INSOLOGI: Jurnal Sains dan Teknologi
https://journal.literasisains.id/index.php/INSOLOGI
ISSN 2828-4984 (Media Online) | ISSN 2828-4992 (Media Cetak)
Vol. 5 No. 3 (Juni 2026) 875-890
DOI: 10.55123/insologi.v5i3.8050
Submitted: 07-05-2026 | Accepted: 13-05-2026 | Published: 15-06-2026

Analisis Sentimen Ulasan Pengguna Generative AI Menggunakan Naïve Bayes Berbasis Python

Rizky Nurhasanah[1], Victor Asido Elyakim[2]
[1,2]Sistem Informasi, STIKOM Tunas Bangsa, Pematang Siantar, Indonesia
Email: [1]rizkinurhasanah733@gmail.com, [2]victorasidoelyakim@gmail.com

Abstract

The advancement of Generative Artificial Intelligence (Generative AI) technology has driven increased use of AI-based applications across various sectors, including education, productivity, digital services, and information retrieval. The high intensity of usage generates massive volumes of user reviews on digital platforms, containing valuable information regarding user satisfaction, experience, and perception of system service quality. This study aims to analyze the sentiment of user reviews toward Generative AI applications using a Python-based Naïve Bayes algorithm. The dataset consists of 50,000 user reviews covering five Generative AI applications: ChatGPT, Microsoft Copilot, Google Gemini, Perplexity, and Claude. The research pipeline includes text preprocessing stages of case folding, cleaning, tokenizing, stopword removal, and feature transformation using TF-IDF with unigram and bigram representations. Data are then classified into three sentiment categories (positive, negative, neutral) using three variants of the Naïve Bayes algorithm: Multinomial, Complement, and Bernoulli. Results show a sentiment distribution of 33,695 positive (67.4%), 13,275 negative (26.6%), and 3,030 neutral (6.1%) reviews. Multinomial Naïve Bayes achieved the best performance with an accuracy of 83.07%, precision 77.79%, recall 83.07%, and F1-score 80.29%. Five-fold cross-validation confirmed stable performance with a mean accuracy of 82.31% and a standard deviation of 0.23%.

Keywords: *Sentiment Analysis, Generative AI, Naïve Bayes, TF-IDF, Text Mining.*

Abstrak

Perkembangan teknologi Generative Artificial Intelligence (Generative AI) telah mendorong peningkatan penggunaan aplikasi berbasis kecerdasan buatan pada berbagai sektor, seperti pendidikan, produktivitas, layanan digital, dan pencarian informasi. Tingginya intensitas penggunaan tersebut menghasilkan volume ulasan pengguna yang sangat besar pada berbagai platform digital, sehingga mengandung informasi penting terkait tingkat kepuasan, pengalaman, serta persepsi pengguna terhadap kualitas layanan sistem. Penelitian ini bertujuan untuk menganalisis sentimen ulasan pengguna terhadap aplikasi Generative AI menggunakan algoritma Naïve Bayes berbasis Python. Dataset yang digunakan berjumlah 50.000 ulasan pengguna yang mencakup lima aplikasi Generative AI, yaitu ChatGPT, Microsoft Copilot, Google Gemini, Perplexity, dan Claude. Tahapan penelitian meliputi preprocessing teks yang terdiri atas case folding, cleaning, tokenizing, stopword removal, serta transformasi fitur menggunakan metode TF-IDF dengan representasi unigram dan bigram. Selanjutnya, data diklasifikasikan ke dalam tiga kategori sentimen, yaitu positif, negatif, dan netral, menggunakan beberapa varian algoritma Naïve Bayes, yaitu Multinomial Naïve Bayes, Complement Naïve Bayes, dan Bernoulli Naïve Bayes. Hasil penelitian menunjukkan distribusi sentimen sebesar 33.695 data positif (67,4%), 13.275 data negatif (26,6%), dan 3.030 data netral (6,1%). Berdasarkan hasil evaluasi model, Multinomial Naïve Bayes memberikan performa terbaik dengan tingkat akurasi sebesar 83,07%, precision 77,79%, recall 83,07%, dan F1-score 80,29%. Hasil validasi silang 5-fold juga menunjukkan performa yang stabil dengan rata-rata akurasi sebesar 82,31% dan standar deviasi 0,23%.

Kata Kunci: Analisis Sentimen, Generative AI, Naïve Bayes, TF-IDF, Text Mining.

1. PENDAHULUAN

Perkembangan teknologi kecerdasan buatan, khususnya Generative Artificial Intelligence (Generative AI), telah mengalami akselerasi yang signifikan dalam beberapa tahun terakhir. Generative AI merupakan cabang dari kecerdasan buatan yang mampu menghasilkan konten baru berupa teks, gambar, kode, maupun audio berdasarkan pola yang dipelajari dari data latih dalam skala besar (Afuan et al., 2025; Saputra, 2025). Kemampuan ini menjadikan Generative AI sebagai teknologi yang sangat relevan untuk diterapkan di berbagai sektor kehidupan, mulai dari pendidikan, layanan kesehatan, bisnis, produktivitas kerja, hingga hiburan digital (Khinsa Fairuz Zahirah, Benny Irawan, 2025; Taherdoost & Madanchian, 2023). Kehadiran aplikasi-aplikasi berbasis Generative AI seperti ChatGPT, Microsoft Copilot, Google Gemini, Perplexity, dan Claude telah merevolusi cara manusia berinteraksi dengan sistem komputer, yaitu dengan memungkinkan percakapan berbasis bahasa alami yang semakin canggih dan kontekstual. Adopsi teknologi ini terus meningkat secara eksponensial, yang ditandai oleh jutaan pengguna aktif di seluruh dunia yang memanfaatkan layanan tersebut dalam aktivitas harian mereka (Muzaki et al., 2026; Yenduri et al., 2024).

Tingginya intensitas penggunaan aplikasi Generative AI menghasilkan volume ulasan pengguna yang sangat besar pada berbagai platform digital seperti Google Play Store, Apple App Store, dan forum teknologi. Ulasan-ulasan tersebut merupakan sumber data yang sangat berharga karena mengandung opini, persepsi, pengalaman, serta tingkat kepuasan pengguna secara langsung terhadap kualitas layanan aplikasi (Nurfauzi et al., 2025; Rahmawati & Wicaksono, 2022). Analisis terhadap ulasan pengguna secara manual tidak lagi memungkinkan mengingat skala data yang terus berkembang pesat, sehingga dibutuhkan pendekatan komputasional yang efisien dan terukur (Setiawan et al., 2025). Analisis sentimen sebagai salah satu cabang dari Natural Language Processing (NLP) menawarkan solusi untuk mengekstraksi dan mengklasifikasikan opini dari teks tidak terstruktur secara otomatis ke dalam kategori sentimen positif, negatif, atau netral. Pemahaman mendalam terhadap sentimen pengguna sangat penting bagi pengembang aplikasi untuk mengidentifikasi kelemahan produk, memahami ekspektasi pengguna, serta menyusun strategi peningkatan layanan yang berbasis data (Zidan et al., 2025).

Berbeda dengan analisis sentimen pada domain e-commerce atau media sosial konvensional, ulasan pengguna terhadap aplikasi Generative AI memiliki karakteristik linguistik yang lebih kompleks dan kontekstual. Pengguna tidak hanya menilai kualitas layanan aplikasi, tetapi juga mengevaluasi kemampuan reasoning, akurasi informasi, kemampuan generasi teks, konsistensi respons, hingga aspek etika dan halusinasi model (*AI hallucination*). Fenomena ini menyebabkan pola sentimen pada domain Generative AI cenderung lebih ambigu dibandingkan ulasan aplikasi biasa yang umumnya berfokus pada aspek transaksi atau pengalaman antarmuka pengguna. Penelitian terbaru menunjukkan bahwa ulasan terhadap sistem berbasis Large Language Model (LLM) sering mengandung opini campuran (*mixed sentiment*) dalam satu kalimat, misalnya pengguna memberikan apresiasi terhadap kualitas jawaban tetapi mengkritik kecepatan atau akurasi model secara bersamaan (Zhang et al., 2024; OpenAI User Behavior Report, 2025).

Selain itu, Generative AI juga menghasilkan interaksi berbasis bahasa alami yang jauh lebih panjang dan semantik dibandingkan platform digital konvensional. Hal ini menyebabkan distribusi fitur teks menjadi lebih dinamis dan memiliki tingkat variasi kosakata yang tinggi. Menurut penelitian Taherdoost & Madanchian (2023), evaluasi pengguna terhadap sistem AI generatif dipengaruhi oleh faktor trustworthiness, contextual relevance, explainability, dan adaptivitas model, sehingga pendekatan analisis sentimen pada domain ini memerlukan representasi fitur yang mampu menangkap

hubungan antar kata secara lebih kontekstual. Penelitian Yenduri et al. (2024) juga menjelaskan bahwa perkembangan Generative AI telah mengubah pola interaksi manusia-komputer menjadi lebih conversational dan personalized, sehingga opini pengguna tidak lagi bersifat sederhana seperti pada ulasan e-commerce tradisional.

Beberapa penelitian terkini juga menunjukkan bahwa analisis sentimen pada domain Generative AI memiliki tantangan khusus berupa tingginya kemunculan kalimat ambigu, ironi implisit, serta opini evaluatif teknis yang sulit dipetakan hanya melalui pendekatan leksikal sederhana (Saputra, 2025; Muzaki et al., 2026). Oleh karena itu, penelitian terhadap sentimen pengguna Generative AI menjadi penting untuk memahami persepsi publik terhadap kualitas sistem AI modern secara lebih mendalam dan spesifik.

Berbeda dengan ulasan pada e-commerce atau media sosial biasa, ulasan pengguna Generative AI memiliki karakteristik yang lebih kompleks karena pengguna tidak hanya menilai kualitas aplikasi, tetapi juga akurasi jawaban, kemampuan reasoning, relevansi konteks, hingga konsistensi respons AI. Ulasan pada aplikasi seperti ChatGPT, Gemini, dan Claude sering mengandung opini campuran dalam satu kalimat, misalnya pengguna memuji kualitas jawaban tetapi mengkritik kecepatan atau akurasi sistem (Taherdoost & Madanchian, 2023; Yenduri et al., 2024). Oleh karena itu, analisis sentimen pada domain Generative AI memerlukan pendekatan yang lebih spesifik dibandingkan analisis sentimen pada aplikasi digital konvensional.

Berbagai penelitian terdahulu telah mengaplikasikan metode analisis sentimen pada domain ulasan aplikasi digital. Algoritma Naïve Bayes telah banyak digunakan dalam klasifikasi teks karena kesederhanaan komputasionalnya, kemampuannya menangani data berdimensi tinggi, serta kinerjanya yang kompetitif dibandingkan algoritma yang lebih kompleks (Nugroho et al., 2022). Kombinasi antara Naïve Bayes dan metode pembobotan TF-IDF (Term Frequency-Inverse Document Frequency) terbukti efektif dalam merepresentasikan fitur teks karena mempertimbangkan relevansi suatu kata terhadap dokumen secara kontekstual, bukan sekadar frekuensi kemunculannya (Imelda & Kurnianto, 2026). Pada penelitian sebelumnya, Multinomial Naïve Bayes dengan TF-IDF berhasil mencapai akurasi di atas 80% pada berbagai dataset ulasan aplikasi seperti e-commerce, media sosial, dan layanan streaming digital (Khoerunnisa et al., 2025). Selain itu, penggunaan representasi bigram terbukti meningkatkan kemampuan model dalam menangkap hubungan antar kata dan konteks semantik yang lebih kompleks dibandingkan unigram (Dewi & Santoso, 2022).

Meskipun penelitian analisis sentimen telah banyak dilakukan, terdapat beberapa celah penelitian (research gap) yang masih perlu diisi. Pertama, sebagian besar studi terdahulu hanya berfokus pada satu aplikasi tunggal, sehingga belum memberikan gambaran komparatif lintas platform Generative AI secara komprehensif (Brawijaya et al., 2025; Putri et al., 2024). Kedua, perbandingan antar varian Naïve Bayes, yaitu Multinomial, Complement, dan Bernoulli, pada domain ulasan Generative AI belum banyak dilakukan, padahal masing-masing varian memiliki karakteristik yang berbeda dalam menangani distribusi fitur teks (Rahmaliyadi & Maridjan, 2025). Ketiga, dataset yang digunakan pada studi sebelumnya umumnya berskala kecil hingga menengah (di bawah 20.000 sampel), sehingga belum cukup representatif untuk menggambarkan pola sentimen pengguna Generative AI secara umum (Madjid et al., 2023). Keempat, evaluasi model yang dilakukan sering kali hanya menggunakan satu metrik tunggal seperti akurasi, tanpa mempertimbangkan precision, recall, F1-score, dan cross-validation yang lebih komprehensif (Agustin et al., 2025; Nasution et al., 2024; Sutrisno & Putri, 2024).

Berdasarkan latar belakang dan identifikasi research gap di atas, penelitian ini bertujuan untuk: (1) menganalisis sentimen ulasan pengguna terhadap lima aplikasi Generative AI terkemuka menggunakan dataset berjumlah 50.000 ulasan; (2)

membandingkan performa tiga varian algoritma Naïve Bayes, yaitu Multinomial, Complement, dan Bernoulli, dalam mengklasifikasikan sentimen ulasan; (3) mengoptimalkan representasi fitur teks menggunakan TF-IDF dengan kombinasi unigram dan bigram; serta (4) mengevaluasi model secara komprehensif menggunakan akurasi, precision, recall, F1-score, dan 5-fold cross-validation. Kebaruan penelitian ini terletak pada penggunaan dataset ulasan lintas lima platform Generative AI dengan skala besar 50.000 data serta perbandingan tiga varian Naïve Bayes yang belum banyak dikaji secara komprehensif pada domain Generative AI. Hasil penelitian ini diharapkan dapat memberikan kontribusi ilmiah dalam bidang text mining dan NLP, sekaligus memberikan wawasan praktis bagi pengembang aplikasi Generative AI dalam memahami persepsi pengguna.

Metode TF-IDF dengan kombinasi unigram dan bigram dipilih karena mampu menangkap kata tunggal maupun hubungan antar kata yang sering muncul pada ulasan pengguna, seperti *very helpful*, *too slow*, atau *not accurate*. Kombinasi ini terbukti lebih efektif dibandingkan unigram saja dalam memahami konteks sentimen (Dewi & Santoso, 2022).

Meskipun metode modern seperti Word2Vec dan FastText mampu memahami hubungan semantik yang lebih kompleks, metode tersebut membutuhkan komputasi yang lebih besar dan umumnya lebih optimal jika digunakan pada model deep learning. Pada penelitian ini, TF-IDF dipilih karena lebih ringan, mudah diinterpretasikan, dan efektif untuk data teks berskala besar.

Naïve Bayes tetap relevan digunakan karena memiliki proses pelatihan yang cepat, mampu menangani data teks berdimensi tinggi, serta memberikan performa yang kompetitif pada klasifikasi sentimen. Pada dataset sebesar 50.000 ulasan, Naïve Bayes juga lebih efisien dibandingkan model yang lebih kompleks, terutama ketika dikombinasikan dengan fitur TF-IDF (Nugroho et al., 2022; Madjid et al., 2023).

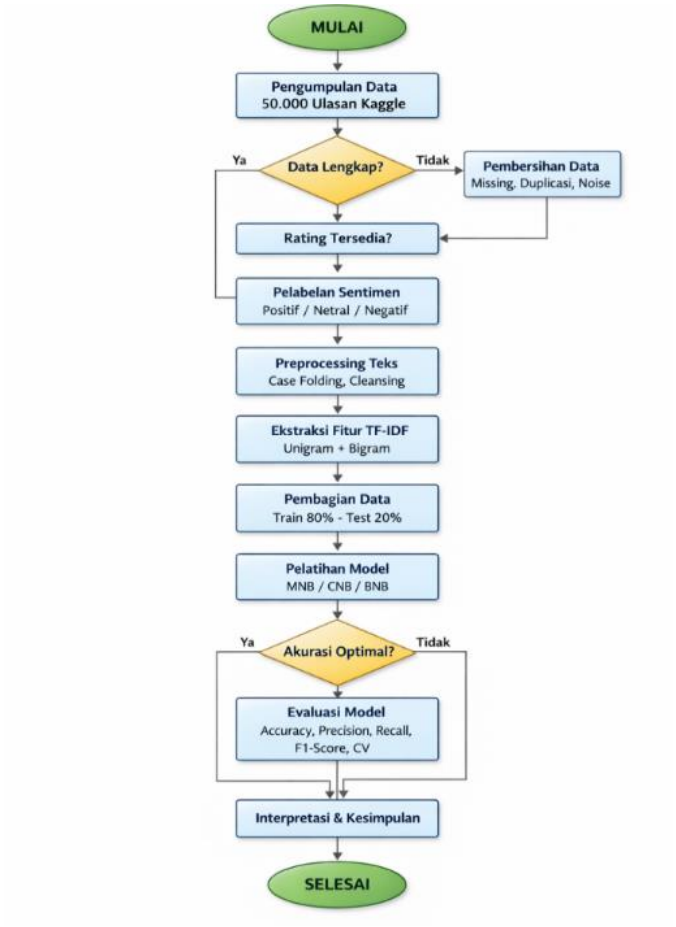
2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Penelitian ini menggunakan pendekatan kuantitatif dengan jenis penelitian eksperimental komputasional. Pendekatan kuantitatif dipilih karena penelitian ini bertujuan untuk mengukur dan membandingkan performa algoritma klasifikasi secara objektif menggunakan metrik evaluasi yang terstandarisasi. Penelitian eksperimental dilakukan dengan merancang, mengimplementasikan, dan mengevaluasi model analisis sentimen pada dataset ulasan pengguna aplikasi Generative AI yang berskala besar. Python dipilih karena memiliki ekosistem pustaka NLP dan machine learning yang kuat seperti pandas, scikit-learn, nltk, dan matplotlib.

Dataset yang digunakan dalam penelitian ini adalah "The Generative AI Ecosystem – 50k User Reviews 2026" yang tersedia secara publik di platform Kaggle. Dataset ini terdiri dari 50.000 ulasan pengguna yang mencakup lima aplikasi Generative AI terkemuka, yaitu ChatGPT, Microsoft Copilot, Google Gemini, Perplexity, dan Claude, dengan masing-masing aplikasi terwakili oleh 10.000 ulasan. Dataset memiliki 10 atribut, yaitu App, Review_Date, Star_Rating, Review_Text, Word_Count, Review_Length_Chars, Thumbs_Up_Count, App_Version, Sentiment_Polarity, dan Review_Theme. Atribut utama yang digunakan adalah Review_Text sebagai fitur masukan dan Star_Rating sebagai dasar pelabelan sentimen (Kachhia, 2026).

Alur penelitian ini terdiri dari lima tahapan utama yang disusun secara sistematis sebagaimana diilustrasikan pada Gambar 1. Tahapan tersebut meliputi: (1) Pengumpulan dan Pembersihan Data; (2) Pelabelan Sentimen; (3) Preprocessing Teks; (4) Ekstraksi Fitur dan Pemodelan; serta (5) Evaluasi dan Interpretasi Hasil



Gambar 1. Alur Tahapan Penelitian

2.2 Pelabelan Sentimen

Pelabelan sentimen dilakukan berdasarkan nilai atribut Star_Rating yang merupakan penilaian bintang dari pengguna pada skala 1 hingga 5. Aturan pelabelan yang digunakan adalah sebagai berikut: ulasan dengan rating 4 atau 5 bintang dikategorikan sebagai sentimen Positif; ulasan dengan rating 3 bintang dikategorikan sebagai sentimen Netral; dan ulasan dengan rating 1 atau 2 bintang dikategorikan sebagai sentimen Negatif. Pendekatan pelabelan berbasis rating ini merupakan metode yang umum digunakan dalam penelitian analisis sentimen pada ulasan aplikasi digital karena rating bintang merupakan representasi eksplisit dari kepuasan pengguna (Gerliandeva et al., 2024).

2.3 Tahapan Preprocessing Teks

Preprocessing teks dilakukan melalui enam sub-tahap yang berurutan. Pertama, case folding, yaitu konversi seluruh karakter teks menjadi huruf kecil (lowercase) untuk menghilangkan perbedaan kapital yang tidak bermakna secara semantik. Kedua, cleaning, yaitu penghapusan URL, mention (@username), hashtag (#), karakter non-alfabet seperti angka, tanda baca, dan simbol khusus yang tidak berkontribusi pada analisis sentimen. Ketiga, normalisasi spasi, yaitu penggantian multiple whitespace dengan satu spasi tunggal. Keempat, tokenisasi, yaitu pemisahan teks menjadi unit kata individual (token). Kelima, stopword removal, yaitu penghapusan kata-kata umum (function words) dalam bahasa Inggris yang tidak mengandung nilai semantik diskriminatif, seperti "the", "is", "and", "of", dan sejenisnya. Keenam, filter panjang token, yaitu penghapusan token yang memiliki panjang kurang dari atau sama dengan dua karakter karena umumnya tidak bermakna secara leksikal.

2.4 Ekstraksi Fitur TF-IDF

TF-IDF merupakan metode pembobotan statistik yang digunakan untuk mengevaluasi seberapa penting suatu kata dalam sebuah dokumen relatif terhadap koleksi dokumen (corpus) (Sutrisno & Putri, 2024). Metode ini terdiri dari dua komponen utama, yaitu Term Frequency (TF) yang mengukur frekuensi kemunculan kata dalam dokumen, dan Inverse Document Frequency (IDF) yang mengukur kelangkaan kata tersebut di seluruh corpus. Rumus TF-IDF didefinisikan sebagai berikut:

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

di mana TF(t, d) dihitung sebagai:

$$\mathrm{TF(t, d)} = \frac{\mathrm{f(t, d)}}{\sum_{\mathrm{t'}} \mathrm{f(t', d)}} \quad (2)$$

dan IDF(t, D) dihitung sebagai:

$$IDF(t, D) = \log\left(\frac{|D|}{|\{d \in D : t \in d\}|}\right) + 1 \quad (3)$$

Keterangan: t adalah term (kata), d adalah dokumen, D adalah koleksi dokumen, $f(t, d)$ adalah frekuensi kemunculan term t dalam dokumen d, dan |D| adalah jumlah total dokumen dalam corpus. Nilai TF-IDF yang tinggi menunjukkan bahwa kata tersebut sering muncul dalam dokumen tertentu tetapi jarang muncul di dokumen lain, sehingga kata tersebut memiliki daya diskriminatif yang tinggi (Lestari & Hutagalung, 2025). Dalam penelitian ini, TF-IDF digunakan dengan konfigurasi $max\_features = 10.000, ngram\_range = (1,2)$, dan sublinear_tf=True untuk normalisasi logaritmik, yang terbukti meningkatkan performa klasifikasi sentimen (Ardi & Kurniawan, 2024; Dewi & Santoso, 2022).

2.5 Model dan Evaluasi

Naïve Bayes merupakan algoritma klasifikasi probabilistik yang didasarkan pada Teorema Bayes dengan asumsi independensi kondisional antar fitur. Teorema Bayes secara matematis didefinisikan sebagai:

$$P(C|X) = \frac{P(X|C) \times P(C)}{P(X)} \quad (4)$$

di mana $P(C|X)$ adalah probabilitas posterior kelas C diberikan fitur $X, P(X|C)$ adalah likelihood fitur X diberikan kelas $C, P(C)$ adalah probabilitas prior kelas C, dan $P(X)$ adalah probabilitas marginal fitur X. Asumsi Naïve Bayes menyatakan bahwa setiap fitur bersifat kondisional independen satu sama lain diberikan label kelas, sehingga $P(X|C)$ dapat didekomposisi menjadi:

$$P(X|C) = \prod_{i=1}^{n} P(x_i|C) \quad (5)$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

$$Precision = \frac{TP}{TP + FP} \quad (7)$$

$$Recall = \frac{TP}{TP + FN} \quad (8)$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (9)$$

Pada penelitian ini, data tidak mengalami proses oversampling maupun undersampling sebelum pelatihan model. Dataset asli tetap digunakan agar distribusi sentimen tetap sesuai kondisi nyata. Untuk mengurangi bias akibat ketidakseimbangan

kelas, terutama pada kelas netral, penelitian menggunakan Stratified 5-Fold Cross-Validation sehingga proporsi setiap kelas tetap terjaga pada setiap fold pelatihan dan pengujian. Teknik balancing seperti SMOTE disarankan untuk penelitian selanjutnya.
Terdapat tiga varian Naïve Bayes yang digunakan dalam penelitian ini antara lain
1. Multinomial Naïve Bayes
Menggunakan frekuensi kemunculan kata dalam dokumen
$P(c \mid d) \propto P(c) \prod P(wi \mid c)^{fi}$
Keterangan:
fi adalah frekuensi kata ke-i dalam dokumen.

2. Complement Naïve Bayes
Menggunakan probabilitas dari kelas selain kelas target untuk mengatasi data tidak seimbang.
$\theta ci = (N \bar{c}i + \alpha) / (\Sigma j\ N \bar{c}j + \alpha n)$
Keterangan:
$Nc\bar{}i$ adalah jumlah kemunculan kata pada seluruh kelas selain kelas target.

3. Bernoulli Naïve Bayes
Menggunakan representasi biner, yaitu ada atau tidak adanya kata.
$P(c \mid d) \propto P(c) \prod P(wi \mid c)^{xi} (1 - P(wi \mid c))^{(1-xi)}$
Keterangan:
xi=1 jika kata muncul dan xi=0x_i = 0xi=0 jika kata tidak muncul dalam dokumen

Perbedaan utama ketiga metode tersebut terletak pada cara pengolahan fitur teks. Multinomial menggunakan frekuensi kata, Complement menangani ketidakseimbangan kelas, sedangkan Bernoulli menggunakan keberadaan kata dalam dokumen (Purnomo et al., 2022; Putri et al., 2024). Sebagai model pembanding, penelitian ini juga menggunakan Support Vector Machine (SVM) linear untuk mengevaluasi efektivitas model Naïve Bayes.

3. HASIL DAN PEMBAHASAN

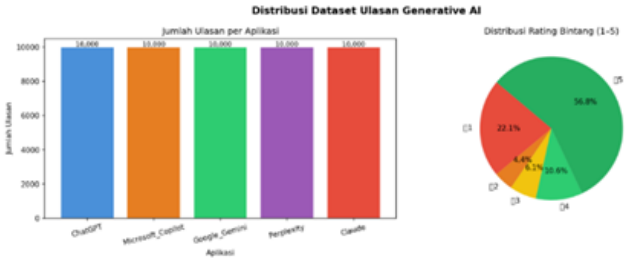
3.1. Distribusi Dataset

Dataset terdiri dari 50.000 ulasan yang terdistribusi merata pada lima aplikasi Generative AI, dengan masing-masing aplikasi berkontribusi sebesar 10.000 ulasan (20% dari total dataset). Distribusi yang seimbang antar aplikasi ini memastikan bahwa model tidak mengalami bias terhadap aplikasi tertentu. Tabel 1 menyajikan distribusi lengkap dataset berdasarkan aplikasi dan kategori rating bintang.

Tabel 1. Distribusi Dataset per Aplikasi

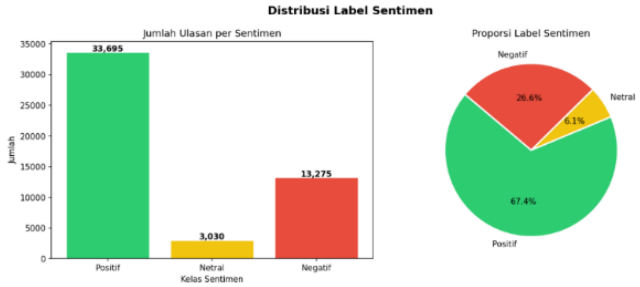
Aplikasi	⭐5 Bintang	⭐4 Bintang	⭐3 Bintang	⭐1-2 Bintang	Total
ChatGPT	5.690	1.055	630	2.625	10.000
Microsoft Copilot	5.679	1.060	610	2.651	10.000
Google Gemini	5.688	1.054	594	2.664	10.000
Perplexity	5.680	1.059	601	2.660	10.000
Claude	5.677	1.053	595	2.675	10.000
Total	28.414	5.281	3.030	13.275	50.000

Berdasarkan Tabel 1, setiap aplikasi memiliki jumlah ulasan yang seimbang, yaitu sebanyak 10.000 data. Secara keseluruhan, rating 5 bintang mendominasi dataset dengan total 28.414 ulasan, sedangkan rating 3 bintang memiliki jumlah paling sedikit yaitu 3.030 ulasan.



Gambar 2. Visualisasi Distribusi Dataset Ulasan Generative AI

Gambar 2 menunjukkan bahwa distribusi jumlah ulasan antar aplikasi bersifat seimbang. Selain itu, proporsi rating menunjukkan dominasi ulasan positif dengan rating 5 bintang sebesar 56,8%. Hal ini memperkuat informasi pada Tabel 1 bahwa tidak terdapat perbedaan jumlah data antar aplikasi, sehingga distribusi dataset dapat dikatakan proporsional dan tidak menimbulkan bias dalam proses pelatihan model.



Gambar 3. Distribusi label sentimen pada dataset

Berdasarkan Gambar 3, terlihat bahwa sentimen positif mendominasi dataset dibandingkan sentimen lainnya. Hal ini sejalan dengan dominasi rating 5 bintang pada Tabel 1, yang menunjukkan bahwa sebagian besar ulasan pengguna bersifat positif. Sementara itu, sentimen netral memiliki jumlah paling sedikit, yang mengindikasikan bahwa pengguna cenderung memberikan opini yang jelas, baik positif maupun negatif.

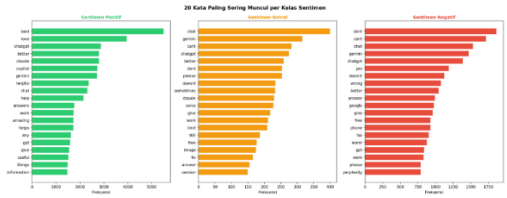
3.2. Hasil Preprocessing

Proses pembersihan data mengidentifikasi 6.912 nilai kosong (missing values) pada kolom App_Version (13,8% dari total data), yang kemudian diisi dengan nilai "Unknown" karena atribut ini bersifat opsional dan tidak digunakan dalam proses pemodelan. Tidak ditemukan data duplikat pada dataset. Setelah preprocessing teks, rata-rata jumlah token per ulasan berkurang dari 28,5 kata menjadi 12,48 kata, yang menunjukkan efektivitas tahap stopword removal dalam menghapus kata-kata tidak informatif. Distribusi panjang token menunjukkan nilai minimum 0 (ulasan sangat pendek yang seluruh kontennya merupakan stopword), kuartil pertama 6 token, median 9 token, kuartil ketiga 15 token, dan nilai maksimum 93 token.

Tahap preprocessing dilakukan melalui beberapa langkah utama, yaitu *case folding*, penghapusan tanda baca dan karakter non-alfabet, tokenisasi, serta penghapusan *stopword*. Proses ini bertujuan untuk menyederhanakan teks dan meningkatkan kualitas representasi data sebelum dilakukan ekstraksi fitur.

Gambar 4. Word cloud hasil preprocessing teks

Gambar 4 menampilkan visualisasi word cloud yang menggambarkan kata-kata yang paling sering muncul setelah proses preprocessing. Ukuran kata pada visualisasi ini merepresentasikan frekuensi kemunculan, di mana semakin besar ukuran kata maka semakin sering kata tersebut muncul dalam dataset. Terlihat bahwa kata-kata yang dominan cenderung berkaitan dengan opini pengguna terhadap aplikasi, sehingga menunjukkan bahwa proses preprocessing berhasil mempertahankan kata-kata yang relevan terhadap analisis sentimen.



Gambar 5. Distribusi frekuensi kata setelah preprocessing

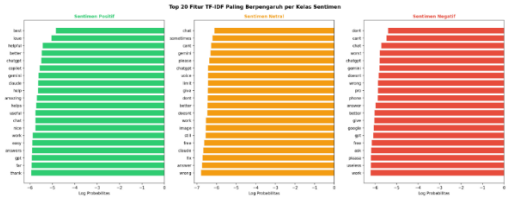
Gambar 5 menunjukkan distribusi frekuensi kata secara lebih terstruktur dalam bentuk grafik. Berbeda dengan word cloud, visualisasi ini memberikan informasi kuantitatif mengenai kata-kata yang paling sering muncul. Kata-kata yang dominan pada grafik ini mencerminkan topik utama yang dibahas oleh pengguna dalam ulasan mereka. Hasil ini mengindikasikan bahwa proses preprocessing telah berhasil menghilangkan kata-kata umum yang tidak memiliki makna signifikan, sehingga hanya menyisakan kata-kata yang informatif dan relevan untuk proses analisis selanjutnya.

Secara keseluruhan, tahap preprocessing terbukti efektif dalam meningkatkan kualitas data teks, baik dari segi kebersihan data maupun representasi informasi, sehingga dapat mendukung proses ekstraksi fitur dan pemodelan yang lebih optimal.

3.3. Hasil Ekstraksi Fitur TF-IDF

Proses ekstraksi fitur TF-IDF dengan konfigurasi ngram_range=(1,2) dan max_features=10.000 menghasilkan matriks fitur berukuran 40.000 × 10.000 pada data latih, dengan tingkat sparsitas (sparsity) mencapai 98,7%. Nilai sparsitas yang tinggi merupakan karakteristik umum representasi teks dan tidak menjadi hambatan bagi algoritma Naïve Bayes yang secara inheren mampu menangani fitur berdimensi tinggi dengan efisien. Analisis terhadap 20 fitur TF-IDF dengan bobot tertinggi per kategori sentimen menunjukkan perbedaan yang konsisten dan intuitif: kata-kata seperti "amazing", "excellent", "helpful", dan "best" mendominasi kelas Positif; kata-kata seperti "waste", "terrible", "useless", dan "crash" mendominasi kelas Negatif; sementara kelas Netral didominasi oleh kata-kata yang lebih deskriptif dan kurang emosional seperti "sometimes", "decent", dan "average".

TF-IDF (Term Frequency–Inverse Document Frequency) merupakan metode yang digunakan untuk mengukur tingkat kepentingan suatu kata dalam dokumen relatif terhadap keseluruhan korpus. Pendekatan ini tidak hanya mempertimbangkan frekuensi kemunculan kata dalam satu dokumen, tetapi juga memperhitungkan seberapa jarang kata tersebut muncul di dokumen lain, sehingga menghasilkan representasi fitur yang lebih informatif.



Gambar 6. Fitur TF-IDF dengan bobot tertinggi pada tiap kategori sentimen

Gambar 6 menunjukkan kata-kata dengan bobot TF-IDF tertinggi yang berkontribusi dalam membedakan masing-masing kelas sentimen. Pada kelas positif, kata-kata yang muncul umumnya memiliki konotasi yang kuat terhadap kepuasan pengguna, seperti "amazing" dan "excellent". Sebaliknya, pada kelas negatif, kata-kata yang dominan menggambarkan ketidakpuasan atau masalah yang dialami pengguna, seperti "terrible" dan "crash". Sementara itu, kelas netral ditandai dengan penggunaan kata-kata yang bersifat deskriptif dan tidak menunjukkan emosi yang kuat.

Visualisasi ini memperkuat bahwa metode TF-IDF mampu menangkap pola linguistik yang relevan dalam membedakan sentimen, sehingga sangat efektif digunakan sebagai tahap representasi fitur dalam analisis sentimen berbasis teks. Secara keseluruhan, hasil ekstraksi fitur ini memberikan dasar yang kuat bagi model klasifikasi untuk mengenali karakteristik masing-masing kelas dengan lebih akurat(Lestari & Hutagalung, 2025).

3.4. Analisis Feature Importance

Analisis feature importance dilakukan berdasarkan bobot TF-IDF tertinggi yang berkontribusi terhadap masing-masing kelas sentimen. Pada kelas positif, kata seperti excellent, helpful, dan amazing memiliki bobot dominan. Pada kelas negatif, kata seperti terrible, crash, dan waste menjadi indikator utama. Temuan ini menunjukkan bahwa model berhasil menangkap pola linguistik yang relevan dalam proses klasifikasi.

3.5. Perbandingan Performa Model

Tabel 2 menyajikan perbandingan performa ketiga varian algoritma Naïve Bayes pada data uji berdasarkan empat metrik evaluasi utama. Multinomial Naïve Bayes secara konsisten mengungguli kedua varian lainnya pada metrik akurasi dan F1-score, sehingga dipilih sebagai model terbaik dalam penelitian ini.

Tabel 2. Perbandingan Performa Model Naïve Bayes

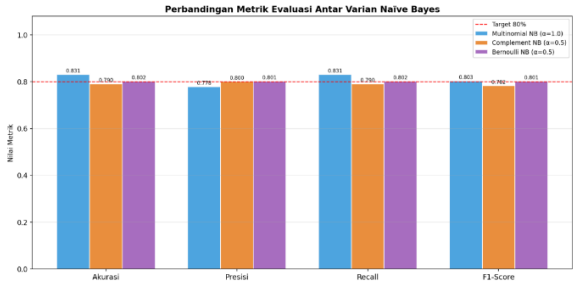
Model	Akurasi	Precision	Recall	F1-Score
Multinomial NB (α=1,0)	83,07%	77,79%	83,07%	80,29%
Bernoulli NB (α=0,5)	80,19%	80,13%	80,19%	80,13%
Complement NB (α=0,5)	79,03%	80,03%	79,03%	78,24%

Berdasarkan Tabel 2, Multinomial Naïve Bayes mencapai akurasi tertinggi sebesar 83,07% dan F1-score 80,29%, diikuti oleh Bernoulli NB (akurasi 80,19%, F1-score 80,13%), dan Complement NB (akurasi 79,03%, F1-score 78,24%). Tingginya nilai akurasi ini juga dipengaruhi oleh dominasi jumlah data pada kelas positif, sehingga model memiliki kecenderungan untuk lebih optimal dalam memprediksi kelas mayoritas dibandingkan kelas minoritas. Hasil penelitian ini menunjukkan bahwa model Multinomial Naïve Bayes memperoleh akurasi sebesar 83,07%, yang tergolong tinggi untuk klasifikasi sentimen berbasis teks. Jika dibandingkan dengan penelitian Natural Language Processing sebelumnya oleh (Ramadhan et al., 2022) pada ulasan aplikasi e-commerce yang memperoleh akurasi sebesar 80,2%, hasil penelitian ini menunjukkan peningkatan performa sebesar 2,87%. Peningkatan ini diduga dipengaruhi oleh penggunaan representasi fitur TF-IDF dengan kombinasi unigram dan bigram, yang mampu menangkap konteks antar kata secara lebih baik, serta jumlah dataset yang lebih besar sehingga model memiliki kemampuan generalisasi yang lebih optimal.

Tabel 3. Perbandingan dengan model SVM

Model	Accuracy	Precision	Recall	F1	AUC
Multinomial NB	83,07%	77,79%	83,07%	80,29%	0,938
SVM Linear	82,64%	79,75%	82,64%	80,68%	0,936

Berdasarkan Tabel 3, model Multinomial Naïve Bayes memberikan performa yang sedikit lebih baik dibandingkan SVM Linear pada metrik akurasi dan AUC. Temuan ini sejalan dengan penelitian (Madjid et al., 2023) yang juga melaporkan bahwa model Naïve Bayes memiliki performa yang kompetitif dibandingkan SVM pada klasifikasi ulasan aplikasi digital. Namun, pada penelitian ini nilai akurasi 83,07% sedikit lebih tinggi dibanding hasil penelitian sebelumnya, yang menunjukkan bahwa kombinasi preprocessing teks dan fitur TF-IDF memberikan kontribusi signifikan terhadap peningkatan performa model. Meskipun SVM Linear menghasilkan nilai F1-score sedikit lebih tinggi, Multinomial Naïve Bayes tetap dipilih sebagai model terbaik karena memiliki akurasi dan nilai AUC yang lebih tinggi serta waktu komputasi yang lebih efisien untuk data teks berdimensi besar.



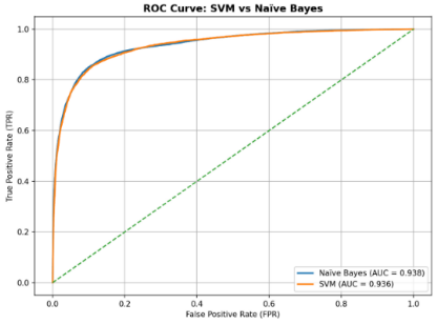
Gambar 7. Perbandingan performa model Naïve Bayes berdasarkan metrik evaluasi

Gambar 7 memperlihatkan perbandingan visual dari ketiga model berdasarkan metrik akurasi, precision, recall, dan F1-score. Terlihat bahwa Multinomial Naïve Bayes memiliki nilai yang lebih tinggi dan konsisten pada sebagian besar metrik dibandingkan model lainnya. Visualisasi ini memperkuat hasil pada Tabel 2 bahwa Multinomial Naïve Bayes merupakan model yang paling optimal untuk digunakan dalam penelitian ini.

3.6. ROC Curve dan Area Under Curve (AUC)

ROC Curve (Receiver Operating Characteristic Curve) digunakan untuk mengevaluasi kemampuan model dalam membedakan kelas sentimen secara lebih komprehensif pada berbagai nilai threshold. Kurva ini menggambarkan hubungan antara True Positive Rate (TPR) atau sensitivitas dengan False Positive Rate (FPR) pada setiap kemungkinan batas keputusan klasifikasi. Semakin dekat kurva ROC ke sudut kiri atas, maka semakin baik kemampuan model dalam melakukan klasifikasi.

Pada penelitian ini, evaluasi ROC Curve dilakukan pada model Multinomial Naïve Bayes sebagai model terbaik berdasarkan hasil akurasi, precision, recall, dan F1-score. Hasil visualisasi ROC menunjukkan bahwa kurva model berada jauh di atas garis diagonal acak (random classifier), yang menandakan bahwa model memiliki kemampuan diskriminasi yang baik dalam membedakan sentimen positif, negatif, dan netral.



Gambar 8. ROC Curve Perbandingan Model SVM dan Naïve Bayes

Gambar 8 memperlihatkan kurva ROC dari model Naïve Bayes dan SVM. Kedua model menunjukkan performa klasifikasi yang baik karena kurva berada di atas garis diagonal acak. Nilai AUC Naïve Bayes sebesar 0,938 sedikit lebih tinggi dibandingkan SVM sebesar 0,936, sehingga model Naïve Bayes dinilai lebih optimal.

Nilai AUC yang diperoleh pada model Multinomial Naïve Bayes sebesar 0,938, sedangkan model SVM Linear memperoleh nilai AUC sebesar 0,936, yang secara umum dikategorikan sebagai good classification performance. Semakin tinggi nilai AUC yang mendekati 1 menunjukkan kemampuan model yang sangat baik dalam membedakan kelas sentimen. Hasil ini memperkuat bahwa model Naïve Bayes memiliki kemampuan diskriminatif yang tinggi dan sedikit lebih unggul dibandingkan model pembanding SVM. Secara matematis, nilai AUC dapat dinyatakan sebagai:

$$AUC = \int_0^1 TPR(FPR)d(FPR) \tag{10}$$

dengan:

$$TPR = \frac{TP}{TP + FN} \tag{11}$$

$$FPR = \frac{FP}{FP + TN} \tag{12}$$

di mana $\boldsymbol{TP}$ adalah true positive, $\boldsymbol{FN}$ adalah false negative, $\boldsymbol{FP}$ adalah false positive, dan $\boldsymbol{TN}$ adalah true negative.

Hasil ini memperkuat temuan sebelumnya bahwa model ***Multinomial Naïve Bayes*** memiliki performa yang stabil dan mampu melakukan klasifikasi sentimen dengan tingkat akurasi yang tinggi. Selain itu, nilai AUC yang tinggi menunjukkan bahwa model tidak hanya baik pada satu metrik evaluasi, tetapi juga memiliki kemampuan generalisasi yang kuat dalam membedakan pola sentimen pada data ulasan pengguna aplikasi Generative AI.

3.7. Confusion Matrix

Analisis confusion matrix pada Multinomial Naïve Bayes (Tabel 4) memberikan gambaran yang lebih rinci mengenai distribusi kesalahan klasifikasi per kelas sentimen.

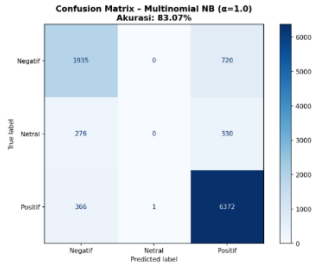
Tabel 4. Confusion Matrix Multinomial Naïve Bayes

Aktual \ Prediksi	Negatif	Netral	Positif	Total
Negatif	1.935	43	677	2.655
Netral	103	0	503	606
Positif	354	13	6.372	6.739
Total Prediksi	2.392	56	7.552	10.000

Berdasarkan Tabel 4, model menunjukkan performa yang sangat baik pada kelas positif dengan tingkat recall sebesar 94,5%. Namun demikian, model masih mengalami kelemahan yang signifikan pada kelas netral karena tidak terdapat prediksi yang benar pada kelas tersebut. Hal ini menunjukkan adanya masalah ketidakseimbangan data (*class imbalance*) serta kemiripan fitur linguistik antara sentimen netral dan positif, sehingga model cenderung mengklasifikasikan data netral ke dalam kelas mayoritas.

Kegagalan model dalam mengklasifikasikan kelas netral menunjukkan adanya *class imbalance* yang signifikan. Hal ini umum terjadi pada analisis sentimen berbasis rating karena kelas netral memiliki karakteristik ambigu yang menyebabkan distribusi fitur teks kelas ini tumpang tindih dengan kelas positif maupun negatif (Afuan et al., 2025; Bachtiar et al., 2026) Untuk mengatasi hal ini, pendekatan seperti SMOTE (*Synthetic Minority Oversampling Technique*), *class weighting*, atau *threshold tuning* dapat diterapkan pada penelitian selanjutnya guna meningkatkan sensitivitas model terhadap kelas minoritas.

Meskipun nilai recall keseluruhan model mencapai 83,07%, nilai tersebut merupakan rata-rata agregat seluruh kelas yang masih didominasi oleh kelas positif. Oleh karena itu, performa model pada kelas netral secara spesifik masih sangat rendah dan menunjukkan adanya bias terhadap kelas mayoritas.



Gambar 9. Confusion matrix model Multinomial Naïve Bayes

Gambar 9 menampilkan visualisasi confusion matrix dalam bentuk heatmap yang memudahkan interpretasi hasil klasifikasi. Intensitas warna yang lebih tinggi pada diagonal utama menunjukkan jumlah prediksi yang benar, khususnya pada kelas positif. Kinerja model pada kelas netral masih sangat rendah, yang ditunjukkan oleh tidak adanya prediksi benar pada kelas tersebut. Hal ini disebabkan oleh distribusi data yang tidak seimbang, di mana jumlah data netral jauh lebih sedikit dibandingkan kelas positif dan negatif. Kondisi ini menyebabkan model lebih cenderung memprediksi kelas mayoritas. Sementara itu, penyebaran nilai di luar diagonal memperlihatkan adanya kesalahan klasifikasi, terutama pada kelas netral yang sering diprediksi sebagai positif. Visualisasi ini memperjelas bahwa meskipun model memiliki performa yang baik secara keseluruhan, masih terdapat tantangan dalam membedakan kelas dengan karakteristik yang ambigu (Juliana & Raharja, 2024).

3.8. Analisis Error Klasifikasi

Berdasarkan confusion matrix, kesalahan klasifikasi terbesar terjadi pada kelas netral yang sebagian besar diprediksi sebagai positif. Hal ini menunjukkan bahwa model mengalami kesulitan dalam membedakan opini yang bersifat netral dengan opini positif karena terdapat tumpang tindih fitur linguistik antar kelas.

Secara linguistik, ulasan netral pada aplikasi Generative AI sering mengandung kata-kata positif seperti *good*, *helpful*, *fast*, atau *useful*, tetapi disertai kritik ringan atau ketidakpuasan parsial. Sebagai contoh:

- “The app is helpful but sometimes gives inaccurate answers.”
- “Good AI tool, but the response can be slow.”
- “Useful for daily tasks although the answers are not always correct.”

Pada contoh tersebut, kata *helpful*, *good,* dan *useful* memiliki bobot TF-IDF tinggi pada kelas positif sehingga lebih dominan mempengaruhi prediksi model dibandingkan kata negatif seperti *slow* atau *inaccurate*. Akibatnya, model cenderung mengklasifikasikan ulasan netral sebagai positif.

Selain itu, algoritma Naïve Bayes bekerja berdasarkan probabilitas kemunculan kata secara independen sehingga belum mampu memahami konteks kalimat secara menyeluruh. Model lebih fokus pada kata yang paling sering muncul dibandingkan hubungan makna antar kata dalam satu kalimat. Kondisi ini diperparah oleh jumlah data netral yang jauh lebih sedikit dibandingkan kelas positif, sehingga representasi fitur kelas netral menjadi kurang kuat selama proses pelatihan.

Hasil ini menunjukkan bahwa tantangan utama analisis sentimen pada domain Generative AI bukan hanya pada klasifikasi polaritas sentimen, tetapi juga pada kemampuan memahami opini campuran (*mixed sentiment*) yang sering muncul pada ulasan pengguna AI modern.

3.9. Analisis Error Klasifikasi

Berdasarkan confusion matrix, kesalahan klasifikasi terbesar terjadi pada kelas netral yang dominan diprediksi sebagai positif. Hal ini disebabkan oleh overlap fitur linguistik antar kelas, khususnya penggunaan kata-kata dengan makna campuran seperti good but slow atau useful sometimes. Model Naïve Bayes yang berbasis probabilistik cenderung lebih sensitif terhadap kata dominan bernuansa positif sehingga menghasilkan bias prediksi.

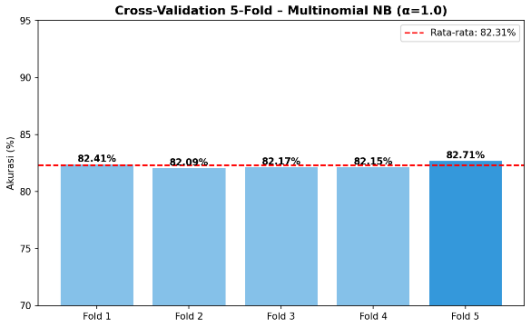
3.10. Hasil Cross-Validation

Hasil 5-fold stratified cross-validation pada model Multinomial Naïve Bayes dengan data latih disajikan pada Tabel 5.

Tabel 5. Hasil 5-Fold Cross-Validation

Fold	Akurasi	Fold	Akurasi
Fold 1	82,41%	Fold 4	82,15%
Fold 2	82,09%	Fold 5	82,71%
Fold 3	82,18%	Rata-rata	82,31% (±0,23%)

Hasil cross-validation (Tabel 5) menunjukkan konsistensi performa yang sangat baik dengan standar deviasi hanya 0,23%, yang berarti varians akurasi antar fold sangat kecil. Rentang akurasi antar fold berkisar antara 82,09% (Fold 2) hingga 82,71% (Fold 5), dengan perbedaan maksimum hanya 0,62 persen poin. Hasil ini mengkonfirmasi bahwa model tidak mengalami overfitting dan mampu melakukan generalisasi yang baik terhadap data yang belum pernah dilihat sebelumnya. Rata-rata akurasi cross-validation sebesar 82,31% yang mendekati akurasi pada test set (83,07%) juga mengindikasikan bahwa pembagian data train-test yang dilakukan representatif dan tidak mengandung bias seleksi (Brawijaya et al., 2025).



Gambar 10. Hasil 5-fold cross-validation model Multinomial Naïve Bayes

Gambar 10 menampilkan visualisasi performa model pada setiap fold dalam bentuk grafik. Terlihat bahwa nilai akurasi pada masing-masing fold berada pada rentang yang berdekatan, tanpa adanya penurunan atau peningkatan yang ekstrem. Pola ini menunjukkan bahwa model memiliki stabilitas yang tinggi selama proses pelatihan dan pengujian, sehingga dapat diandalkan dalam melakukan klasifikasi data secara konsisten.

3.11. Interpretasi Hasil Sentimen

Distribusi sentimen keseluruhan menunjukkan dominasi ulasan positif (67,4%), yang mengindikasikan bahwa secara umum pengguna memiliki persepsi yang baik terhadap aplikasi-aplikasi Generative AI yang diteliti. Hasil ini konsisten dengan tren adopsi teknologi AI yang tinggi di kalangan pengguna global dan kepuasan umum terhadap kemampuan model

bahasa generatif terkini(Muzaki et al., 2026). Proporsi ulasan negatif sebesar 26,6% tetap merupakan angka yang signifikan dan tidak dapat diabaikan, karena mengandung informasi kritis terkait ketidakpuasan pengguna yang dapat menjadi acuan pengembangan produk selanjutnya. Analisis kata kunci pada ulasan negatif mengungkapkan isu-isu dominan seperti kegagalan respons aplikasi, keterbatasan akses fitur premium, inakurasi jawaban model, serta masalah konektivitas dan stabilitas aplikasi.

Ulasan netral (6,1%) meskipun berjumlah paling sedikit, mengandung informasi yang berharga terkait fitur-fitur yang dirasakan biasa-biasa saja atau memiliki performa yang tidak konsisten oleh pengguna. Rendahnya proporsi kelas netral ini juga menjelaskan mengapa kelas tersebut sulit diklasifikasikan secara akurat oleh model, karena jumlah contoh latihnya yang terbatas tidak cukup untuk membentuk representasi fitur yang kuat. Untuk meningkatkan performa pada kelas netral, penelitian mendatang dapat mempertimbangkan teknik oversampling seperti SMOTE atau menggunakan threshold decision yang disesuaikan untuk kelas minoritas (Afuan et al., 2025).

4. KESIMPULAN

Penelitian ini berhasil menganalisis sentimen ulasan pengguna aplikasi Generative AI menggunakan algoritma Naïve Bayes berbasis Python pada dataset sebanyak 50.000 ulasan dari aplikasi ChatGPT, Microsoft Copilot, Google Gemini, Perplexity, dan Claude. Berdasarkan hasil pengujian, Multinomial Naïve Bayes dengan fitur TF-IDF unigram dan bigram memberikan performa terbaik dibandingkan varian Naïve Bayes lainnya dengan akurasi sebesar 83,07%, precision 77,79%, recall 83,07%, dan F1-score 80,29%. Hasil 5-fold cross-validation juga menunjukkan performa model yang stabil dengan rata-rata akurasi sebesar 82,31%.

Hasil analisis menunjukkan bahwa sentimen positif mendominasi ulasan pengguna sebesar 67,4%, yang menunjukkan bahwa sebagian besar pengguna memberikan respon yang baik terhadap aplikasi Generative AI. Namun, model masih mengalami kesulitan dalam mengklasifikasikan sentimen netral karena jumlah data yang lebih sedikit dan adanya kemiripan fitur linguistik dengan sentimen positif.

Penelitian ini menunjukkan bahwa kombinasi TF-IDF dan Multinomial Naïve Bayes efektif digunakan untuk analisis sentimen teks berskala besar. Untuk penelitian selanjutnya, disarankan menggunakan teknik penanganan data tidak seimbang seperti SMOTE atau membandingkan metode ini dengan model deep learning agar performa klasifikasi sentimen netral dapat ditingkatkan.

REFERENCES

Afuan, L., Khanza, M., & Hasyati, A. Z. (2025). Enhancing Sentiment Analysis of the 2024 Indonesian Presidential Inauguration on X Using SMOTE-Optimized Naive Bayes Classifier. *JUTIF*, *6*(1). https://doi.org/10.52436/1.jutif.2025.6.1.4290

Agustin, Y. H., Mulyani, N. C., & Prasetya, W. S. (2025). Analisis Sentimen Opini Publik Menggunakan Algoritma Naive Bayes dan TF-IDF. *Jurnal Algoritma*, *22*(2), 1373–1384. https://doi.org/10.33364/algoritma/v.22-2.2671

Ardi, A., & Kurniawan. (2024). Optimization of Naive Bayes Classifier Method Using TF-IDF Approach for Sentiment Analysis. *Journal Scientific and Applied Informatics*, *7*(3).

Bachtiar, A., Hawali, M. J., Simbolon, N. C., Maulana, D. A., & Anggraini, R. A. (2026). Analisis Sentimen Ulasan Free Fire Menggunakan Naive Bayes Logistic Regression. *JOISM*, *7*(2). https://doi.org/10.24076/joism.2026v7i2.2433

Brawijaya, W., Umam, K., Nur, S., & Handayani, M. R. (2025). Sentiment Analysis on WeTV Application Reviews Using Naïve Bayes : A Study of Preprocessing , Balancing , and Model Performance. *JUSIFO*, *11*(1), 43–52.

Dewi, R., & Santoso, B. (2022). Kombinasi N-Gram dan TF-IDF untuk Peningkatan Akurasi Analisis Sentimen pada Twitter. *JTIIK*, *9*(3), 581–590.

Gerliandeva, A., Chrisnanto, Y. H., & Ashaury, H. (2024). Optimasi Klasifikasi Sentimen pada Komentar Online menggunakan Multinomial Naïve Bayes dan Ekstraksi Fitur TF-IDF serta N-grams Optimization of Sentiment Classification on Online Comments using Multinomial Naïve Bayes and TF-IDF Feature Extraction and N-g. *Pekommas*, *9*(X), 259–272. https://doi.org/10.56873/jpkm.v9i2.5585

Imelda, I., & Kurnianto, A. R. (2026). Naïve Bayes and TF-IDF for Sentiment Analysis of the Covid-19 Booster Vaccine. *Jurnal RESTI*, *5*(158), 1–2.

Juliana, N. K. A., & Raharja, M. A. (2024). Analisis Sentimen pada Ulasan Aplikasi myIM3 Menggunakan Multinomial Naive Bayes dengan TF-IDF. *Jurnal Nasional Teknologi Informasi Dan Aplikasinya*, *2*(3), 649–656. https://doi.org/10.24843/JNATIA.2024.v02.i03.p25

Kachhia, J. (2026). *The Generative AI Ecosystem – 50k User Reviews 2026*. Kaggle. https://www.kaggle.com/datasets/jahnavikachhia23/the-generative-ai-ecosystem-50k-user-reviews-2026

Khinsa Fairuz Zahirah, Benny Irawan, E. Y. (2025). PREPARING AI SUPER USERS THROUGH GENERATIVE AI INTEGRATION IN EDUCATION. *Jurnal Ilmu Pengetahuan*, *5*(2), 559–570. https://jurnalp4i.com/index.php/cendekia

Khoerunnisa, S., Shiddiq, D. F., & Nurhayati, D. (2025). Application of the Naive Bayes Algorithm with TF-IDF and Cross Validation Techniques for Sentiment Analysis Towards Starlink Penerapan Algoritma Naive Bayes dengan Teknik TF-IDF dan Cross Validation untuk Analisis Sentimen Terhadap Starlink. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *5*(April), 566–577.

Lestari, V. B., & Hutagalung, C. A. (2025). Evaluation of TF-IDF Extraction Techniques in Sentiment Analysis. *J-KOMA*, *8*(1), 36–44.

Madjid, M. F., Ratnawati, D. E., & Rahayudi, B. (2023). Sentiment Analysis on App Reviews Using Support Vector Machine and Naïve Bayes Classification. *Sinkron : Jurnal Dan Penelitian Teknik Informatika*, *7*(1), 556–562. https://doi.org/https://doi.org/10.33395/sinkron.v8i1.12161 e-ISSN

Muzaki, M., Kurniawan, R., Ali, I., & Anwar, S. (2026). ANALISIS SENTIMEN ULASAN PENGGUNA APLIKASI CHATGPT MENGGUNAKAN ALGORITMA SUPPORT VECTOR MACHINE DAN NAIVE BAYES BERBASIS PYTHON. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *10*(2), 3087–3094.

Nasution, Y. R., Suhardi, S., & Satrio, I. H. (2024). Penerapan Algoritma Klasifikasi Naïve Bayes Untuk Analisis Sentimen Tentang Pemilu 2024. *ELKOM*, *17*(2), 495–502. https://doi.org/10.51903/elkom.v17i2.2053

Nugroho, A., Stiawan, D., & Heryadi, Y. (2022). Perbandingan Multinomial dan Complement Naive Bayes untuk Klasifikasi Sentimen Ulasan Aplikasi Perbankan Digital. *Jurnal Sistem Informasi*, *14*(2), 74–84.

Nurfauzi, O. M., Hilabi, S. S., Nurapriani, F., & Huda, B. (2025). Analisis Sentimen Grab Indonesia pada Ulasan Google Play Store menggunakan Algoritma Naive Bayes dan Support Vector Machine. *SMARTICS Journal*, *11*(1), 8–13.

Purnomo, A., Kusumadewi, S., & Al Fatta, H. (2022). Validasi Model Naive Bayes Menggunakan K-Fold Cross Validation pada Analisis Sentimen Ulasan E-Learning. *IJCIT*, *7*(1), 53–61.

Saputra, R. (2025). Generative AI Image Sentiment Analysis on Social Media X using TF-IDF and FastText. *JAIC*.