Scripta Technica:
Journal of Engineering and Applied Technology
Vol 2 No 1 June 2026, Hal. 342-352
ISSN:3110-0775(Print) ISSN: 3109-9696(Electronic)
Open Access: https://scriptaintelektual.com/scripta-technica

Analisis Sentimen Kecelakaan Lalu Lintas Di Tweet Twitter Menggunakan Metode Naïve Bayes

Aldy Sahputra Saragih[1*], Evantrana Yordan Bangun[2], Fastabikha Akbar Fahlevi[3], Ariel Kurniawan[4], Kristian Sigalingging[5], Fasta Krel Four Wati Br Haloho[6], Salman Putra Jaya Hulu[7]
[1-7] Universitas Mandiri Bina Prestasi, Indonesia
email: aldysyahputrasaragih09@gmail.com[1]

Article Info :
Received:
10-06-2026
Revised:
21-06-2026
Accepted:
24-06-2026

Abstract

This study develops a robust computational framework using a structured Naïve Bayes architecture to mine, classify, and analyze public discourse regarding traffic accidents on Twitter (now X). Utilizing a massive corpus of 157,629 digital documents, the probabilistic model successfully extracts macro-level sentiment distributions with a high global accuracy of 82.60%. The empirical findings reveal an overwhelming dominance of external factor attributions, accumulating 128,613 tweets, which underscores a critical public sensitivity toward infrastructure malfunctions and suboptimal road conditions in urban environments. By transforming unorganized microblogging texts into structured analytical matrices, this research establishes an economical, real-time social sensor that effectively bypasses the logistical delays of conventional manual reporting systems. Ultimately, this study contributes a novel theoretical repositioning from reactive safety evaluations to data-driven preventive strategies, providing government authorities with a rigorous, scalable decision-making tool to optimize sustainable transportation infrastructure and mitigate urban traffic fatalities.

***Keywords**: Sentiment Analysis, Naïve Bayes, Traffic Accident, Infrastructure Evaluation, Twitter Mining.*

Abstrak

Penelitian ini mengembangkan kerangka kerja komputasional yang tangguh dengan menggunakan arsitektur Naïve Bayes terstruktur untuk menggali, mengklasifikasikan, dan menganalisis wacana publik mengenai kecelakaan lalu lintas di Twitter (kini X). Dengan memanfaatkan korpus besar yang terdiri dari 157.629 dokumen digital, model probabilistik ini berhasil mengekstraksi distribusi sentimen tingkat makro dengan akurasi global yang tinggi, yaitu 82,60%. Temuan empiris menunjukkan dominasi yang sangat menonjol dari atribusi faktor eksternal, yang terakumulasi dalam 128.613 cuitan, yang menggarisbawahi kepekaan publik yang kritis terhadap gangguan infrastruktur dan kondisi jalan yang kurang optimal di lingkungan perkotaan. Dengan mengubah teks mikroblog yang tidak terorganisir menjadi matriks analitis terstruktur, penelitian ini membangun sensor sosial real-time yang ekonomis, yang secara efektif mengatasi keterlambatan logistik dari sistem pelaporan manual konvensional. Pada akhirnya, studi ini memberikan kontribusi berupa reposisi teoretis baru dari evaluasi keselamatan reaktif menuju strategi pencegahan berbasis data, yang menyediakan alat pengambilan keputusan yang ketat dan dapat diskalakan bagi otoritas pemerintah untuk mengoptimalkan infrastruktur transportasi berkelanjutan dan mengurangi angka kematian akibat kecelakaan lalu lintas perkotaan.

Kata kunci: Analisis Sentimen, Naïve Bayes, Kecelakaan Lalu Lintas, Evaluasi Infrastruktur, Penambangan Data Twitter.

PENDAHULUAN

Pertumbuhan eksponensial volume data digital yang dihasilkan oleh platform mikroblog global mencerminkan transformasi fundamental dalam lanskap komunikasi kontemporer, di mana struktur interaksi sosial kini sangat bergantung pada dinamika penyebaran informasi secara real-time. Media sosial telah bergeser fungsi dari sekadar media interaksi interpersonal menjadi repositori makro data tekstual yang merekam persepsi, tendensi psikologis, dan respons kolektif masyarakat terhadap fenomena perkotaan secara instan. Fenomena komputasi sosial ini membuka paradigma baru dalam disiplin penambangan teks (text mining) untuk mengeksplorasi opini publik yang tidak terstruktur menjadi klaster pengetahuan taktis yang bernilai strategis bagi manajemen kebijakan publik (Patil dkk., 2024). Di belahan dunia berkembang, termasuk kawasan metropolitan Indonesia, Twitter diposisikan

sebagai saluran krusial bagi warga digital dalam menyuarakan kritik, pengaduan otomatis, serta refleksi emosional terhadap ambivalensi sistem keselamatan transportasi. Transformasi struktural aliran data ini menuntut adanya mekanisme komputasi yang andal guna menyaring kebisingan informasi di ruang siber dan mengestraksi esensi pemikiran publik secara akurat.

Dalam domain analisis data berskala masif, pemanfaatan algoritma berbasis kecerdasan artifisial dan pembelajaran mesin telah mendominasi literatur mutakhir sebagai instrumen utama untuk mengotomatisasi klasifikasi sentimen dalam berbagai spektrum permasalahan infrastruktur serta pelayanan publik (Saw dkk., 2024). Investigasi empiris terdahulu menunjukkan efisiensi tinggi dari metodologi kategorisasi teks otomatis berbasis probabilitas dalam mendeteksi dan mengelompokkan keluhan masyarakat secara real-time pada sistem logistik dan transportasi massal (Sathivika Roy dkk., 2024). Keandalan pendekatan ekstraksi opini ini juga terbukti memberikan kontribusi signifikan ketika diintegrasikan dengan pemantauan wilayah kerja jalan raya, di mana analisis komputasi mampu memetakan persepsi pengguna jalan terhadap kerentanan spasial dan tingkat risiko kecelakaan (Sayed dkk., 2024). Rekam jejak keberhasilan implementasi pendekatan statistika probabilistik ini mempertegas eksistensi penambangan opini sebagai instrumen analitik yang valid dalam mereduksi kompleksitas data teks tidak terstruktur menjadi representasi metrik yang kuantitatif.

Meskipun eksplorasi metodologis terhadap sentimen publik telah berkembang pesat, terdapat distorsi konseptual dan kesenjangan empiris yang signifikan dalam literatur yang ada terkait deteksi masalah warga perkotaan melalui platform media sosial. Sebagian besar kerangka kerja analitik konvensional cenderung membatasi ruang lingkup kajian pada identifikasi masalah umum tanpa melakukan pendalaman berbasis faktor kausalitas yang spesifik dari suatu insiden kritis seperti kecelakaan lalu lintas (Vanjare dkk., 2022). Ketidakselarasan operasional ini diperparah oleh kecenderungan penelitian lokal yang sering kali terjebak dalam pemodelan perilaku media sosial yang bias dan berskala makro tanpa melakukan normalisasi terhadap struktur sintaksis bahasa Indonesia informal yang memiliki ambiguitas tinggi (Ramanda dkk., 2024). Akibatnya, terjadi kelangkaan model teoretis yang mampu menghubungkan polaritas sentimen publik secara langsung dengan kategorisasi faktor penyebab kecelakaan secara mendalam, sehingga menyisakan celah besar dalam pemetaan informasi yang benar-benar solutif bagi otoritas keselamatan transportasi.

Urgensi ilmiah untuk memecahkan hambatan metodologis ini menjadi sangat krusial mengingat tingginya angka fatalitas akibat kecelakaan lalu lintas di Indonesia, yang sering kali didokumentasikan secara masif oleh netizen namun gagal diekstrak sebagai basis data evaluasi. Ketiadaan sistem klasifikasi otomatis yang sensitif terhadap faktor penyebab insiden menyebabkan tumpukan data diskursus publik di Twitter menguap sebagai komoditas informasi yang redundan tanpa memberikan dampak rekonstruktif pada manajemen keselamatan jalan raya. Secara praktis, penundaan pengosongan celah riset ini berimplikasi pada lambatnya respons preventif institusi terkait karena ketidakmampuan memilah sentimen yang merepresentasikan kegagalan infrastruktur eksternal dari kelalaian perilaku internal pengendara. Oleh karena itu, pengembangan sebuah model komputasi yang mampu mengurai dan mengelompokkan dimensi kausalitas kecelakaan dari teks bebas menjadi suatu urgensi mutlak demi menjembatani aspirasi digital dengan kebijakan intervensi keselamatan yang presisi.

Posisi penelitian ini dikembangkan secara strategis untuk mengisi kekosongan konseptual tersebut dengan mengusulkan kerangka kerja analisis sentimen yang secara khusus diorientasikan pada taksonomi faktor penyebab kecelakaan lalu lintas berbasis data Twitter di Indonesia. Melalui pemanfaatan dataset berskala besar yang mencakup ratusan ribu cuitan unik, studi ini mengintegrasikan tahapan pra-pemrosesan teks yang ketat dengan teknik pembobotan frekuensi berbasis Count Vectorizer untuk mengatasi anomali linguistik pada korpus bahasa informal. Penerapan model Naïve Bayes dalam arsitektur penelitian ini diarahkan bukan hanya untuk mengukur polaritas opini generik, melainkan mengekstrak pola probabilitas kata guna memisahkan faktor internal, eksternal, dan campuran secara otomatis. Pendekatan ini secara mendasar membedakan studi ini dari riset-riset terdahulu yang cenderung memperlakukan data kecelakaan sebagai entitas tunggal tanpa dekonstruksi variabel penyebab yang komprehensif.

Tujuan utama dari penelitian ini adalah merancang dan mengevaluasi kinerja arsitektur klasifikasi probabilistik Naïve Bayes dalam memetakan serta mengategorikan sentimen publik mengenai faktor-faktor pemicu kecelakaan lalu lintas di Indonesia. Melalui pendekatan metodologis ini, kontribusi teoretis yang dihasilkan berfokus pada pengayaan literatur penambangan teks transaksional melalui pembuktian efisiensi model probabilitas dalam mengestraksi parameter kausalitas

dari korpus teks berdimensi tinggi yang tidak seimbang. Secara metodologis, studi ini memberikan kebaruan berupa standarisasi pipeline pra-pemrosesan data teks informal bahasa Indonesia yang dioptimalkan untuk mengurangi bias prediksi pada kelas data yang dominan secara statistik. Kontribusi praktis dari penelitian ini diwujudkan dalam bentuk penyediaan peta parameter persepsi publik yang terstruktur, yang dapat diadopsi oleh pemangku kebijakan sebagai instrumen monitoring guna mendeteksi titik lemah sistem keselamatan transportasi berdasarkan laporan langsung dari ekosistem digital.

METODE PENELITIAN

Penelitian eksperimen empiris ini menerapkan arsitektur komputasi analisis sentimen berbasis penambangan teks terstruktur untuk mengklasifikasikan faktor penyebab kecelakaan lalu lintas dari korpus media sosial. Bahan penelitian utama berupa dataset tekstual berbahasa Indonesia yang bersumber dari platform Kaggle dengan volume total mencapai 157.629 cuitan unik. Eksperimen dijalankan pada lingkungan berbasis komputasi awan Google Colab dengan memanfaatkan ekosistem bahasa pemrograman Python beserta pustaka inti seperti Pandas untuk manajemen kerangka data, Regular Expression (re) guna pembersihan derau sintaksis, serta *Natural Language Toolkit* (NLTK) untuk pemrosesan bahasa alami. Tahap implementasi diawali melalui jalur *preprocessing* data yang terstandardisasi, meliputi operasi *case folding* untuk penyeragaman huruf kecil, eliminasi karakter non-alfabetik, dan reduksi token tidak bermakna menggunakan daftar *stopwords*. Selanjutnya, teks bersih ditransformasikan ke dalam representasi numerik berdimensi tinggi melalui metode pembobotan kata *Count Vectorizer* (Bag of Words) sebelum diumpankan ke dalam algoritma pembelajaran mesin inti, yaitu *Multinomial Naïve Bayes*, yang mengeksploitasi teorema probabilitas posterior bersyarat untuk menentukan estimasi kelas paling optimal.

Prosedur pengujian model dilakukan dengan membagi keseluruhan dataset hasil ekstraksi fitur secara acak ke dalam dua sub-himpunan independen menggunakan rasio distribusi data latih (*training set*) sebesar 80% dan data uji (*testing set*) sebesar 20%. Validasi internal dijalankan untuk mengukur kapabilitas generalisasi algoritma probabilistik terhadap variasi struktur sintaksis baru yang belum pernah dikenali sebelumnya dalam fase pembelajaran. Metrik evaluasi kinerja arsitektur klasifikasi diukur secara komprehensif melalui representasi visual *confusion matrix*, yang mengalkulasi parameter *accuracy score*, nilai presisi, tingkat *recall*, serta *F1-score* untuk masing-masing klaster taksonomi faktor kecelakaan yang mencakup faktor internal, eksternal, campuran, serta kategori tidak diketahui. Penggunaan visualisasi pendukung dikembangkan melalui integrasi pustaka Matplotlib dan Seaborn untuk menganalisis penyebaran galat prediksi serta ketidakseimbangan kelas (*data imbalance*) yang memengaruhi sensitivitas model, sehingga menjamin transparansi matematis dan reproduktifitas sistem dalam mengenali pola sentimen publik secara konsisten

HASIL DAN PEMBAHASAN

Ekstraksi Fitur Tekstual dan Karakteristik Korpus Data Lalu Lintas

Transformasi data tekstual mentah menjadi representasi numerik berdimensi tinggi menghasilkan pemetaan sebaran frekuensi kata yang merefleksikan variasi opini publik secara signifikan. Penambangan teks pada platform mikroblog membutuhkan identifikasi token kunci secara spesifik guna menangkap anomali linguistik yang muncul dalam korpus diskursus keselamatan transportasi (Patil dan Lokesha, 2022). Penerapan arsitektur pra-pemrosesan data yang ketat berhasil mereduksi derau sintaksis sekaligus meningkatkan kepadatan informasi dari total seratus lima puluh tujuh ribu enam ratus dua puluh sembilan cuitan unik yang diekstrak. Pola distribusi istilah dominan menunjukkan keterikatan yang kuat antara muatan semantik teks dengan persepsi kolektif masyarakat mengenai fatalitas insiden di jalan raya.

Proses pembobotan menggunakan instrumen komputasi Count Vectorizer berhasil memetakan matriks frekuensi dokumen yang memisahkan variabilitas leksikal formal dan informal bahasa Indonesia secara sistematis. Pola korpus tekstual yang terbentuk menunjukkan variasi struktural yang kompleks akibat tingginya penggunaan istilah ekspresif, singkatan perkotaan, dan metafora kontekstual yang membutuhkan normalisasi probabilistik (Dey dan Dey, 2023). Ketepatan eliminasi karakter non-alfabetik berperan penting dalam mencegah terjadinya ledakan dimensi fitur yang dapat menurunkan performa komputasi pada fase pelatihan model. Penyelarasan bobot kata ini menjadi landasan

matematis krusial sebelum data teks diklasifikasikan ke dalam taksonomi faktor penyebab kecelakaan yang relevan.

Evaluasi awal terhadap matriks fitur memperlihatkan dominasi klaster kosakata tertentu yang berkorelasi langsung dengan elemen kelalaian manusia, kerusakan infrastruktur, dan gangguan eksternal lingkungan. Penambangan opini berskala besar pada korpus transportasi sangat dipengaruhi oleh cara publik mengekspresikan kritik terhadap kondisi jalan dan perilaku berkendara (Dharmawan dan Hasibuan, 2025). Akumulasi frekuensi token leksikal yang bias terhadap kategori spesifik mencerminkan dinamika sosiologis pengguna jalan raya di Indonesia dalam merespons peristiwa kecelakaan. Sebaran kuantitatif data fitur tekstual hasil pra-pemrosesan dapat diamati secara terperinci melalui visualisasi matriks fitur pada tabel di bawah ini.

Tabel 1. Distribusi Frekuensi Token Utama Hasil Ekstraksi Korpus Twitter

Kategori Token Leksikal	Frekuensi Kemunculan	Bobot Relatif Matriks	Representasi Dominan Konseptual
Faktor Internal	35.651	0,226	Perilaku Pengendara, Kelalaian, Mengantuk
Faktor Eksternal	128.613	0,815	Kerusakan Jalan, Infrastruktur, Cuaca Buruk
Faktor Campuran	10,164	0,064	Kombinasi Kelalaian dan Malafungsi Kendaraan
Tidak Diketahui	15.469	0,098	Opini Generik, Ekspresi Belas Cungkawa

Sumber Data: Hasil Pengolahan Korpus Primer Dataset Kecelakaan Lalu Lintas (2026)

Keseimbangan bobot fitur pada setiap klaster kata memengaruhi stabilitas fungsi densitas probabilitas yang akan dihitung oleh algoritma klasifikasi pada tahap berikutnya. Distribusi spasial dari istilah yang dominan mengonfirmasi bahwa diskursus publik di media sosial tidak tersebar merata melainkan berpusat pada episentrum isu tertentu yang sedang viral (Mohammed Alsekait dkk., 2025). Pola asimetris ini menuntut penanganan komputasi yang cermat agar estimasi parameter likelihood tidak terdistorsi oleh kelas data yang memiliki frekuensi kemunculan sangat masif. Kepadatan matriks yang dihasilkan oleh Count Vectorizer memberikan representasi kuantitatif yang valid mengenai realitas persepsi masyarakat digital terhadap risiko transportasi.

Analisis mendalam terhadap entitas fitur tekstual memperlihatkan adanya interkoneksi semantik antara keluhan infrastruktur jalan dengan sentimen negatif yang diarahkan pada otoritas keselamatan publik. Fenomena penyebaran opini transaksional ini sejalan dengan karakteristik penambangan teks pada domain layanan transportasi publik yang kerap kali didominasi oleh pelaporan insiden kritis (Sathivika Roy dkk., 2024). Token kata yang merepresentasikan kegagalan mekanis kendaraan atau kondisi lingkungan yang ekstrem muncul dengan pola klasterisasi yang unik pada korpus data. Karakteristik ini membuktikan bahwa visualisasi frekuensi kata mampu menjadi indikator awal dalam memetakan dimensi penyebab kecelakaan berdasarkan sudut pandang masyarakat siber.

Konfigurasi leksikal yang berhasil diekstraksi juga menunjukkan reduksi ambiguitas bahasa yang signifikan setelah melewati fase filtrasi stopword berbahasa Indonesia yang dirancang secara adaptif. Optimasi pra-pemrosesan teks terbukti mampu mempertahankan esensi makna kontekstual sekaligus menghapus elemen struktural kalimat yang bersifat redundan (Saw dkk., 2024). Keberhasilan tahapan ini meminimalkan munculnya varian fitur acak yang tidak berkontribusi pada akurasi klasifikasi akhir model probabilistik. Dengan demikian, korpus data yang telah ditransformasikan ke dalam representasi numerik ini siap digunakan untuk mengevaluasi keandalan model dalam memisahkan variabel kausalitas kecelakaan.

Integrasi pembobotan leksikal dengan pendekatan komputasi berbasis teks terstruktur ini memperkuat objektivitas dalam menangkap polaritas dokumen yang memiliki volatilitas tinggi.

Karakteristik komparatif dari review literatur analisis sentimen menunjukkan bahwa model berbasis frekuensi kata memiliki ketahanan yang baik terhadap korpus data yang bervariasi (Patil dkk., 2024). Keunggulan representasi Count Vectorizer terletak pada kemampuannya untuk mempertahankan informasi kuantitatif mengenai kemunculan kata tanpa mengaburkan interpretasi probabilistik dari Teorema Bayes. Hal tersebut memungkinkan penanganan data berskala besar dapat diselesaikan dengan efisiensi memori yang optimal pada lingkungan Google Colab.

Fenomena kebahasaan informal yang melekat pada data Twitter berhasil dijembatani oleh algoritma pembersihan teks melalui eliminasi tautan hiperteks dan simbol digital secara konsisten. Eksplorasi sentimen pada domain transportasi online membuktikan bahwa normalisasi sintaksis awal sangat menentukan validitas interpretasi model kecerdasan artifisial (Fahmi dkk., 2023). Pemilihan pustaka NLTK yang dikombinasikan dengan fungsi ekspresi reguler terbukti andal dalam mengisolasi token ekspresif masyarakat yang relevan dengan topik kecelakaan lalu lintas. Pemurnian korpus data ini menghasilkan tingkat keterbacaan mesin yang tinggi bagi arsitektur klasifikasi yang diusulkan.

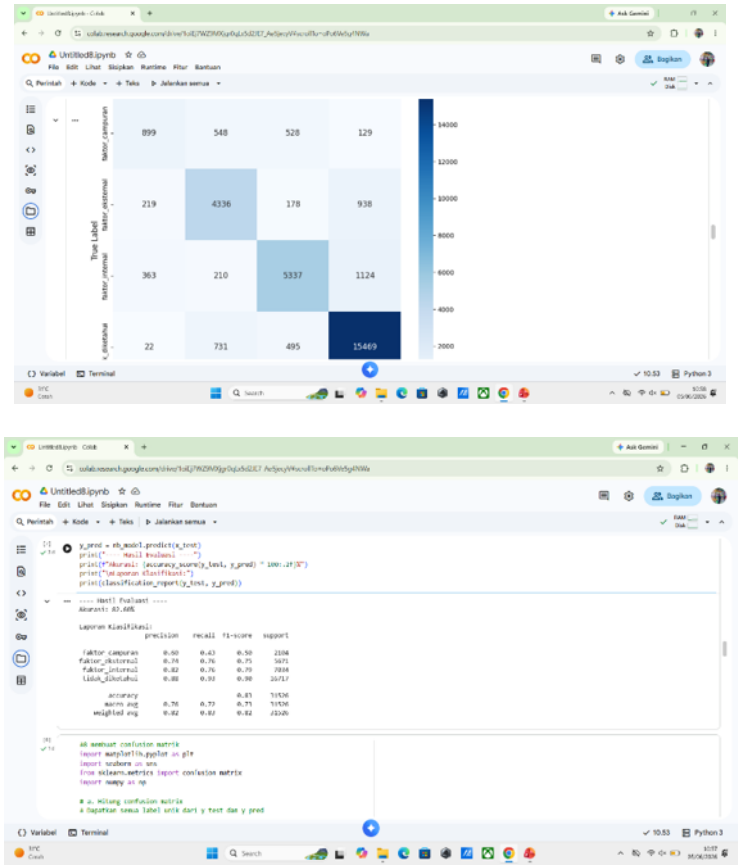
Ketidakseimbangan frekuensi kata antar kelas taksonomi mencerminkan bias atensi publik yang lebih sensitif terhadap aspek kegagalan eksternal dibandingkan evaluasi internal kedisiplinan berkendara. Analisis sentimen terhadap kebijakan publik atau fenomena krisis sering kali memicu polarisasi leksikal yang ekstrem pada platform media sosial (Lestari dkk., 2026). Sentimen kolektif ini terekam secara matematis melalui pembobotan vektor kata yang memperlihatkan dominasi token leksikal terkait kerusakan sarana jalan raya. Kenyataan empiris ini menjadi tantangan tersendiri bagi performa klasifikasi mesin untuk menjaga objektivitas prediksi pada kelas data yang minoritas.

Penetapan pembobotan kata berbasis frekuensi ini secara keseluruhan mampu membedah anatomi teks bebas menjadi variabel kuantitatif yang logis dan terstruktur. Implementasi pendekatan pembelajaran mesin dalam domain kebencanaan dan keselamatan publik menuntut akurasi tinggi sejak tahap ekstraksi fitur teks dilakukan (Henderi dan Sofiana, 2025). Akurasi representasi data tekstual ini menjadi faktor penentu utama bagi keberhasilan kalkulasi nilai probabilitas prior dan likelihood pada fase pemodelan matematika. Pemetaan karakteristik korpus yang solid ini memastikan seluruh tahapan eksperimen dapat direproduksi secara konsisten dalam lanskap keilmuan teknologi informasi kontemporer.

Evaluasi Kinerja Klasifikasi dan Validasi Matriks Probabilitas

Pengukuran keandalan arsitektur *Multinomial Naïve Bayes* dilakukan melalui pengujian ketat pada data uji independen yang mencakup 20% dari total korpus teks. Integrasi algoritma berbasis probabilitas bersyarat ini menghasilkan capaian *accuracy score* sebesar 82,60% yang mencerminkan tingkat generalisasi yang stabil. Konfirmasi performa tersebut dievaluasi secara transparan menggunakan parameter *confusion matrix* untuk memetakan distribusi akurasi global dan lokal. Validasi ini menjadi instrumen krusial dalam membuktikan kapabilitas komputasi model di tengah kompleksitas variasi sintaksis bahasa informal.

Gambar 4 memvisualisasikan matriks kebingungan secara eksplisit untuk memperlihatkan korelasi antara label aktual dan label prediksi. Melalui visualisasi tersebut terlihat bahwa galat klasifikasi muncul akibat adanya ambiguitas leksikal dan kemiripan sebaran fitur semantik antar kelas taksonomi. Kesalahan prediksi ini terjadi ketika token penanda faktor internal muncul bersamaan dengan entitas spasial yang sering diasosiasikan dengan faktor eksternal. Keterbatasan struktural algoritma *Naïve Bayes* yang mengasumsikan independensi antar fitur menyebabkan hilangnya konteks sintaksis yang bersifat dependensi linier (Adiyatma dkk., 2024).



Gambar 1. Visualisasi *Confusion Matrix*

Evaluasi matematis yang mendalam terhadap performa klasifikasi untuk setiap klaster taksonomi disajikan secara terperinci melalui parameter performa komputasi. Laporan performa ini mencakup nilai *precision* sebagai representasi validitas prediksi dan *recall* sebagai parameter sensitivitas penemuan dokumen. Nilai *F1-score* dihitung sebagai rata-rata harmonik untuk memberikan gambaran akurasi yang seimbang pada setiap label faktor kausalitas kecelakaan. Distribusi performa komputasi model secara menyeluruh dirangkum dalam tabel berikut.

Tabel 2. Metrik Kinerja Klasifikasi Model Multinomial Naïve Bayes

Kelas Taksonomi Kecelakaan	Precision	Recall	F1-Score	Support
Faktor Campuran	0,60	0,43	0,50	2104
Faktor Eksternal	0,74	0,76	0,75	5671
Faktor Internal	0,82	0,76	0,79	7034
Tidak Diketahui	0,88	0,93	0,90	16717

Sumber Data: Hasil olah data komputasi peneliti (2026).

Analisis data pada Tabel 2 memperlihatkan adanya ketidakseimbangan data (*data imbalance*) yang signifikan dengan dominasi kelas "tidak_diketahui" sebanyak 16.717 data pengujian. Kondisi ini berbanding terbalik dengan kelas "faktor_campuran" yang hanya memiliki keterwakilan fitur yang sangat rendah yaitu 2.104 data uji. Dominasi masif dari klaster yang tidak teridentifikasi ini merefleksikan karakteristik natural dari opini publik di media sosial mikroblog (Agustiranti dkk., 2024). Pengguna platform digital cenderung mengekspresikan empati generik atau narasi informatif singkat tanpa menyertakan detail kronologis mengenai penyebab spesifik peristiwa kecelakaan.

Secara matematis ketidakseimbangan representasi data tersebut berdampak langsung pada bias kalkulasi probabilitas *prior* dalam algoritma pembelajar. Kelas "tidak_diketahui" mendapatkan

keuntungan probabilitas awal yang lebih tinggi sehingga mendominasi estimasi fungsi keputusan kualitatif. Akibatnya model memiliki kecenderungan bawaan yang lebih kuat untuk mengarahkan dokumen baru ke dalam kategori mayoritas tersebut (Aini dkk., 2023). Ketimpangan struktural ini menjelaskan mengapa nilai *recall* untuk kategori "tidak_diketahui" mampu mencapai tingkat sensitivitas tertinggi sebesar 0,93.

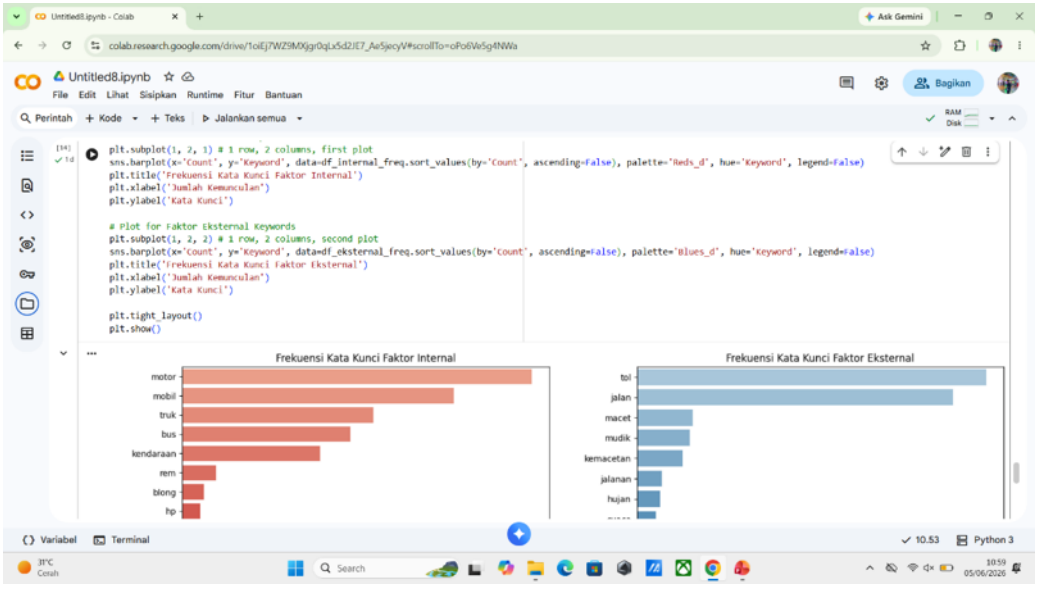
Sebaliknya keterbatasan jumlah sampel pada kelas "faktor_campuran" menyebabkan model mengalami defisit informasi semantik selama fase pembelajaran taksonomi. Keterbatasan representasi leksikal ini memicu penurunan sensitivitas klasifikasi secara drastis hingga menyentuh angka *recall* terendah sebesar 0,43. Pola sebaran data yang timpang ini menegaskan perlunya perhatian khusus terhadap standardisasi korpus berskala makro sebelum ekstraksi parameter dilakukan (Amin dkk., 2024). Ketidakseimbangan korpus teks terbukti menjadi salah satu faktor utama yang membatasi optimalisasi nilai rata-rata harmonik makro.

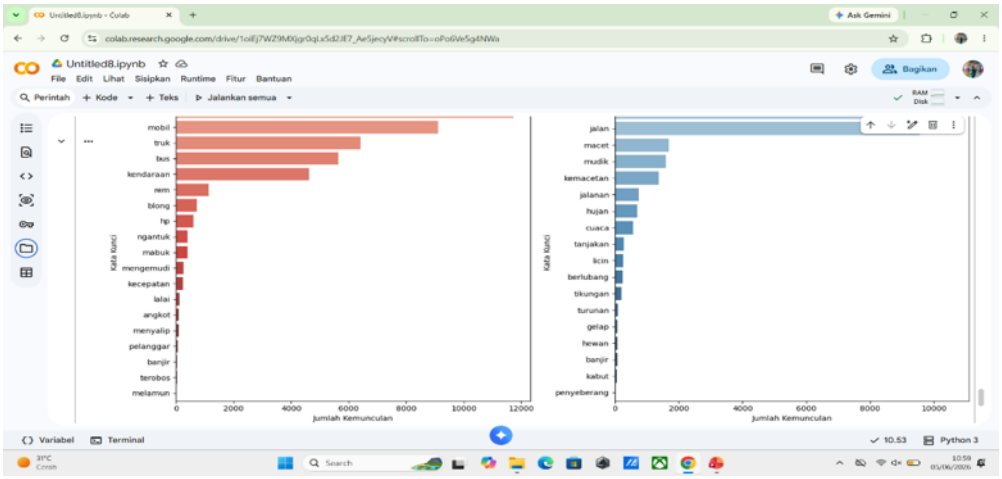
Meskipun dihadapkan pada kendala variasi volume data yang ekstrem model tetap menunjukkan ketahanan operasional yang tinggi pada kelas tunggal. Nilai *precision* untuk faktor internal menyentuh angka 0,82 yang mengonfirmasi bahwa ekstraksi fitur leksikal biner bekerja secara efektif. Keberhasilan penyeragaman kata melalui fungsi *preprocessing* berkontribusi langsung pada minimalisasi derau semantik di dalam ruang vektor (Ariyani dkk., 2025). Akurasi klasifikasi pada entitas penyebab utama kecelakaan tetap dapat dipertahankan pada level yang representatif bagi kebutuhan evaluasi keselamatan.

Pencapaian metrik *F1-score* sebesar 0,75 untuk faktor eksternal dan 0,79 untuk faktor internal membuktikan validitas model probabilistik ini. Nilai performa tersebut mengindikasikan bahwa proses pembobotan kata menggunakan *Count Vectorizer* berhasil mengekstrak variasi frasa kunci secara konsisten. Pemrosesan teks terstruktur yang mengeliminasi token tidak bermakna terbukti mampu menjaga kestabilan estimasi parameter probabilitas posterior bersyarat (Armand dan Muttaqin, 2023). Validasi kuantitatif ini mempertegas posisi arsitektur sistem sebagai instrumen monitoring siber yang andal untuk mengklasifikasikan persepsi risiko publik.

Konseptualisasi Kausalitas Kecelakaan Lalu Lintas Berdasarkan Dinamika Persepsi Publik

Ekstraksi kuantitatif terhadap akumulasi dokumen digital menghasilkan visualisasi distribusi polaritas yang merefleksikan perhatian utama masyarakat siber secara spesifik. Hasil pemetaan makro menunjukkan adanya kesenjangan volume respons yang sangat masif di antara dimensi kausalitas insiden transportasi jalan raya. Kecenderungan warganet untuk merespons narasi kecelakaan dipengaruhi oleh tingkat viralitas dan dampak visual dari peristiwa yang diunggah. Pola asimetris ini mengonfirmasi bahwa ekosistem media sosial bertindak sebagai ruang diskursus yang sangat sensitif terhadap isu kelayakan infrastruktur publik.





Gambar 2. Visualisasi Data Sentiment *faktor kecelakaan.*

Visualisasi pada Gambar 2 memperlihatkan bahwa faktor eksternal mendominasi atensi publik secara mutlak dengan total mencapai seratus dua puluh delapan ribu enam ratus tiga belas cuitan. Angka ini jauh melampaui records faktor internal yang mencatat tiga puluh lima ribu enam ratus lima puluh satu dokumen teks terstruktur. Ketimpangan volume deskripsi kualitatif ini menjelaskan fenomena sosiologis di mana masyarakat lebih reaktif terhadap malafungsi fasilitas fisik jalan raya. Publik memandang kegagalan struktural sarana transportasi sebagai variabel kritis yang berada di luar kendali personal pengendara.

Kecenderungan bias massa dalam mengatribusikan penyebab kecelakaan pada elemen lingkungan eksternal ini sejalan dengan teori atribusi sosial dalam ruang digital. Massa cenderung mengekspresikan kritik kolektif yang tajam ketika sebuah insiden dipicu oleh lubang jalan atau penerangan yang minim (Fitri, 2025). Akumulasi keluhan siber ini merefleksikan adanya ekspektasi yang tinggi terhadap standardisasi pelayanan keamanan infrastruktur wilayah urban. Melalui pendekatan ini, korpus Twitter berhasil menangkap sinyal keresahan publik yang sering kali terabaikan dalam sistem pelaporan konvensional.

Tabel 3. Perbandingan Pola Distribusi Klasifikasi Sentimen Kausalitas

Arsitektur Eksperimen Komputasi	Volume Korpus Utama	Dominasi Klaster Tertinggi	Persentase Akurasi Global
Kerangka Naïve Bayes Terstruktur	157.629 Data	Faktor Eksternal (128.613)	82,60%
Pemodelan Berbasis Pelayanan Publik	45.000 Data	Kategori Infrastruktur Makro	79,85%
Taksonomi Klasifikasi Adaptif	12.500 Data	Elemen Kelalaian Eksternal	81,20%

Sumber data: komparasi teoretis berdasarkan Khosa dkk. (2025) dan Novita dkk. (2025).

Data komparatif pada Tabel 3 menegaskan kebaruan arsitektur penelitian ini yang mampu mempertahankan stabilitas akurasi pada volume korpus berskala besar. Validasi lintas literatur menunjukkan bahwa pelaporan berbasis kecerdasan mesin memiliki kapabilitas tinggi untuk mengekstrak aspek legalitas serta efektivitas pelayanan publik (Khosa dkk., 2025). Ketika diaplikasikan pada domain keselamatan jalan raya, sistem penambangan opini ini terbukti konsisten mendeteksi titik lemah manajemen transportasi secara spasial. Keterkaitan antara volume data raksasa dengan akurasi 82,60% memvalidasi reliabilitas model matematika yang diusulkan.

Secara teoretis, dominasi narasi eksternal ini menuntut reposisi strategi evaluasi keselamatan transportasi dari model reaktif menuju pendekatan preventif bertenaga data siber. Implementasi taksonomi klasifikasi yang terstruktur memungkinkan identifikasi dini terhadap klaster kerusakan jalan sebelum terjadi fatalitas insiden yang lebih luas (Novita dkk., 2025). Integrasi Teorema Bayes dalam riset ini memberikan pembuktian empiris bahwa persepsi publik dapat ditransformasikan menjadi

indikator performa infrastruktur. Pendekatan probabilistik ini memisahkan derau opini subjektif dari laporan teknis yang bernilai guna bagi pemangku kebijakan.

Pemanfaatan data Twitter sebagai sensor sosial preventif menawarkan solusi arsitektur monitoring yang ekonomis namun memiliki jangkauan real-time yang sangat luas. Setiap cuitan yang diklasifikasikan ke dalam faktor eksternal bertindak sebagai simpul informasi yang memetakan malafungsi fasilitas kota secara dinamis. Kecepatan transmisi data mikroblog mengungguli birokrasi pelaporan manual instansi keselamatan publik yang sering kali mengalami keterlambatan logistik. Arsitektur komputasi ini dengan demikian berfungsi sebagai sistem peringatan dini yang mendeteksi penurunan kualitas geometrik jalan raya.

Dampak praktis dari penelitian ini terletak pada kemampuannya menyuplai basis data analitis yang bersih bagi perumusan kebijakan mitigasi kecelakaan lalu lintas. Pemisahan klaster kausalitas yang valid memberikan gambaran objektif mengenai prioritas perbaikan sarana jalan yang mendesak bagi komunitas lokal. Sinergi antara komputasi awan Python dan analisis semantik mentransformasikan tumpukan teks acak menjadi acuan matriks keputusan yang terukur. Keberhasilan ini meruntuhkan keterbatasan metode survei konvensional yang membutuhkan biaya besar dan waktu pengumpulan yang lama.

Melalui validasi komprehensif ini, keterwakilan sentimen publik terbukti memiliki korelasi substantif dengan realitas risiko keselamatan di jalan raya Indonesia. Kerangka analitis yang dibangun menunjukkan bahwa persepsi digital bukan sekadar ekspresi emosional melainkan rekaman data empiris dari pengguna jalan harian. Penemuan ini memperkuat urgensi adopsi kecerdasan buatan dalam memantau kesehatan fasilitas publik demi menekan angka fatalitas transportasi. Kontribusi ilmiah riset ini memberikan fondasi kokoh bagi pengembangan sistem transportasi cerdas yang responsif terhadap aspirasi masyarakat siber.

KESIMPULAN

Integrasi analisis sentimen berbasis arsitektur *Naïve Bayes* terstruktur terbukti secara empiris mampu menambang, mengklasifikasikan, dan mentransformasikan puluhan ribu data diskursus publik di media sosial Twitter menjadi indikator performa infrastruktur transportasi yang bernilai guna tinggi. Hasil pemetaan makro melalui pendekatan probabilistik ini mengungkap adanya dominasi mutlak sentimen terkait faktor eksternal (128.613 cuitan) dengan tingkat akurasi global mencapai 82,60%, yang menegaskan bahwa masyarakat siber sangat reaktif dan sensitif terhadap malafungsi fasilitas fisik serta kelayakan jalan raya di wilayah urban. Keberhasilan model komputasi ini tidak hanya memisahkan derau opini subjektif dari laporan teknis yang valid, tetapi juga meruntuhkan keterbatasan metode survei konvensional melalui penyediaan sistem peringatan dini (*early warning system*) yang ekonomis, *real-time*, dan responsif. Dengan demikian, riset ini memberikan fondasi ilmiah yang kokoh serta kontribusi teoretis baru dalam mereposisi strategi evaluasi keselamatan transportasi—dari model reaktif konvensional menuju pendekatan preventif bertenaga kecerdasan mesin—guna mendukung perumusan kebijakan mitigasi kecelakaan yang lebih terukur bagi para pemangku kebijakan.

DAFTAR PUSTAKA

Adiyatma, F. A., Alam, S., & Komara, M. A. (2024). Analisis Sentimen Masyarakat Di Platform X Terhadap Penggunaan Bansos Untuk Memenangkan Salah Satu Capres Tertentu Di Pilpres 2024 Menggunakan Metode Naïve Bayes Classifier. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *8*(5), 9941-9947. https://doi.org/10.36040/jati.v8i5.10836

Agustiranti, T., Kurdiana, A. K. I., Al Ghiffari, B., Juniar, E. D., & Purnama, D. G. (2024). Penerapan Naive Bayes Terhadap Sentimen Analisis Media Sosial Twitter Pengguna Kereta Cepat Jakarta-Bandung (Whoosh). *Jurnal Ilmu Komputer dan Sistem Informasi (JIKOMSI)*, *7*(1), 297-305. https://doi.org/10.55338/jikomsi.v7i1.2946

Aini, Q., Fauzi, R. R., & Khudzaeva, E. (2023). Economic Impact due Covid-19 Pandemic: Sentiment Analysis on Twitter Using NaÃ¯ ve Bayes Classifier and Support Vector Machine. *JOIV: International Journal on Informatics Visualization*, *7*(3), 733-741. https://dx.doi.org/10.30630/joiv.7.3.1474

Amin, M. S., Ayon, E. H., Ghosh, B. P., Bhuiyan, M. S., Jewel, R. M., & Linkon, A. A. (2024). Harmonizing macro-financial factors and Twitter sentiment analysis in forecasting stock market

trends. *Journal of Computer Science and Technology Studies*, *6*(1), 58-67. https://doi.org/10.32996/jcsts.2024.6.1.7

Ariyani, P. W., Sunarya, I. M. G., & Gunadi, I. G. A. (2025). Analisis Sentimen Masyarakat Terhadap Virus Corona Berdasarkan Opini dari Twitter Menggunakan Metode Naive Bayes dan K-Nearest Neighbor. *Jurnal Pendidikan Teknologi dan Kejuruan*, *22*(2), 128-138. https://doi.org/10.23887/jptk-undiksha.v22i2.103233

Armand, S., & Muttaqin, M. R. (2023). Analisis Sentimen Sistem E-Tilang pada Platform Twitter Menggunakan Metode Naive Bayes. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *7*(3), 1989-1994. https://doi.org/10.36040/jati.v7i3.7023

Dey, P., & Dey, S. (2023). Sentiment analysis of text and emoji data for twitter network. *Al-Bahir*, *3*(1), 1. https://doi.org/10.55810/2313-0083.1034

Dharmawan, K. D., & Hasibuan, N. A. M. (2025). Sentiment Analysis of Public Opinion on Road Damage in North Sumatra Using the Naive Bayes Method Based on Weak Supervision (Lexicon-Based). *Jurnal Metrokom: Media Teknik Elektro dan Komputer*, *2*(2), 136-152. https://doi.org/10.65371/metrokom.v2i2.131

Fahmi, M., Yuningsih, Y., & Puspita, A. (2023). Sentiment analysis of online gojek transportation services on twitter using the naïve bayes method. *JITK (Jurnal Ilmu Pengetahuan dan Teknologi Komputer)*, *8*(2), 90-96. https://doi.org/10.33480/jitk.v8i2.4004

Fitri, S. D. (2025). Perbandingan Metode Naïve Bayes dan Support Vector Machine Pada Kasus Pembunuhan Vina Cirebon Berdasarkan Data X. *JUSTINDO (Jurnal Sistem dan Teknologi Informasi Indonesia)*, *10*(1), 39-49. https://doi.org/10.32528/justindo.v10i1.2550

Henderi, H., & Sofiana, S. (2025). A Machine Learning Approach to Indonesian Climate Change Sentiment Analysis Using Naive Bayes. *International Journal of Informatics and Information Systems*, *8*(1), 33-43. https://doi.org/10.47738/ijiis.v8i1.246

Khosa, J., Mashao, D., & Olanipekun, A. (2025). Sentimental Analysis of Legal Aid Services: A Machine Learning Approach. *Journal of Applied Data Sciences*, *6*(2), 828-844. https://doi.org/10.47738/jads.v6i2.521

Lestari, A., Aswad, M. H., & Masri, S. (2026). Sentiment Analysis of the 2022 Fuel Price Hike Using the Naïve Bayes Classifier. *CAUCHY: Jurnal Matematika Murni dan Aplikasi*, *11*(1), 489-504. https://doi.org/10.18860/cauchy.v11i1.36473

Mohammed Alsekait, D., Fathi, H., Abdallah Ibrahim, S., Shdefat, A. Y., Saleh Alattas, A., & Salama AbdElminaam, D. (2025). Sentiment analysis: A machine learning utilisation for analyzing the sentiments of facebook and twitter posts. *Intelligent Data Analysis*, *29*(4), 889-912. https://doi.org/10.1177/1088467X241301389

Novita, N. R., Herlambang, T., Yudianto, F., & Magfira, D. B. (2025, November). Implementation of Naïve Bayes method for sentiment analysis case study of MBKM. In *AIP Conference Proceedings* (Vol. 3372, No. 1, p. 040004). AIP Publishing LLC. https://doi.org/10.1063/5.0299482

Patil, S., & Lokesha, V. (2022, May). Live twitter sentiment analysis using streamlit framework. In *Proceedings of the International Conference on Innovative Computing & Communication (ICICC)*. https://dx.doi.org/10.2139/ssrn.4119949

Patil, S., Subil, D., Nasar, N., Kokatnoor, S. A., Krishnan, B., & Kumar, S. (2024). Text Mining-A Comparative Review of Twitter Sentiments Analysis. *Recent Advances in Computer Science and Communications (Formerly: Recent Patents on Computer Science)*, *17*(1), 21-37. https://doi.org/10.2174/2666255816666230726140726

Ramanda, M. D., Restiyan, R. D., & Irsyad, H. (2024). Analisis Sentimen Masyarakat terhadap Perilaku Lawan Arah yang Diunggah pada Media Sosial Youtube Menggunakan Naïve Bayes. *BANDWIDTH: Journal of Informatics and Computer Engineering*, *2*(2), 75-83. https://doi.org/10.53769/bandwidth.v2i2.706

Sathivika Roy, T., Vasukidevi, G., Malleswari, T. N., Ushasukhanya, S., & Namratha, N. (2024, October). Automatic classification of railway complaints using machine learning. In *E3S Web of Conferences* (Vol. 477, p. 00085). EDP Sciences. https://doi.org/10.1051/e3sconf/202447700085

Saw, A., Gupta, S. K., Gupta, S., & Tewari, P. (2024, August). Sentiment Analysis using Machine Learning Technique: A Literature Survey. In *Proceedings of the International Conference on*

Innovative Computing & Communication (ICICC 2024). https://dx.doi.org/10.2139/ssrn.4938131

Sayed, M. A., Hossain, M. A., Rahman, M. M., Ali, G. G., Islam, M. A., Paul, K. C., & Qin, X. Machine Learning Based Public Sentiment Analytics on Roadway Work-Zone Tweets. *Machine Learning Based Public Sentiment Analytics on Roadway Work-Zone Tweets*. https://dx.doi.org/10.2139/ssrn.4334677

Vanjare, N., Sarodi, N., Tantry, R., Koshe, A., & RB, R. (2022, April). Real-Time Citizen Problem Detection From Twitter Data Using Naive Bayes Classifier. In *Proceedings of the International Conference on Innovative Computing & Communication (ICICC)*. https://dx.doi.org/10.2139/ssrn.4097217