturan Presiden Nomor 40 Tahun 2018, Pengembangan Coretax merupakan bagian dari Proyek Pembaruan Sistem Administrasi Pajak Inti (PSIAP) dengan tujuan memodernisasi sistem administrasi pajak. Studi ini mengisi celah (gap) dengan mengkaji sentimen pengguna terhadap aplikasi Coretax secara khusus, yang belum banyak dieksplorasi dalam penelitian sejenis. Proses klasifikasi mencakup pengumpulan data, pra-proses teks, pelabelan sentimen, hingga evaluasi model.

Dua penelitian terbaru pada tahun 2024 yang membahas aplikasi layanan pemerintah menunjukkan pendekatan serupa dalam analisis sentimen, yaitu dengan menggunakan algoritma Naïve Bayes terhadap data ulasan pengguna di Google Play Store. Seiring dengan perkembangan teknologi, masyarakat tidak hanya memberikan umpan balik melalui aplikasi resmi, tetapi juga mendiskusikannya di media sosial. Platform seperti X berperan penting dalam memahami sentimen publik terhadap aplikasi yang digunakan, sehingga penelitian ini relevan untuk mengkaji sentimen masyarakat terhadap aplikasi perpajakan Coretax. Penelitian oleh Deni Wijaya dkk. yang menganalisis sentimen terhadap aplikasi Samsat Digital menggunakan algoritma Naïve Bayes menunjukkan akurasi sebesar 63,61% [6]. Sementara itu, studi oleh Putra dan Febriawan terhadap aplikasi Digital Korlantas POLRI dengan algoritma yang sama menghasilkan akurasi sebesar 70,33% [7]. Berbeda dari kedua penelitian tersebut, studi ini memfokuskan analisis pada sistem Coretax sebagai infrastruktur strategis dalam sistem perpajakan nasional. Selain itu, penelitian ini juga menerapkan teknik optimasi berupa SMOTE dan penggunaan leksikon khusus, yang terbukti mampu meningkatkan akurasi model hingga mencapai 93,07%.

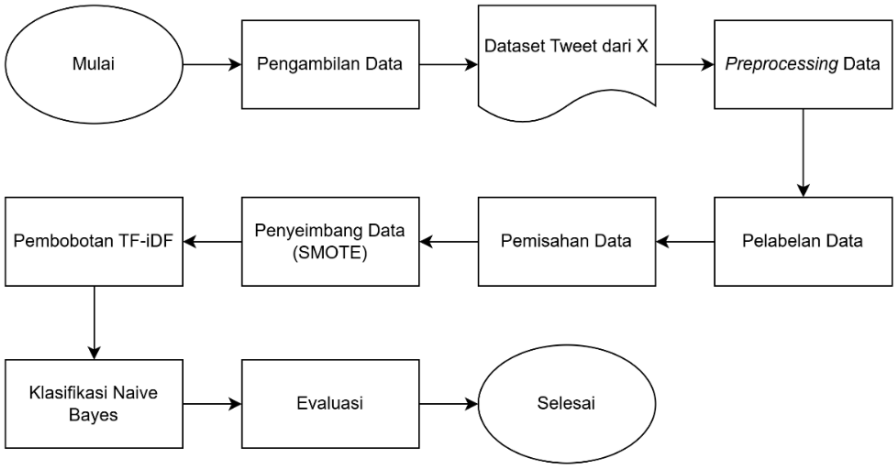
Agar penelitian ini tetap fokus dan terarah, beberapa batasan yang diterapkan mencakup sumber data yang digunakan, yaitu unggahan media sosial dari platform X yang membahas Coretax dalam rentang waktu tertentu. Algoritma yang diterapkan dalam klasifikasi sentimen adalah *Naïve Bayes*, tanpa melakukan perbandingan dengan metode lainnya. Selain itu, kategori sentimen yang dianalisis hanya terbatas pada sentimen positif dan negatif, tanpa mempertimbangkan netralitas atau konteks emosional yang lebih kompleks dalam unggahan pengguna. Proses analisis juga dibatasi pada pendekatan *lexicon-based* dan penghitungan bobot fitur menggunakan TF-IDF.

Dalam konteks identifikasi polaritas sikap pengguna terhadap Coretax di aplikasi X, pemanfaatan algoritma Naïve Bayes menghadirkan pendekatan yang menarik untuk memahami persepsi masyarakat terhadap kebijakan pajak, sehingga penelitian ini akan menjawab beberapa pertanyaan, yaitu: bagaimana akurasi algoritma Naïve Bayes dalam mengklasifikasikan sentimen ulasan Coretax di aplikasi X, seberapa akurat hasil klasifikasi sentimen menggunakan algoritma Naïve Bayes, dan sejauh mana analisis sentimen dapat digunakan untuk mengidentifikasi tren dan pola pada ulasan Coretax di aplikasi X.

Merujuk pada latar belakang yang telah dipaparkan, fokus dari penelitian ini diarahkan pada penggalian persepsi publik untuk mengidentifikasi pola dan tren dalam merespon pengguna. Melalui analisis yang mendalam dan berbasis data, hasil dari penelitian ini diharapkan dapat menjadi dasar pengambilan keputusan yang lebih tepat bagi pengelola layanan serta pihak terkait dalam upaya pengembangan produk kedepannya.

2. METODOLOGI PENELITIAN

Tahapan ini menjelaskan langkah-langkah penelitian yang terlibat dalam klasifikasi sentimen pengguna terhadap aplikasi X dan sistem Coretax di Indonesia dengan menerapkan algoritma *machine learning Naïve Bayes*. Proses penelitian ini terdiri dari delapan tahapan utama yang saling terkait. Secara visual, alur proses tersebut dapat dilihat pada Gambar 1 berikut:



Gambar 1. Alur Penelitian

JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 3, Juni 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i3.8548
Hal 161-171
http://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Gambar 1 mengilustrasikan alur metodologis penelitian yang diawali dengan akuisisi data tekstual dari platform X. Tahap selanjutnya meliputi: *pre-processing* data untuk standardisasi input teks melalui normalisasi dan pembersihan, anotasi sentimen menggunakan leksikon yang telah divalidasi, partisi dataset menjadi subset training dan testing dengan rasio tertentu. Untuk mengatasi masalah ketidakseimbangan kelas, teknik *Synthetic Minority Over-sampling Technique* (SMOTE) diaplikasikan secara selektif pada subset training. Representasi fitur tekstual kemudian diubah menjadi vektor numerik melalui *Term Frequency-Inverse Document Frequency* (TF-IDF). Model klasifikasi dibangun dengan pendekatan Naive Bayes, yang selanjutnya dievaluasi menggunakan berbagai metrik performa. Desain alur ini secara khusus memperhatikan aspek preventif terhadap data leakage dengan membatasi aplikasi SMOTE hanya pada data training, sekaligus menjaga orisinalitas data testing untuk pengujian yang tidak bias.

2.1 Pengambilan Data

Tahap pertama dalam penelitian ini adalah pengumpulan data publik dari platform media sosial, yaitu X (sebelumnya Twitter), dengan memanfaatkan kata kunci 'Coretax'. Teknik *web crawling* diimplementasikan melalui *Application Programming Interface* (API) X dengan menggunakan pustaka khusus untuk mengekstrak data secara otomatis.

2.2 *Preprocessing Data*

Data *tweet* yang dikumpulkan mengandung sejumlah besar informasi yang tidak relevan, sehingga perlu dilakukan *preprocessing* teks untuk meningkatkan keakuratan identifikasi polaritas sikap. Proses ini memerlukan beberapa langkah yang memiliki tujuan untuk memproduksi data yang telah terstandarisasi untuk kebutuhan analisis [8], [9].

a. *Cleaning* melibatkan penghapusan bagian teks yang tidak relevan.
b. *Case folding* melibatkan konversi seluruh teks menjadi huruf kecil dan penghapusan elemen asing seperti tautan, skrip, dan simbol lainnya.
c. *Tokenization* merupakan proses yang melibatkan pemecahan kalimat menjadi serangkaian kata, dengan menghapus tanda baca dan karakter non-huruf.
d. *Stopword removal* bertujuan untuk menyaring berbagai jenis kata fungsional yang tidak memberikan makna substantif dalam analisis.
e. *Stemming* melibatkan pengubahan kata majemuk menjadi bentuk dasarnya, sehingga meningkatkan efektivitas analisis teks.

2.3 Pelabelan Data

Tahap selanjutnya adalah penerapan metode *Lexicon* Bahasa Indonesia untuk meningkatkan ketepatan identifikasi polaritas sikap. Setelah tahap *preprocessing* teks, data teks dianalisis berdasarkan kamus *lexicon* yang berisi kata-kata positif dan negatif. Item-item *lexicon* yang teridentifikasi dalam kamus kemudian akan dihitung berdasarkan jumlah kemunculannya dalam setiap teks atau kalimat [10]. Perhitungan ini akan menentukan apakah *tweet* tersebut positif atau negatif.

$$S_{positive} \sum_{i \in t}^{n} positive\ score_i \quad (1)$$

$$S_{negative} \sum_{i \in t}^{n} negative\ score_i \quad (2)$$

Istilah ($Spositive$) merujuk pada bobot sebuah kalimat yang dihitung berdasarkan penjumlahan dari n skor polaritas kata-kata opini yang bersifat positif. Sebaliknya, ($Snegative$) merepresentasikan bobot kalimat yang berasal dari akumulasi n skor polaritas kata opini negatif. Nilai sentimen dari sebuah kalimat, sebagaimana dijelaskan dalam Persamaan 3, ditentukan dengan membandingkan total skor dari kata-kata positif, negatif, dan netral untuk mengetahui kecenderungan sentimennya.

$$Sentence_{sentiment} \begin{Bmatrix} positive\ if\ S_{positive} > S_{negative} \\ positive\ if\ S_{positive} = S_{negative} \\ positive\ if\ S_{positive} < S_{negative} \end{Bmatrix} \quad (3)$$

Jika sebuah teks mengandung lebih banyak kata positif daripada kata negatif, maka data teks tersebut akan diberi label sentimen positif. Di sisi lain, jika teks mengandung lebih banyak kata negatif daripada kata positif, data teks akan diberi label dengan sentimen negatif. Sementara itu, jika teks mengandung jumlah kata positif dan negatif yang setara, teks tersebut akan dilabeli dengan sentimen netral.

Selanjutnya, penentuan skor polaritas dari kata-kata sentimen tersebut didasarkan pada *lexicon*. Nilai sentiment skor yang kurang < 0 mengindikasikan sentimen negatif. Sebaliknya, jika skor sentimen > 0, maka diklasifikasikan memiliki sentimen positif. Penting untuk dicatat bahwa sentimen netral jika skor sentimen setara dengan 0.

2.4 Pembagian Data

Dalam tahap ini, data dibagi ke dalam dua kategori utama, yaitu data pelatihan dan data pengujian. Sebanyak 80% dari keseluruhan data dialokasikan untuk proses pelatihan guna membentuk model atau pola klasifikasi oleh algoritma, sementara 20% sisanya digunakan sebagai data pengujian untuk mengevaluasi kinerja model yang telah dikembangkan.

2.5 Penyeimbang Data dengan SMOTE

JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 3, Juni 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i3.8548
Hal 161-171
http://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Tahap selanjutnya adalah menyeimbangkan distribusi kelas menggunakan teknik SMOTE *(Synthetic Minority Over-sampling Technique).* Berbeda dengan *oversampling* acak yang hanya mereplikasi data [11], SMOTE secara cerdas menghasilkan sampel sintetis baru di ruang fitur, sehingga mengurangi risiko *overfitting* [12]. Teknik ini mengubah distribusi data latih menjadi seimbang untuk semua kelas sebelum pelatihan model *Naïve Bayes*.

2.6 Pembobotan TF-IDF

Pembobotan istilah (*term weighting*) merupakan proses pemberian nilai pada setiap kata dalam teks dengan tujuan meningkatkan efektivitas pemetaan sikap pengguna dalam konteks *text mining*. *Term Frequency* mencerminkan seberapa penting sebuah kata berdasarkan frekuensi kemunculannya dalam suatu dokumen, sedangkan *Inverse Document Frequency* (IDF) berfungsi sebagai teknik pembobotan yang mengukur sejauh mana suatu token muncul secara keseluruhan dalam kumpulan dokumen [13].

2.7 Klasifikasi *Naïve Bayes*

Pada tahap ini, sentimen data *tweet* yang telah melalui proses transformasi diklasifikasikan menggunakan algoritma *Naïve Bayes*, yaitu metode klasifikasi yang menerapkan aturan *Bayes* dan perhitungan probabilitas [14]. Algoritma ini memerlukan sejumlah parameter, yang disebut atribut, untuk mengidentifikasi kelas yang sesuai dari sampel yang akan dianalisis [15]. Teorema ini menyatakan bahwa:

$$P(A|B) = \frac{P(B|A).P(A)}{P(B)} \quad (4)$$

Dengan:
B = Data yang belum diketahui kelasnya;
A = Hipotesis bahwa data B termasuk ke dalam suatu kelas tertentu;
P(A|B) = Probabilitas dari hipotesis A berdasarkan kondisi data B yang diketahui (*posteriori probability*);
P(A): Probabilitas awal dari hipotesis A (*prior probability*);
P(B|A) = Probabilitas terjadinya data B, dengan anggapan bahwa hipotesis A benar;
P(B) = Probabilistik awal dari hipotesis B
Naïve Bayes menggunakan asumsi bahwa setiap atribut independen satu sama lain.

2.8 Evaluasi

Pada tahap akhir, evaluasi dilakukan terhadap model yang telah dikembangkan untuk mengukur tingkat akurasi dalam klasifikasi menggunakan algoritma *Naïve Bayes*. Penilaian yang dilakukan mencakup komponen *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN). Metodologi ini memastikan keakuratan hasil klasifikasi.

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

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

$$Recall = \frac{TP}{TP+FN} \quad (7)$$

$$F - Measure = \frac{2\,(recall\ .precision)}{(recall + precision)} \quad (8)$$

3. HASIL DAN PEMBAHASAN

Pada bagian ini, tahap-tahap selanjutnya mengenai prosedur yang diterapkan dalam upaya penelitian akan dijelaskan. Prosedur-prosedur ini telah diuraikan dengan cermat dalam metodologi penelitian yang telah diuraikan sebelumnya.

3.1 Pengumpulan Data

Data dikumpulkan menggunakan metode *web scraping* yang diimplementasikan menggunakan Python, mengekstrak tanggapan pengguna aplikasi Coretax dari platform X dalam periode Januari-Maret 2025. Pencarian data dilakukan dengan menggunakan kata kunci "Coretax", dengan 1.628 kejadian yang tercatat. Penelitian ini menggunakan *library tweet-harvest* untuk mengambil data *tweet.* Tabel 1 menyajikan hasil *web scraping* yang dilakukan dengan hanya mengakses konten publik yang tersedia secara terbuka.

Tabel 1. Sampel Data

No	Tweet Pengguna
1	KPK Kepolisian atau Kejaksaan sampai hari ini ga ada yg bergerak memeriksa @KemenkeuRI atas kegagalan Coretax ini? Brp kerugian negara di Januari 2025 krn pajak yg tidak terpungut? Belum masalah dari Coretax itu sendiri.

JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 3, Juni 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i3.8548
Hal 161-171
http://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

2	@kring_pajak halo saat saya sedang menambah orang terkait di coretax. Saya dpt issue ini. Kiranya bagaimana ya? Mohon bantuannya min https://t.co/Aadr8T2Vha
3	@MorphoMenelausX Gimana mau punya uang mau bayar pajak aja dipersulit coretax tolol aja tolol buat coretax malah nyusahin bayar pajak ibaratnya ada orang mau ngasi lu duit gratis lu suruh jalan jongkok keliling lapangan bola dulu 10x mana ada yg mau.
...	...
1628	@Antonius7088 Hai Kak. Untuk pembuatan billing PPh pasal 25 OP pada Coretax silakan pilih Pembayaran – Layanan Mandiri Kode Billing klik Lanjut lalu pilih KAP-KJS: 411125-100 pilih masa dan tahun pajaknya. Apabila Alamat WP yang tertera masih tidak sesuai dengan kondisi sesungguhnya

Tabel 1 menyajikan subset representatif dari 1.628 *tweet* yang berhasil dikumpulkan selama periode Januari-Maret 2025. Data diperoleh melalui API resmi platform X dengan *rate limiting* 500 *request*/hari.

3.2 *Preprocessing Data*

Sebelum proses pelabelan, data yang terkumpul melalui tahap *preprocessing* untuk meningkatkan akurasi pemodelan polaritas sikap. Setiap tahap dalam data *preprocessing*, data memainkan peranan yang krusial dalam menangani berbagai kendala dan menghilangkan potensi hambatan, sehingga dapat meningkatkan akurasi hasil analisis [16].

3.2.1 Data Cleaning

Data *cleaning* adalah tahap penting di mana data mentah seringkali mengandung informasi yang tidak relevan. Salah satu tantangan dalam proses pembersihan data adalah adanya nilai yang hilang, yaitu keadaan di mana data tidak tersedia atau tidak dapat diakses [17]. Tabel 2 memperlihatkan perbandingan hasil teks sebelum dan setelah proses pembersihan data.

Tabel 2. Hasil Cleaning Data

Tweet Pengguna	*Text Clean*
KPK Kepolisian atau Kejaksaan sampai hari ini ga ada yg bergerak memeriksa @KemenkeuRI atas kegagalan Coretax ini? Brp kerugian negara di Januari 2025 krn pajak yg tidak terpungut? Belum masalah dari Coretax itu sendiri.	KPK Kepolisian atau Kejaksaan sampai hari ini ga ada yg bergerak memeriksa atas kegagalan Coretax ini Brp kerugian negara di Januari krn pajak yg tidak terpungut Belum masalah dari Coretax itu sendiri

Pada Tabel 2 ditampilkan hasil proses pembersihan *(cleaning)* data terhadap *tweet* pengguna. Proses yang dilakukan meliputi penghapusan *mention* (@username), *hashtag* (#...), tautan *(link)*, *retweet* (RT), angka, serta karakter baris baru. Proses pembersihan teks juga melibatkan penghapusan seluruh tanda baca untuk memastikan fokus analisis hanya pada leksem bermakna. Selain itu, normalisasi spasi dilakukan dengan menghilangkan spasi berlebih di awal maupun akhir string teks, sehingga menghasilkan data yang lebih konsisten.

3.2.2 *Case Folding*

Case Folding merupakan langkah penting dalam proses *preprocessing* yang bertujuan untuk melakukan standarisasi penulisan dengan mengkonversi semua huruf kapital menjadi bentuk *lowercase*, sehingga menciptakan keseragaman dalam data teks. Tabel 3 memperlihatkan perbandingan hasil teks sebelum dan setelah proses *case folding*.

Tabel 3. Hasil Case Folding

Text Clean	*Case Folding*
KPK Kepolisian atau Kejaksaan sampai hari ini ga ada yg bergerak memeriksa atas kegagalan Coretax ini Brp kerugian negara di Januari krn pajak yg tidak terpungut Belum masalah dari Coretax itu sendiri	kpk kepolisian atau kejaksaan sampai hari ini ga ada yg bergerak memeriksa atas kegagalan coretax ini brp kerugian negara di januari krn pajak yg tidak terpungut belum masalah dari coretax itu sendiri

Dari Tabel 3 terlihat bahwa seluruh karakter yang sebelumnya menggunakan huruf kapital telah berubah menjadi huruf kecil. Proses ini penting untuk memastikan konsistensi dalam pengolahan teks. Sebagai contoh, kata 'KPK' dan 'kpk' sekarang dianggap identik oleh sistem setelah melalui tahap *case folding*.

3.2.3 Tokenization

Tokenization adalah proses memecah sebuah kalimat atau teks menjadi unit-unit linguistik terkecil yang disebut token, seperti kata-kata terpisah. Selain itu, tokenisasi juga mencakup langkah untuk menghapus simbol atau angka yang dapat mempengaruhi pemrosesan teks [18]. Tabel 4 memperlihatkan perbandingan hasil teks sebelum dan setelah proses tokenisasi.

Tabel 4. Hasil Tokenization

Case Folding	*Tokenization*
kpk kepolisian atau kejaksaan sampai hari ini ga ada yg bergerak memeriksa atas kegagalan coretax ini	['kpk', 'kepolisian', 'atau', 'kejaksaan', 'sampai', 'hari', 'ini', 'ga', 'ada', 'yg', 'bergerak', 'memeriksa', 'atas', 'kegagalan', 'coretax',

JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 3, Juni 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i3.8548
Hal 161-171
http://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

brp kerugian negara di januari krn pajak yg tidak terpungut belum masalah dari coretax itu sendiri	'ini', 'brp', 'kerugian', 'negara', 'di', 'januari', 'krn', 'pajak', 'yg', 'tidak', 'terpungut', 'belum', 'masalah', 'dari', 'coretax', 'itu', 'sendiri'],

Dapat diamati pada Tabel 4 proses tokenisasi berhasil memecah kalimat utuh menjadi token-token individual. Proses ini mempermudah analisis lanjutan karena teks telah disusun dalam bentuk yang lebih terstruktur.

3.2.4 Stopword Removal

Tahap *stopword removal* dilakukan dengan memanfaatkan filter *stopword* untuk mengeliminasi kata-kata yang tidak memiliki nilai informatif, mengacu pada daftar kata penghubung baku dalam bahasa Indonesia. Dengan demikian, proses ini berfungsi untuk menyaring kata-kata agar lebih bermakna [19]. Tabel 5 memperlihatkan perbandingan hasil teks sebelum dan setelah penerapan *stopword removal*.

Tabel 5. Hasil Stopword Removal

Text Clean	*Stopword Removal*
KPK Kepolisian atau Kejaksaan sampai hari ini ga ada yg bergerak memeriksa atas kegagalan Coretax ini Brp kerugian negara di Januari krn pajak yg tidak terpungut Belum masalah dari Coretax itu sendiri	['kpk', 'kepolisian', 'kejaksaan', 'sampai', 'hari', 'ga', 'bergerak', 'memeriksa', 'kegagalan', 'coretax', 'brp', 'kerugian', 'negara', 'januari', 'krn', 'pajak', 'terpungut', 'belum', 'masalah', 'coretax', 'sendiri']

Pada Tabel 5, dapat diamati bahwa kata-kata fungsi (*stopwords*) seperti 'atau', 'di', dan 'yang' telah dihilangkan. Hal ini mengurangi *noise* pada data dan meningkatkan efisiensi pemrosesan teks.

3.2.5 Stemming

Stemming adalah tahap dalam pemrosesan teks yang bertujuan untuk mereduksi suatu kata menjadi bentuk dasarnya, tanpa tambahan imbuhan di awal maupun di akhir, serta tanpa sisipan atau pengulangan dalam penggunaan kata tersebut [18]. Tabel 6 memperlihatkan perbandingan hasil teks sebelum dan setelah proses *stemming*.

Tabel 6. Hasil Stemming

Stopword Removal	*Stemming*
['kpk', 'kepolisian', 'kejaksaan', 'sampai', 'hari', 'ga', 'bergerak', 'memeriksa', 'kegagalan', 'coretax', 'brp', 'kerugian', 'negara', 'januari', 'krn', 'pajak', 'terpungut', 'belum', 'masalah', 'coretax', 'sendiri'],	['kpk', 'polisi', 'jaksa', 'sampai', 'hari', 'ga', 'gerak', 'periksa', 'gagal', 'coretax', 'brp', 'rugi', 'negara', 'januari', 'krn', 'pajak', 'pungut', 'belum', 'masalah', 'coretax', 'diri']

Proses *stemming* yang ditampilkan pada Tabel 6 berhasil menormalisasi variasi kata ke bentuk dasarnya. Contohnya dapat dilihat pada transformasi kata 'kepolisian' menjadi bentuk dasar 'polisi'. Meskipun akurasi *stemming* tidak sempurna, teknik ini membantu standardisasi kata dengan akar yang sama.

3.2.6 Normalisasi

Normalisasi membuat standar data guna mempermudah proses analisis lebih lanjut. Dalam penelitian ini, normalisasi dicapai melalui penggunaan kamus yang dikurasi secara manual. Tabel 7 memperlihatkan perbandingan hasil teks sebelum dan setelah dilakukan normalisasi.

Tabel 7. Hasil Normalisasi

Stemming	*Normalisasi*
['kpk', 'polisi', 'jaksa', 'sampai', 'hari', 'ga', 'gerak', 'periksa', 'gagal', 'coretax', 'brp', 'rugi', 'negara', 'januari', 'krn', 'pajak', 'pungut', 'belum', 'masalah', 'coretax', 'diri']	['kpk', 'polis', 'jaksa', 'sampai', 'hari', 'tidak', 'gerak', 'periksa', 'gagal', 'coretax', 'berapa', 'rugi', 'negara', 'januari', 'karena', 'pajak', 'pungut', 'belum', 'masalah', 'coretax', 'diri']

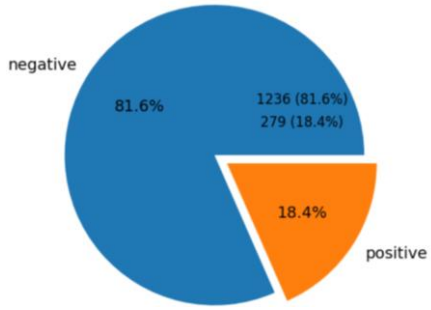
Variabel hasil *preprocessing* yang ditampilkan pada Tabel 7 adalah hasil dari proses normalisasi, yang menjadi dasar dalam proses pelabelan sentimen. Normalisasi dilakukan dengan mengganti kata tidak baku dan singkatan menjadi bentuk baku agar sesuai dengan entri dalam kamus *Lexicon*. Misalnya, kata 'ga' dinormalisasi menjadi 'tidak', dan 'brp' menjadi 'berapa'. Proses ini meningkatkan konsistensi data untuk analisis sentimen.

3.3 Pelabelan Data

Tahap pelabelan dilakukan dengan mengakumulasikan nilai bobot setiap kata dalam tweet berdasarkan daftar kata yang terdapat dalam leksikon bahasa Indonesia. Penelitian ini memanfaatkan leksikon Inset yang dikembangkan oleh Fajri Koto dan Gemal Y. Rahmaningtyas dalam publikasinya berjudul *"Inset Lexicon : Evaluation of a Word List for Indonesian Sentiment Analysis in Microblogs"*, di mana kata-kata dalam leksikon tersebut dikumpulkan dari Twitter [20]. Metode ini beroperasi melalui mekanisme yang cukup sederhana. Pertama, setiap kata dalam kumpulan data dicocokkan dengan entri dalam kamus leksikal yang telah dilengkapi dengan sistem pembobotan dalam rentang -5 (paling negatif) hingga +5 (paling positif). Setelah nilai bobot seluruh kata berhasil diidentifikasi, dilakukan penjumlahan terhadap semua

JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 3, Juni 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i3.8548
Hal 161-171
http://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

bobot kata yang berasal dari tweet yang sama untuk memperoleh skor polaritas setiap unggahan. Skor polaritas ini kemudian diklasifikasikan ke dalam dua kategori sentimen: (1) negatif apabila total bobot bernilai di bawah nol, dan (2) positif bila total bobot menunjukkan nilai di atas nol. Distribusi label sentimen divisualisasikan pada Gambar 2 untuk menunjukkan sebaran data sebelum pemodelan.



Gambar 2. Hasil Pelabelan dengan Inset Lexicon

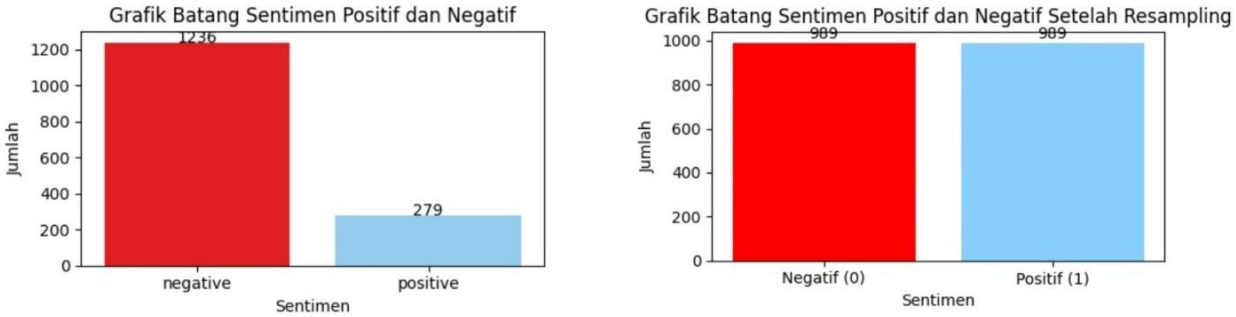
Analisis distribusi kelas pada Gambar 2 menunjukkan ketidakseimbangan antara kelas sentiment, dengan proposi negative 81.6% dan positif 18.4%. Dominasi kelas negatif ini dapat menyebabkan bias pada model klasifikasi, dimana prediksi akan cenderung mengarah ke kelas mayoritas (negatif) jika tidak dilakukan penanganan khusus.

3.4 Pembagian Data

Setelah pelabelan data berhasil dilakukan, langkah selanjutnya adalah memisahkan dataset menjadi dua komponen yang berbeda. Di mana sebanyak 80% data dimanfaatkan untuk proses pelatihan model, sedangkan 20% sisanya digunakan dalam tahap evaluasi kinerja model. Penerapan metode pengambilan sampel bertingkat memastikan bahwa distribusi label tetap seimbang di kedua subset data.

3.5 Penyemibang Data dengan SMOTE

Teknik SMOTE mengatasi ketidakseimbangan kelas. Berbeda dengan *oversampling* konvensional yang hanya menduplikasi *instance* minoritas, teknik ini secara cerdas menghasilkan contoh sintetik baru melalui interpolasi linier antar titik data minoritas yang berdekatan dalam ruang fitur [21]. Implementasi ini berhasil meningkatkan jumlah sampel kelas minoritas secara signifikan dari 279 menjadi 989, sehingga menciptakan distribusi data yang lebih seimbang untuk pelatihan model. Gambar 3 memvisualisasikan perbandingan distribusi kelas sebelum dan setelah augmentasi data dengan SMOTE.



Gambar 3. Perbandingan Distribusi Kelas Sebelum dan Sesudah SMOTE

Sebagai solusi untuk menangani ketidakseimbangan kelas dimana sampel kelas minoritas (sentimen positif) jumlahnya signifikan lebih sedikit, diimplementasikan teknik SMOTE (*Synthetic Minority Over-sampling Technique*). Hasil implementasi yang divisualisasikan dalam Gambar 3 menunjukkan distribusi kelas yang lebih seimbang setelah augmentasi data sintetik dilakukan.

3.6 Pembobotan TF-IDF

Pada tahap TF-IDF, teks dikonversi ke dalam bentuk numerik. Pembobotan TF-IDF menghasilkan representasi numerik yang lebih informatif dari teks, sehingga dapat dimanfaatkan secara optimal oleh algoritma klasifikasi [22]. Tabel 8 mempresentasikan 10 term dengan nilai TF-IDF tertinggi dari kumpulan dokumen teks yang telah melalui proses tahap pra-pemrosesan.

Tabel 8. Representasi TF-IDF

No.	Term	TF-IDF
1	pajak	101.387428

JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 3, Juni 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i3.8548
Hal 161-171
http://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

2	coretax	90.839502
3	ya	52.660620
4	min	50.728848
5	lapor	47.765519
6	bayar	41.237209
7	spt	41.003940
8	faktur	40.745694
9	yg	38.738737
10	kak	38.291824

Dari hasil analisis pada Tabel 8 teridentifikasi bahwa kata-kata seperti 'pajak' (skor TF-IDF: 101.39) dan 'coretax' (skor TF-IDF: 90.83) memiliki skor yang tinggi, menunjukkan bahwa kata-kata tersebut muncul cukup sering dalam dokumen-dokumen teks tertentu.

3.7 Klasifikasi *Naïve Bayes*

Selanjutnya, klasifikasi dilakukan menggunakan metode *Naïve Bayes*. Metode ini bekerja dengan asumsi bahwa setiap atribut bersifat independen [23]. Selain itu, kata-kata yang telah diberi bobot digunakan dalam fase pelatihan guna meningkatkan akurasi model *Naïve Bayes*. Gambar 4 memaparkan hasil evaluasi model pada data latih dan data uji.

```
Train report
              precision    recall  f1-score   support

    negative       1.00      0.97      0.99       989
    positive       0.97      1.00      0.99       989

    accuracy                           0.99      1978
   macro avg       0.99      0.99      0.99      1978
weighted avg       0.99      0.99      0.99      1978


Test report
              precision    recall  f1-score   support

    negative       0.95      0.96      0.96       247
    positive       0.83      0.79      0.81        56

    accuracy                           0.93       303
   macro avg       0.89      0.87      0.88       303
weighted avg       0.93      0.93      0.93       303
```

Gambar 4. Hasil Modeling Naïve Bayes

Seperti yang ditunjukkan pada Gambar 4, hasil modeling menunjukkan tingkat akurasi yang tinggi, dengan akurasi 99% pada data pelatihan dan 93% pada data pengujian. Namun, kelas positif masih kurang optimal, terutama dalam hal *recall*, hanya mencapai 79% pada data pengujian. Sebaliknya, kelas negatif menunjukkan nilai *recall* sangat tinggi sebesar 96%. Perbedaan ini kemungkinan besar bersumber dari distribusi data yang tidak proporsional antar kelas, mengakibatkan kecenderungan model untuk memprioritaskan identifikasi kelas yang lebih dominan.

3.8 Evaluasi

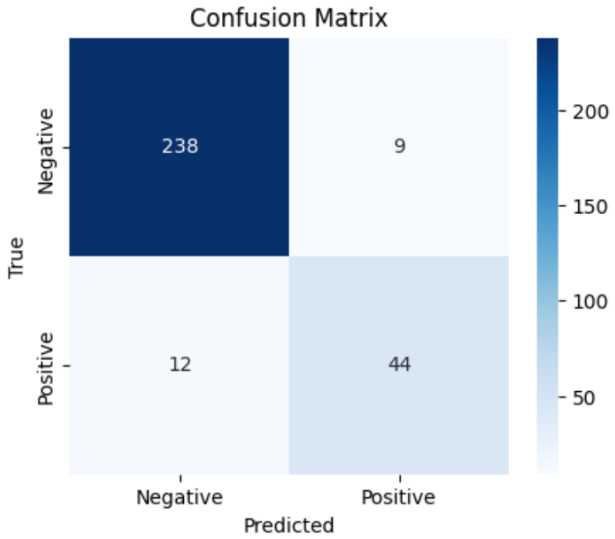
Proses evaluasi bertujuan untuk mengukur efektivitas dan kualitas model yang diterapkan [24]. Hasil dari evaluasi ini akan membantu dalam menentukan skenario yang menghasilkan nilai Akurasi, Presisi dan *Recall* yang optimal [25]. Proses pengujian dalam data mining klasifikasi dilakukan untuk memprediksi keakuratan suatu objek. Dalam pengujian ini, digunakan *confusion matrix* sebagai alat evaluasi. Pada matriks tersebut, kelas prediksi diletakkan pada bagian atas, sedangkan kelas aktual yang diamati ditempatkan di sisi kiri matriks [26]. Gambar 5 merupakan representasi *Confusion Matrix* yang menjadi dasar perhitungan akurasi, presisi, *recall*, dan *F1-Score*.

```
Test Accuracy: 0.9307
Classification Report:
              precision    recall  f1-score   support

    negative       0.95      0.96      0.96       247
    positive       0.83      0.79      0.81        56

    accuracy                           0.93       303
   macro avg       0.89      0.87      0.88       303
weighted avg       0.93      0.93      0.93       303
```

Gambar 5. Hasil Evaluasi Model

JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 3, Juni 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i3.8548
Hal 161-171
http://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Berdasarkan Gambar 5, setelah dilakukan evaluasi menggunakan data uji, diperoleh akurasi sebesar 93.07%. Kelas "negative" memiliki performa yang sangat baik dengan nilai *precision* 0.95, *recall* 0.96, dan *F1-score* 0.96. Sementara itu, kelas "positive" menunjukkan performa yang sedikit lebih rendah, dengan nilai *precision* 0.83, *recall* 0.79, dan *F1-score* 0.81. Nilai *weighted average* yang konsisten pada ketiga metrik utama 0.93 mencerminkan performa model yang stabil dalam mengklasifikasikan data secara keseluruhan. Visualisasi *confusion matrix* pada Gambar 6 memudahkan interpretasi performa model.



Gambar 6. Visualisasi Confusion Matrix

Gambar 6 memvisualisasikan confusion matrix yang menggambarkan performa model klasifikasi melalui empat komponen evaluasi utama. Dominasi warna gelap pada elemen diagonal, yaitu true positive (TP) dan true negative (TN), mengindikasikan bahwa model memiliki tingkat akurasi yang tinggi serta kemampuan klasifikasi yang tepat. Tabel 9 menyajikan *confusion matrix* yang digunakan dalam pengukuran kinerja model.

Tabel 9. Confusion Matrix

	Predicted Negative	Predicted Positive
Actual Negative	238	9
Actual Positive	12	44

Berdasarkan Tabel 9, diperoleh bahwa model menghasilkan 238 *True Negative* (TN), 9 *False Positive* (FP), 12 *False Negative* (FN), dan 44 *True Positive* (TP). Nilai-nilai ini digunakan untuk menghitung metrik evaluasi sebagai berikut:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} = \frac{44+238}{44+9+238+12} = \frac{282}{303} = 0.9307$$

$$Precission = \frac{TP}{TP+FP} = \frac{44}{44+9} = \frac{44}{53} = 0.83$$

$$Recall = \frac{TP}{TP+FN} = \frac{44}{44+12} = \frac{44}{56} = 0.79$$

$$F-Measure = \frac{2\,(recall\,.precision)}{(recall\ +\ precision)} = 2 \times \frac{0.83 \times 0.79}{0.83+0.79} = \frac{0.3114}{1.62} = 0.81$$

Berdasarkan analisis terhadap sikap pengguna aplikasi, model klasifikasi yang dikembangkan berhasil mencapai akurasi keseluruhan 93.07%. Pemantauan lebih mendalam melalui confusion matrix mengungkapkan performa yang sangat baik dalam mengidentifikasi sentimen negatif, dengan *precision* 95%, *recall* 96%, dan *f1-score* 96%. Angka-angka ini menunjukkan konsistensi dan reliabilitas model dalam mengenali pola-pola sentimen negatif.

Sementara itu, pada sentimen positif, model mencapai *precision* 83%, *recall* 79%, dan *f1-score* 81%. Nilai *F1-score* yang diperoleh menunjukkan keseimbangan yang relatif baik antara presisi dan recall dalam mengidentifikasi sentimen positif. Meski demikian, performa model untuk kelas positif masih tertinggal sedikit dibandingkan dengan kelas negatif. Kondisi ini mengisyaratkan perlunya optimasi lebih lanjut, terutama dalam meningkatkan sensitivitas model terhadap berbagai ekspresi sentimen positif agar tingkat pendeteksiannya *(recall)* dapat ditingkatkan.

JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 3, Juni 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i3.8548
Hal 161-171
http://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

4. KESIMPULAN

Berdasarkan hasil perhitungan metrik evaluasi, penerapan metode *Naïve Bayes* dalam klasifikasi sentimen menunjukkan kinerja yang sangat memuaskan. Akurasi tertinggi dicapai pada skenario dengan pembagian data 80% untuk pelatihan dan 20% untuk pengujian. Pada skenario ini, model berhasil mencapai akurasi 93.07%, yang tergolong sangat baik untuk sebuah model *machine learning*. Hasil ini menunjukkan bahwa model mampu belajar secara efektif dari data pelatihan dan memberikan prediksi yang akurat pada data uji. Analisis lebih lanjut menggunakan metrik klasifikasi mengungkapkan bahwa model unggul dalam mengidentifikasi sentimen negatif. Namun, performanya dalam mendeteksi sentimen positif masih belum maksimal. Ketidakseimbangan ini diduga disebabkan oleh beberapa faktor, seperti representasi fitur yang kurang optimal atau kompleksitas pola dalam data sentimen positif. Untuk meningkatkan performa model, eksplorasi terhadap teknik ekstraksi fitur yang lebih komprehensif perlu dilakukan. Selain itu, penggunaan algoritma yang lebih kompleks seperti *Support Vector Machine* (SVM), *Random Forest*, atau pendekatan berbasis *deep learning* juga dapat menjadi alternatif yang layak dipertimbangkan. Diharapkan dengan penerapan teknik-teknik tersebut, model dapat mencapai performa yang lebih seimbang dan *robust* di berbagai skenario klasifikasi. Pengembangan model selanjutnya juga harus memperhatikan aspek interpretabilitas. Dalam penerapan dunia nyata, kemampuan untuk menjelaskan alasan di balik suatu prediksi tidak kalah penting dibandingkan akurasi itu sendiri. Oleh karena itu, model klasifikasi sentimen yang dikembangkan tidak hanya harus akurat, tetapi juga mudah dipahami dan dapat dijelaskan secara logis dalam berbagai konteks aplikasi.

REFERENCES

[1] R. Harun, R. Ishak, and S. Panna, "Analisis Sentimen Opini Publik Pengguna Twitter Terhadap Kenaikan Harga BBM Menggunakan Algoritma Naïve Bayes," *J. Ilm. Ilmu Komput. Banthayo Lo Komput.*, vol. 2, no. 1, pp. 26–33, 2023, doi: 10.37195/balok.v2i1.414.

[2] P. Arsi and R. Waluyo, "Analisis Sentimen Wacana Pemindahan Ibu Kota Indonesia Menggunakan Algoritma Support Vector Machine (SVM)," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 8, no. 1, p. 147, 2021, doi: 10.25126/jtiik.0813944.

[3] D. D. Tarigan, S. Iskandar, and A. Idrus, "Journal of Informatics and Data Science (J-IDS) Sentiment Analysis of Twitter Users Regarding Taxation Topics in," vol. 03, no. 01, 2024, doi: 10.24114/j-ids.xxxxx.

[4] H. B. Rochmanto, G. Anuraga, M. Athoillah, A. Azies, and M. Naufal, "Klasifikasi Opini Publik terhadap Kenaikan PPN 12 % di Platform X menggunakan Multinomial Naïve Bayes," UJMC, vol. 10, pp. 57–66, 2024.

[5] S. H. Ramadhani and M. I. Wahyudin, "Analisis Sentimen Terhadap Vaksinasi Astra Zeneca pada Twitter Menggunakan Metode Naïve Bayes dan K-NN," *J. JTIK (Jurnal Teknol. Inf. dan Komunikasi)*, vol. 6, no. 4, pp. 526–534, 2022, doi: 10.35870/jtik.v6i4.530.

[6] F. I. Deni Wijaya, Rizki Adi Saputra, "Analisis Sentimen Ulasan Aplikasi Samsat Digital Nasional Pada Google Playstore Menggunakan Algoritma Naïve Bayes," *KLIK Kaji. Ilm. Inform. dan Komput.*, vol. 4, pp. 2369–2380, 2024, doi: DOI 10.30865/klik.v4i4.1738.

[7] S. H. W. Putra and D. Febriawan, "Analisis Sentimen Ulasan Aplikasi Digital Korlantas POLRI Menggunakan Naïve Bayes pada Google Play Store," *KLIK Kaji. Ilm. Inform. dan Komput.*, vol. 4, no. 4, pp. 1962–1971, 2024, doi: 10.30865/klik.v4i4.1600.

[8] R. Firdaus, I. Asror, and A. Herdiani, "Lexicon-Based Sentiment Analysis of Indonesian Language Student Feedback Evaluation," *Indones. J. Comput.*, vol. 6, no. 1, pp. 1–12, 2021, doi: 10.34818/indojc.2021.6.1.408.

[9] E. R. Subhiyakto, Y. P. Astuti, N. Alexander, and E. Kartikadarma, "Analisis Sentimen Menggunakan Metode Naïve Bayes Untuk Mengetahui Respon Masyarakat Terhadap Vaksinasi," *J. Ilm. Intech Inf. Technol. J. UMUS*, vol. 4, no. 02, pp. 179–188, 2022, doi: 10.46772/intech.v4i02.864.

[10] A. R. Ismail and Raden Bagus Fajriya Hakim, "Implementasi Lexicon Based Untuk Analisis Sentimen Dalam Menentukan Rekomendasi Pantai Di DI Yogyakarta Berdasarkan Data Twitter," *Emerg. Stat. Data Sci. J.*, vol. 1, no. 1, pp. 37–46, 2023, doi: 10.20885/esds.vol1.iss.1.art5.

[11] Q. Gao *et al.*, "Identification of Orphan Genes in Unbalanced Datasets Based on Ensemble Learning," *Front. Genet.*, vol. 11, Oct. 2020, doi: 10.3389/fgene.2020.00820.

[12] A. S. Almajid, "Multilayer Perceptron Optimization on Imbalanced Data Using SVM-SMOTE and One-Hot Encoding for Credit Card Default Prediction," *J. Adv. Inf. Syst. Technol.*, vol. 3, no. 2, pp. 67–74, Sep. 2022, doi: 10.15294/jaist.v3i2.57061.

[13] 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.

[14] K. V. S. Toy, Y. A. Sari, and I. Cholissodin, "Analisis Sentimen Twitter menggunakan Metode Naive Bayes dengan Relevance Frequency Feature Selection (Studi Kasus: Opini Masyarakat mengenai Kebijakan New Normal)," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 11, pp. 5068–5074, 2021, [Online]. Available: http://j-ptiik.ub.ac.id

[15] P. Simanjuntak, H. Pangaribuan, and M. T. Syastra, "Data Mining Rekomendasi Pemakaian Skincare," *MEANS (Media Inf. Anal. dan Sist.*, no. February, pp. 80–83, 2021, doi: 10.54367/means.v6i1.1224.

[16] Q. A. A'yuniyah and M. Reza, "Penerapan Algoritma K-Nearest Neighbor Untuk Klasifikasi Jurusan Siswa Di Sma Negeri 15 Pekanbaru," *Indones. J. Inform. Res. Softw. Eng.*, vol. 3, no. 1, pp. 39–45, 2023, doi: 10.57152/ijirse.v3i1.484.

[17] A. A. Nopebrian, R. N. Shofa, and S. Yuliyanti, "Perbandingan Algoritma Pendekatan Supervised Learning Menggunakan Seleksi Fitur Chi-Square untuk Klasifikasi Status Kesehatan Jemaah Haji Comparison of Algorithms Supervised Learning Approach Using Chi-Square Feature Selection for Classification of Hajj P," vol. 13, no. 1, pp. 166–172, 2025, doi: 10.26418/justin.v13i1.86639.

[18] F. A. Irawan and D. A. Rochmah, "Penerapan Algoritma CNN Untuk Mengetahui Sentimen Masyarakat Terhadap Kebijakan Vaksin Covid-19," *J. Inform.*, vol. 9, no. 2, pp. 148–158, 2022, doi: 10.31294/inf.v9i2.13257.

[19] N. F. Hilmi and F. Irwiensyah, "Analisis Sentimen Terhadap Aplikasi Tiktok Dari Ulasan Pada Google Playstore Menggunakan Metode Naïve Bayes," *Smatika J.*, vol. 14, no. 01, pp. 146–156, 2024, doi: 10.32664/smatika.v14i01.1210.

[20] F. Koto and G. Y. Rahmaningtyas, "Inset lexicon: Evaluation of a word list for Indonesian sentiment analysis in microblogs," *Proc. 2017 Int. Conf. Asian Lang. Process. IALP 2017*, vol. 2018-Janua, no. December, pp. 391–394, 2017, doi: 10.1109/IALP.2017.8300625.

[21] A. N. Rais, "Integrasi SMOTE Dan Ensemble AdaBoost Untuk Mengatasi Imbalance Class Pada Data Bank Direct Marketing," *J. Inform.*, vol. 6, no. 2, pp. 278–285, Sep. 2019, doi: 10.31311/ji.v6i2.6186.

[22] M. Apriliyani, M. I. Musyaffaq, S. Nur'Aini, M. R. Handayani, and K. Umam, "Implementasi analisis sentimen pada ulasan aplikasi Duolingo di Google Playstore menggunakan algoritma Naïve Bayes," *AITI*, vol. 21, no. 2, pp. 298–311, Sep. 2024, doi: 10.24246/aiti.v21i2.298-311.

[23] J. A. Rieuwpassa, S. Sugito, and T. Widiharih, "Implementasi Metode Naive Bayes Classifier Untuk Klasifikasi Sentimen Ulasan Pengguna Aplikasi Netflix Pada Google Play," *J. Gaussian*, vol. 12, no. 3, pp. 362–371, 2024, doi: 10.14710/j.gauss.12.3.362-371.

[24] F. N. A. A. H. W. R. S. S. N. A. D. P. Y. A. F. E. A. I. H. I. Y. S. Z. G. C. P. D. G. E. S. N. Ni Luh Wiwik Sri Rahayu Ginantra, *1.2 FullBook Data Mining dan Penerapan Algoritma*. Yayasan Kita Menulis, 2021.

[25] B. Z. Ramadhan, R. I. Adam, and I. Maulana, "Analisis Sentimen Ulasan pada Aplikasi E-Commerce dengan Menggunakan Algoritma Naïve Bayes," *J. Appl. Informatics Comput.*, vol. 6, no. 2, pp. 220–225, Dec. 2022, doi: 10.30871/jaic.v6i2.4725.

[26] M. A. Muslim *et al.*, *Data Mining Algoritma C4.5 Disertai contoh kasus dan penerapannya dengan program computer*, Pertama. Semarang, 2019.