Jurnal Minfo Polgan (JMP)
Volume 15, Nomor 1, Maret 2026
DOI : https://doi.org/10.33395/jmp.v12i1.xxx
e-ISSN : 2797-3298
p-ISSN : 2089-9424

Analisis Sentimen Komentar Toxic pada Video Musik YouTube Menggunakan Metode Naive Bayes

[1*]Romindo, [2]Kevin Bastian Sirait, [3]Valentino Riffan, [4]Chailine Garcia Wijaya
Program Studi Sistem Informasi, Fakultas *Artificial Intelligence and Data Science,*
Universitas Pelita Harapan, Medan, Indonesia
**Korespondensi:* romindo@uph.edu

Submit : 02 Feb 2026 | **Diterima** : 13 Mar 2026 | **Terbit** : 16 Mar 2026

ABSTRACT

The YouTube video-sharing platform has evolved into a massive digital interaction space; however, alongside its rapid growth, the phenomenon of toxic or negative comments has become increasingly prevalent in popular content of a controversial nature. This study aims to analyze the sentiment patterns of toxic comments on the music video "Hozier - Take Me to Church" and to measure the performance of the Naive Bayes algorithm in automatically classifying sentiment. Data were collected using the YouTube API, yielding a total of 100,000 English-language comments, which after undergoing a cleaning process produced 51,348 valid data entries. The text preprocessing pipeline encompassed lowercasing, removal of URLs, punctuation, numbers, and emoticons, followed by contraction expansion, tokenization, stopword removal, and lemmatization using the Natural Language Toolkit (NLTK) library. The data were subsequently labeled into three sentiment classes positive, negative, and neutral before being classified using the Naive Bayes algorithm. Model performance evaluation was conducted using a confusion matrix, yielding an overall accuracy of 74%. Precision for the negative class reached 96%, neutral 97%, and positive 64%. Recall values for the negative class were 57%, neutral 49%, and positive 99%. Meanwhile, the F1-score for the negative class was 71%, neutral 65%, and positive 78%. Of the total processed data, 8,475 toxic (negative) comments were identified, indicating that the majority of the audience responded to the video positively. These findings demonstrate that the Naive Bayes algorithm possesses adequate capability in classifying sentiment from social media comments.

Keywords: *Sentiment Analysis, Naive Bayes, Toxic Comment, YouTube API, Confusion Matrix*

ABSTRAK

Platform berbagi video YouTube telah menjadi ruang interaksi digital yang masif, namun seiring pertumbuhannya, fenomena komentar toxic atau negatif semakin marak ditemukan pada konten-konten populer yang bersifat kontroversial. Penelitian ini bertujuan untuk menganalisis pola sentimen komentar toxic pada video musik "Hozier - *Take Me to Church*" serta mengukur performa algoritma Naive Bayes dalam melakukan klasifikasi sentimen secara otomatis. Data dikumpulkan menggunakan YouTube API dengan total 100.000 data komentar berbahasa Inggris, yang setelah melalui proses pembersihan menghasilkan 51.348 data valid. Proses text preprocessing mencakup lowercasing, penghapusan URL, tanda baca, angka, dan emotikon, diikuti tahap contraction, tokenization, penghapusan stopwords, serta lemmatization menggunakan *library Natural Language Toolkit* (NLTK). Data kemudian dilabeli ke dalam tiga kelas sentimen: positif, negatif, dan netral, sebelum diklasifikasi menggunakan algoritma Naive Bayes. Evaluasi performa model dilakukan dengan confusion matrix yang menghasilkan nilai akurasi sebesar 74%. Presisi untuk kelas negatif mencapai 96%, netral 97%, dan positif 64%. Nilai recall kelas negatif sebesar 57%, netral 49%, dan positif 99%. Sedangkan F1-score kelas negatif sebesar 71%, netral 65%, dan positif 78%. Dari total data yang diproses, ditemukan 8.475 komentar toxic (negatif), menunjukkan bahwa sebagian besar audiens merespons video tersebut secara positif. Hasil ini membuktikan bahwa algoritma Naive Bayes memiliki kemampuan yang memadai dalam klasifikasi sentimen komentar media sosial.

Kata Kunci: Analisis Sentimen, *Naive Bayes*, Komentar *Toxic*, YouTube API, *Confusion Matrix*

PENDAHULUAN

Perkembangan platform media sosial berbasis video, khususnya *YouTube*, telah mengubah lanskap komunikasi digital secara fundamental. Dengan lebih dari dua miliar pengguna aktif di seluruh dunia, *YouTube* tidak hanya berfungsi sebagai media hiburan, tetapi juga sebagai ruang ekspresi publik yang memungkinkan pengguna berinteraksi melalui kolom komentar secara bebas dan masif (Zhang et al., 2018). Namun, kebebasan berekspresi tersebut kerap disalahgunakan dalam bentuk komentar *toxic* yang mencakup ujaran kebencian (*hate speech*), diskriminasi, provokasi, dan konten negatif lainnya yang berpotensi merusak ekosistem digital (Fortuna & Nunes, 2019). Fenomena ini semakin mengkhawatirkan mengingat volume komentar pada konten-konten populer dapat mencapai ratusan ribu dalam rentang waktu yang sangat singkat, sehingga moderasi secara manual menjadi tidak efektif, tidak efisien, dan tidak skalabel (Rupapara et al., 2021). Oleh karena itu, pendekatan otomatis berbasis *machine learning* dan *Natural Language Processing* (NLP) menjadi solusi yang sangat dibutuhkan untuk mendeteksi, mengklasifikasikan, dan memahami pola sentimen komentar secara akurat dan real-time (Bordoloi & Biswas, 2023a). Dalam konteks ini, analisis sentimen (*sentiment analysis*) hadir sebagai teknik komputasional yang mampu mengekstraksi opini, perasaan, dan penilaian dari data teks tidak terstruktur, sehingga dapat mengkategorikan komentar ke dalam kelas-kelas seperti positif, negatif, dan netral (Medhat et al., 2014). Salah satu video musik yang memunculkan fenomena komentar kontroversial secara masif adalah "Take Me to Church" karya Hozier, yang mengangkat tema keagamaan dan isu sosial sensitif sehingga memancing beragam reaksi global mulai dari apresiasi artistik hingga komentar bernada kebencian. Kondisi ini menjadikan video tersebut sebagai objek penelitian yang sangat relevan untuk mengeksplorasi pola sentimen publik menggunakan pendekatan algoritmik. Penelitian ini mengajukan *Naive Bayes* sebagai metode klasifikasi utama karena terbukti efisien dalam pemrosesan data teks berdimensi tinggi dengan kompleksitas komputasi rendah, serta mampu memberikan performa yang kompetitif pada berbagai kasus klasifikasi teks (Talaat, 2023).

Berbagai penelitian terdahulu telah mengeksplorasi analisis sentimen dan deteksi komentar *toxic* pada platform media sosial dengan beragam pendekatan dan metode. Pertama, Wankhade et al. (Wankhade et al., 2022) menyajikan survei komprehensif mengenai metode, aplikasi, dan tantangan analisis sentimen, mencakup pendekatan berbasis leksikon, *machine learning* tradisional, hingga *deep learning*. Penelitian tersebut menegaskan bahwa tidak ada satu metode tunggal yang unggul di semua konteks, dan pemilihan metode harus mempertimbangkan karakteristik data serta tujuan analisis. Perbedaan dengan penelitian ini terletak pada fokus spesifik terhadap komentar *toxic* pada platform *YouTube* dengan skala data yang jauh lebih besar dan evaluasi menggunakan *confusion matrix* tiga kelas secara komprehensif. Kedua, Rupapara et al. (Rupapara et al., 2021) mengusulkan model RVVC untuk klasifikasi komentar *toxic* dengan menggabungkan teknik *SMOTE* untuk mengatasi ketidakseimbangan kelas (*class imbalance*) pada dataset teks, menghasilkan performa yang signifikan lebih baik dibandingkan model tanpa *oversampling*. Meskipun relevan, penelitian tersebut tidak menerapkan *Naive Bayes* sebagai metode utama dan tidak berfokus pada analisis sentimen tiga kelas secara serentak pada dataset komentar *YouTube* yang bersifat kontroversial. Ketiga, Raza Ali et al. (Ali et al., 2022) menerapkan *transfer learning* dengan model BERT untuk deteksi ujaran kebencian (*hate speech*) di *Twitter*, menghasilkan akurasi yang sangat tinggi. Kendati demikian, penelitian tersebut terbatas pada platform *Twitter* dengan karakteristik teks yang berbeda dari *YouTube*, serta tidak mencakup evaluasi performa pada klasifikasi tiga kelas dengan *confusion matrix* yang terperinci. Keempat, Dharil et al.(Patel et al., 2025) melakukan evaluasi ekstensif terhadap berbagai teknik *machine learning* untuk deteksi komentar *toxic* di media sosial, membandingkan *Naive Bayes*, *Support Vector Machine* (SVM), *Random Forest*, dan model berbasis *deep learning*. Hasil penelitian tersebut menunjukkan bahwa *Naive Bayes* tetap kompetitif pada dataset teks berukuran sedang, namun penelitian tersebut tidak mengaplikasikan metodologinya secara spesifik pada komentar *YouTube* berbahasa Inggris dengan konten kontroversial berskala besar. Kelima, Dhiaa A et al (Musleh et al., 2023) menganalisis komentar *YouTube* berbahasa Arab menggunakan pendekatan NLP berbasis *machine learning*, termasuk *Naive Bayes* dan SVM, dengan fokus pada evaluasi konten berbasis sentimen. Penelitian tersebut membuktikan efektivitas *Naive Bayes* untuk klasifikasi komentar *YouTube*, namun terbatas pada bahasa Arab dan tidak mengeksplorasi komentar *toxic* pada konten musik kontroversial berbahasa Inggris. Keenam, Bordoloi dan Biswas (Bordoloi & Biswas, 2023b) menyajikan survei menyeluruh mengenai kerangka desain, aplikasi,

dan prospek masa depan analisis sentimen, menyoroti pentingnya tahapan *preprocessing* teks dalam meningkatkan akurasi model. Temuan tersebut memperkuat justifikasi penelitian ini dalam mengembangkan *pipeline preprocessing* yang lebih lengkap, mencakup *contraction expansion* yang belum banyak diterapkan dalam penelitian-penelitian terdahulu.

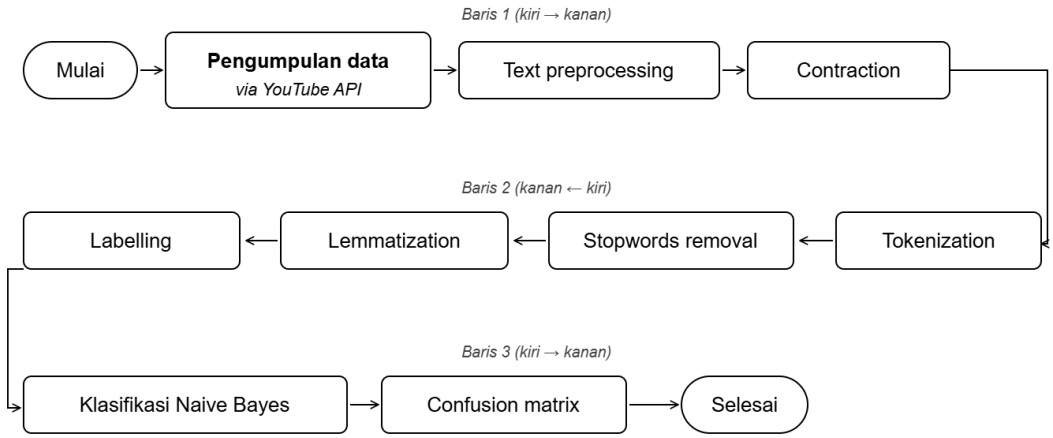
Berdasarkan tinjauan terhadap penelitian-penelitian terdahulu tersebut, teridentifikasi beberapa celah (*gap*) yang signifikan. Pertama, belum ada penelitian yang secara spesifik menganalisis komentar *toxic* pada video musik kontroversial di *YouTube* berbahasa Inggris dengan skala data melebihi 50.000 data valid menggunakan algoritma *Naive Bayes* (Mazari et al., 2024). Kedua, sebagian besar penelitian terdahulu hanya menggunakan klasifikasi dua kelas (positif/negatif), sementara penelitian ini menerapkan klasifikasi tiga kelas (positif, negatif, netral) secara serentak dengan evaluasi *confusion matrix* yang komprehensif (Pookpanich & Siriborvornratanakul, 2024). Ketiga, tahapan *contraction expansion* sebagai komponen krusial dalam *preprocessing* teks informal berbahasa Inggris masih jarang diimplementasikan secara eksplisit dalam penelitian analisis sentimen komentar *YouTube* (Hickman et al., 2022). Keempat, mayoritas penelitian terdahulu menggunakan dataset berukuran kecil hingga menengah (di bawah 10.000 data), sehingga generalisabilitas hasil evaluasi performa model menjadi terbatas (ONAN, 2021). Penelitian ini hadir untuk mengisi seluruh celah tersebut dengan mengintegrasikan *pipeline preprocessing* yang komprehensif meliputi *lowercasing*, penghapusan URL, tanda baca, angka, emotikon, *contraction expansion*, *tokenization*, *stopwords removal*, dan *lemmatization* menggunakan *library* NLTK pada dataset berskala besar yang dikumpulkan melalui *YouTube API* (Toraman et al., 2023).

Tujuan utama penelitian ini adalah: (1) menganalisis pola sentimen komentar *toxic* pada video musik "Take Me to Church" di *YouTube* menggunakan 100.000 data komentar berbahasa Inggris yang dikumpulkan melalui *YouTube API*; dan (2) mengukur performa algoritma *Naive Bayes* dalam klasifikasi sentimen tiga kelas melalui metrik akurasi, presisi, *recall*, dan *F1-score* berbasis *confusion matrix* (Ting, 2017). Penelitian ini diharapkan dapat memberikan kontribusi ilmiah dalam pengembangan sistem deteksi konten *toxic* secara otomatis pada platform video digital, sekaligus menjadi referensi metodologis bagi para peneliti NLP dalam menangani klasifikasi sentimen multikelas pada dataset komentar media sosial berskala besar (Cambria et al., 2017).

METODE PENELITIAN

Tahapan Penelitian

Penelitian ini menggunakan pendekatan kuantitatif berbasis *machine learning* dengan menerapkan algoritma *Naive Bayes* untuk klasifikasi sentimen komentar *toxic* pada video musik *YouTube* (Friedman et al., 1997). Secara keseluruhan, tahapan penelitian dirancang secara sistematis dan terstruktur sebagaimana digambarkan pada Gambar 1.



Gambar 1. Diagram Flow

Tahapan penelitian secara rinci diuraikan sebagai berikut (Thelwall, 2018; Toraman et al., 2023):

a. Pengumpulan Data Data komentar diperoleh langsung dari kolom komentar video musik "Take Me to Church" oleh Hozier melalui *YouTube API* dengan ID video *PVjiKRfKpPI*. Total data yang berhasil dikumpulkan adalah 100.000 komentar berbahasa Inggris. Setelah

melewati seluruh proses *preprocessing*, data yang valid dan dapat digunakan untuk analisis berjumlah 51.348 data.

b. Text Preprocessing Tahap ini merupakan proses pembersihan dan normalisasi data teks mentah agar siap diproses pada tahapan klasifikasi. Proses *preprocessing* terdiri dari delapan sub-tahapan berurutan, yaitu: *lowercasing*, penghapusan URL dan karakter khusus, penghapusan angka, penghapusan tanda baca, penghapusan emotikon, penghapusan spasi berlebih, penghapusan duplikat, dan penghapusan komentar non-Bahasa Inggris.
c. Contraction Expansion *Contraction* adalah proses mengurai bentuk singkat kata yang menggunakan tanda petik tunggal dalam Bahasa Inggris menjadi bentuk penuh dan formal. Contohnya: *"don't"* diubah menjadi *"do not"*, *"we'll"* menjadi *"we will"*, dan *"he's"* menjadi *"he is"*. Tahapan ini penting untuk memastikan konsistensi representasi kata dalam proses klasifikasi.
d. Tokenization *Tokenization* adalah proses pemecahan kalimat menjadi unit-unit kata (*token*) yang terpisah. Setiap kata dalam kalimat diperlakukan sebagai satu entitas data yang independen untuk memudahkan pemrosesan lebih lanjut . Proses ini dilakukan menggunakan *library* NLTK (*Natural Language Toolkit*) pada bahasa pemrograman *Python*.
e. Stopwords Removal *Stopwords* adalah kata-kata yang tidak memiliki makna signifikan dalam analisis teks, seperti kata penghubung dan partikel. Pada tahap ini, kata-kata tersebut dihapus dari token hasil *tokenization* untuk mereduksi dimensi data dan meningkatkan relevansi fitur yang tersisa .
f. Lemmatization *Lemmatization* merupakan proses transformasi kata ke bentuk dasarnya (*lemma*) dengan mempertimbangkan konteks gramatikal. Berbeda dengan *stemming*, *lemmatization* menghasilkan kata dasar yang valid secara linguistik. Proses ini dilakukan menggunakan *WordNetLemmatizer* dari *library* NLTK.
g. Labelling Data yang telah melalui seluruh tahap *preprocessing* kemudian dilabeli secara manual ke dalam tiga kelas sentimen: Positif, Negatif, dan Netral. Proses *labelling* ini berfungsi sebagai dasar pelatihan model *machine learning* untuk mengenali pola sentimen pada data baru.
h. Klasifikasi Naive Bayes Algoritma *Naive Bayes* diterapkan untuk melatih model dan mengklasifikasikan sentimen komentar. Proses ini mencakup pembagian data menjadi *data latih* dan *data uji*, dilanjutkan dengan evaluasi menggunakan *confusion matrix*.

Metode Naive Bayes dan Teknik Evaluasi

Algoritma *Naive Bayes* merupakan metode klasifikasi probabilistik yang didasarkan pada Teorema Bayes dengan asumsi independensi kondisional antar fitur (*naive assumption*) (Xu, 2018). Metode ini dipilih karena terbukti efektif dalam klasifikasi teks berdimensi tinggi dengan kompleksitas komputasi yang rendah. Rumus dasar yang digunakan adalah sebagai berikut (Luque et al., 2019; Wickramasinghe & Kalutarage, 2021):

$$P(C_k \mid x) = \frac{P(C_k) \cdot P(x|C_k)}{P(x)} \quad (1)$$

Untuk menghindari nilai probabilitas nol pada kata yang tidak muncul dalam data latih, diterapkan teknik *Laplace Smoothing* pada perhitungan *likelihood* setiap kata sebagai berikut:

$$P(w \mid C_k) = \frac{n_k + 1}{n_{C_k} + |V|} \quad (2)$$

Distribusi data latih (*training*) dan data uji (*testing*) dalam penelitian ini menggunakan rasio 80:20. Spesifikasi pembagian data disajikan pada Tabel 1.

Tabel 1. Distribusi Data Latih dan Data Uji

Kelas Sentimen	Data Latih (80%)	Data Uji (20%)	Total
Positif	18.472	4.618	23.09
Negatif	11.436	2.859	14.295
Netral	11.17	2.793	13.963
Total	41.078	10.27	51.348

Evaluasi performa model dilakukan menggunakan *confusion matrix* tiga kelas yang menghasilkan empat metrik utama, yaitu akurasi, presisi, *recall*, dan *F1-score*. Skema *confusion matrix* tiga kelas yang digunakan dalam penelitian ini disajikan pada Tabel 2.

Tabel 2. Skema *Confusion Matrix* Tiga Kelas

	Predicted Negatif	Predicted Netral	Predicted Positif
Actual Negatif	TP_Negatif	FN_Neg→Net	FN_Neg→Pos
Actual Netral	FP_Net→Neg	TP_Netral	FN_Net→Pos
Actual Positif	FP_Pos→Neg	FP_Pos→Net	TP_Positif

Rumus perhitungan setiap metrik evaluasi dijabarkan sebagai berikut (Romano et al., 2024; Rustam et al., 2021):
Akurasi:

$$Akurasi = \frac{TP_{Negatif}+TP_{Netral}+TP_{Positif}}{Total\ Data} \quad (3)$$

Presisi per kelas:

$$Presisi_{Negatif} = \frac{TP_{Negatif}}{TP_{Negatif}+FP_{Net\rightarrow Neg}+FP_{Pos\rightarrow Neg}} \quad (4)$$

Recall per kelas:

$$Recall_{Negatif} = \frac{TP_{Negatif}}{TP_{Negatif}+FN_{Neg\rightarrow Net}+FN_{Neg\rightarrow Pos}} \quad (5)$$

F1-Score per kelas:

$$F1\text{-}Score = 2 \times \frac{Presisi \times Recall}{Presisi+Recall} \quad (6)$$

Seluruh proses komputasi dalam penelitian ini diimplementasikan menggunakan bahasa pemrograman *Python* versi 3.x pada platform *Google Colaboratory*, dengan memanfaatkan *library* NLTK untuk *preprocessing*, *Scikit-learn* untuk implementasi *Naive Bayes*, serta *Matplotlib* dan *Seaborn* untuk visualisasi hasil *confusion matrix*. Spesifikasi *library* yang digunakan secara lengkap disajikan pada Tabel 3.

Tabel 3. Spesifikasi *Library* Python yang Digunakan

No.	*Library*	Fungsi	Versi
1	NLTK	*Tokenization*, *Stopwords*, *Lemmatization*	3.8
2	Scikit-learn	Klasifikasi *Naive Bayes*, Evaluasi	1.2
3	Pandas	Manajemen dan manipulasi data	1.5
4	Matplotlib	Visualisasi grafik dan plot	3.6
5	Google API Client	Pengambilan data via *YouTube API*	2

HASIL DAN PEMBAHASAN

Hasil Pengumpulan dan Preprocessing Data

Pengumpulan Data

Proses pengumpulan data dilakukan menggunakan *YouTube API* yang mengakses seluruh komentar pada video musik "Take Me to Church" oleh Hozier dengan ID video *PVjiKRfKpPI*. Total data mentah yang berhasil dikumpulkan adalah 100.000 komentar. Data mentah tersebut memuat informasi berupa nama pengguna (*author*), waktu publikasi (*published_at*), waktu pembaruan (*updated_at*), jumlah *like* (*like_count*), serta isi komentar (*text*). Contoh struktur data mentah hasil pengumpulan disajikan pada Tabel 4.

Tabel 4. Sampel Data Mentah Hasil *YouTube API*

No.	*Author*	*Published At*	*Like Count*	*Text*
1	@nightsky1328	5/19/2024	0	Anyone here in may 2024
2	@Lia-zl6sw	5/19/2024	0	What's wrong with gay they're human...
3	@user-lm2um	5/19/2024	0	My lover's got humor she's the giggle...
4	@EditorMexicano	5/18/2024	2	Corinthians and do you not know...
5	@TheAbdul307	8/11/2016	0	Why he's saying 'she' and 'her'?!

Setelah melalui seluruh rangkaian proses *text preprocessing*, data yang tersisa dan valid untuk digunakan dalam analisis berjumlah 51.348 data. Berkurangnya jumlah data dari 100.000 menjadi 51.348 disebabkan oleh beberapa faktor, antara lain:

a. Penghapusan komentar duplikat atau komentar yang berulang secara identik.
b. Penghapusan komentar yang tidak berbahasa Inggris karena model dikembangkan untuk komentar berbahasa Inggris.
c. Penghapusan komentar yang seluruh kontennya berupa URL, emotikon, atau simbol tanpa kandungan teks bermakna.
d. Penghapusan komentar kosong (*null*) setelah proses pembersihan dilakukan.

Hasil Text Preprocessing

Tahapan *text preprocessing* dilakukan secara berurutan dan bertahap untuk memastikan kualitas data yang optimal sebelum memasuki proses klasifikasi. Hasil dari setiap sub-tahapan *preprocessing* diuraikan sebagai berikut.

a. Lowercasing dan Penghapusan Karakter Tidak Relevan
   Pada tahap awal, seluruh teks komentar dikonversi menjadi huruf kecil untuk menyeragamkan representasi kata. Selanjutnya, dilakukan penghapusan URL, karakter khusus, angka, tanda baca, emotikon, dan spasi berlebih. Tabel 5 menyajikan perbandingan data sebelum dan sesudah tahap *preprocessing* awal.

Tabel 5. Perbandingan Data Sebelum dan Sesudah *Preprocessing* Awal

Komentar Asli	Hasil *Preprocessing*
He's making to many masterpiece..... But don't let em stop it	*he's making to many masterpiece but don't let em stop it*
This song is still gay	*this song is still gay*
Beautiful song, lyrics, and video. 🎵	*beautiful song lyrics and video*
Who's here in 2066 ??	*who's here in*
Really loved that last scene 🥹🥹	*really loved that last scene*

b. Contraction Expansion
   Tahapan *contraction* mengurai bentuk singkat kata Bahasa Inggris menjadi bentuk penuh dan formal. Proses ini sangat krusial mengingat karakteristik komentar *YouTube* yang bersifat informal dan banyak menggunakan singkatan. Tabel 6 menampilkan sampel hasil *contraction expansion*.

Tabel 6. Sampel Hasil *Contraction Expansion*

Sebelum *Contraction*	Sesudah *Contraction*
he's making to many masterpiece but don't let em stop it	*he is making to many masterpiece but do not let them stop it*
i am so confused is is a religious song or is it the opposite	*i am so confused is is a religious song or is it the opposite*
who's here in	*who is here in*
it's still gay	*it is still gay*

c. Tokenization
Proses *tokenization* memecah setiap kalimat menjadi unit-unit kata (*token*) individual. Hasil *tokenization* menggunakan *library* NLTK menghasilkan daftar *token* terpisah untuk setiap komentar. Contoh hasil *tokenization* disajikan pada Tabel 7.

Tabel 7. Sampel Hasil *Tokenization*

Komentar (Setelah *Contraction*)	Hasil *Tokenization*
he is making to many masterpiece but do not let them stop it	[he, is, making, to, many, masterpiece, but, do, not, let, them, stop, it]
this song is still gay	[this, song, is, still, gay]
beautiful song lyrics and video	[beautiful, song, lyrics, and, video]
really loved that last scene	[really, loved, that, last, scene]

a. Stopwords Removal
Kata-kata yang tidak memiliki kontribusi makna signifikan dihapus menggunakan daftar *stopwords* Bahasa Inggris dari NLTK. Kata-kata seperti *"is"*, *"the"*, *"and"*, *"to"*, *"it"*, dan sejenisnya dieliminasi dari daftar *token*. Tabel 8 menampilkan perbandingan sebelum dan sesudah penghapusan *stopwords*.

Tabel 8. Sampel Hasil *Stopwords Removal*

Sebelum *Stopwords Removal*	Sesudah *Stopwords Removal*
[he, is, making, to, many, masterpiece, but, do, not, let, them, stop, it]	[making, many, masterpiece, let, stop]
[this, song, is, still, gay]	[song, still, gay]
[beautiful, song, lyrics, and, video]	[beautiful, song, lyrics, video]
[really, loved, that, last, scene]	[really, loved, last, scene]

e. Lemmatization
Tahap terakhir *preprocessing* adalah *lemmatization*, yaitu transformasi setiap kata ke bentuk dasarnya (*lemma*) menggunakan *WordNetLemmatizer* dari NLTK. Hasil *lemmatization* memastikan kata-kata dengan variasi bentuk gramatikal direpresentasikan secara seragam. Tabel 9 menyajikan sampel hasil *lemmatization*.

Tabel 9. Sampel Hasil *Lemmatization*

Sebelum *Lemmatization*	Sesudah *Lemmatization*
[making, many, masterpiece, let, stop]	making many masterpiece let stop
[song, still, gay]	song still gay
[beautiful, song, lyrics, video]	beautiful song lyric video
[really, loved, last, scene]	really loved last scene
[reminds, song, relative, way, besides, I, agenda, tries, describe]	reminds song relative way besides I agenda try describe

Implementasi Klasifikasi Naive Bayes
Proses Labelling dan Pembobotan
Setelah seluruh tahap *preprocessing* selesai dilakukan, data dilabeli ke dalam tiga kelas sentimen: Positif, Negatif, dan Netral. Distribusi hasil *labelling* pada sampel data yang digunakan untuk demonstrasi perhitungan manual disajikan pada Tabel 10.

Tabel 10. Distribusi Label Sentimen Sampel Data

Komentar (Setelah *Lemmatization*)	Sentimen
making many masterpiece let stop	Positif
song still gay	Negatif

Komentar (Setelah *Lemmatization*)	Sentimen
beautiful song lyric video	Positif
so confused religious song opposite making fun christianity	Positif
whose idea stop caring kind something guide right principle	Positif
really loved last scene	Positif
reminds song relative way besides I agenda try describe	Netral

Pembobotan kata dilakukan menggunakan metode *Term Frequency* (TF), yakni menghitung total kemunculan setiap kosakata pada masing-masing kelas. Dari 37 kosakata unik yang diperoleh, distribusi pembobotan menunjukkan kelas Positif memiliki total frekuensi kata terbesar yaitu **30**, diikuti kelas Netral sebesar **9**, dan kelas Negatif sebesar **3**. Distribusi ini mencerminkan dominasi komentar bernada positif pada sampel data yang dianalisis.

Perhitungan Prior Probability dan Likelihood

Prior Probability dihitung berdasarkan proporsi jumlah dokumen pada setiap kelas terhadap total dokumen dalam data latih:

a. $P(Positif) = \frac{5}{7} = 0.7143$

b. $P(Negatif) = \frac{1}{7} = 0.1429$

c. $P(Netral) = \frac{1}{7} = 0.1429$

Perhitungan likelihood setiap kata menggunakan *Laplace Smoothing* dengan formula:

$P(w \mid C_k) = \frac{n_k+1}{n_{C_k}+|V|}$

dimana $|V| = 37$(total kosakata unik). Tabel 11 menyajikan contoh nilai *likelihood* untuk kata-kata kunci yang paling berpengaruh dalam klasifikasi.

Tabel 11. Nilai *Likelihood* Kata-Kata Kunci

Kata	P(w|Positif)	P(w|Netral)	P(w|Negatif)
making	0,0448	0,0217	0,0250
song	0,0448	0,0435	0,0500
gay	0,0149	0,0435	0,0250
masterpiece	0,0299	0,0217	0,0250
stop	0,0448	0,0217	0,0250
beautiful	0,0299	0,0217	0,0250
reminds	0,0149	0,0217	0,0500
agenda	0,0149	0,0217	0,0500

Klasifikasi Data Uji

Proses klasifikasi data uji dilakukan dengan mengalikan *prior probability* kelas dengan seluruh nilai *likelihood* kata-kata yang terdapat dalam kalimat uji. Hasil klasifikasi dua komentar uji diuraikan pada Tabel 12.

Tabel 12. Hasil Klasifikasi Data Uji

Komentar Uji	P(|Positif)	P(|Netral)	P(|Negatif)	Prediksi
"Stop play this song it is gay"	0,0000214	0,0000059	0,0000045	Positif
"This song is a masterpiece"	0,0009547	0,0001350	0,0001786	Positif

Pada komentar uji pertama *"Stop play this song it is gay"*, kata-kata yang masuk dalam kosakata data latih adalah *"stop"*, *"song"*, dan *"gay"*. Nilai probabilitas tertinggi diperoleh pada kelas Positif sebesar 0,0000214, sehingga komentar tersebut diklasifikasikan sebagai Positif. Pada komentar uji kedua *"This song is a masterpiece"*, kata yang relevan adalah *"song"* dan *"masterpiece"*, dengan nilai probabilitas Positif tertinggi sebesar 0,0009547, sehingga juga diklasifikasikan sebagai Positif.

Pembahasan

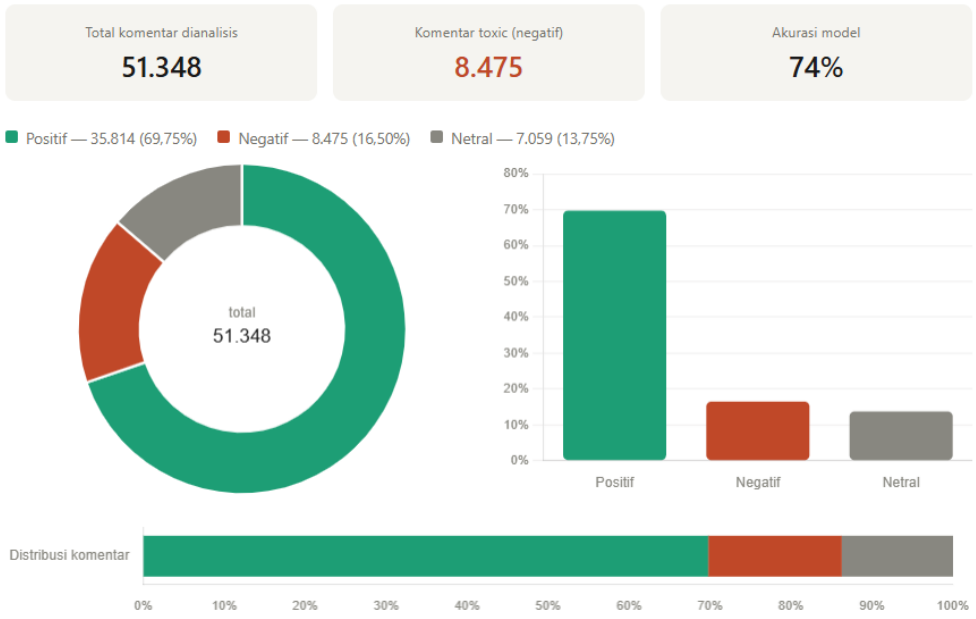
Hasil Analisis Sentimen Keseluruhan

Penerapan algoritma *Naive Bayes* pada 51.348 data komentar yang telah diproses menghasilkan distribusi prediksi sentimen sebagaimana disajikan pada Tabel 13 dan Gambar 2.

Tabel 13. Distribusi Hasil Prediksi Sentimen

Kelas Sentimen	Jumlah Prediksi	Persentase
Positif	35.814	69,75%
Negatif	8.475	16,50%
Netral	7.059	13,75%
Total	51.348	100%

Hasil analisis menunjukkan bahwa mayoritas komentar pada video musik "Take Me to Church" bersifat Positif dengan jumlah 35.814 komentar (69,75%), diikuti komentar Negatif sebanyak 8.475 komentar (16,50%), dan komentar Netral sebanyak 7.059 komentar (13,75%). Temuan ini mengindikasikan bahwa meskipun video musik tersebut bersifat kontroversial, sebagian besar audiens merespons secara positif, kemungkinan besar didorong oleh apresiasi terhadap kualitas musik dan pesan artistik yang disampaikan.



Gambar 2. Diagram Distribusi Hasil Prediksi Sentimen Komentar

Hasil Evaluasi Confusion Matrix

Evaluasi performa model *Naive Bayes* menggunakan *confusion matrix* tiga kelas menghasilkan data sebagaimana disajikan pada Tabel 14.

Tabel 14. Hasil *Confusion Matrix* Klasifikasi Sentimen

	Pred. Negatif	Pred. Netral	Pred. Positif
Actual Negatif	8.096	101	6.098
Actual Netral	237	6.859	6.867
Actual Positif	142	99	22.849

Berdasarkan Tabel 14, interpretasi hasil *confusion matrix* adalah sebagai berikut:

a. Kelas Negatif: sebanyak 8.096 sampel diprediksi dengan benar sebagai Negatif (*True Positive*), namun 101 sampel salah diprediksi sebagai Netral dan 6.098 sampel salah diprediksi sebagai Positif. Tingginya *misclassification* ke kelas Positif mengindikasikan bahwa model kesulitan membedakan komentar negatif yang menggunakan diksi ambigu.

b. Kelas Netral: sebanyak 6.859 sampel diprediksi dengan benar, tetapi 237 sampel salah diprediksi sebagai Negatif dan 6.867 sampel salah diprediksi sebagai Positif. Tingkat kesalahan klasifikasi kelas Netral yang tinggi mencerminkan tantangan inherent dalam mengenali komentar yang bersifat deskriptif tanpa muatan emosi yang jelas.
c. Kelas Positif: sebanyak 22.849 sampel diprediksi dengan benar, sementara hanya 142 sampel salah diprediksi sebagai Negatif dan 99 sampel salah diprediksi sebagai Netral. Kelas Positif memiliki performa terbaik karena didukung oleh jumlah data latih yang paling besar.

Hasil Perhitungan Metrik Evaluasi

Berdasarkan data *confusion matrix* pada Tabel 14, perhitungan keempat metrik evaluasi menghasilkan nilai sebagaimana disajikan pada Tabel 15.

Tabel 15. Rekap Hasil Metrik Evaluasi Model

Kelas	Presisi	*Recall*	*F1-Score*	*Support*
Negatif	0,96 (96%)	0,57 (57%)	0,71 (71%)	14.295
Netral	0,97 (97%)	0,49 (49%)	0,65 (65%)	13.963
Positif	0,64 (64%)	0,99 (99%)	0,78 (78%)	23.09
Akurasi			0,74 (74%)	51.348

Analisis terhadap hasil metrik evaluasi menghasilkan beberapa temuan penting:

a. Akurasi keseluruhan model sebesar 74% menunjukkan bahwa algoritma *Naive Bayes* mampu mengklasifikasikan komentar secara benar pada tiga dari empat kasus secara rata-rata, yang tergolong performa memadai untuk klasifikasi teks tiga kelas dengan skala data besar.
b. Presisi kelas Negatif (96%) dan Netral (97%) sangat tinggi, mengindikasikan bahwa ketika model memprediksi sebuah komentar sebagai Negatif atau Netral, prediksi tersebut hampir selalu benar. Sebaliknya, presisi kelas Positif (64%) yang lebih rendah menunjukkan adanya kecenderungan model untuk *over-predict* kelas Positif.
c. *Recall* kelas Positif yang sangat tinggi (99%) membuktikan bahwa model sangat sensitif dalam mendeteksi komentar positif dan hampir tidak melewatkan satu pun komentar positif dalam dataset. Namun, *recall* kelas Negatif (57%) dan Netral (49%) yang relatif rendah mengindikasikan bahwa masih banyak komentar negatif dan netral yang salah diklasifikasikan sebagai positif.
d. *F1-Score* tertinggi diraih kelas Positif (78%), diikuti Negatif (71%), dan Netral (65%). Ketimpangan ini sejalan dengan distribusi data yang tidak seimbang (*imbalanced*), di mana kelas Positif memiliki jumlah sampel terbesar sehingga model lebih terlatih untuk mengenali pola sentimen positif.

KESIMPULAN

Penelitian ini telah berhasil menganalisis sentimen komentar toxic pada video musik "Take Me to Church" oleh Hozier di platform YouTube menggunakan algoritma Naive Bayes dengan pipeline text preprocessing yang komprehensif. Dari total 100.000 data komentar yang dikumpulkan melalui YouTube API, sebanyak 51.348 data valid berhasil diproses setelah melewati tahapan lowercasing, penghapusan karakter tidak *relevan, contraction expansion, tokenization, stopwords removal,* dan *lemmatization* menggunakan library NLTK. Hasil klasifikasi menunjukkan bahwa sebanyak 35.814 komentar (69,75%) tergolong Positif, 8.475 komentar (16,50%) tergolong Negatif atau toxic, dan 7.059 komentar (13,75%) tergolong Netral. Temuan ini secara empiris membuktikan bahwa meskipun video musik tersebut mengandung konten yang kontroversial, mayoritas audiens global merespons konten tersebut dengan sentimen yang positif, didorong oleh apresiasi terhadap kualitas musikal dan nilai artistik yang terkandung di dalamnya. Evaluasi performa algoritma Naive Bayes menggunakan confusion matrix tiga kelas menghasilkan nilai akurasi keseluruhan sebesar 74%. Secara lebih rinci, presisi kelas Negatif mencapai 96%, presisi kelas Netral sebesar 97%, dan presisi kelas Positif sebesar 64%. Nilai recall kelas Negatif sebesar 57%, recall kelas Netral sebesar 49%, dan recall kelas

Positif sebesar 99%. Sementara F1-score kelas Negatif mencapai 71%, kelas Netral 65%, dan kelas Positif 78%. Hasil-hasil ini membuktikan bahwa algoritma Naive Bayes memiliki kemampuan yang memadai dalam menangani klasifikasi sentimen teks tiga kelas pada skala data yang besar, khususnya dalam mendeteksi komentar bernada positif dengan sensitivitas yang sangat tinggi. Meskipun demikian, penelitian ini memiliki beberapa keterbatasan yang perlu diakui. Pertama, rendahnya nilai recall pada kelas Negatif (57%) dan Netral (49%) mengindikasikan adanya ketidakseimbangan distribusi kelas (*class imbalance*) yang berdampak pada kemampuan model dalam mengenali komentar toxic dan netral secara optimal. Kedua, model yang dikembangkan hanya mencakup komentar berbahasa Inggris, sehingga komentar dalam bahasa lain yang turut muncul pada video tersebut tidak dapat dianalisis. Ketiga, proses labelling yang dilakukan secara manual berpotensi mengandung subjektivitas yang dapat memengaruhi kualitas data latih.

REFERENSI

Ali, R., Farooq, U., Arshad, U., Shahzad, W., & Beg, M. O. (2022). Hate speech detection on Twitter using transfer learning. *Computer Speech & Language*, *74*, 101365. https://doi.org/10.1016/j.csl.2022.101365

Bordoloi, M., & Biswas, S. K. (2023). Sentiment analysis: A survey on design framework, applications and future scopes. *Artificial Intelligence Review*, *56*(11), 12505–12560. https://doi.org/10.1007/s10462-023-10442-2

Bordoloi, M., & Biswas, S. K. (2023b). Sentiment analysis: A survey on design framework, applications and future scopes. *Artificial Intelligence Review*, *56*(11), 12505–12560. https://doi.org/10.1007/s10462-023-10442-2

Cambria, E., Das, D., Bandyopadhyay, S., & Feraco, A. (2017). *A Practical Guide to Sentiment Analysis* (Vol. 5). Springer International Publishing. https://doi.org/10.1007/978-3-319-55394-8

Fortuna, P., & Nunes, S. (2019). A Survey on Automatic Detection of Hate Speech in Text. *ACM Computing Surveys*, *51*(4), 1–30. https://doi.org/10.1145/3232676

Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian Network Classifiers. *Machine Learning*, *29*(2–3), 131–163. https://doi.org/10.1023/A:1007465528199

Hickman, L., Thapa, S., Tay, L., Cao, M., & Srinivasan, P. (2022). Text Preprocessing for Text Mining in Organizational Research: Review and Recommendations. *Organizational Research Methods*, *25*(1), 114–146. https://doi.org/10.1177/1094428120971683

Luque, A., Carrasco, A., Martín, A., & de las Heras, A. (2019). The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition*, *91*, 216–231. https://doi.org/10.1016/j.patcog.2019.02.023

Mazari, A. C., Boudoukhani, N., & Djeffal, A. (2024). BERT-based ensemble learning for multi-aspect hate speech detection. *Cluster Computing*, *27*(1), 325–339. https://doi.org/10.1007/s10586-022-03956-x

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, *5*(4), 1093–1113. https://doi.org/10.1016/j.asej.2014.04.011

Musleh, D. A., Alkhwaja, I., Alkhwaja, A., Alghamdi, M., Abahussain, H., Alfawaz, F., Min-Allah, N., & Abdulqader, M. M. (2023). Arabic Sentiment Analysis of YouTube Comments: NLP-Based Machine Learning Approaches for Content Evaluation. *Big Data and Cognitive Computing*, *7*(3), 127. https://doi.org/10.3390/bdcc7030127

ONAN, A. (2021). Sentiment analysis on massive open online course evaluations: A text mining and deep learning approach. *Computer Applications in Engineering Education*, *29*(3), 572–589. https://doi.org/10.1002/cae.22253

Patel, D., Pramanik, P. K. D., Suryawanshi, C., & Pareek, P. (2025). Detecting toxic comments on social media: an extensive evaluation of machine learning techniques. *Journal of Computational Social Science*, *8*(1), 20. https://doi.org/10.1007/s42001-024-00349-5

Pookpanich, P., & Siriborvornratanakul, T. (2024). Offensive language and hate speech detection using deep learning in football news live streaming chat on YouTube in Thailand.

Social Network Analysis and Mining, *14*(1), 18. https://doi.org/10.1007/s13278-023-01183-9

Romano, M., Zammarchi, G., & Conversano, C. (2024). Iterative threshold-based Naïve bayes classifier. *Statistical Methods & Applications*, *33*(1), 235–265. https://doi.org/10.1007/s10260-023-00721-1

Rupapara, V., Rustam, F., Shahzad, H. F., Mehmood, A., Ashraf, I., & Choi, G. S. (2021). Impact of SMOTE on Imbalanced Text Features for Toxic Comments Classification Using RVVC Model. *IEEE Access*, *9*, 78621–78634. https://doi.org/10.1109/ACCESS.2021.3083638

Rustam, F., Khalid, M., Aslam, W., Rupapara, V., Mehmood, A., & Choi, G. S. (2021). A performance comparison of supervised machine learning models for Covid-19 tweets sentiment analysis. *PLOS ONE*, *16*(2), e0245909. https://doi.org/10.1371/journal.pone.0245909

Talaat, A. S. (2023). Sentiment analysis classification system using hybrid BERT models. *Journal of Big Data*, *10*(1), 110. https://doi.org/10.1186/s40537-023-00781-w

Thelwall, M. (2018). Social media analytics for YouTube comments: potential and limitations. *International Journal of Social Research Methodology*, *21*(3), 303–316. https://doi.org/10.1080/13645579.2017.1381821

Ting, K. M. (2017). Confusion Matrix. In *Encyclopedia of Machine Learning and Data Mining* (pp. 260–260). Springer US. https://doi.org/10.1007/978-1-4899-7687-1_50

Toraman, C., Yilmaz, E. H., Şahinuç, F., & Ozcelik, O. (2023). Impact of Tokenization on Language Models: An Analysis for Turkish. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *22*(4), 1–21. https://doi.org/10.1145/3578707

Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, *55*(7), 5731–5780. https://doi.org/10.1007/s10462-022-10144-1

Wickramasinghe, I., & Kalutarage, H. (2021). Naive Bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation. *Soft Computing*, *25*(3), 2277–2293. https://doi.org/10.1007/s00500-020-05297-6

Xu, S. (2018). Bayesian Naïve Bayes classifiers to text classification. *Journal of Information Science*, *44*(1), 48–59. https://doi.org/10.1177/0165551516677946

Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. *WIREs Data Mining and Knowledge Discovery*, *8*(4). https://doi.org/10.1002/widm.1253