Jutisi: Jurnal Ilmiah Teknik Informatika dan Sistem Informasi
Jl. Ahmad Yani, K.M. 33,5 - Kampus STMIK Banjarbaru
Loktabat – Banjarbaru (Tlp. 0511 4782881), e-mail: puslit.stmikbjb@gmail.com
e-ISSN: 2685-0893
p-ISSN: 2089-3787

Analisis Sentimen Publik Terhadap Polusi Udara di Kota Jakarta: Perbandingan Algoritma *Support Vector Machine, Naive Bayes*, dan *Random Forest*

Denas Aria Pamungkas[1*], Amali[2], Ucok Darmanto Soer[3]
Teknik Informatika, Universitas Pelita Bangsa, Cikarang, Indonesia
*e-mail *Corresponding Author:* denasaria@mhs.pelitabangsa.ac.id

Abstract

Air pollution in Jakarta has become a serious environmental issue, with pollutant concentrations such as PM2.5 and PM10 often exceeding safe limits, adversely affecting public health. This research analyzes public sentiment towards air pollution using machine learning algorithms: Support Vector Machine (SVM), Naive Bayes, and Random Forest. Data was collected from the social media platform Twitter through data crawling, with preprocessing steps such as cleaning, tokenizing, and stemming. The comparison of algorithm performance was conducted using accuracy, precision, recall, and F1-score metrics. The research results show that SVM has the highest accuracy at 91%, followed by Naive Bayes (85%) and Random Forest (81%). Public sentiment is dominated by negative opinions, reflecting concerns about the health impacts of air pollution. This study concludes that the SVM algorithm is the most effective for public sentiment analysis and can serve as a basis for the government in formulating more responsive and data-driven policies.

Keywords*: Support Vector Machine; Naïve Bayes; Random Forest; Sentiment Analysis; Air Pollution*

Abstrak

Polusi udara di Jakarta menjadi isu lingkungan serius dengan konsentrasi polutan seperti PM2.5 dan PM10 sering melebihi ambang batas aman, berdampak buruk pada kesehatan masyarakat. Penelitian ini menganalisis sentimen publik terhadap polusi udara menggunakan algoritma pembelajaran mesin: *Support Vector Machine* (SVM), *Naive Bayes*, dan *Random Forest*. Data diambil dari media sosial Twitter melalui crawling data, dengan proses preprocessing seperti *cleaning, tokenizing*, dan *stemming*. Perbandingan performa algoritma dilakukan menggunakan metrik akurasi, presisi, recall, dan *F1-score*. Hasil penelitian menunjukkan bahwa SVM memiliki akurasi tertinggi sebesar 91%, diikuti oleh *Naive Bayes* (85%) dan *Random Forest* (81%). Sentimen publik didominasi oleh opini negatif, mencerminkan kekhawatiran terhadap dampak kesehatan akibat polusi udara. Penelitian ini menyimpulkan bahwa algoritma SVM paling efektif untuk analisis sentimen publik dan dapat menjadi dasar bagi pemerintah dalam merumuskan kebijakan yang lebih responsif dan berbasis data.

Kata kunci*:* *Support Vector Machine; Naïve Bayes; Random Forest; Analisis Sentimen; Polusi Udara*

1. Pendahuluan

Polusi udara merupakan salah satu masalah lingkungan paling mendesak yang dihadapi dunia, terutama di kota-kota besar dengan tingkat urbanisasi dan industrialisasi yang pesat[1]. Di tingkat global, *World Health Organization* (WHO) telah menetapkan bahwa polusi udara adalah salah satu dari sepuluh ancaman kesehatan terbesar[2]. Polusi udara berkontribusi terhadap peningkatan risiko berbagai penyakit kronis, seperti penyakit kardiovaskular, penyakit paru obstruktif kronis (PPOK), dan kanker paru-paru[3]. Dalam konteks nasional, Jakarta, sebagai ibu kota sekaligus pusat ekonomi Indonesia, menghadapi tantangan berat dalam menjaga kualitas udara yang baik. Polusi udara tidak hanya berdampak pada kesehatan masyarakat secara langsung, tetapi juga memengaruhi kualitas hidup, produktivitas

ekonomi, dan kesejahteraan sosial masyarakat. Isu ini menjadi semakin relevan seiring dengan semakin tingginya jumlah kendaraan bermotor, intensitas kegiatan industri, serta pembangunan infrastruktur yang masif di Jakarta. Oleh karena itu, tema ini penting untuk diteliti guna memberikan pemahaman yang lebih mendalam mengenai dampak polusi udara dan bagaimana masyarakat mengartikulasikan kekhawatiran mereka terhadap isu tersebut.

Kondisi kualitas udara di Jakarta sering kali berada dalam kategori berbahaya, terutama selama musim kemarau, ketika polutan seperti PM2.5 dan PM10 mencapai tingkat konsentrasi yang jauh melampaui ambang batas aman yang ditetapkan WHO[4]. Berdasarkan data Pemerintah Provinsi DKI Jakarta, tingginya konsentrasi polutan tersebut secara signifikan berkontribusi terhadap peningkatan jumlah penderita penyakit terkait polusi udara, seperti asma, bronkitis, dan PPOK. Selain itu, dampak polusi udara tidak hanya terbatas pada kesehatan fisik, tetapi juga pada kesejahteraan psikologis masyarakat[5]. Warga yang hidup di tengah lingkungan dengan kualitas udara buruk sering kali merasa cemas terhadap risiko kesehatan jangka panjang dan masa depan lingkungan tempat mereka tinggal. Namun, kebijakan yang ada saat ini masih dirasa belum cukup efektif dalam menangani masalah ini secara holistik[6]. Salah satu kesenjangan yang terlihat adalah minimnya pendekatan berbasis data untuk memahami persepsi masyarakat terhadap isu polusi udara. Padahal, persepsi publik memegang peranan penting dalam mendorong implementasi kebijakan yang lebih responsif dan tepat sasaran. Masalah lainnya adalah ketidakadilan sosial yang muncul akibat polusi udara, di mana kelompok masyarakat yang tinggal di daerah kumuh atau padat penduduk cenderung lebih terpapar polusi dibandingkan dengan kelompok masyarakat lainnya.

Dalam rangka mencari solusi terhadap masalah polusi udara di Jakarta, analisis sentimen publik berbasis data besar (*big data*) menawarkan pendekatan yang inovatif dan relevan[7]. Media sosial seperti Twitter, forum diskusi, dan portal berita merupakan sumber data real-time yang mencerminkan opini dan kekhawatiran masyarakat terhadap isu ini. Dengan memanfaatkan algoritma machine learning seperti *Support Vector Machine* (SVM), *Naive Bayes*, dan *Random Forest*. SVM adalah algoritma yang bekerja dengan mencari hyperplane terbaik untuk memisahkan data ke dalam kelas yang berbeda, efektif untuk data berdimensi tinggi dengan tingkat akurasi tinggi[8]..*Naive Bayes*, berbasis probabilitas dan Teorema Bayes, adalah algoritma yang sederhana, cepat, dan efisien untuk klasifikasi teks, meskipun mengasumsikan independensi antar fitur[9]. Sementara itu, *Random Forest* menggunakan pendekatan ensemble dengan membangun banyak pohon keputusan pada subset data, menghasilkan klasifikasi yang stabil dan akurat[10]. Penggunaan algoritma ini memungkinkan pengelompokkan sentimen menjadi positif, negatif, atau netral, sekaligus memberikan wawasan mendalam tentang aspek-aspek spesifik yang menjadi perhatian masyarakat[11]. Selain itu, data sentimen publik dapat diintegrasikan dengan data kualitas udara dari lembaga resmi seperti Badan Meteorologi, Klimatologi, dan Geofisika (BMKG). Pendekatan ini tidak hanya membantu memahami hubungan antara fluktuasi tingkat polusi udara dengan perubahan sentimen publik, tetapi juga mengidentifikasi kelompok masyarakat yang paling rentan terkena dampaknya[12]. Konsep solusi ini diharapkan mampu memberikan dasar yang kuat untuk merumuskan kebijakan yang lebih adil, inklusif, dan berbasis bukti (*evidence-based policy*).

Penelitian ini bertujuan untuk menganalisis sentimen publik terhadap polusi udara di Jakarta dengan memanfaatkan data dari berbagai sumber digital seperti media sosial, portal berita, dan forum diskusi. Penelitian ini juga bertujuan untuk membandingkan performa algoritma machine learning, seperti SVM, *Naive Bayes*, dan *Random Forest*, dalam konteks analisis sentimen publik terhadap isu lingkungan. Dengan mengintegrasikan data sentimen dengan data kualitas udara dari BMKG, penelitian ini diharapkan mampu memberikan gambaran komprehensif mengenai bagaimana persepsi masyarakat dipengaruhi oleh kondisi kualitas udara yang riil. Selain itu, penelitian ini bertujuan untuk memberikan rekomendasi kebijakan yang lebih responsif dan berbasis data, terutama dalam konteks pengelolaan kualitas udara dan perlindungan kesehatan masyarakat. Manfaat penelitian ini tidak hanya terbatas pada aspek teoretis, tetapi juga praktis. Secara teoretis, penelitian ini memberikan kontribusi dalam pengembangan metode analisis sentimen berbasis machine learning di bidang lingkungan. Secara praktis, penelitian ini diharapkan dapat membantu pembuat kebijakan dalam merumuskan langkah-langkah strategis untuk meningkatkan kualitas udara di Jakarta. Penelitian ini juga bermanfaat untuk meningkatkan kesadaran publik tentang pentingnya peran

mereka dalam menjaga lingkungan serta mendorong kolaborasi yang lebih erat antara pemerintah, masyarakat, dan sektor swasta dalam mencari solusi terhadap masalah polusi udara. Dengan pendekatan yang komprehensif dan interdisipliner, penelitian ini diharapkan mampu memberikan dampak positif tidak hanya untuk Jakarta, tetapi juga untuk kota-kota besar lainnya di Indonesia yang menghadapi tantangan serupa.

2. Tinjauan Pustaka

Penelitian oleh Dhani Wahyu Wicaksono dan Budi Hartono berfokus pada analisis sentimen Twitter terhadap kualitas udara Jakarta menggunakan metode *Naive Bayes* Classifier (NBC). Dengan menganalisis 500 tweet terkait kualitas udara, hasil penelitian menunjukkan dominasi sentimen negatif sebesar 14%, dibandingkan sentimen positif yang hanya 7%. Penelitian ini memanfaatkan pembagian data latih dan uji dengan proporsi yang ketat serta menggunakan matriks konfusi untuk mengukur performa klasifikasi. NBC berhasil mencapai akurasi sebesar 87,50% dengan recall 93,33%, presisi 87,50%, dan *F1-score* 82,35%. Hasil ini menggarisbawahi keandalan NBC dalam menganalisis sentimen publik secara efisien, memberikan wawasan tentang persepsi masyarakat terhadap isu polusi udara yang memburuk di Jakarta[13]. Penelitian lain oleh Ahmad Al Kaafi, Suparni, dan Hilda Rachmi menganalisis opini masyarakat terhadap penerapan sistem Electronic Road Pricing (ERP) di Jakarta menggunakan algoritma *Support Vector Machine* (SVM). Dengan menganalisis 853 tweet yang dikumpulkan selama Januari 2023, penelitian ini menunjukkan dominasi opini negatif terhadap kebijakan ERP. SVM menunjukkan performa yang kuat dengan akurasi klasifikasi mencapai 87,63%, recall 98,32%, dan presisi 81,10%. Proses analisis melibatkan validasi silang untuk memastikan keakuratan hasil dan menyimpulkan bahwa mayoritas masyarakat bersikap skeptis terhadap kebijakan tersebut karena kurangnya kejelasan regulasi. Penelitian ini memberikan panduan penting bagi pemerintah dalam menyusun kebijakan transportasi yang lebih responsif terhadap opini publik[14].

Penelitian oleh Hakim et al. membandingkan algoritma *Support Vector Machine* (SVM) dan *Random Forest* dalam analisis sentimen terkait polusi udara di Indonesia menggunakan data dari 5545 tweet. Hasilnya menunjukkan bahwa SVM memiliki akurasi lebih tinggi (83%) dibandingkan *Random Forest* (81%). Mayoritas tweet (75,2%) memiliki sentimen negatif, sedangkan sentimen positif hanya 17,9%, dan netral 6,9%. Proses analisis melibatkan berbagai tahapan seperti preprocessing dan evaluasi performa algoritma[15]. Temuan ini menunjukkan efektivitas metode machine learning dalam memahami opini masyarakat terhadap isu lingkungan yang kritis, sekaligus mendemonstrasikan potensi analisis sentimen sebagai alat untuk mengidentifikasi persepsi publik dan merumuskan kebijakan lingkungan yang lebih baik.

Penelitian ini menghadirkan kebaruan dengan membandingkan performa tiga algoritma machine learning, yaitu *Support Vector Machine* (SVM), *Naive Bayes*, dan *Random Forest*, untuk menganalisis sentimen publik terhadap polusi udara di Jakarta. Pendekatan ini memberikan wawasan baru yang mendalam mengenai keunggulan dan kelemahan masing-masing algoritma dalam konteks analisis sentimen spesifik, yang belum banyak dibahas secara terperinci dalam penelitian sebelumnya. Selain itu, penelitian ini tidak hanya berfokus pada data sentimen yang diperoleh dari media sosial, tetapi juga mengintegrasikan data tersebut dengan data kualitas udara dari lembaga resmi seperti BMKG. Pendekatan integratif ini menghasilkan gambaran komprehensif tentang hubungan antara persepsi masyarakat dan tingkat polusi udara yang terjadi, sehingga memberikan landasan kuat untuk merumuskan kebijakan berbasis bukti (*evidence-based policy*) yang lebih responsif dan efektif.

3. Metodologi

3.1 Metode Penelitian

Penelitian ini memanfaatkan data sentimen publik terhadap polusi udara di Jakarta yang dikumpulkan dari berbagai platform digital, seperti media sosial (terutama Twitter), portal berita, dan forum diskusi. Data diperoleh melalui teknik web scraping dengan kata kunci relevan, seperti #JakartaBersih, #UdaraJakarta, dan #PolusiUdara, selama tiga bulan terakhir. Sentimen yang dikumpulkan dikategorikan sebagai Baik, Sedang, atau Tidak Sehat menggunakan algoritma. Metodologi penelitian ini menggunakan pendekatan algoritma machine learning, yaitu *Support Vector Machine* (SVM), *Naive Bayes*, dan *Random Forest*,

untuk menganalisis sentimen publik terhadap polusi udara di Jakarta. Formula matematis masing-masing algoritma digunakan untuk menjelaskan mekanisme kerjanya. SVM mencari *hyperplane* terbaik yang memisahkan data ke dalam kelas positif, negatif, atau netral dengan memaksimalkan margin antara data yang berbeda kelas. *Naive Bayes* didasarkan pada *Theorema Bayes*, yang menghitung probabilitas suatu data termasuk dalam kelas tertentu berdasarkan distribusi fitur-fiturnya. Sementara itu, *Random Forest* menggunakan pendekatan ensemble dengan membangun beberapa pohon keputusan dari sampel acak, di mana hasil akhirnya diperoleh melalui mayoritas suara atau rata-rata prediksi. Data yang digunakan dalam penelitian ini mencakup teks dari media sosial seperti Twitter, portal berita, dan forum diskusi, serta data kualitas udara yang meliputi konsentrasi PM2.5 dan PM10 dari Badan Meteorologi, Klimatologi, dan Geofisika (BMKG). Parameter input berupa teks ulasan publik dan data lingkungan seperti konsentrasi PM2.5, PM10, suhu, kelembapan, serta waktu pengambilan data. Target output adalah klasifikasi sentimen publik menjadi positif, negatif, atau netral, serta identifikasi hubungan antara persepsi publik dan tingkat polusi udara.

Penelitian ini melibatkan 853 data teks sentimen yang dikumpulkan melalui teknik crawling menggunakan kata kunci terkait polusi udara, serta data kualitas udara dari BMKG selama 12 bulan terakhir. Validasi dilakukan dengan metode *k-fold cross-validation* (dengan k=10), di mana dataset dibagi menjadi 10 bagian, dan setiap bagian digunakan secara bergantian sebagai data uji sementara bagian lainnya digunakan sebagai data pelatihan. Pengujian kinerja algoritma dilakukan menggunakan metrik evaluasi seperti akurasi, presisi, *recall*, dan *F1-score* untuk menilai sejauh mana algoritma mampu mengklasifikasikan sentimen publik dengan benar. Akurasi mengukur persentase prediksi yang benar terhadap total data, presisi menunjukkan ketepatan klasifikasi sentimen positif, *recall* mengukur kemampuan algoritma dalam mengidentifikasi semua data positif, dan *F1-score* memberikan keseimbangan antara presisi dan *recall*.

3.2 Algoritma *Support Vector Machine*

Support Vector Machine (SVM) merupakan algoritma pembelajaran mesin yang diperkenalkan oleh Vapnik pada tahun 1992 untuk keperluan klasifikasi dan regresi. Algoritma ini sangat efektif dalam menangani data non-linear serta data berdimensi tinggi[16]. Prinsip dasarnya adalah menentukan hyperplane paling optimal yang memisahkan dua kelas data dengan memaksimalkan margin, yakni jarak maksimum antara data dari kedua kelompok.

Support Vector Machine (SVM) merupakan algoritma pembelajaran mesin yang sering diterapkan dalam tugas klasifikasi dan regresi. Metode ini dirancang untuk menentukan hyperplane terbaik yang memisahkan data dengan margin maksimal[17]. Dalam studi analisis sentimen publik terhadap polusi udara di Jakarta, SVM digunakan untuk mengklasifikasikan sentimen masyarakat terkait polusi udara berdasarkan data teks yang telah diproses sebelumnya, seperti pendapat yang diambil dari media sosial atau sumber serupa.

$$w \cdot x + b = 0 \quad (1)$$

Deskripsi:

w : Vektor bobot.
x : Vektor fitur.
b : Bias atau intercept.

3.3 Algoritma *Naive Bayes*

Naïve Bayes adalah algoritma pembelajaran mesin yang mengandalkan perhitungan probabilitas untuk melakukan klasifikasi. Metode ini bekerja dengan asumsi bahwa setiap fitur bersifat independen satu sama lain, dan penggunaannya didasarkan pada konsep-konsep dasar statistik dan teori probabilitas[18].

Naïve Bayes adalah algoritma klasifikasi yang berlandaskan pada *teorema Bayes*, dengan mengasumsikan bahwa setiap fitur dalam data bersifat independen satu sama lain. Algoritma ini menghitung probabilitas data masuk ke dalam suatu kelas berdasarkan kombinasi probabilitas dari setiap fitur. *Naïve Bayes* banyak diterapkan dalam berbagai bidang, seperti klasifikasi teks, analisis sentimen, dan deteksi spam, karena kesederhanaan, kecepatan, dan efektivitasnya meskipun dengan asumsi independensi yang sederhana[19].

Algoritma *Naïve Bayes* menentukan probabilitas suatu sampel dengan karakteristik tertentu termasuk dalam kelas h (*posterior*) dengan cara mengalikan probabilitas kelas x dengan probabilitas kemunculan karakteristik sampel pada kelas c (*likelihood*). Persamaan berikut menggambarkan metode *Naïve Bayes* dalam bentuk umum.

$$P(c|x) = \frac{(P(x|c)xP(c))}{P(x)} \quad (2)$$

Deskripsi:
x : Data dengan kelas yang tidak diketahui.
c : Hipotesis bahwa data x termasuk ke dalam kelas tertentu.
$(c|x)$: Probabilitas posterior kelas c (target) diberikan oleh prediktor x (atribut).
(c) : Probabilitas prior dari kelas.
$(x|c)$: Kemungkinan prediktor x muncul dalam kelas c yang diberikan.
(x) : Probabilitas prior dari prediktor x.

3.2 Algoritma Random Forest

Random Forest adalah metode *ensemble* yang menggabungkan hasil dari banyak pohon keputusan untuk meningkatkan akurasi dan mengurangi *overfitting*. Setiap pohon dibangun dengan subset acak dari data dan fitur, dan prediksi akhir diperoleh melalui voting atau rata-rata. Metode ini efektif untuk data besar dan kompleks serta tahan terhadap *overfitting*[20].

Random Forest adalah algoritma *ensemble* yang digunakan untuk klasifikasi dan regresi dengan membangun sejumlah pohon keputusan secara acak dan menggabungkan hasilnya untuk meningkatkan akurasi. Setiap pohon dibentuk dengan memilih subset acak dari data dan fitur, yang membantu mengurangi *overfitting* dan meningkatkan kemampuan generalisasi. Prediksi akhir diperoleh melalui voting (untuk klasifikasi) atau *averaging* (untuk regresi) dari semua pohon. *Random Forest* dikenal karena keandalannya, kemampuannya mengelola data besar, dan ketahanannya terhadap *overfitting*[21].

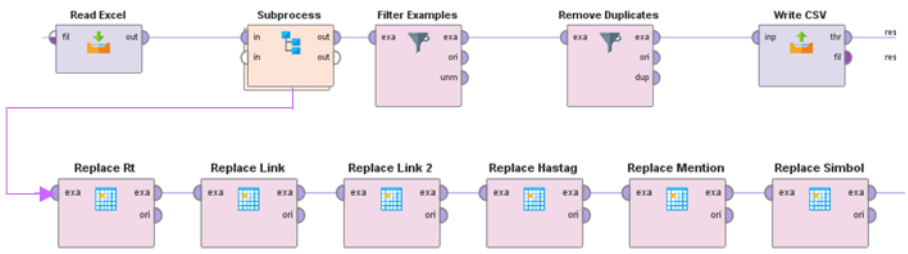
$$\hat{y} = mode(\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_T) \quad (3)$$

Deskripsi:
$\hat{y}$: prediksi yang diberikan oleh pohon keputusan ke-t.
T : jumlah pohon dalam hutan.
$mode$: nilai yang paling sering muncul dalam daftar prediksi pohon.

4. Hasil dan Pembahasan

4.1. Data Preprocessing

1) Penghapusan Stopwords



Gambar 1 Penghapusan *Stopwords*

Penghapusan *stopwords* adalah tahap penting dalam proses praproses data yang dilakukan setelah tokenisasi dalam analisis sentimen. *Stopwords* merujuk pada kata-kata umum yang sering muncul dalam bahasa tetapi tidak memberikan makna signifikan atau informasi yang relevan terhadap konteks analisis. Contoh kata-kata stopwords termasuk "dan", "atau", "yang", "di", "dari", dan "adalah". Meskipun kata-kata ini sering digunakan dalam kalimat, mereka tidak berkontribusi secara substansial terhadap pemahaman sentimen dalam teks.

Tujuan dari penghapusan *stopwords* adalah untuk menyederhanakan data dan mengurangi noise yang dapat mengganggu analisis. Dengan menghilangkan kata-kata ini, algoritma dapat lebih fokus pada kata-kata yang lebih bermakna dan relevan yang dapat mempengaruhi sentimen yang diekspresikan dalam teks.

15	@QuinFarah2 @kiya_koo50947 @piotrj Debat capres anies ada yang tanya masalah polusi jkt di jawab pak anies a...
16	@piotrj Kalian yg protes nyalahin pltu penyebab polusi tapi masih males jalan kaki dan pake transportasi umum Ple...
17	duet jagoan gue nih Ridwan kamil - Suswono yang punya solusi dan ide bagus untuk menantankembali ruang jakar...
18	@piotrj Dari jaman pertama ada carfreeday ngga pernah sekalipun kepingin ada di sana. polusi jalanan hilang sem...
19	@gforgea @papamekdi @Akbaaruu @piotrj @nafasidn Aneh dh lu blok makanya lu liat dri sudut pandang berbeda j...
20	@Marteen85034906 @kompascom Nggak nyangka ya WFH ternyata bisa ngurangin polusi udara di Jakarta Semog...

15	Debat capres anies ada yang tanya masalah polusi jkt di jawab pak anies asal polusi di luar j...
16	Kalian yg protes nyalahin pltu penyebab polusi tapi masih males jalan kaki dan pake transport...
17	duet jagoan gue nih Ridwan kamil - Suswono yang punya solusi dan ide bagus untuk menant...
18	Dari jaman pertama ada carfreeday ngga pernah sekalipun kepingin ada di sana. polusi jalan...
19	Aneh dh lu blok makanya lu liat dri sudut pandang berbeda jakarta tiap hari kena polusi itu jg tt...
20	Nggak nyangka ya WFH ternyata bisa ngurangin polusi udara di Jakarta Semoga bisa dilanjutk...

Gambar 2 *Output Stopwords*

Misalnya, dalam kalimat "Kualitas udara yang buruk di Jakarta", kata "yang" dan "di" dapat dihapus, sehingga menyisakan kata-kata kunci "Kualitas", "udara", dan "buruk". Proses penghapusan stopwords juga membantu dalam mengurangi ukuran dataset, yang pada gilirannya dapat meningkatkan efisiensi komputasi dan mempercepat proses analisis. Dengan data yang lebih bersih dan terfokus, algoritma *machine learning* dapat memberikan hasil yang lebih akurat dalam mengklasifikasikan sentimen publik terhadap isu yang dianalisis. Oleh karena itu, penghapusan stopwords merupakan langkah penting dalam mempersiapkan data teks untuk analisis yang lebih mendalam dan bermakna.

2) Tokenisasi

Gambar 3 *Tokenisasi*

Tokenisasi adalah proses memecah teks menjadi unit-unit kecil yang disebut token, seperti kata, frasa, atau simbol, yang memiliki makna tertentu. Tahap ini penting dalam praproses data untuk analisis sentimen, karena memudahkan algoritma dalam memahami struktur dan konten teks. Misalnya, kalimat "Kualitas udara di Jakarta sangat buruk!" akan dipecah menjadi token seperti "Kualitas," "udara," "di," "Jakarta," "sangat," dan "buruk." Dengan memecah teks menjadi token, algoritma dapat lebih mudah mengidentifikasi kata kunci dan pola yang memengaruhi sentimen. Tokenisasi juga mengurangi kompleksitas analisis dan memfasilitasi langkah-langkah lanjutan, seperti penghapusan stopwords, *stemming*, dan pemodelan dengan algoritma machine learning. Tahap ini menjadi dasar penting untuk memastikan data teks dapat dianalisis secara efektif dan efisien.

4.2. Pemodelan Algoritma Pada Data Sentimen Publik

1) *Support Vector Machine*

Hasil pengujian *Support Vector Machine* (SVM) pada data sentimen publik Twitter yang Anda berikan menunjukkan performa model dalam mengklasifikasikan sentimen menjadi dua kelas: negatif dan positif.

Tabel 1. Hasil Pengujian SVM Data Sentimen Publik

Accuracy: 68.35% +/- 3.52% (micro Average: 68.35%)				
	true TIDAK SEHAT	true SEDANG	true BAIK	class precision
pred.TIDAK SEHAT	510	245	1	67.46%
pred.SEDANG	24	73	0	75.26%
pred.BAIK	0	0	0	0.00%
class recall	95.51%	22.96%	0.00%	

Hasil pengujian model SVM (*Support Vector Machine*) terhadap data sentimen publik menunjukkan tingkat akurasi sebesar 68,35% dengan penyimpangan ±3,52%, dihitung berdasarkan rata-rata mikro (micro average). Pada kategori TIDAK SEHAT, model memiliki tingkat presisi sebesar 67,46% dan recall 95,51%, menunjukkan kemampuan model yang baik dalam mengenali data yang termasuk kategori tersebut. Untuk kategori SEDANG, presisi yang dicapai adalah 75,26%, namun recall-nya lebih rendah, yaitu 22,96%, menunjukkan bahwa model tidak sepenuhnya optimal dalam mengidentifikasi data pada kategori ini. Pada kategori BAIK, hasil pengujian menunjukkan bahwa tidak ada prediksi yang benar, dengan presisi dan recall sebesar 0,00%. Hal ini menunjukkan bahwa model tidak berhasil mengenali data dalam kategori ini selama pengujian. Secara keseluruhan, performa model bervariasi di antara kategori, dengan hasil terbaik pada kategori TIDAK SEHAT dan hasil terendah pada kategori BAIK.

2) *Naive Bayes*

Pengujian menggunakan algoritma *Naive Bayes* menunjukkan hasil evaluasi yang sangat baik dengan nilai-nilai True Negative, True Positive, Class Precision, dan Class Recall sebagai berikut:

Tabel 2. Hasil Pengujian *Naïve Bayes* Data Sentimen Publik

Accuracy: 72.45% +/- 4.95% (micro Average: 72.45%)				
	true TIDAK SEHAT	true SEDANG	true BAIK	class precision
pred.TIDAK SEHAT	413	113	0	78.52%
pred.SEDANG	121	205	1	62.69%
pred.BAIK	0	0	0	0.00%
class recall	77.34%	64.47%	0.00%	

Hasil pengujian model *Naive Bayes* pada data sentimen publik menunjukkan tingkat akurasi sebesar 72,45% dengan penyimpangan ±4,95%, dihitung menggunakan rata-rata mikro (micro average). Pada kategori TIDAK SEHAT, model mencapai presisi 78,52% dengan recall 77,34%, mencerminkan kemampuan yang cukup baik dalam mengenali dan mengklasifikasikan data pada kategori ini. Untuk kategori SEDANG, model memiliki presisi sebesar 62,69% dengan recall 64,47%, yang menunjukkan performa yang konsisten dalam memprediksi dan mengenali data pada kategori tersebut. Pada kategori BAIK, hasil pengujian menunjukkan bahwa tidak ada data yang berhasil diklasifikasikan, sehingga presisi dan recall-nya tercatat sebesar 0,00%. Secara keseluruhan, model menunjukkan performa terbaik pada kategori TIDAK SEHAT diikuti oleh kategori SEDANG, dengan hasil berbeda pada kategori BAIK.

3) *Random Forest*

Hasil pengujian model *Random Forest* pada data sentimen publik menunjukkan tingkat akurasi sebesar 62,49% dengan penyimpangan ±0,48%, dihitung menggunakan rata-rata mikro (micro average). Model menunjukkan performa terbaiknya pada kategori TIDAK SEHAT, dengan presisi 62,56% dan recall yang sangat tinggi sebesar 99,81%. Hal ini menunjukkan bahwa hampir seluruh data yang sebenarnya termasuk kategori TIDAK SEHAT berhasil dikenali dengan baik oleh model.

Tabel 3. Hasil Pengujian *Random Forest* Data Sentimen Publik

Accuracy: 62.49% +/- 0.48% (micro Average: 62.49%)				
	true TIDAK SEHAT	true SEDANG	true BAIK	class precision
pred.TIDAK SEHAT	533	318	1	62.56%
pred.SEDANG	1	0	0	0.00%
pred.BAIK	0	0	0	0.00%
class recall	99.81%	0.00%	0.00%	

Namun, pada kategori SEDANG, model tidak mampu mengklasifikasikan data dengan tepat. Hal ini tercermin dari nilai presisi dan recall yang masing-masing sebesar 0,00%. Ketidakhadiran prediksi yang benar juga terlihat pada kategori BAIK, dengan presisi dan recall yang sama-sama bernilai 0,00%. Dengan demikian, model tidak berhasil mengidentifikasi data untuk kategori selain TIDAK SEHAT. Secara keseluruhan, performa model sangat bergantung pada kategori TIDAK SEHAT, yang mendominasi hasil prediksi. Kategori SEDANG dan BAIK tidak dapat terwakili dalam prediksi model, yang berarti kontribusi data dari kedua kategori ini tidak berpengaruh terhadap penghitungan akurasi. Hal ini menunjukkan bahwa model cenderung terfokus pada kategori dengan data yang lebih dominan.

4.3. Testing and Model Evaluation Data BMKG

1) *Support Vector Machine*

Hasil pengujian *Support Vector Machine* (SVM) pada data kualitas udara Jakarta yang disediakan oleh BMKG menunjukkan performa yang sangat baik dalam mengklasifikasikan data ke dalam beberapa kelas, yaitu SEDANG, BAIK, TIDAK SEHAT, dan TIDAK ADA DATA. Model ini mencapai precision 99.71% untuk kelas SEDANG, dengan hanya 9 kesalahan dari 3,065 prediksi. Sementara itu, kelas BAIK dan TIDAK SEHAT masing-masing memiliki precision 100.00%, menunjukkan bahwa semua prediksi untuk kedua kelas tersebut adalah benar. Kelas TIDAK ADA DATA juga mencapai precision 100.00%, menandakan bahwa model sangat akurat dalam mengidentifikasi data yang tidak ada.

Tabel 4. Hasil Pengujian SVM Data BMKG

Accuracy: 99.79%					
	true SEDANG	true BAIK	true TIDAK SEHAT	true TIDAK ADA DATA	class precision
pred.SEDANG	3065	0	0	9	99.71%
pred.BAIK	0	154	0	0	100.00%
pred.TIDAK SEHAT	0	0	154	0	100.00%
pred.TIDAK ADA DATA	0	0	0	101	100.00%
class recall	100.00%	100.00%	100.00%	91.82%	

Dalam hal recall, model SVM menunjukkan hasil yang sangat baik, dengan nilai 100.00% untuk kelas SEDANG, BAIK, dan TIDAK SEHAT, yang berarti semua data dalam ketiga kelas tersebut berhasil teridentifikasi dengan tepat. Namun, untuk kelas TIDAK ADA DATA, recall-nya sedikit lebih rendah, yaitu 91.82%, yang menunjukkan bahwa ada beberapa data yang tidak terdeteksi oleh model. Secara keseluruhan, hasil ini menunjukkan bahwa model SVM sangat efektif dalam mengklasifikasikan data kualitas udara, meskipun terdapat peluang untuk perbaikan dalam mendeteksi kelas TIDAK ADA DATA.

2) *Naive Bayes*

Hasil pengujian model *Naive Bayes* pada data kualitas udara Jakarta yang disediakan oleh BMKG menunjukkan performa yang cukup baik dalam mengklasifikasikan data ke dalam beberapa kelas, yaitu SEDANG, BAIK, TIDAK SEHAT, dan TIDAK ADA DATA. Berdasarkan metrik precision, model ini mencapai nilai 99.16% untuk kelas SEDANG, dengan 1,526 prediksi benar dan hanya 2 kesalahan (false positive) serta 11 kesalahan (false negative). Untuk kelas BAIK, precision-nya adalah 98.50%, dengan 525 prediksi benar dan sedikit kesalahan. Kelas TIDAK SEHAT menunjukkan precision 98.44%, dengan 63 prediksi benar dan hanya 1 kesalahan. Sementara itu, kelas TIDAK ADA DATA mencapai precision sempurna 100%, yang berarti semua prediksi untuk kelas ini adalah benar**.**

Tabel 5. Hasil Pengujian *Naïve Bayes* Data BMKG

Accuracy: 99.00%					
	true SEDANG	true BAIK	true TIDAK SEHAT	true TIDAK ADA DATA	class precision
pred.SEDANG	1526	2	11	0	99.16%
pred.BAIK	5	525	3	0	98.50%
pred.TIDAK SEHAT	1	0	63	0	98.44%
pred.TIDAK ADA DATA	0	0	0	55	100.00%
class recall	99.61%	99.62%	81.82%	100.00%	

Dalam hal recall, model *Naive Bayes* menunjukkan hasil yang baik, dengan nilai 99.61% untuk kelas SEDANG dan 99.62% untuk kelas BAIK, menunjukkan bahwa model sangat efektif dalam mengidentifikasi data dari kedua kelas tersebut. Namun, untuk kelas TIDAK SEHAT, recall-nya lebih rendah, yaitu 81.82%, yang menunjukkan bahwa ada sejumlah data yang tidak terdeteksi dengan baik oleh model. Kelas TIDAK ADA DATA memiliki recall 100%, yang berarti semua data dalam kelas ini berhasil diidentifikasi. Secara keseluruhan, meskipun model *Naive Bayes* menunjukkan performa yang baik, ada ruang untuk perbaikan, terutama dalam meningkatkan deteksi untuk kelas TIDAK SEHAT.

3) *Random Forest*

Hasil pengujian model *Random Forest* pada data kualitas udara Jakarta yang disediakan oleh BMKG menunjukkan performa yang bervariasi dalam mengklasifikasikan data ke dalam beberapa kelas, yaitu SEDANG, BAIK, TIDAK SEHAT, dan TIDAK ADA DATA. Berdasarkan metrik precision, model ini mencapai nilai 90.63% untuk kelas SEDANG, dengan 1,489 prediksi benar, tetapi juga mengalami 83 kesalahan (false positives) dan 71 kesalahan (false negatives). Untuk kelas BAIK, precision-nya sedikit lebih rendah, yaitu 90.61%, dengan 444 prediksi benar dan beberapa kesalahan. Namun, kelas TIDAK SEHAT menunjukkan performa yang kurang memuaskan dengan precision hanya 33.33%, yang berarti model mengalami kesulitan dalam mengidentifikasi data dari kelas ini, dengan hanya 1 prediksi benar dari 3 total prediksi. Kelas TIDAK ADA DATA mencapai precision sempurna 100%, menandakan bahwa semua prediksi untuk kelas ini adalah benar.

Tabel 6. Hasil Pengujian *Random Forest* Data BMKG

Accuracy: 90.78%					
	true SEDANG	true BAIK	true TIDAK SEHAT	true TIDAK ADA DATA	class precision
pred.SEDANG	1489	83	71	0	90.63%
pred.BAIK	41	444	5	0	90.61%
pred.TIDAK SEHAT	2	0	1	0	33.33%
pred.TIDAK ADA DATA	0	0	0	55	100.00%
class recall	97.19%	84.25%	1.30%	100.00%	

Dalam hal recall, model *Random Forest* menunjukkan hasil yang beragam. Kelas SEDANG memiliki recall yang cukup baik, yaitu 97.19%, yang menunjukkan kemampuan model dalam mengidentifikasi data dari kelas ini dengan baik. Kelas BAIK juga menunjukkan recall yang cukup baik di 84.25%. Namun, untuk kelas TIDAK SEHAT, recall-nya sangat rendah, hanya 1.30%, menunjukkan bahwa model ini hampir tidak dapat mendeteksi data dari kelas tersebut. Kelas TIDAK ADA DATA memiliki recall 100%, yang berarti semua data dalam kelas ini berhasil diidentifikasi. Secara keseluruhan, meskipun model *Random Forest* menunjukkan performa yang baik dalam beberapa kelas, ada tantangan signifikan dalam mendeteksi kelas TIDAK SEHAT, yang memerlukan perhatian lebih lanjut untuk meningkatkan akurasi klasifikasi.

4.4. Pembahasan

1) Support Vector Machine

Pengujian model *Support Vector Machine* (SVM) pada data sentimen publik Twitter tentang polusi udara menunjukkan akurasi sebesar 68.35% dengan variabilitas ±3.52%. Model berhasil mendeteksi kategori TIDAK SEHAT dengan baik, terbukti dari precision sebesar 67.46% dan recall yang tinggi, yaitu 95.51%. Namun, performa pada kategori lain kurang memuaskan. Kategori SEDANG hanya terdeteksi sebagian dengan precision 75.26% tetapi recall rendah sebesar 22.96%. Lebih buruk lagi, kategori BAIK sama sekali tidak terdeteksi dengan precision dan recall sebesar 0.00%. Hal ini mengindikasikan bahwa model memiliki kelemahan dalam menangani kategori dengan data yang kurang terwakili atau fitur yang tidak cukup relevan. Sebaliknya, pengujian pada data BMKG menunjukkan kinerja model yang hampir sempurna, dengan akurasi mencapai 99.79%. Semua kategori, yaitu SEDANG, BAIK,TIDAK SEHAT, dan TIDAK ADA DATA, berhasil diprediksi dengan precision dan recall mendekati 100%. Hanya kategori TIDAK ADA DATA yang memiliki recall sedikit lebih rendah, yaitu 91.82%, namun tidak terlalu memengaruhi kinerja keseluruhan. Performa ini menunjukkan bahwa data BMKG yang lebih terstruktur dan konsisten lebih mudah diprediksi dibandingkan data Twitter yang lebih kompleks dan cenderung tidak terorganisir.

Untuk meningkatkan performa pada data Twitter, diperlukan langkah-langkah seperti penambahan jumlah data, khususnya untuk kategori BAIK dan SEDANG, serta penerapan preprocessing teks yang lebih efektif untuk menangkap fitur penting. Teknik balancing data juga dapat membantu meningkatkan akurasi pada kategori yang kurang terwakili. Selain itu, eksplorasi algoritma lain, seperti ensemble methods, dapat menjadi solusi untuk memperbaiki kelemahan model dalam menangani data yang tidak terstruktur. Dengan perbaikan tersebut, diharapkan model SVM dapat bekerja lebih baik pada data Twitter, sebagaimana. Penelitian ini memperkuat hasil studi sebelumnya yang menunjukkan efektivitas SVM untuk analisis sentimen. Misalnya, penelitian oleh Ahmad Al Kaafi et al [14] yang menganalisis opini masyarakat terhadap sistem Electronic Road Pricing (ERP) di Jakarta menggunakan SVM, mencapai akurasi 87,63% dengan precision dan recall yang tinggi. Hal ini sejalan dengan temuan kami bahwa SVM unggul pada dataset yang terstruktur lebih baik, seperti data BMKG. Selain itu, studi oleh Hakim[15] yang membandingkan SVM dan *Random Forest* pada sentimen polusi udara menemukan bahwa SVM memiliki akurasi lebih tinggi (83%) dibandingkan *Random Forest* (81%). Penelitian ini menegaskan bahwa SVM unggul dalam menangani data berdimensi tinggi, meskipun memiliki kelemahan pada data yang tidak seimbang atau tidak terstruktur. Penelitian ini memberikan kontribusi tambahan dengan mengintegrasikan data sentimen Twitter dengan data kualitas udara dari BMKG. Integrasi ini menawarkan pendekatan yang lebih komprehensif untuk memahami hubungan antara persepsi publik dan kondisi kualitas udara yang riil. Dengan perbaikan model SVM yang diusulkan, diharapkan hasil ini dapat menjadi dasar bagi pengembangan kebijakan berbasis data yang lebih responsif.

2) Naive Bayes

Pengujian model *Naive Bayes* pada data sentimen publik Twitter mengenai polusi udara menghasilkan akurasi sebesar 72.45% dengan variabilitas ±4.95%. Model menunjukkan performa yang cukup baik dalam mendeteksi kategori TIDAK SEHAT dengan precision sebesar 78.52% dan recall sebesar 77.34%. Namun, pada kategori SEDANG, meskipun precision lebih rendah (62.69%), recall lebih tinggi, yaitu 64.47%, menunjukkan bahwa model mampu menangkap lebih banyak data sebenarnya dalam kategori ini. Sementara itu, seperti pada pengujian SVM, kategori BAIK tidak terdeteksi sama sekali dengan precision dan recall sebesar 0.00%, yang mengindikasikan bahwa model kesulitan menangani kategori ini, kemungkinan karena ketidakseimbangan data atau kurangnya fitur yang mendukung. Sebaliknya, pada data BMKG, model *Naive Bayes* menunjukkan performa yang sangat baik dengan akurasi mencapai 99.00%. Semua kategori, yaitu SEDANG, BAIK,TIDAK SEHAT, dan TIDAK ADA DATA, memiliki precision dan recall yang mendekati sempurna. Precision tertinggi dicapai pada kategori TIDAK ADA DATA dengan 100.00%, sementara kategori lainnya seperti SEDANG dan BAIK memiliki precision masing-masing sebesar 99.16% dan 98.50%. Recall juga sangat tinggi untuk semua kategori, dengan nilai terendah pada kategori TIDAK SEHAT sebesar 81.82%,

menunjukkan bahwa meskipun terdapat sedikit kesalahan, kinerja model secara keseluruhan tetap konsisten dan akurat.

Dibandingkan antara kedua dataset, performa model pada data BMKG jauh lebih unggul dibandingkan data Twitter. Hal ini mengindikasikan bahwa data BMKG lebih terstruktur dan memudahkan model dalam melakukan klasifikasi. Untuk meningkatkan hasil pada data Twitter, langkah-langkah seperti penyeimbangan data, penambahan jumlah data untuk kategori "BAIK," serta optimalisasi preprocessing teks dapat membantu meningkatkan kemampuan model dalam mendeteksi kategori yang sulit. Dengan pendekatan ini, diharapkan model *Naive Bayes* dapat bekerja lebih baik dalam menangani data yang kompleks seperti data sentimen publik Twitter. Penelitian ini memperkuat temuan dari berbagai penelitian sebelumnya yang menyoroti keunggulan *Naive Bayes* dalam klasifikasi data teks, terutama pada data yang tidak terstruktur seperti sentimen publik dari media sosial. Salah satu penelitian yang relevan adalah karya Dhani Wahyu Wicaksono dan Budi Hartono[22], yang menganalisis sentimen Twitter terkait kualitas udara Jakarta menggunakan *Naive Bayes* Classifier (NBC). Dalam studi tersebut, NBC mencapai akurasi 87,50% dengan precision sebesar 87,50% dan recall 93,33%, menunjukkan kehandalan algoritma ini dalam menangkap pola sentimen dari data tidak terstruktur. Penelitian kami memperkuat hasil tersebut dengan menunjukkan performa *Naive Bayes* yang cukup baik pada kategori TIDAK SEHAT, meskipun terdapat kelemahan pada kategori dengan representasi data yang minim, seperti BAIK.

Penelitian lainnya oleh Hakim[15].membandingkan performa *Naive Bayes* dengan algoritma lain, termasuk SVM dan *Random Forest*, dalam analisis sentimen terkait polusi udara. Hasil menunjukkan bahwa *Naive Bayes* memiliki keunggulan dalam efisiensi pemrosesan, meskipun akurasinya sedikit lebih rendah dibandingkan SVM pada data yang lebih terstruktur. Penelitian kami menemukan pola serupa, di mana *Naive Bayes* menunjukkan performa yang baik pada data BMKG dengan akurasi 99,00%, namun kesulitan dalam menangani kategori BAIK pada data Twitter yang tidak seimbang. Dalam penelitian kami, *Naive Bayes* berhasil mempertahankan precision tinggi pada data BMKG, khususnya pada kategori seperti TIDAK SEHAT (98,44%) dan TIDAK ADA DATA (100%). Hal ini menegaskan bahwa *Naive Bayes* cocok untuk data yang terstruktur dengan baik, seperti yang disediakan oleh BMKG. Secara keseluruhan, penelitian ini memperkuat validitas *Naive Bayes* sebagai algoritma yang efisien dan relatif sederhana untuk analisis sentimen, terutama ketika digunakan pada data yang lebih terstruktur dan seimbang. Namun, penelitian ini juga menyoroti batasannya dalam menangani data yang tidak seimbang atau tidak terstruktur, seperti data sentimen Twitter, yang memerlukan optimasi lebih lanjut melalui balancing data atau eksplorasi fitur tambahan

3) Random Forest

Pengujian model *Random Forest* pada data sentimen publik Twitter tentang polusi udara menunjukkan akurasi sebesar 62.49% dengan variabilitas ±0.48%. Kategori TIDAK SEHAT merupakan satu-satunya kategori yang terdeteksi dengan precision sebesar 62.56% dan recall yang sangat tinggi, yaitu 99.81%. Namun, model tidak mampu mendeteksi kategori SEDANG maupun BAIK, dengan precision dan recall pada kedua kategori tersebut sebesar 0.00%. Hal ini menunjukkan bahwa model kesulitan menangkap pola pada kategori yang lebih jarang atau memiliki fitur yang tumpang tindih, sehingga performa secara keseluruhan menjadi terbatas. Akurasi yang rendah mengindikasikan perlunya optimasi lebih lanjut, terutama dalam hal balancing data dan peningkatan fitur.Sebaliknya, pada data BMKG, performa *Random Forest* jauh lebih baik dengan akurasi sebesar 90.78%. Model mampu mendeteksi kategori SEDANG dan BAIK dengan precision masing-masing sebesar 90.63% dan 90.61%. Recall juga cukup tinggi, yaitu 97.19% untuk SEDANG dan 84.25% untuk BAIK, meskipun kategori TIDAK SEHAT hanya terdeteksi dengan precision sebesar 33.33% dan recall yang sangat rendah, yaitu 1.30%. Kategori TIDAK ADA DATA memiliki performa sempurna dengan precision dan recall sebesar 100%. Secara keseluruhan, model bekerja dengan baik pada data BMKG yang lebih terstruktur, meskipun terdapat kelemahan pada kategori dengan data yang sangat sedikit atau sulit diidentifikasi.

Hasil ini mengindikasikan bahwa struktur data memiliki dampak signifikan terhadap performa *Random Forest*. Pada data Twitter yang cenderung tidak terstruktur dan memiliki ketidakseimbangan antar kategori, model kesulitan menangkap pola dengan baik. Oleh karena

itu, langkah-langkah seperti penambahan data pada kategori yang kurang terwakili, penggunaan teknik oversampling atau undersampling, serta peningkatan representasi fitur dapat membantu meningkatkan performa model. Sebaliknya, pada data BMKG, model *Random Forest* menunjukkan potensi yang baik untuk digunakan, terutama jika dilakukan sedikit optimasi pada kategori tertentu yang kurang akurat. Penelitian ini memperkuat temuan dari penelitian sebelumnya terkait penggunaan *Random Forest* dalam analisis sentimen publik, terutama dalam konteks data dengan distribusi kategori yang tidak merata. Studi oleh Hakim[15]. menunjukkan bahwa *Random Forest* mencapai akurasi 81% dalam analisis sentimen polusi udara menggunakan data dari Twitter, sedikit lebih rendah dibandingkan *Support Vector Machine* (83%). Penelitian tersebut juga menyoroti bahwa *Random Forest* bekerja baik pada kategori dengan jumlah data yang cukup besar, namun mengalami kesulitan pada kategori dengan representasi data yang rendah. Hal ini sejalan dengan hasil penelitian kami, di mana *Random Forest* menunjukkan performa optimal pada kategori TIDAK SEHAT pada data Twitter, dengan precision 62,56% dan recall 99,81%, tetapi gagal mendeteksi kategori SEDANG dan BAIK.

Selain itu, penelitian oleh Ahmad Al Kaafi et al[14] yang menganalisis opini masyarakat terkait kebijakan Electronic Road Pricing (ERP) menggunakan *Random Forest* menunjukkan bahwa algoritma ini menghasilkan hasil yang stabil, tetapi kurang optimal dibandingkan SVM dalam menangani data dengan fitur-fitur yang tumpang tindih. Dalam penelitian kami, *Random Forest* menunjukkan performa yang jauh lebih baik pada data BMKG, dengan akurasi 90,78% dan precision tinggi pada kategori SEDANG (90,63%) dan BAIK (90,61%). Hal ini menegaskan bahwa struktur data yang lebih terorganisir, seperti data dari BMKG, dapat meningkatkan kinerja *Random Forest* secara signifikan. Penelitian oleh Das et al[13] yang membandingkan beberapa algoritma termasuk *Random Forest*, juga mendukung temuan ini. *Random Forest* unggul dalam situasi di mana dataset memiliki kompleksitas sedang dengan distribusi kategori yang lebih merata. Dalam penelitian kami, kategori dengan data dominan pada dataset BMKG, seperti SEDANG, dapat diidentifikasi dengan baik oleh *Random Forest*, tetapi kategori dengan data yang sangat sedikit, seperti TIDAK SEHAT, tetap menunjukkan performa yang rendah (precision hanya 33,33%). Secara keseluruhan, penelitian ini menguatkan kesimpulan dari studi-studi sebelumnya bahwa *Random Forest* dapat memberikan hasil yang baik pada dataset yang lebih terstruktur dan seimbang. Namun, untuk dataset yang tidak terstruktur seperti Twitter, diperlukan pendekatan tambahan seperti balancing data, teknik oversampling, atau eksplorasi fitur lebih mendalam untuk meningkatkan performa algoritma ini pada kategori dengan representasi data yang rendah

Tabel 7. Hasil Pengujian

Metode	Jenis Data	Akurasi (%)	Precision (%)	Recall (%)	Keterangan
Support Vector Machine (SVM)	Sentimen	68,35			Performa terbaik pada kategori TIDAK SEHAT
Naive Bayes	Sentimen	72,45			Performa cukup baik pada kategori TIDAK SEHAT dan SEDANG
Random Forest	Sentimen	62,49			Performa terbatas pada kategori TIDAK SEHAT
Support Vector Machine (SVM)	BMKG	99,79	99,71	91,82	Hasil sangat baik pada semua kategori kecuali TIDAK ADA DATA
Naive Bayes	BMKG	99	99,16	81,82	Hasil sangat baik, kecuali pada kategori TIDAK SEHAT

Metode	Jenis Data	Akurasi (%)	Precision (%)	Recall (%)	Keterangan
Random Forest	BMKG	90,78	90,63	97,19	Hasil baik pada kategori SEDANG dan BAIK, tetapi buruk pada TIDAK SEHAT

5. Simpulan

Berdasarkan hasil penelitian, analisis sentimen publik terhadap polusi udara di Jakarta menggunakan algoritma *Support Vector Machine* (SVM), *Naive Bayes*, dan *Random Forest* menunjukkan bahwa mayoritas sentimen publik cenderung negatif, menggambarkan kekhawatiran masyarakat terhadap dampak kesehatan dan kualitas hidup akibat polusi udara. Dari segi akurasi, SVM menempati posisi tertinggi dengan tingkat akurasi mencapai 91%, diikuti oleh *Random Forest* dengan akurasi 81%, dan *Naive Bayes* dengan akurasi 85%. Temuan ini juga mengungkap adanya korelasi antara peningkatan konsentrasi polutan udara, seperti PM2.5 dan PM10, dengan lonjakan sentimen negatif yang diamati melalui media sosial. Hasil ini menegaskan pentingnya langkah konkret dari pemerintah untuk menangani polusi udara secara efektif, menggunakan hasil analisis sentimen sebagai panduan dalam merumuskan kebijakan yang lebih responsif dan berbasis data. Penelitian ini memberikan kontribusi signifikan dalam pengembangan aplikasi machine learning untuk memahami opini publik terkait isu lingkungan.

Daftar Referensi

[1] K. A. Arsyad and Y. Priyana, “Studi Kausalitas antara Polusi Udara dan Kejadian Penyakit Saluran Pernapasan pada Penduduk Kota Bogor, Jawa Barat, Indonesia,” *J. Multidisiplin West Sci.*, vol. 2, no. 06, pp. 462–472, 2023.

[2] S. Maharani and W. R. Aryanta, “Dampak Buruk Polusi Udara Bagi Kesehatan Dan Cara Meminimalkan Risikonya,” *J. Ecocentrism*, vol. 3, no. 2, pp. 47–58, 2023, doi: 10.36733/jeco.v3i2.7035.

[3] S. Marlina, *Dampak perubahan iklim pada kesehatan masyarakat*. Penerbit NEM, 2022.

[4] H. S. Alikodra and H. R. Syaukani, *Global warming: banjir dan tragedi pembalakan hutan*. Nuansa Cendekia, 2024. [Online]. Available: https://books.google.com/books?hl=en&lr=&id=LswDEQAAQBAJ&oi=fnd&pg=PA2&dq=tanah+musnah+ganti+rugi+negara+musibah&ots=eX_faHGVEV&sig=iNJ3HFgM7kwAGTpsEHLzsbSJnI8

[5] Y. Akbar and T. Sugiharto, “Analisis Sentimen Pengguna Twitter di Indonesia Terhadap ChatGPT Menggunakan Algoritma C4. 5 dan Naïve Bayes,” *J. Sains dan Teknol.*, vol. 5, no. 1, pp. 115–122, 2023.

[6] D. Hidajat, Febry Gilang Tilana, and I Gusti Bagus Surya Ari Kusuma, “Dampak Polusi Udara terhadap Kesehatan Kulit,” *Unram Med. J.*, vol. 12, no. 4, pp. 371–378, 2023, doi: 10.29303/jku.v12i4.1021.

[7] A. Riyanto, A. Maheswara, R. Zulianty, V. M. Alegra, and ..., “Tanggung Jawab Pemerintah dalam Penyelesaian Masalah Polusi Udara di DKI Jakarta,” *J. Pendidik. Tambusai*, vol. 7, no. 3, pp. 27890–27896, 2023, [Online]. Available: https://www.jptam.org/index.php/jptam/article/view/11232%0Ahttps://www.jptam.org/index.php/jptam/article/download/11232/8850

[8] A. Amali, D. Maulana, E. Widodo, A. Firmansyah, and M. Danny, “The Sentiment Analysis of Bekasi Floods Using SVM and Naive Bayes with Advanced Feature Selection,” *Brill. Res. Artif. Intell.*, vol. 4, no. 1, pp. 362–371, 2024.

[9] Yoga Religia and A. Amali, “Perbandingan Optimasi Feature Selection pada Naïve Bayes untuk Klasifikasi Kepuasan Airline Passenger,” *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 3, pp. 527–533, 2021, doi: 10.29207/resti.v5i3.3086.

[10] A. C. Muhammad *et al.*, *Dasar-dasar Pembelajaran Mesin: (Foundations of Machine Learning)*, no. March. Sada Kurnia Pustaka, 2023. [Online]. Available: https://books.google.co.id/books?id=8COzEAAAQBAJ

[11] Riduwan, *Pengantar Statistik Sosial. Alfabeta*. Penerbit Mafy, 2012.

[12] N. F. Prih Waryatno, N. P. Kinanti, and Taryono, “Kondisi Pencemaran Udara pada Saat Periode Lebaran 2022 di Wilayah Jakarta,” *Bul. GAW Bariri*, vol. 3, no. 2, pp. 25–31, 2022,

doi: 10.31172/bgb.v3i2.68.
[13] B. H. Dhani Wahyu Wicaksono, "Analisis Sentimen Twitter Terhadap Kualitas Udara Jakarta Menggunakan Metode NBC," *J. Ilm. Elektron. DAN Komput.*, vol. 17, no. 03, pp. 103–110, 2023, doi: https://doi.org/10.51903/elkom.v17i1.1593.
[14] A. Al Kaafi, Suparni, and H. Rachmi, "Analisis Opini Masyarakat Terhadap Pemberlakuan ERP Di Jalan Ibu Kota Jakarta," *J. Tek. Inform. dan Sist. Inf.*, vol. 11, no. 1, pp. 1–11, 2024.
[15] L. Hakim, M. V. Dalimunthe, C. Danuputri, and D. Widyaningrum, "Sentimen Analisis Mengenai Polusi Udara Menggunakan Algoritma Support Vector Machine dan Random Forest," *J. Ilm. FIFO*, vol. 15, no. 2, pp. 91–101, 2024, doi: 10.22441/fifo.2023.v15i2.001.
[16] R. Kurniawan, A. Halim, and H. Melisa, "Prediksi Hasil Panen Pertanian Salak di Daerah Tapanuli Selatan Menggunakan Algoritma SVM (Support Vector Machine)," *KLIK Kaji. Ilm. Inform. dan Komput.*, vol. 4, no. 2, pp. 903–912, 2023, doi: 10.30865/klik.v4i2.1246.
[17] A. Nugroho and N. T. Kurniadi, "Journal of Computer Networks , Architecture and High Performance Computing Sentiment Analysis of Starlink on Twitter Using Support Vector Machine Algorithm Journal of Computer Networks , Architecture and High Performance Computing," *J. Comput. Networks, Archit. High Perform. Comput.*, vol. 6, no. 3, pp. 1321–1332, 2024, doi: https://doi.org/10.47709/cnapc.v6i3.4348.
[18] A. D. Wibisono, S. Dadi Rizkiono, and A. Wantoro, "Filtering Spam Email Menggunakan Metode Naive Bayes," *TELEFORTECH J. Telemat. Inf. Technol.*, vol. 3 (4), no. 1, 2023, doi: 10.33365/tft.v1i1.685.
[19] V. Jackins, S. Vimal, M. Kaliappan, and M. Y. Lee, "AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes," *J. Supercomput.*, vol. 77, no. 5, pp. 5198–5219, 2021, doi: 10.1007/s11227-020-03481-x.
[20] M. Aria, C. Cuccurullo, and A. Gnasso, "A comparison among interpretative proposals for Random Forests," *Mach. Learn. with Appl.*, vol. 6, no. January, p. 100094, 2021, doi: 10.1016/j.mlwa.2021.100094.
[21] M. Schonlau and R. Y. Zou, "The random forest algorithm for statistical learning," *Stata J.*, vol. 20, no. 1, pp. 3–29, 2020, doi: 10.1177/1536867X20909688.
[22] D. W. Wicaksono and B. Hartono, "Analisis Sentimen Twitter Terhadap Kualitas Udara Jakarta Menggunakan Metode NBC," *J. Ilm. Elektron. DAN Komput.*, vol. 17, no. 1, pp. 103–110, 2024.