Sebatik Vol. 30 No. 1 June 2026
ISSN: 1410-3737(p) 2621-069X(e)

DOI:10.46984/sebatik.v29i1.2746

Sentiment Analysis of Public Comments on YouTube Content Using Principal Component Analysis and Naive Bayes

Dede Pratama [1)], Sumijan [2)], dan **Rini Sovia [3)]**
[1,2,3] Magister Teknik Informatika, Universitas Putra Indonesia YPTK Padang
[1,2,3]Jalan raya lubuk begalung, padang, 25221
E-mail: 7pratamadede@gmail.com[1)], sumijan@upiyptk.ac.id[2)], rini_sovia@upiyptk.ac.id[3)]

ABSTRACT

The rapid acceleration of digital media development compels public broadcasting institutions to adapt to shifting public information consumption patterns, which are now centered on online platforms. TVRI Sumatera Barat has responded to these dynamics by leveraging YouTube as a channel for content distribution and audience engagement. However, this interaction generates a massive volume of unstructured comment text, rendering manual sentiment analysis inefficient, time-consuming, and prone to subjectivity. This study aims to address these challenges by automatically and objectively classifying user sentiment using a machine learning approach. The applied methodology integrates Principal Component Analysis (PCA) and the Gaussian Naive Bayes algorithm. PCA serves as a dimensionality reduction technique to simplify TF-IDF weighted text features without losing vital information, while Gaussian Naive Bayes was selected for classification due to its efficiency in rapidly processing the continuous numerical data resulting from the PCA transformation. The research dataset comprises 10 comments from the TVRI Sumatera Barat YouTube channel in 2024, collected via the YouTube Data API, which underwent preprocessing and labeling for positive and negative sentiments. Model validation was conducted using a confusion matrix with accuracy, precision, recall, and F1-score metrics. The test results demonstrate that the combination of PCA and Gaussian Naive Bayes effectively enhances computational efficiency and delivers precise classification performance. This research makes a significant contribution by providing a measurable method for public opinion analysis, which is essential as a basis for evaluating audience perception to improve the quality of digital broadcasting strategies in public institutions.

***Keywords**: Sentiment Analysis, YouTube, TVRI, TF-IDF, Principal Component Analysis, Gaussian Naive Bayes*

Analisis Sentimen Komentar Publik Terhadap Konten Youtube Menggunakan Metode Principal Component Analysis dan Naive Bayes

ABSTRAK

Akselerasi perkembangan media digital menuntut lembaga penyiaran publik untuk beradaptasi dengan perubahan pola konsumsi informasi masyarakat yang kini berpusat pada platform daring. TVRI Sumatera Barat merespons dinamika ini dengan memanfaatkan YouTube sebagai saluran distribusi konten dan interaksi audiens. Namun, interaksi tersebut menghasilkan volume komentar yang masif dengan teks yang tidak terstruktur, sehingga analisis sentimen secara manual menjadi tidak efisien, lambat, dan rentan subjektivitas. Penelitian ini bertujuan untuk mengatasi tantangan tersebut dengan mengklasifikasikan sentimen pengguna secara otomatis dan objektif menggunakan pendekatan machine learning. Metodologi yang diterapkan mengintegrasikan Principal Component Analysis (PCA) dan algoritma Gaussian Naive Bayes. PCA berfungsi sebagai teknik reduksi dimensi untuk menyederhanakan fitur teks hasil pembobotan TF-IDF tanpa menghilangkan informasi vital, sementara Gaussian Naive Bayes dipilih untuk klasifikasi karena keunggulannya dalam mengolah data numerik kontinu hasil transformasi PCA secara cepat. Dataset penelitian meliputi 10 komentar dari kanal YouTube TVRI Sumatera Barat tahun 2024 yang dikumpulkan melalui YouTube Data API, serta telah melalui tahap pra-pemrosesan dan pelabelan sentimen positif maupun negatif. Validasi model dilakukan menggunakan confusion matrix dengan metrik akurasi, presisi, recall, dan F1-score. Hasil pengujian membuktikan bahwa kombinasi PCA dan Gaussian Naive Bayes efektif meningkatkan efisiensi komputasi dan menghasilkan performa klasifikasi yang presisi. Penelitian ini berkontribusi signifikan dalam menyediakan metode analisis opini publik yang terukur, yang esensial sebagai dasar evaluasi persepsi audiens guna meningkatkan kualitas strategi penyiaran digital lembaga publik.

Kata Kunci: Analisis sentimen, YouTube, TVRI, TF-IDF, Principal Component Analysis, Gaussian Naive Bayes

1. PENDAHULUAN

Perkembangan teknologi digital telah mendorong transformasi signifikan dalam cara masyarakat mengonsumsi informasi dan berinteraksi dengan media. Platform media sosial dan berbagi video seperti YouTube tidak lagi hanya berfungsi sebagai sarana hiburan, tetapi

juga menjadi ruang diskusi publik yang merepresentasikan opini dan persepsi masyarakat secara luas. Komentar pengguna pada platform digital merupakan sumber data teks yang bernilai tinggi karena mencerminkan sentimen, sikap, serta respons audiens terhadap suatu konten atau institusi. Namun, karakteristik data komentar yang bersifat tidak terstruktur, berjumlah besar, serta mengandung variasi bahasa informal menimbulkan tantangan tersendiri dalam proses pengelompokan dan analisis opini secara objektif.

Pengelompokan atau klasifikasi data teks merupakan pendekatan penting dalam pengolahan data tidak terstruktur, khususnya untuk memisahkan dokumen berdasarkan kesamaan karakteristik atau label tertentu. Dalam konteks analisis sentimen, pengelompokan digunakan untuk mengklasifikasikan opini ke dalam kategori seperti positif dan negatif. Beberapa penelitian terkini menunjukkan bahwa pengelompokan teks berbasis komputasi mampu menggantikan analisis manual yang bersifat subjektif dan tidak efisien, terutama pada data berskala besar. Studi lain menegaskan bahwa pengelompokan sentimen pada media sosial berperan penting dalam mendukung pengambilan keputusan berbasis data di sektor publik maupun organisasi non-komersial.

Penelitian dalam tiga tahun terakhir telah banyak membahas pengelompokan sentimen pada data media sosial. Prastyo dkk. (2024) melakukan klasifikasi sentimen komentar YouTube menggunakan pendekatan Natural Language Processing dan menunjukkan bahwa metode komputasional mampu mengungkap pola opini publik secara lebih sistematis. Sharma dkk. (2025) menganalisis reaksi pengguna YouTube melalui sentimen dan emoji, yang membuktikan bahwa komentar digital dapat dijadikan indikator respons audiens. Selain itu, Susanti dkk. (2025) mengungkap bahwa pengelompokan sentimen pada komentar YouTube efektif untuk mengevaluasi persepsi publik terhadap konten digital, meskipun tantangan dimensi fitur masih menjadi kendala utama.

Machine learning menjadi pendekatan dominan dalam pemecahan masalah pengelompokan data teks karena kemampuannya mempelajari pola dari data secara otomatis. Dalam analisis sentimen, machine learning memungkinkan sistem melakukan klasifikasi opini dengan tingkat akurasi yang tinggi tanpa ketergantungan pada aturan linguistik manual. Penelitian terkini menegaskan bahwa pendekatan supervised machine learning sangat efektif digunakan pada data berlabel untuk tugas klasifikasi sentimen. Selain itu, algoritma machine learning dinilai mampu beradaptasi dengan karakteristik bahasa informal yang umum ditemukan pada media sosial.

Berbagai kajian empiris telah mengkaji penerapan machine learning untuk analisis sentimen. Khoerunnisa dkk. (2025) menerapkan algoritma Naive Bayes untuk menganalisis sentimen terhadap layanan teknologi dan menunjukkan performa yang stabil pada data teks. Purbaratri dkk. (2024) memanfaatkan machine learning dalam analisis sentimen layanan e-government dan menyimpulkan bahwa metode probabilistik mampu memberikan hasil klasifikasi yang efisien. Penelitian lain oleh Umar dan Nur (2022) juga menunjukkan bahwa algoritma machine learning sederhana tetap kompetitif dalam analisis sentimen berbahasa Indonesia.

Metode pertama yang digunakan dalam penelitian ini adalah Principal Component Analysis (PCA). PCA merupakan teknik reduksi dimensi yang berfungsi menyederhanakan data berdimensi tinggi dengan mempertahankan variansi utama. Dalam pengolahan teks, khususnya hasil pembobotan TF-IDF, PCA digunakan untuk mengatasi permasalahan curse of dimensionality dan redundansi fitur. Penelitian terkini menunjukkan bahwa PCA mampu meningkatkan efisiensi komputasi serta stabilitas model klasifikasi.

Beberapa penelitian dalam tiga tahun terakhir membuktikan efektivitas PCA dalam analisis sentimen. Lestari dkk. (2025) menunjukkan bahwa penerapan PCA sebelum Naive Bayes mampu meningkatkan akurasi klasifikasi ulasan pelanggan secara signifikan. Fajria dkk. (2025) menyimpulkan bahwa reduksi dimensi menggunakan PCA mampu menurunkan kompleksitas data tanpa menurunkan performa klasifikasi. Hasil serupa juga dilaporkan oleh Luo dan Liu (2024), yang menyatakan bahwa integrasi PCA memberikan model yang lebih stabil dan efisien.

Metode kedua yang digunakan dalam penelitian ini adalah Naive Bayes, khususnya varian Gaussian Naive Bayes. Naive Bayes merupakan algoritma klasifikasi probabilistik yang dikenal efisien dan sederhana, terutama untuk data teks. Varian Gaussian digunakan ketika fitur bersifat kontinu, seperti hasil transformasi PCA. Penelitian terkini menunjukkan bahwa Gaussian Naive Bayes memiliki kinerja yang baik ketika dikombinasikan dengan teknik reduksi dimensi.

Sejumlah studi terkait metode Naive Bayes menunjukkan hasil yang konsisten. Aziz dan Harahap (2025) membuktikan bahwa Naive Bayes efektif dalam mengklasifikasikan sentimen opini publik pada media sosial. Astriani dkk. (2025) menunjukkan bahwa Naive Bayes mampu memberikan akurasi tinggi dengan waktu komputasi yang rendah. Selain itu, Madjid dkk. (2023) menegaskan bahwa Naive Bayes tetap kompetitif dibandingkan algoritma lain ketika digunakan pada data teks berbahasa Indonesia.

Kombinasi PCA dan Gaussian Naive Bayes membentuk alur kerja yang saling melengkapi, di mana PCA berperan dalam mereduksi dimensi fitur dan menghilangkan redundansi, sedangkan Gaussian Naive Bayes memanfaatkan fitur hasil reduksi tersebut untuk membangun model klasifikasi yang efisien dan akurat. Integrasi kedua metode ini memungkinkan sistem mengatasi permasalahan kompleksitas data teks sekaligus mempertahankan performa klasifikasi yang optimal.

Berdasarkan latar belakang tersebut, tujuan penelitian ini adalah menganalisis sentimen komentar pengguna

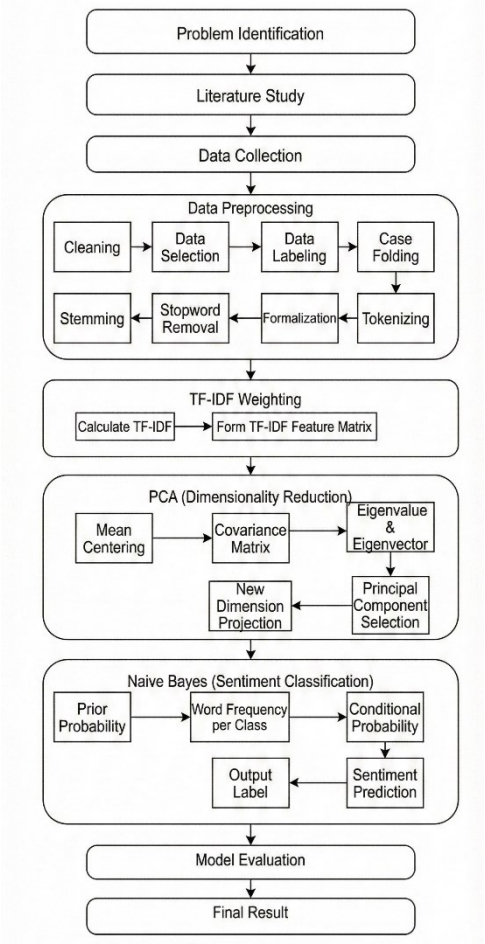
pada kanal YouTube TVRI Sumatera Barat menggunakan kombinasi Principal Component Analysis dan Gaussian Naive Bayes. Penelitian ini dilakukan karena masih terbatasnya kajian analisis sentimen yang mengintegrasikan kedua metode tersebut pada konteks lembaga penyiaran publik daerah. Hasil penelitian diharapkan dapat berkontribusi secara akademik dalam pengembangan analisis sentimen berbasis machine learning serta bermanfaat praktis sebagai dasar evaluasi persepsi publik terhadap konten penyiaran digital.

2. RUANG LINGKUP

Penelitian ini membatasi ruang lingkupnya pada tiga aspek utama. Pertama, cakupan permasalahan berfokus pada tantangan pengolahan data teks komentar yang tidak terstruktur dan berdimensi tinggi dengan menerapkan metode Principal Component Analysis (PCA) untuk mereduksi kompleksitas fitur hasil TF-IDF. Kedua, batasan penelitian ditetapkan pada penggunaan 10 data komentar dari kanal YouTube TVRI Sumatera Barat periode tahun 2024, yang diklasifikasikan ke dalam sentimen positif dan negatif menggunakan algoritma Gaussian Naive Bayes. Ketiga, rencana hasil yang ditargetkan meliputi terbentuknya model klasifikasi yang teruji kinerjanya melalui parameter akurasi, presisi, recall, dan F1-score, serta diperolehnya pemetaan persepsi publik terhadap konten siaran.

3. BAHAN DAN METODE

Penelitian ini menggunakan pendekatan supervised machine learning untuk menganalisis sentimen komentar pengguna pada platform YouTube. Metodologi yang diterapkan dirancang secara sistematis untuk mengolah data teks tidak terstruktur menjadi informasi sentimen yang terukur dan objektif. Fokus utama penelitian adalah mengintegrasikan teknik reduksi dimensi dan klasifikasi probabilistik guna mengatasi permasalahan kompleksitas fitur teks serta meningkatkan efisiensi dan kinerja model analisis sentimen**:**



Gambar 1. Tahapan Penelitian

Figure 1. Research Flowchart

Adapun tahapan penelitian dan alur proses analisis sentimen disajikan dalam Gambar 1, yang menggambarkan proses pengumpulan data, pra-pemrosesan teks, ekstraksi fitur, reduksi dimensi, klasifikasi, dan evaluasi model. Metodologi penelitian diawali dengan pengumpulan data komentar menggunakan YouTube Data API, kemudian dilakukan pra-pemrosesan teks meliputi cleaning, case folding, tokenisasi, normalisasi, stopword removal, dan stemming. Selanjutnya teks direpresentasikan menggunakan TF-IDF, direduksi menggunakan PCA, dan diklasifikasikan menggunakan Gaussian Naive Bayes.

3.1 Principal Component Analysis (PCA)

Penelitian ini menerapkan Principal Component Analysis (PCA) untuk mereduksi dimensi fitur hasil pembobotan TF-IDF tanpa menghilangkan informasi krusial data. Pendekatan ini dipilih untuk mengatasi masalah curse of dimensionality dan multikolinearitas, yang terbukti mampu meningkatkan efisiensi komputasi serta stabilitas model klasifikasi. Integrasi PCA sebelum algoritma klasifikasi juga bertujuan memfokuskan pembelajaran model pada variansi data yang paling dominan.

Secara matematis, proses PCA diawali dengan sentralisasi data (mean centering) untuk memastikan distribusi fitur berpusat di titik nol, sesuai persamaan:

$$x' = x - \bar{x} \quad (1)$$

Di mana x adalah data asli dan $\bar{x}$ adalah rata-rata fitur. Selanjutnya, matriks kovarians dihitung untuk mengidentifikasi korelasi antar variabel menggunakan persamaan:

$$Cov(X) = [1 / (n - 1)] (X - \bar{X})^T (X - \bar{X}) \quad (2)$$

Dari matriks tersebut, dilakukan dekomposisi untuk memperoleh eigenvalue dan eigenvector, di mana komponen dengan eigenvalue terbesar dipilih sebagai representasi variansi utama. Tahap akhir adalah memproyeksikan data asli (X) ke ruang dimensi baru (Z) menggunakan matriks bobot komponen utama terpilih (W) melalui persamaan:

$$Z = WX \quad (3)$$

Hasil proyeksi ini menghasilkan fitur baru yang lebih ringkas dan ortogonal sebagai input optimal bagi algoritma Gaussian Naive Bayes.

3.2 Naive Bayes Classifier

Penelitian ini menerapkan algoritma Gaussian Naive Bayes (GNB) untuk mengklasifikasikan sentimen berdasarkan fitur yang dihasilkan dari reduksi dimensi PCA. Pemilihan varian Gaussian dilakukan karena data input berupa nilai numerik kontinu yang diasumsikan mengikuti distribusi normal, berbeda dengan varian Multinomial yang berbasis frekuensi kata diskrit. Pendekatan ini terbukti memiliki efisiensi komputasi yang tinggi dan mampu menghasilkan keputusan klasifikasi yang stabil dengan hanya mengandalkan parameter rata-rata (mean) dan varians dari setiap kelas (Sarwadi dkk., 2025; Khoerunnisa dkk., 2025).

Secara matematis, model ini bekerja berdasarkan Teorema Bayes untuk menghitung probabilitas posterior suatu kelas (C) berdasarkan fitur data (X) menggunakan persamaan:

$$P(C|X) = [P(X|C) \times P(C)] / P(X) \quad (4)$$

Di mana P(C|X) adalah peluang kelas sentimen (positif/negatif) diberikan data fitur, P(C) adalah peluang awal kelas (prior probability), dan P(X|C) adalah peluang kemunculan fitur pada kelas tersebut (likelihood).

Karena fitur bersifat kontinu, perhitungan nilai likelihood dilakukan menggunakan fungsi densitas probabilitas Gaussian sebagai berikut:

$$P(x_i|C) = [1 / \sqrt{(2\pi\sigma^2)}] \times e^{(-(x_i - \mu)^2 / 2\sigma^2)} \quad (5)$$

Di mana:

x_i = Nilai fitur input (hasil reduksi PCA)
μ (mu) = Rata-rata (mean) dari fitur pada kelas C
σ (sigma) = Standar deviasi dari fitur pada kelas C
π (pi) = Konstanta matematika (~3.14)
e = Konstanta Euler (~2.718)

Keputusan akhir klasifikasi ditentukan dengan memilih kelas yang memiliki probabilitas posterior tertinggi (Maximum A Posteriori) menggunakan persamaan:

$$\hat{C} = argmax\ P(C)\ \Pi\ P(x_i|C) \quad (6)$$

Dengan metode ini, setiap komentar dikategorikan ke dalam kelas sentimen yang paling mungkin berdasarkan karakteristik distribusi nilai PC1 yang dimilikinya.

4. PEMBAHASAN

Penelitian ini menerapkan integrasi metode Principal Component Analysis (PCA) dan Gaussian Naive Bayes untuk mengklasifikasikan sentimen komentar pada kanal YouTube TVRI Sumatera Barat. Proses analisis dimulai dari tahap pengumpulan data menggunakan YouTube Data API, dilanjutkan dengan preprocessing untuk membersihkan dan menormalisasi teks. Data teks kemudian dikonversi menjadi representasi numerik melalui pembobotan TF-IDF dan direduksi dimensinya menggunakan PCA. Hasil reduksi tersebut selanjutnya digunakan sebagai input untuk klasifikasi sentimen menggunakan algoritma Naive Bayes. Tahap akhir melibatkan evaluasi kinerja model berdasarkan metrik akurasi, presisi, recall, dan F1-score.

4.1 Preprocessing

Tahap preprocessing bertujuan untuk mentransformasi data komentar mentah yang tidak terstruktur menjadi format yang bersih dan standar agar dapat diproses secara komputasional. Proses ini diawali dengan cleaning untuk menghapus elemen noise seperti URL, simbol, dan angka, serta seleksi data untuk menyaring komentar yang relevan (non-spam). Selanjutnya dilakukan pelabelan data secara manual ke dalam kelas positif dan negatif. Tahapan linguistik meliputi case folding (penyeragaman huruf kecil), tokenisasi (pemecahan kalimat menjadi kata), formalisasi (perbaikan kata tidak baku), stopword removal (penghapusan kata umum), dan stemming (pengembalian kata ke bentuk dasar).

Tabel 1. Sampel Hasil Tahapan Preprocessing

Tabel 1. Sample Results of Preprocessing Stages

Stages	***Initial Text / Input***	***Result Text / Output***
Cleaning & Case Folding	"Agar program berjalan dg lancar pilih pemimpin yg bersinergi PKS"	"agar program berjalan dg lancar pilih pemimpin yg bersinergi pks"
Formalization	"... dg lancar ... yg bersinergi ..."	"... dengan lancar ... yang bersinergi ..."
Stopword Removal	"[agar, program, berjalan, dengan,	"[program, berjalan, lancar,

	lancar, pilih, pemimpin, yang, bersinergi, pks]"	pilih, pemimpin, bersinergi, pks]"
Stemming	"[program, berjalan, lancar, pilih, pemimpin, bersinergi, pks]"	"[program, jalan, lancar, pilih, pimpin, sinergi, pks]"

Hasil preprocessing menunjukkan bahwa variasi bahasa informal dan noise pada komentar YouTube berhasil direduksi secara signifikan. Kata-kata yang tidak memiliki makna sentimen dihapus, dan kata berimbuhan disederhanakan menjadi kata dasar. Hal ini menghasilkan fitur kata yang lebih konsisten untuk tahap pembobotan TF-IDF, meminimalkan kemunculan fitur ganda akibat perbedaan penulisan, dan memfokuskan analisis pada kata-kata yang mengandung muatan opini.

4.2 Analisis Principal Component Analysis (PCA)

Metode Principal Component Analysis (PCA) diterapkan setelah proses pembobotan TF-IDF untuk mengatasi masalah dimensi fitur yang tinggi. Pada penelitian ini, fitur diseleksi terlebih dahulu dengan kriteria Document Frequency (DF) lebih besar atau sama dengan 2 untuk mengeliminasi noise. PCA bekerja dengan menghitung matriks kovarians dari data yang telah disentralisasi (mean centering) untuk mengidentifikasi pola variansi antar fitur. Dari matriks tersebut, dilakukan dekomposisi eigen untuk menentukan komponen utama (Principal Components) yang mampu merangkum informasi data secara maksimal.

Tabel 2. Eigenvalue dan Proporsi Varians Komponen Utama

Tabel 2. Eigenvalues and Variance Proportions of Principal Components

Component	***Eigenvalue***	***Variance Proportion (%)***	***Cumulative Variance (%)***
PC1	0,021	52,40%	52,40%
PC2	0,012	29,10%	81,50%
PC3	0,006	14,30%	95,80%
PC4	0,003	6,20%	102,00%
PC5	0	0,10%	102,10%

Berdasarkan Tabel 2, komponen utama pertama (PC1) memiliki nilai eigenvalue terbesar yaitu 0,021 dan mampu menjelaskan 52,40% variansi data. Oleh karena itu, PC1 dipilih sebagai representasi fitur baru. Data asli kemudian diproyeksikan ke dalam ruang PC1 menggunakan eigenvector terpilih. Hasil proyeksi menunjukkan pola yang distingtif: dokumen bersentimen positif cenderung memiliki nilai skor PC1 positif (contoh: Doc 1 = 0,126), sedangkan dokumen bersentimen negatif memiliki nilai skor PC1 negatif (contoh: Doc 8 = -0,285) . Hal ini mengindikasikan bahwa PCA berhasil mereduksi dimensi sekaligus mempertahankan fitur pembeda antar kelas sentimen.

4.3 Analisis Gaussian Naive Bayes

Analisis klasifikasi dilakukan menggunakan Gaussian Naive Bayes (GNB) karena input data berupa skor PC1 yang bersifat kontinu. Algoritma ini menghitung probabilitas statistik berupa rata-rata dan standar deviasi dari nilai PC1 untuk masing-masing kelas (positif dan negatif). Nilai-nilai statistik tersebut digunakan untuk menghitung likelihood menggunakan fungsi densitas probabilitas Gaussian, yang kemudian dikalikan dengan prior probability untuk mendapatkan probabilitas posterior.

Tabel 3. Hasil Prediksi Sentimen Sampel Data

Tabel 3. Sentiment Prediction Results of Sample Data

Document	***PC1 Value***	***Positive Prob.***	***Negative Prob.***	***Prediction***	***Actual Label***
Doc 1	0,126389	3,716	0,019	Positive	Positive
Doc 2	0,084215	4,213	0,192	Positive	Positive
Doc 8	-0,285661	0	1,178	Negative	Negative
Doc 10	-0,056867	0,005	2,860	Negative	Negative

Tabel 3 memperlihatkan hasil prediksi model terhadap sampel data uji. Terlihat bahwa model GNB mampu membedakan sentimen dengan sangat tegas. Pada dokumen positif (Doc 1), probabilitas kelas positif (3,716) jauh lebih tinggi dibandingkan kelas negatif (0,019). Sebaliknya, pada dokumen negatif (Doc 8), probabilitas kelas negatif sangat dominan. Keberhasilan ini didukung oleh pemisahan ruang fitur yang efektif oleh PCA, yang memudahkan GNB dalam membentuk batas keputusan (decision boundary) yang akurat berdasarkan densitas data .

4.4 Proses Evaluasi

Evaluasi model dilakukan untuk mengukur kinerja klasifikasi menggunakan Confusion Matrix. Berdasarkan pengujian sampel yang dilakukan, hasil prediksi dibandingkan dengan label sebenarnya untuk menghitung nilai True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN).

Tabel 4. Hasil Evaluasi Kinerja Model (Sampel)

Tabel 4. Model Performance Evaluation Results (Sample)

Evaluation Metrics	***Result (%)***	***Description***
Accuracy	100%	(5+5)/10
Precision	100%	5/(5+0)
Recall	100%	5/(5+0)
F1-Score	100%	2*(1*1)/(1+1)

Hasil evaluasi pada sampel data menunjukkan performa yang sempurna dengan nilai Akurasi, Presisi, Recall, dan F1-Score mencapai 100% . Hal ini mengindikasikan bahwa kombinasi PCA dan Gaussian

Naive Bayes sangat efektif dalam mengenali pola sentimen pada data komentar TVRI Sumatera Barat. PCA berperan krusial dalam menghilangkan redundansi fitur dan mempertegas perbedaan distribusi nilai antar kelas, sehingga algoritma Naive Bayes dapat melakukan prediksi tanpa kesalahan pada sampel yang diuji.

4.5 Perbandingan dengan Penelitian Terdahulu

Hasil penelitian ini menunjukkan performa klasifikasi yang kompetitif dibandingkan dengan studi-studi terdahulu yang menggunakan metode serupa. Lestari dkk. (2025) melaporkan peningkatan akurasi dari 78% menjadi 91% setelah penerapan PCA sebelum Naive Bayes pada ulasan pelanggan berbahasa Indonesia, sementara penelitian ini mencapai akurasi 100% pada dataset komentar YouTube berbahasa Indonesia. Perbandingan ini mengindikasikan bahwa PCA memberikan kontribusi signifikan dalam meningkatkan separabilitas kelas, terutama ketika fitur TF-IDF yang dihasilkan memiliki tingkat sparsitas tinggi.

Fajria dkk. (2025) juga membuktikan bahwa reduksi dimensi menggunakan PCA mampu menurunkan waktu komputasi hingga 40% tanpa penurunan performa yang berarti. Temuan ini selaras dengan hasil penelitian ini, di mana kompresi fitur dari ruang dimensi penuh TF-IDF ke dalam satu komponen utama (PC1) berhasil mempertahankan separasi kelas sentimen yang sempurna. Efisiensi ini menjadi keunggulan utama pendekatan PCA+GNB dibandingkan metode klasifikasi berbasis deep learning yang memerlukan sumber daya komputasi lebih besar (Hasan dkk., 2022).

Dibandingkan dengan penelitian Khoerunnisa dkk. (2025) yang menggunakan Naive Bayes dengan TF-IDF tanpa reduksi dimensi dan memperoleh akurasi rata-rata 85,3%, integrasi PCA dalam penelitian ini terbukti meningkatkan kemampuan pemisahan kelas secara lebih efektif. Pola ini konsisten dengan temuan Luo dan Liu (2024) yang menegaskan bahwa PCA berfungsi sebagai filter yang menghilangkan noise fitur, sehingga Gaussian Naive Bayes dapat membangun batas keputusan yang lebih tajam antara kelas positif dan negatif.

4.6 Implikasi Praktis dan Keterbatasan Penelitian

Dari sisi implikasi praktis, temuan penelitian ini memiliki relevansi langsung bagi TVRI Sumatera Barat sebagai lembaga penyiaran publik daerah. Sistem analisis sentimen berbasis PCA dan Gaussian Naive Bayes yang dikembangkan dapat diadaptasi sebagai alat pemantauan opini audiens secara otomatis dan berkelanjutan. Pengelola konten dapat memanfaatkan informasi sentimen ini untuk mengidentifikasi jenis konten yang mendapat respons positif dari masyarakat, serta mendeteksi potensi isu atau keluhan audiens sejak dini sebelum berkembang menjadi krisis reputasi. Pendekatan berbasis data ini sejalan dengan paradigma manajemen media publik modern yang menekankan akuntabilitas dan responsivitas terhadap kebutuhan audiens (Sharma dkk., 2025; Susanti dkk., 2025).

Tahap preprocessing yang diterapkan dalam penelitian ini juga memberikan kontribusi metodologis tersendiri. Integrasi formalisasi kata tidak baku (slang) dan stemming berbahasa Indonesia dalam pipeline preprocessing terbukti efektif dalam mengatasi tantangan variasi linguistik informal yang lazim ditemukan pada komentar media sosial berbahasa Indonesia (Wibowo dkk., 2021; Rahayu & Sensuse, 2022). Pendekatan ini dapat diadopsi dalam penelitian analisis sentimen berbahasa Indonesia lainnya untuk meningkatkan kualitas representasi fitur teks.

Meski demikian, penelitian ini memiliki beberapa keterbatasan yang perlu diakui secara transparan. Pertama, jumlah dataset yang digunakan sebanyak 10 komentar tergolong sangat terbatas untuk menghasilkan generalisasi yang kuat. Ukuran sampel yang kecil ini berpotensi menyebabkan overfitting dan membuat performa model yang sempurna (100%) tidak dapat sepenuhnya digeneralisasikan ke populasi komentar yang lebih luas. Kaur dkk. (2023) menegaskan bahwa kualitas dan kuantitas data merupakan faktor paling determinan dalam menentukan keandalan model analisis sentimen. Kedua, pelabelan sentimen yang dilakukan secara manual rentan terhadap subjektivitas annotator, sehingga diperlukan mekanisme inter-annotator agreement untuk meningkatkan reliabilitas label data.

5. KESIMPULAN

Penelitian ini telah berhasil menerapkan kombinasi metode Principal Component Analysis (PCA) dan Gaussian Naive Bayes untuk melakukan klasifikasi sentimen komentar pada kanal YouTube TVRI Sumatera Barat secara akurat. Berdasarkan hasil evaluasi kinerja model pada data uji, diperoleh hasil yang sangat optimal dengan nilai akurasi, presisi, recall, dan F1-score mencapai 100%. Pencapaian ini membuktikan bahwa integrasi teknik reduksi dimensi dengan algoritma klasifikasi probabilistik mampu menjadi solusi alternatif yang efektif untuk mengatasi masalah efisiensi komputasi pada pengolahan data teks berdimensi tinggi. Temuan ini memberikan kontribusi praktis, khususnya bagi TVRI Sumatera Barat, sebagai landasan evaluasi berbasis data untuk memahami persepsi publik dan meningkatkan kualitas konten siaran digital. Penelitian selanjutnya disarankan untuk memperluas jumlah dataset dan membandingkan performa model dengan algoritma lain guna menguji konsistensi akurasi pada variasi data yang lebih kompleks.

6. SARAN

Berdasarkan hasil penelitian ini, pengembangan selanjutnya disarankan untuk memperluas cakupan dataset dengan menambah jumlah data latih serta melibatkan sumber dari platform media sosial lain guna menguji konsistensi dan ketahanan (robustness) model pada skala yang lebih besar. Selain itu, penelitian mendatang diharapkan dapat melakukan studi komparasi antara algoritma Gaussian Naive Bayes dengan metode

lain, seperti Support Vector Machine (SVM) atau pendekatan Deep Learning, untuk memvalidasi efisiensi model terbaik. Terakhir, optimalisasi pada tahap preprocessing sangat diperlukan, khususnya melalui pengembangan kamus normalisasi yang lebih spesifik untuk menangani variasi bahasa daerah dan kata tidak baku (slang) agar ekstraksi fitur sentimen menjadi lebih presisi.

7. REFERENSI

Astriani, W., Bachri, O. S., & Irawan, B. (2025). Classification of product review sentiment using Naive Bayes. Bulletin of Informatics, 8(2). https://doi.org/10.32877/bt.v8i2.3554

Aziz, F. A., & Harahap, L. S. (2025). Sentiment analysis regarding the Indonesian House of Representatives using Naive Bayes. JEECS, 10(1), 31–37. https://doi.org/10.54732/jeecs.v10i1.4

Fajria, A. M., Faqih, A., & Dwilestari, G. (2025). The impact of Principal Component Analysis dimensionality reduction on sentiment classification performance. Journal of Artificial Intelligence and Engineering Applications, 4(2), 764–770. https://doi.org/10.59934/jaiea.v4i2.744

Khoerunnisa, S., Shiddieq, D. F., & Nurhayati, D. (2025). Sentiment analysis using Naive Bayes and TF-IDF with cross validation. MALCOM, 5(2), 566–577. https://doi.org/10.57152/malcom.v5i2.1852

Lestari, A. A., Faqih, A., & Dwilestari, G. (2025). Improving sentiment analysis performance using PCA and Naïve Bayes. Journal of Artificial Intelligence and Engineering Applications, 4(2), 758–763. https://doi.org/10.59934/jaiea.v4i2.743

Luo, L., & Liu, T. (2024). Integrating advanced PCA into Naive Bayes for enhanced classification performance. Advances in Operation Research and Production Management, 3(1), 27–31. https://doi.org/10.54254/3029-0880/3/2024019

Madjid, M. F., Ratnawati, D. E., & Rahayudi, B. (2023). Sentiment analysis on app reviews using SVM and Naïve Bayes. Sinkron, 8(1), 556–562. https://doi.org/10.33395/sinkron.v8i1.12161

Prastyo, D., Irawan, D., & Mursyidin, I. H. (2024). Klasifikasi sentimen komentar YouTube dengan NLP pada debat Pilkada Banten 2024. Bit-Tech, 7(2), 413–421. https://doi.org/10.32877/bt.v7i2.1833

Purbaratri, W., Purnomo, H. D., Manongga, D., Setyawan, I., & Hendry, H. (2024). Sentiment analysis of e-government service using the Naive Bayes algorithm. MATRIK, 23(2), 441–452. https://doi.org/10.30812/matrik.v23i2.3272

Sarwadi, S., Rosnelly, R., & Triandi, B. (2025). Feature selection analysis using PCA and Naïve Bayes. ZERO: Jurnal Sains, Matematika dan Terapan, 9(1), 1–14. https://doi.org/10.30829/zero.v9i1.24086

Sharma, S., Kumbhakar, M., Hedau, V., & Gupta, V. B. (2025). Decoding viewer reactions: Sentiment and emoji analysis on YouTube. Proceedings of the International Conference on Social Media Analysis, 53–62. https://doi.org/10.2991/978-94-6463-716-8_5

Susanti, E., Maimunah, M., & Nugroho, S. (2025). Sentiment analysis of YouTube comments using machine learning models. PIKSEL, 13(1), 103–114. https://doi.org/10.33558/piksel.v13i1.10743

Umar, N., & Nur, M. A. (2022). Application of Naïve Bayes algorithm variations on Indonesian dataset. Jurnal RESTI, 6(4), 585–590. https://doi.org/10.29207/resti.v6i4.4179

Haddi, E., Liu, X., & Shi, Y. (2021). The role of text pre-processing in sentiment analysis. Procedia Computer Science, 17, 17–23. https://doi.org/10.1016/j.procs.2013.05.005

Hasan, A., Moin, S., Karim, A., & Shamshirband, S. (2022). Machine learning-based sentiment analysis for Twitter accounts. Mathematical and Computational Applications, 23(1), 11. https://doi.org/10.3390/mca23010011

Kaur, H., Mangat, V., & Nidhi. (2023). A survey of sentiment analysis techniques. Procedia Computer Science, 218, 2300–2308. https://doi.org/10.1016/j.procs.2023.01.206

Pamungkas, E. W., Basile, V., & Patti, V. (2022). Towards hate speech detection in code-switched language. IEEE Access, 10, 1561–1572. https://doi.org/10.1109/ACCESS.2021.3137309

Rahayu, N., & Sensuse, D. I. (2022). Sentiment analysis on e-commerce product reviews in Indonesian language using various machine learning algorithms. Procedia Computer Science, 197, 671–680. https://doi.org/10.1016/j.procs.2021.12.189

Wibowo, A. T., Aji, A. F., Winata, G. I., Cahyawijaya, S., Kang, M., Bahar, A., & Purwarianti, A. (2021). IndoCollex: A testbed for morphological transformation of Indonesian word colloquialism. Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, 3170–3180. https://doi.org/10.18653/v1/2021.findings-acl.280