**Journal of Information System Research (JOSH)**
Volume 7, No. 4, Juli 2026, pp 1057−1066
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i4.10574

# Sentiment Analysis Kepercayaan Publik Terhadap Pertamina Pada Media Berita di YouTube

**Alrafi Syammajaya*, Daniel H. F. Manongga**

Fakultas Teknologi Informasi, Teknik Informatika, Universitas Kristen Satya Wacana, Salatiga
Jl. Dr. O. Notohamidjodjo Blotongan, Sidorejo, Kota Salatiga, Indonesia
Email: [1,*]rafipedia3@email.com, [2]fti.dekan@uksw.edu
Email Penulis Korespondensi: rafipedia3@email.com
Submitted: **02/07/2026**; Accepted: **18/07/2026**; Published: **20/07/2026**

**Abstrak**−Kepercayaan publik terhadap PT. Pertamina (Persero) tercermin dari opini masyarakat yang berkembang di media sosial, salah satunya melalui kolom komentar pada video berita di YouTube. Penelitian ini bertujuan untuk menganalisis sentimen publik terhadap Pertamina berdasarkan komentar pada video berita YouTube guna mengukur kecenderungan opini masyarakat, baik positif, negatif, maupun netral. Data dikumpulkan melalui teknik crawling menggunakan YouTube Data API, kemudian diolah melalui tahapan text preprocessing (cleaning, case folding, normalisasi kata, tokenization, dan stopword removal). Pelabelan sentimen dilakukan secara otomatis menggunakan metode VADER Sentiment ke dalam tiga kelas, yaitu positif, negatif, dan netral, dilanjutkan dengan ekstraksi fitur menggunakan TF-IDF. Proses klasifikasi dilakukan dengan membandingkan tiga varian algoritma Naïve Bayes, yaitu Gaussian, Multinomial, dan Bernoulli. Dari 56.868 komentar yang berhasil di-crawling, 22.527 komentar yang berhasil dikumpulkan, hasil pelabelan menunjukkan dominasi sentimen netral sebesar 95,89%, diikuti sentimen positif 2,86% dan sentimen negatif 1,25%. Hasil pengujian menunjukkan bahwa Naïve Bayes Multinomial memberikan performa terbaik dengan akurasi sekitar 95%, mengungguli Bernoulli Naïve Bayes (±92%) dan Gaussian Naïve Bayes (±79%), karena kesesuaiannya dengan representasi fitur TF-IDF yang berbasis frekuensi kata. Hasil penelitian ini diharapkan dapat menjadi bahan pertimbangan bagi Pertamina dan pemangku kebijakan dalam merumuskan strategi komunikasi publik yang lebih responsif.

**Kata Kunci:** Analisis Sentimen; Naïve Bayes; YouTube; VADER Sentiment; Pertamina

**Abstract**−Public trust in PT Pertamina (Persero) is reflected in public opinion expressed on social media, particularly through the comment sections of news videos on YouTube. This study aims to analyze public sentiment toward Pertamina based on comments posted on YouTube news videos in order to identify the distribution of positive, negative, and neutral opinions. The data were collected using the YouTube Data API through a web crawling process and subsequently processed using several text preprocessing techniques, including cleaning, case folding, word normalization, tokenization, and stopword removal. Sentiment labeling was performed automatically using the VADER Sentiment method, which classified the comments into three categories: positive, negative, and neutral. Feature extraction was then conducted using the Term Frequency–Inverse Document Frequency (TF-IDF) method. The classification process compared the performance of three Naïve Bayes variants, namely Gaussian Naïve Bayes, Multinomial Naïve Bayes, and Bernoulli Naïve Bayes. Of the 56,868 comments that were successfully crawled, 22,527 comments were successfully collected. The sentiment labeling results revealed that neutral sentiment dominated the dataset, accounting for 95.89% of all comments, followed by positive sentiment at 2.86% and negative sentiment at 1.25%. The experimental results demonstrated that Multinomial Naïve Bayes achieved the best performance with an accuracy of approximately 95%, outperforming Bernoulli Naïve Bayes (approximately 92%) and Gaussian Naïve Bayes (approximately 79%). This superior performance is attributed to the compatibility of the Multinomial Naïve Bayes algorithm with TF-IDF feature representation, which is based on word frequency. The findings of this study are expected to provide valuable insights for Pertamina and policymakers in formulating more responsive and effective public communication strategies.

**Keywords**: Sentiment Analysis; Naïve Bayes; YouTube; VADER Sentiment; Pertamina

## 1. PENDAHULUAN

Dalam era digital saat ini, media sosial telah menjadi platform utama bagi masyarakat untuk mengekspresikan opini dan persepsi terhadap kebijakan maupun kinerja perusahaan milik negara seperti PT. Pertamina (Persero). YouTube, sebagai salah satu platform berbagi video terbesar, tidak hanya menyajikan konten visual tetapi juga menyediakan ruang interaksi melalui kolom komentar, yang mencerminkan sentimen publik secara real-time. Kolom komentar pada video berita terkait Pertamina memuat berbagai keluhan, dukungan, maupun kritik masyarakat terhadap isu bahan bakar minyak, subsidi, tata kelola perusahaan, hingga tokoh-tokoh yang terlibat di dalamnya. Fenomena ini menunjukkan bahwa analisis sentimen terhadap komentar-komentar tersebut dapat menjadi indikator penting dalam memahami tingkat kepercayaan publik terhadap Pertamina, sekaligus menjadi masalah penelitian yang melatarbelakangi studi ini, yaitu belum terpetakannya secara sistematis bagaimana kecenderungan opini publik (positif, negatif, atau netral) terhadap Pertamina pada media berita di YouTube.

Sebagai metode pembanding, sebagian besar penelitian sejenis menggunakan algoritma Naïve Bayes karena kesederhanaan komputasi dan performanya yang baik pada klasifikasi teks berbahasa Indonesia. Naïve Bayes dipilih dan dibandingkan dengan pendekatan lain, misalnya Random Forest dan Support Vector Machine (SVM) [1], yang pada beberapa kasus data ulasan berbasis aspek terbukti mampu memberikan akurasi yang kompetitif atau bahkan lebih unggul dibandingkan Naïve Bayes. Perbandingan ini penting untuk menunjukkan posisi Naïve Bayes di antara metode klasifikasi teks lain, sekaligus menjadi dasar pemilihan tiga varian Naïve

Journal of Information System Research (JOSH)
Volume 7, No. 4, Juli 2026, pp 1057−1066
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i4.10574

Bayes (Gaussian, Multinomial, dan Bernoulli) yang diuji dalam penelitian ini untuk melihat varian mana yang paling sesuai dengan karakteristik data komentar YouTube.

Beberapa penelitian terdahulu telah mengkaji sentimen publik terhadap Pertamina maupun topik sejenis melalui media sosial. Analisis sentimen masyarakat terhadap kebijakan subsidi BBM berbasis QR Code menggunakan komentar YouTube dan algoritma Naive Bayes, memperoleh akurasi 79,40% dengan dominasi sentimen negatif akibat proses pendaftaran yang rumit [2]. Studi dalam [3] menganalisis 1.000 komentar video berita tentang penolakan serikat pekerja Pertamina terhadap penunjukan Ahok sebagai komisaris utama. Pengujian tersebut menghasilkan tingkat akurasi model yang sangat tinggi hingga mencapai 98%, walau penerapannya terbatas pada satu peristiwa spesifik. Pengembangan sistem berbasis web untuk menganalisis sentimen ulasan aplikasi MyPertamina menggunakan Random Forest [4]. Penelitian lain [5] berfokus pada isu kenaikan harga BBM dengan Gaussian Naïve Bayes pada 3.053 komentar dan akurasi 74%,Kajian sentimen komentar YouTube pada topik non-korporat, seperti tayangan berita televisi [6], ceramah keagamaan [7], dan isu Covid-19 [8], juga turut membuktikan konsistensi Naive Bayes dalam mengklasifikasikan opini publik dari komentar berbahasa Indonesia. Dari uraian tersebut terlihat gap (kesenjangan) penelitian yang jelas: penelitian-penelitian sebelumnya umumnya hanya berfokus pada satu isu spesifik atau satu peristiwa tunggal terkait Pertamina, sehingga belum mencerminkan keseluruhan persepsi publik yang mencakup beragam isu dan kebijakan perusahaan secara lebih menyeluruh. Penelitian ini berupaya mengisi kesenjangan tersebut dengan menganalisis komentar dari berbagai video berita Pertamina, bukan hanya satu topik atau peristiwa saja.

Berdasarkan uraian masalah dan gap penelitian di atas, penelitian ini bertujuan untuk: (1) Mengumpulkan data komentar pada berbagai video berita YouTube terkait pertamina, kemudian melabelinya secara otomatis ke dalam tiga kelas sentimen (positif, negatif, dan netral) menggunakan metode VADER Sentiment, (2) Membangun model klasifikasi sentimen menggunakan tiga varian algoritma Naïve Bayes (Gaussian, Multinomial, dan Bernoulli) yang dikombinasikan dengan ekstraksi fitur TF-IDF; dan (3) Membandingkan performa ketiga varian tersebut untuk menentukan algoritma yang paling sesuai dalam mengukur kecenderungan kepercayaan publik terhadap Pertamina secara menyeluruh.

Penelitian ini menggunakan beberapa landasan teori utama. Pertama, VADER (Valence Aware Dictionary and Sentiment Reasoner) digunakan sebagai metode pelabelan sentimen otomatis berbasis leksikon yang telah banyak diterapkan untuk klasifikasi opini ke dalam kategori positif, negatif, dan netral, sebagaimana diterapkan pada analisis komentar video wisata di YouTube [7]. TF-IDF (Term Frequency–Inverse Document Frequency) digunakan sebagai metode ekstraksi fitur yang merepresentasikan bobot kata berdasarkan frekuensi kemunculannya dalam dokumen, dan telah terbukti efektif dikombinasikan dengan model klasifikasi berbasis frekuensi kata pada analisis sentimen ulasan film [9] serta ulasan opini kenaikan harga BBM [5]. Adapun Naïve Bayes merupakan algoritma klasifikasi probabilistik berbasis Teorema Bayes dengan asumsi independensi antar fitur, yang dalam penelitian ini diuji dalam tiga varian, yaitu Gaussian (mengasumsikan distribusi normal pada fitur), Multinomial (dirancang untuk data frekuensi kata), dan Bernoulli (berbasis data biner ada/tidaknya kata), sebagaimana masing-masing telah diterapkan pada klasifikasi sentimen komentar YouTube terkait isu Covid-19 [8] dan kenaikan harga BBM [3], [4].

Kontribusi utama penelitian ini adalah menghadirkan analisis sentimen kepercayaan publik terhadap Pertamina yang lebih komprehensif karena mencakup berbagai video berita dan isu, tidak terbatas pada satu peristiwa atau kebijakan tunggal seperti penelitian-penelitian sebelumnya. Selain itu, penelitian ini memberikan perbandingan empiris tiga varian Naïve Bayes pada data komentar YouTube berbahasa Indonesia berskala besar (22.527 komentar), sehingga hasilnya dapat menjadi rujukan metodologis bagi penelitian analisis sentimen sejenis, sekaligus menjadi bahan pertimbangan praktis bagi Pertamina dan pemangku kebijakan dalam merumuskan strategi komunikasi publik yang lebih responsif.

## 2. METODOLOGI PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif untuk menganalisis sentimen kepercayaan publik terhadap Pertamina pada media berita di YouTube. Data yang digunakan dalam penelitian ini berupa data teks komentar yang diperoleh dari platform YouTube. Pengumpulan data dilakukan menggunakan teknik web scraping atau pengambilan data otomatis melalui Application Programming Interface (API). Studi dalam [10] menyatakan bahwa “data ulasan aplikasi MyPertamina diperoleh menggunakan teknik Web Scraping menggunakan API Google-Play-Scraper”. Pendekatan ini memungkinkan pengumpulan data dalam jumlah besar secara efisien serta memastikan data yang diperoleh relevan dengan topik penelitian.

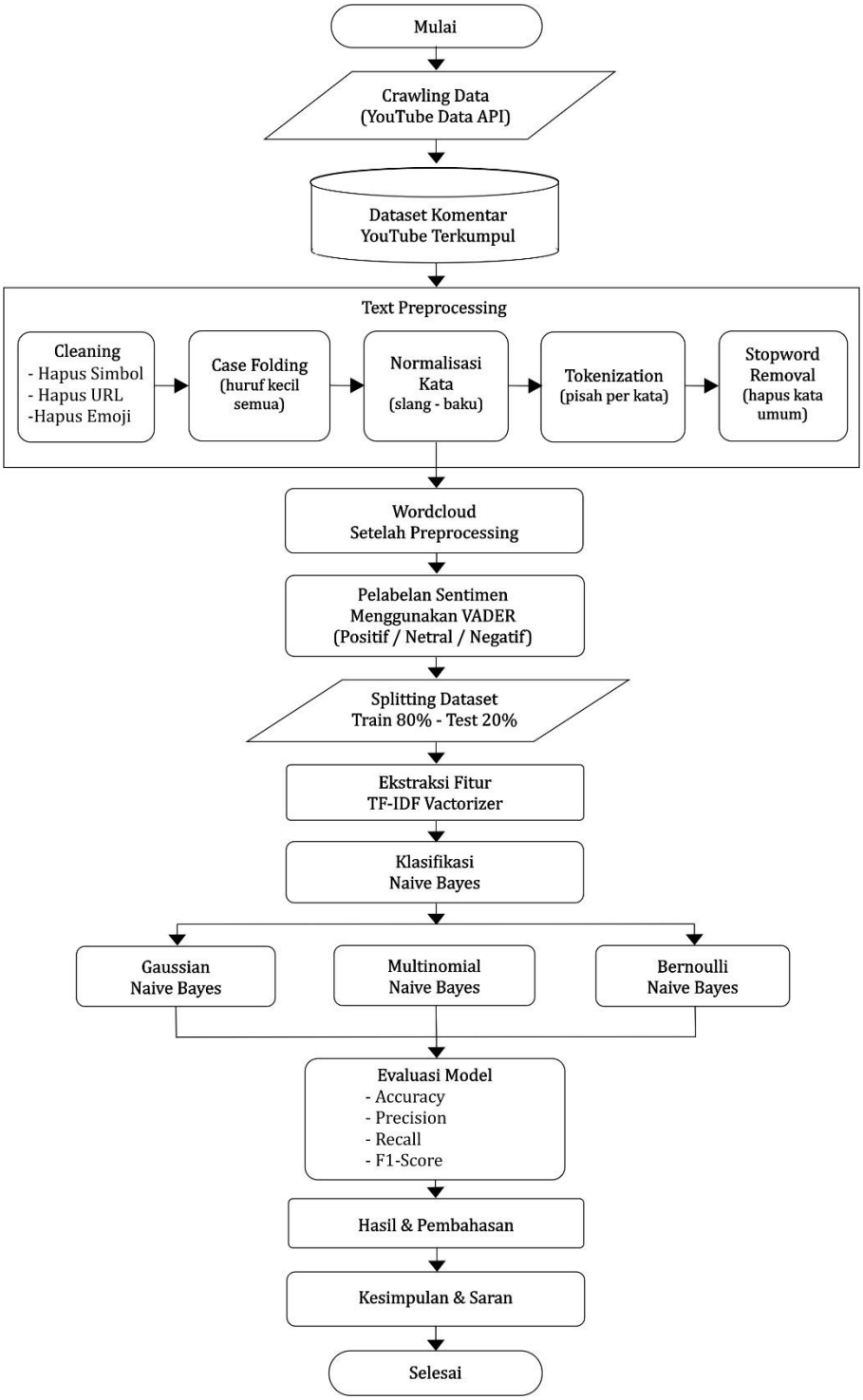
Alur penelitian secara keseluruhan mencakup seluruh tahapan penelitian, mulai dari proses crawling data, text preprocessing, pelabelan sentimen menggunakan VADER, splitting dataset, ekstraksi fitur TF-IDF, klasifikasi Naïve Bayes (Gaussian, Multinomial, dan Bernoulli), evaluasi model, hingga penarikan kesimpulan.

### 2.1 Tahapan Penelitian

Alur penelitian secara keseluruhan mencakup seluruh tahapan penelitian, mulai dari proses crawling data, text preprocessing, pelabelan sentimen menggunakan VADER, splitting dataset, ekstraksi fitur TF-IDF, klasifikasi

Journal of Information System Research (JOSH)
Volume 7, No. 4, Juli 2026, pp 1057−1066
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i4.10574

Naïve Bayes (Gaussian, Multinomial, dan Bernoulli), evaluasi model, hingga penarikan kesimpulan. Diagram alir proses penelitian disajikan pada Gambar 1.




**Gambar 1.** Diagram Alir Penelitian

### 2.1.1 Crawling Data

Tahap pertama dalam penelitian ini adalah crawling data, yaitu proses pengumpulan data komentar dari video berita yang berkaitan dengan Pertamina di platform YouTube. Proses crawling dilakukan menggunakan bahasa pemrograman Python dengan memanfaatkan YouTube Data API untuk memperoleh data komentar secara otomatis. Data yang berhasil dikumpulkan kemudian disimpan dalam bentuk dataset yang selanjutnya akan digunakan dalam proses analisis. Pada penelitian ini, proses crawling berhasil mengumpulkan sebanyak 56.868 komentar YouTube yang berasal dari beberapa video berita mengenai Pertamina. Selanjutnya, data tersebut diseleksi melalui beberapa tahapan, yaitu penghapusan komentar duplikat, komentar kosong, komentar yang tidak menggunakan bahasa Indonesia, serta komentar yang tidak relevan dengan topik penelitian. Setelah melalui proses

Journal of Information System Research (JOSH)
Volume 7, No. 4, Juli 2026, pp 1057−1066
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i4.10574

penyaringan tersebut, diperoleh 22.527 komentar yang memenuhi kriteria dan layak digunakan sebagai dataset pada tahapan analisis sentimen. Proses seleksi data ini bertujuan untuk meningkatkan kualitas dataset sehingga analisis yang dilakukan dapat menghasilkan informasi yang lebih akurat dan representatif terhadap opini publik mengenai Pertamina. Pengumpulan data secara otomatis melalui pemanfaatan YouTube Data API memungkinkan diperolehnya data dalam jumlah besar secara efisien sehingga mendukung pelaksanaan penelitian analisis sentimen.

Data mentah yang diperoleh dari hasil crawling memiliki struktur sebagai berikut: setiap baris mewakili satu komentar dengan kolom-kolom yang mencakup (1) comment_id: ID unik komentar, (2) author: nama pengguna yang memberikan komentar, (3) comment_text: isi teks komentar, (4) like_count: jumlah like pada komentar, (5) published_at: tanggal dan waktu komentar diunggah. Contoh dari struktur data mentah ini dapat dilihat pada Tabel 1 di bawah ini.

**Tabel 1.** Contoh Struktur Data Mentah Hasil Crawling

| id | author | comment_text | like_count | published_at |
|---|---|---|---|---|
| 0 | @ronimaulana73 | Mengapa harus inpor padahal kita punya banyak cadangan minyak | 0 | 2026-01-26T06:36:31Z |
| 1 | @JokoSupadno | Rakyat tidak pernah hutang saat beli bbm kenapa pertamina merugi | 0 | 2026-01-26T06:20:05Z |
| 2 | @Hadirieric | Data mana data | 0 | 2026-01-05T00:05:38Z |
| 3 | @spotweld9870 | Kenapa uud perampasan asset belum disahkan | 0 | 2025-10-15T15:32:05Z |

### 2.1.2 Text Preprocessing

Setelah proses pengumpulan data selesai, tahap selanjutnya adalah text preprocessing, yang bertujuan untuk membersihkan dan menyiapkan data teks agar siap digunakan dalam proses analisis sentimen. Tahap preprocessing merupakan langkah penting dalam text mining untuk meningkatkan kualitas data serta mengurangi kesalahan dalam proses analisis. Menurut penelitian [11], "Proses filtering dilakukan setelah tahap pra-pemrosesan dalam text mining, dan merupakan salah satu langkah penting untuk mempersiapkan data sebelum masuk ke tahap selanjutnya".

Tahapan text preprocessing juga banyak diterapkan pada penelitian klasifikasi teks menggunakan algoritma Naïve Bayes karena mampu meningkatkan kualitas data sebelum proses ekstraksi fitur dan klasifikasi dilakukan. Penerapan tahapan tersebut terbukti membantu menghasilkan representasi data yang lebih baik sehingga dapat meningkatkan performa model klasifikasi [12].

Selain diterapkan pada analisis sentimen, tahapan text preprocessing juga digunakan pada penelitian klasifikasi berbasis Naïve Bayes di bidang prediksi penyebaran Covid-19. Tahapan tersebut bertujuan meningkatkan kualitas data sehingga proses pembentukan model menjadi lebih optimal [13]. Tahapan preprocessing yang dilakukan dalam penelitian ini meliputi beberapa proses sebagai berikut:

#### 2.1.2.1 Cleaning

Proses cleaning merupakan proses pembersihan data teks dari karakter yang tidak diperlukan, seperti tanda baca, angka, simbol, URL, emoji, dan karakter khusus lainnya. Proses ini bertujuan untuk menghasilkan data teks yang lebih bersih dan terstruktur sehingga dapat diproses dengan lebih akurat pada tahap analisis selanjutnya.

#### 2.1.2.2 Case Folding

Proses case folding dilakukan dengan mengubah seluruh huruf dalam teks menjadi huruf kecil (lowercase). Tujuan dari proses ini adalah untuk menyamakan format penulisan kata dan menghindari perbedaan makna akibat penggunaan huruf besar dan huruf kecil.

#### 2.1.2.3 Normalisasi Kata

Proses normalisasi kata dilakukan untuk mengubah kata tidak baku, singkatan, atau bahasa informal menjadi kata baku sesuai dengan kaidah bahasa Indonesia. Tahap ini penting karena komentar pada media sosial umumnya menggunakan bahasa tidak formal, sehingga normalisasi kata dapat membantu meningkatkan akurasi.

#### 2.1.2.4 Tokenization

Proses tokenization merupakan proses pemecahan kalimat menjadi unit-unit kata atau token. Tokenization bertujuan untuk memudahkan sistem dalam menganalisis setiap kata dalam dokumen teks.

#### 2.1.2.5 Stopword Removal

Proses stopword removal dilakukan untuk menghapus kata-kata umum yang tidak memiliki makna penting dalam analisis sentimen, seperti kata penghubung atau kata yang sering muncul namun tidak memberikan informasi signifikan terhadap sentimen.

Journal of Information System Research (JOSH)
Volume 7, No. 4, Juli 2026, pp 1057−1066
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i4.10574

#### 2.1.2.6 Wordcloud Setelah Preprocessing

Wordcloud setelah preprocessing digunakan untuk memvisualisasikan kata-kata yang paling sering muncul setelah seluruh tahapan preprocessing selesai dilakukan. Visualisasi ini membantu menggambarkan topik utama yang paling banyak dibahas oleh pengguna YouTube mengenai Pertamina sehingga memudahkan interpretasi hasil analisis sentimen.

### 2.2 Pelabelan Data Menggunakan VADER Sentiment (3 Class)

Tahap berikutnya dalam penelitian ini adalah pelabelan data sentimen. Pada tahap ini, setiap komentar dalam dataset akan diberi label sentimen secara otomatis menggunakan metode VADER Sentiment. Metode ini digunakan untuk mengklasifikasikan komentar ke dalam tiga kategori sentimen, yaitu positif, negatif, dan netral. Penggunaan metode pelabelan sentimen secara otomatis juga telah diterapkan pada penelitian analisis sentimen komentar YouTube menggunakan metode VADER. Pendekatan tersebut mampu mempercepat proses anotasi data dalam jumlah besar sekaligus memberikan dasar yang baik untuk proses klasifikasi menggunakan algoritma machine learning [14], sebagaimana efektivitas pendekatan VADER juga dibuktikan dalam mengukur respon kebijakan publik di Indonesia [15].

### 2.3 Splitting Dataset

Setelah proses pelabelan data selesai, dataset kemudian masuk ke tahap splitting dataset, yaitu proses pembagian data menjadi dua bagian, yaitu data pelatihan (training data) dan data pengujian (testing data). Pada penelitian ini, pembagian dataset dilakukan dengan proporsi 80% data pelatihan dan 20% data pengujian. Pembagian dataset ini bertujuan untuk melatih model klasifikasi menggunakan data pelatihan serta menguji performa model menggunakan data pengujian. Pembagian dataset menjadi data pelatihan dan data pengujian merupakan langkah penting dalam pengembangan model klasifikasi untuk memastikan model dapat diuji secara objektif dan menghasilkan performa yang optimal [11].

## 3. HASIL DAN PEMBAHASAN

Pada penelitian ini, proses pengumpulan data dilakukan menggunakan teknik crawling data atau web scraping dari platform YouTube. Data yang dikumpulkan berupa komentar dari video berita yang berkaitan dengan Pertamina. Proses pengambilan data dilakukan menggunakan bahasa pemrograman Python dengan memanfaatkan YouTube Data API.

### 3.1 Hasil Pengumpulan Data (Crawling Data)

Pengumpulan data komentar dari platform YouTube melalui proses crawling telah banyak digunakan dalam penelitian analisis sentimen karena mampu menyediakan data dalam jumlah besar yang merepresentasikan opini masyarakat terhadap suatu isu. Pendekatan serupa juga diterapkan pada penelitian klasifikasi komentar YouTube berbasis Naïve Bayes [16]. Pendekatan analisis sentimen menggunakan algoritma Naïve Bayes juga telah diterapkan pada data Twitter untuk menganalisis ulasan masyarakat terhadap layanan BMKG Nasional. Hasil penelitian tersebut menunjukkan bahwa algoritma Naïve Bayes mampu memberikan performa yang baik pada berbagai jenis data media sosial [17].

### 3.2 Hasil Text Preprocessing

Tahap text preprocessing dilakukan untuk membersihkan dan menyiapkan data teks agar siap digunakan dalam proses analisis sentimen. Tahapan preprocessing yang dilakukan meliputi: Cleaning, Case Folding, Normalisasi Kata, Tokenization, Stopword Removal. Proses preprocessing bertujuan untuk menghilangkan data yang tidak relevan dan meningkatkan akurasi model klasifikasi [18]. Hasil penelitian ini juga sejalan dengan penelitian komparatif yang menunjukkan bahwa algoritma Naïve Bayes mampu memberikan performa yang kompetitif pada analisis sentimen, meskipun pada karakteristik dataset tertentu Logistic Regression dapat memberikan hasil yang sedikit lebih baik. Oleh karena itu, pemilihan algoritma perlu disesuaikan dengan karakteristik data yang digunakan [19].

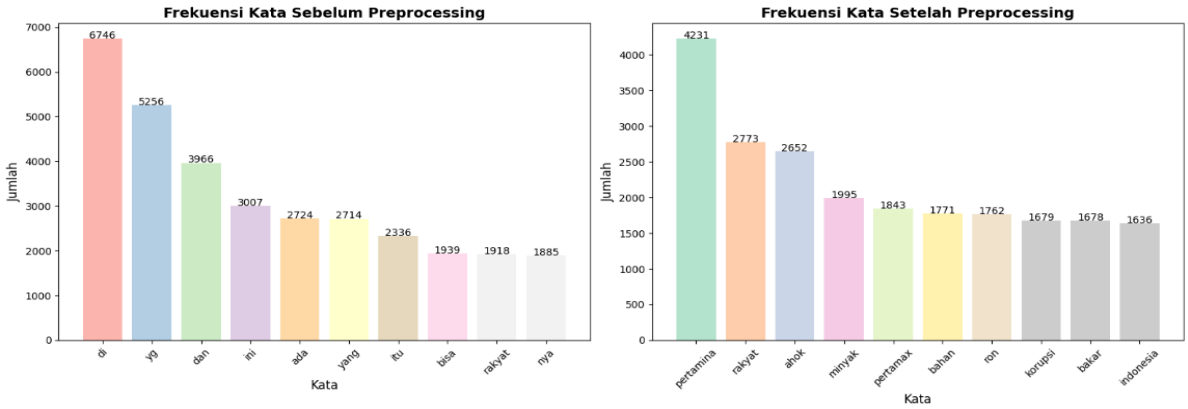
#### 3.2.1 Hasil Frekuensi Kata Sebelum dan Setelah Preprocessing

Setelah dilakukan proses preprocessing, dilakukan analisis frekuensi kata untuk mengetahui perubahan jumlah kata sebelum dan setelah proses pembersihan data.

Perbandingan frekuensi kemunculan kata sebelum dan setelah proses preprocessing disajiakan pada Gambar 2. Berdasarkan visualisasi tersebut, frekuensi kata sebelum preprocessing masih mengandung banyak kata tidak penting (stopwords), seperti tanda baca, kata umum seperti "yang", "ini", "di", serta kata tidak baku. Setelah dilakukan preprocessing, jumlah kata menjadi lebih terstruktur dan hanya menyisakan kata yang memiliki makna penting dalam analisis sentimen. Kata-kata yang paling sering muncul setelah preprocessing antara lain: "pertamina" (4.231), "rakyat" (2.773), "minyak" (2.652), "bahan" (1.995), "bakar" (1.843), dan "negara" (1.771).

Journal of Information System Research (JOSH)
Volume 7, No. 4, Juli 2026, pp 1057−1066
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i4.10574

Hal ini menunjukkan bahwa topik utama yang dibahas masyarakat berkaitan dengan isu bahan bakar minyak dan peran negara.




**Gambar 2.** Perbandingan Frekuensi Kata Sebelum dan Setelah Preprocessing

### 3.2.2 Hasil Wordcloud Setelah Preprocessing

Representasi visual dari distribusi kata yang paling sering muncul sebelum dan setelah proses pembersihan data disajikan dalam wordcloud pada Gambar 3. Berdasarkan visualisasi tersebut, wordcloud sebelum preprocessing masih didominasi oleh kata-kata umum (stopwords) yang kurang bermakna untuk analisis sentimen, seperti "yang", "di", "dan", "ini", serta singkatan tidak baku. Setelah preprocessing, wordcloud menunjukkan kata-kata bermakna yang lebih dominan, antara lain "pertamina", "bahan bakar", "minyak", "rakyat", "korupsi", dan "ahok". Dominasi kata-kata tersebut mencerminkan topik utama yang menjadi perhatian masyarakat dalam komentar video berita terkait Pertamina, yaitu isu bahan bakar minyak, tata kelola perusahaan, serta tokoh-tokoh yang terkait.

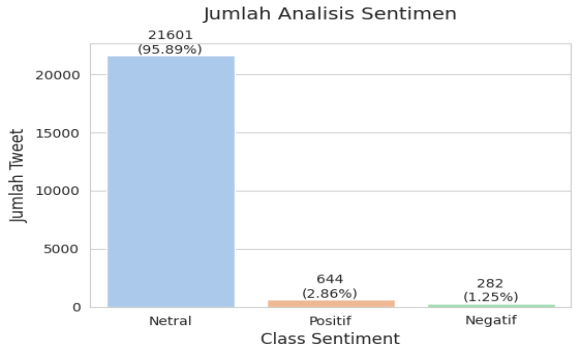

**Gambar 3.** Wordcloud Sebelum dan Setelah Preprocessing

### 3.2.3 Hasil Pelabelan Data Sentimen

Setelah proses preprocessing selesai, tahap selanjutnya adalah pelabelan data sentimen menggunakan metode VADER Sentiment. Setiap komentar dalam dataset diklasifikasikan ke dalam tiga kategori sentimen, yaitu positif, negatif, dan netral.Dominasi salah satu kelas sentimen pada data media sosial juga ditemukan pada berbagai penelitian analisis sentimen terhadap opini masyarakat. Kondisi tersebut menunjukkan bahwa distribusi sentimen pada media sosial sering kali tidak seimbang sehingga perlu diperhatikan dalam proses pembangunan model klasifikasi [19], [20].

Distribusi hasil pelabelan sentimen pada dataset yang dianalisis disajikan pada gambar 4. Berdasarkan visualisasi tersebut, sentimen netral mendominasi dataset dengan jumlah 21.601 komentar (95,89%), diikuti oleh sentimen positif sebanyak 644 komentar (2,86%), dan sentimen negatif sebanyak 282 komentar (1,25%). Dominasi sentimen netral mengindikasikan bahwa sebagian besar komentar masyarakat bersifat deskriptif atau informatif tanpa ekspresi emosi yang kuat, sementara terdapat ketidakseimbangan kelas (class imbalance) yang perlu diperhatikan dalam proses klasifikasi. Kondisi ketidakseimbangan kelas semacam ini juga umum ditemukan pada dataset sentimen berskala besar lainnya, misalnya pada klasifikasi sentimen kebijakan vaksin Covid-19 di Twitter yang turut menyoroti perlunya strategi khusus dalam menangani dominasi salah satu kelas agar performa model tetap optimal [21].

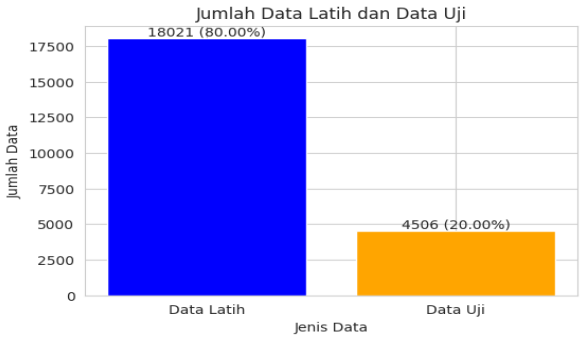
Journal of Information System Research (JOSH)
Volume 7, No. 4, Juli 2026, pp 1057−1066
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i4.10574




**Gambar 4.** Grafik Distribusi Hasil Pelabelan Data Sentimen

### 3.2.4 Hasil Splitting Dataset

Setelah proses pelabelan data selesai, dataset kemudian dibagi menjadi dua bagian: data training dan data testing. Pada penelitian ini, pembagian dataset dilakukan dengan proporsi perbandingan 80% data latih dan 20% data uji, sebagaimana disajikan pada Gambar 5. Pembagian dataset dengan skema ini bertujuan untuk melatih model klasifikasi menggunakan data training serta menguji performa model menggunakan data testing secara objektif dan tidak bias.



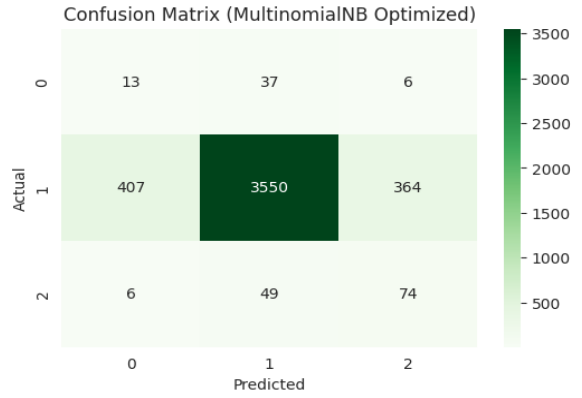

**Gambar 5.** Grafik Distribusi Pembagian Dataset (Data Latih dan Data Uji)

### 3.3 Hasil Klasifikasi Menggunakan Naïve Bayes

Pada tahap ini, dilakukan proses klasifikasi sentimen menggunakan algoritma Naïve Bayes. Dalam penelitian ini digunakan tiga jenis algoritma Naïve Bayes, yaitu: Naïve Bayes Gaussian, Naïve Bayes Multinomial serta Naïve Bayes Bernoulli. Penggunaan tiga jenis algoritma ini bertujuan untuk membandingkan performa masing-masing metode dalam mengklasifikasikan sentimen komentar.

### 3.3.1 Hasil Klasifikasi Naïve Bayes Gaussian

Pada metode Naïve Bayes Gaussian, model digunakan untuk mengklasifikasikan data berdasarkan distribusi normal (Gaussian). Metode ini umumnya digunakan untuk data numerik yang mengikuti distribusi normal. Pada data teks yang telah direpresentasikan menggunakan TF-IDF, Gaussian Naïve Bayes kurang optimal karena representasi TF-IDF tidak selalu mengikuti distribusi normal. Confusion matrix hasil klasifikasi Naïve Bayes Gaussian ditampilkan pada Gambar 6.

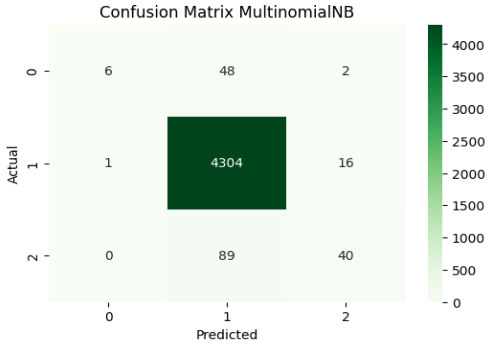



**Gambar 6.**Confusion Matrix Hasil Klasifikasi Naïve Bayes Gaussian (Optimized)

Journal of Information System Research (JOSH)
Volume 7, No. 4, Juli 2026, pp 1057−1066
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i4.10574

### 3.3.2 Hasil Klasifikasi Naïve Bayes Multinomial

Metode Naïve Bayes Multinomial merupakan metode yang paling umum digunakan dalam analisis sentimen berbasis teks karena mampu menangani data frekuensi kata dengan baik. Metode ini bekerja dengan baik ketika dikombinasikan dengan pembobotan TF-IDF yang merepresentasikan frekuensi kemunculan kata. Confusion matrix menunjukkan bahwa model Multinomial Naïve Bayes menghasilkan performa yang lebih baik dibandingkan Gaussian, dengan nilai pada diagonal utama yang lebih tinggi, terutama untuk kelas netral (nilai 4.304). Hal ini menunjukkan kemampuan Multinomial NB yang lebih baik dalam menangani data teks berbasis frekuensi kata, sejalan dengan efektivitas kombinasi TF-IDF dan model klasifikasi berbasis frekuensi kata yang dilaporkan pada penelitian sentimen ulasan film [22].

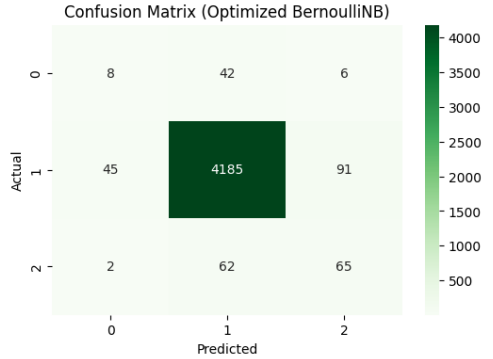



**Gambar 7.** Confusion Matrix Hasil Klasifikasi Naïve Bayes Multinomial

Berdasarkan Gambar 7, confusion matrix menunjukkan bahwa model Multinomial Naïve Bayes menghasilkan performa yang lebih baik dibandingkan Gaussian, dengan nilai pada diagonal utama yang lebih tinggi, terutama untuk kelas netral (nilai 4.304). Hal ini menunjukkan kemampuan Multinomial NB yang lebih baik dalam menangani data teks berbasis frekuensi kata.

### 3.3.3 Hasil Klasifikasi Naïve Bayes Bernoulli

Metode Naïve Bayes Bernoulli digunakan untuk data biner, yaitu data yang hanya memiliki dua kemungkinan nilai (ada atau tidak adanya suatu kata dalam dokumen). Metode ini cocok untuk teks pendek atau kasus di mana kehadiran/ketidakhadiran suatu kata lebih relevan daripada frekuensinya. Confusion matrix hasil klasifikasi Naïve Bayes Bernoulli ditampilkan pada Gambar 8.




**Gambar 8.** Confusion Matrix Hasil Klasifikasi Naïve Bayes Bernoulli

Berdasarkan Gambar 8, confusion matrix Bernoulli Naïve Bayes (setelah optimasi) menunjukkan bahwa model mampu mengklasifikasikan sentimen dengan nilai diagonal utama 4.185 untuk kelas netral. Dibandingkan Multinomial NB, performa Bernoulli NB sedikit lebih rendah, yang mengindikasikan bahwa untuk dataset komentar YouTube berbahasa Indonesia ini, informasi frekuensi kata (yang dimanfaatkan oleh Multinomial NB) lebih informatif daripada sekadar kehadiran/ketidakhadiran kata.yang mengindikasikan bahwa untuk dataset komentar YouTube berbahasa Indonesia ini, informasi frekuensi kata (yang dimanfaatkan oleh Multinomial NB) lebih informatif daripada sekadar kehadiran/ketidakhadiran kata

### 3.4 Pembahasan

Berdasarkan hasil penelitian yang telah dilakukan, perlu dilakukan diskusi mendalam mengenai perbandingan ketiga varian algoritma Naïve Bayes yang digunakan, yaitu Gaussian, Multinomial, dan Bernoulli. Secara umum, Gaussian Naïve Bayes mencapai akurasi sekitar 79%, Bernoulli Naïve Bayes sekitar 92%, dan Multinomial Naïve Bayes sekitar 95%, dengan Multinomial NB unggul pada seluruh metrik precision, recall, dan f1-score.

Journal of Information System Research (JOSH)
Volume 7, No. 4, Juli 2026, pp 1057−1066
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i4.10574

Naïve Bayes Multinomial adalah varian yang paling unggul dalam penelitian ini, dengan alasan sebagai berikut. Pertama, dari sisi kesesuaian dengan jenis data: data pada penelitian ini berupa teks yang direpresentasikan menggunakan TF-IDF, yang menghasilkan nilai frekuensi (bukan biner dan bukan distribusi normal). Multinomial NB secara desain dirancang untuk menangani distribusi multinomial, yaitu distribusi yang merepresentasikan frekuensi kemunculan fitur (kata). Hal ini sesuai dengan karakteristik data TF-IDF yang digunakan. Kedua, dari sisi hasil evaluasi: confusion matrix Multinomial NB menunjukkan jumlah prediksi benar (diagonal utama) yang paling tinggi dibandingkan Gaussian dan Bernoulli, yaitu dengan nilai 4.304 untuk kelas netral, yang menghasilkan akurasi tertinggi di antara ketiga metode. Ketiga, dari sisi konsistensi: berbagai penelitian terdahulu, termasuk [12] pada dataset komentar YouTube G20, juga mengonfirmasi bahwa Multinomial NB memberikan hasil terbaik untuk analisis sentimen berbasis teks, sebagaimana juga tampak konsisten pada berbagai studi klasifikasi sentimen komentar YouTube lainnya [6], [7], [8], [22]. Keunggulan Naïve Bayes juga tidak bersifat mutlak di semua kasus; pada data ulasan berbasis aspek dengan jumlah komentar besar, [1] SVM dapat sedikit lebih unggul dibandingkan Naïve Bayes, sehingga pemilihan algoritma tetap perlu mempertimbangkan karakteristik data yang dianalisis.

Naïve Bayes Gaussian menunjukkan performa yang kurang optimal karena metode ini mengasumsikan bahwa data fitur mengikuti distribusi normal (Gaussian). Representasi TF-IDF pada data teks umumnya tidak memenuhi asumsi tersebut, sehingga estimasi probabilitas yang dihasilkan kurang akurat, terutama untuk kelas minoritas (positif dan negatif) yang memiliki jumlah sampel jauh lebih sedikit akibat ketidakseimbangan kelas pada dataset [21].

Naïve Bayes Bernoulli memberikan performa yang cukup baik namun masih di bawah Multinomial NB. Hal ini karena Bernoulli NB hanya mempertimbangkan ada-tidaknya suatu kata (binary: 0 atau 1), sehingga informasi mengenai seberapa sering suatu kata muncul tidak dimanfaatkan. Untuk komentar YouTube yang memiliki variasi panjang teks yang signifikan, frekuensi kemunculan kata merupakan informasi yang lebih kaya dibandingkan sekadar kehadiran kata [23].

Dengan demikian, untuk penelitian analisis sentimen berbasis komentar teks di media sosial, khususnya YouTube dengan bahasa Indonesia, Naïve Bayes Multinomial merupakan pilihan algoritma yang paling direkomendasikan karena kesesuaiannya dengan karakteristik data teks berfrekuensi, hasil evaluasi yang superior, serta dukungan dari berbagai literatur terkini.

## 4. KESIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan mengenai analisis sentimen kepercayaan publik terhadap Pertamina pada media berita di YouTube, dapat disimpulkan bahwa proses pengumpulan data menggunakan teknik crawling melalui YouTube Data API berhasil memperoleh dataset komentar pengguna yang relevan dengan topik penelitian. Dari total 56.868 komentar yang berhasil dikumpulkan, dilakukan proses seleksi sehingga diperoleh 22.527 komentar yang memenuhi kriteria untuk dianalisis. Data tersebut kemudian diolah melalui tahapan text preprocessing yang meliputi cleaning, case folding, normalisasi kata, tokenization, dan stopword removal sehingga menghasilkan data yang lebih bersih dan siap digunakan dalam proses klasifikasi sentimen.
Hasil klasifikasi menunjukkan bahwa algoritma Multinomial Naïve Bayes memberikan performa terbaik dibandingkan Gaussian Naïve Bayes dan Bernoulli Naïve Bayes. Hal ini menunjukkan bahwa Multinomial Naïve Bayes lebih sesuai untuk mengklasifikasikan data komentar YouTube yang direpresentasikan menggunakan pembobotan TF-IDF. Selain itu, visualisasi menggunakan wordcloud setelah proses preprocessing berhasil menampilkan kata-kata yang paling sering muncul sehingga memberikan gambaran mengenai topik utama yang menjadi perhatian masyarakat terhadap Pertamina. Berdasarkan hasil penelitian tersebut, penelitian selanjutnya disarankan untuk menggunakan jumlah dataset yang lebih besar agar hasil analisis sentimen menjadi lebih akurat dan representatif terhadap opini masyarakat. Selain itu, penelitian selanjutnya dapat membandingkan performa algoritma Naïve Bayes dengan metode klasifikasi lain, seperti Support Vector Machine (SVM), Random Forest, maupun metode Deep Learning seperti Long Short-Term Memory (LSTM) dan Convolutional Neural Network (CNN). Penelitian selanjutnya juga dapat menerapkan teknik penanganan ketidakseimbangan kelas (class imbalance), seperti oversampling atau undersampling, serta menambahkan tahapan stemming atau lemmatization pada proses preprocessing untuk meningkatkan kualitas data dan performa model klasifikasi. Pengembangan penelitian juga dapat dilakukan dengan memanfaatkan data dari berbagai platform media sosial, seperti Twitter (X), Instagram, dan TikTok, sehingga mampu memberikan gambaran opini publik yang lebih luas. Hasil penelitian ini diharapkan dapat menjadi bahan pertimbangan bagi Pertamina maupun pihak terkait dalam memahami persepsi masyarakat serta menyusun strategi komunikasi publik yang lebih efektif.

## REFERENCES

[1] Chely Aulia Misrun, E. Haerani, M. Fikry, and E. Budianita, "Analisis sentimen komentar youtube terhadap Anies Baswedan sebagai bakal calon presiden 2024 menggunakan metode naive bayes classifier," J. CoSciTech (Computer Sci. Inf. Technol.), vol. 4, no. 1, pp. 207–215, 2023, doi: 10.37859/coscitech.v4i1.4790.

[2] T. Putri, S. Nurhaliza, and D. Vionanda, "Sentiment Analysis of Using the YouTube Application Using the Naïve Bayes Method," J. Edukasi dan Penelitian Inform. (JEPIN), vol. 3, pp. 60–66, 2025, doi: https://doi.org/10.24036/ujsds/vol3-

iss1/343.
[3] M. S. P. Putra, Sri Dharma Wati, and Imam Solikin, "Analisis Sentimen Serikat Pekerja Pertamina Tolak Ahok Pada Media Sosial Youtube Menggunakan Algoritma Naive Bayes," J. Ilm. Betrik, vol. 12, no. 2, pp. 99–105, 2021, doi: 10.36050/betrik.v12i2.219.
[4] C. G. Indrayanto, D. E. Ratnawati, and B. Rahayudi, "Analisis Sentimen Data Ulasan Pengguna Aplikasi MyPertamina di Indonesia pada Google Play Store menggunakan Metode Random Forest," J. Pengemb. Teknol. Inf. dan Ilmu Komput., vol. 7, no. 3, pp. 1131–1139, 2023, [Online]. Available: http://j-ptiik.ub.ac.id
[5] S. Mujahidin, B. Prasetio, and M. C. C. Utomo, "Implementasi Analisis Sentimen Masyarakat Mengenai Kenaikan Harga BBM Pada Komentar Youtube Dengan Metode Gaussian naïve bayes," Voteteknika (Vocational Tek. Elektron. dan Inform., vol. 10, no. 3, p. 17, 2022, doi: 10.24036/voteteknika.v10i3.118299.
[6] M. Hudha, E. Supriyati, and T. Listyorini, "Analisis Sentimen Pengguna YouTube Terhadap Tayangan #MataNajwaMenantiTerawan dengan Metode Naïve Bayes Classifier," JIKO (Jurnal Informatika dan Komputer), vol. 5, no. 1, pp. 1–6, 2022, doi: 10.33387/jiko.v5i1.3376.
[7] H. A. R. Harpizon, R. Kurniawan, I. Iskandar, R. Salambue, E. Budianita, and F. Syafria, "Analisis Sentimen Komentar Di YouTube Tentang Ceramah Ustadz Abdul Somad Menggunakan Algoritma Naïve Bayes," JNKTI (Jurnal Nasional Komputasi dan Teknologi Informasi), vol. 5, no. 1, pp. 131–140, 2022, doi: 10.32672/jnkti.v5i1.4008.
[8] M. I. Ahmadi, D. Gustian, and F. Sembiring, "Analisis Sentimen Masyarakat terhadap Kasus Covid-19 pada Media Sosial YouTube dengan Metode Naive Bayes," J-SAKTI (Jurnal Sains Komputer dan Informatika), vol. 5, no. 2, pp. 807–814, 2021, doi: 10.30645/j-sakti.v5i2.378.
[9] O. I. Gifari, M. Adha, I. R. Hendrawan, and F. F. S. Durrand, "Analisis Sentimen Review Film Menggunakan TF-IDF dan Support Vector Machine," JIFOTECH (Journal of Information Technology), vol. 2, no. 1, pp. 36–40, 2022, doi: 10.46229/jifotech.v2i1.330.
[10] S. Mujahidin, B. Prasetio, and M. C. C. Utomo, "Implementasi Analisis Sentimen Masyarakat Mengenai Kenaikan Harga BBM pada Komentar YouTube dengan Metode Gaussian Naïve Bayes," Voteteknika (Vocational Teknik Elektronika dan Informatika), vol. 10, no. 3, pp. 17–24, 2022, doi: 10.24036/voteteknika.v10i3.118299.
[11] A. Liawati, R. Narasati, D. Solihudin, C. Lukman Rohmat, and S. Eka Permana, "Analisis Sentimen Komentar Politik Di Media Sosial X Dengan Pendekataan Deep Learning," JATI (Jurnal Mhs. Tek. Inform., vol. 7, no. 6, pp. 3557–3563, 2024, doi: 10.36040/jati.v7i6.8248.
[12] R. A. Firsttama, A. A. Arifiyanti, and D. S. Y. Kartika, "Analisis Sentimen Komentar Youtube Konferensi Tingkat Tinggi G20 Menggunakan Metode Naive Bayes," J. Teknol. Dan Sist. Inf. Bisnis, vol. 6, no. 2, pp. 282–285, 2024, doi: 10.47233/jteksis.v6i2.1263.
[13] Rayuwati, Husna Gemasih, and Irma Nizar, "Implementasi Algoritma Naive Bayes Untuk Memprediksi Tingkat Penyebaran Covid," Jural Ris. Rumpun Ilmu Tek., vol. 1, no. 1, pp. 38–46, 2022, doi: 10.55606/jurritek.v1i1.127.
[14] B. Santoso, S. Prawiradiredja, D. Abror, S. A. Sufa, A. Raharja, and P. T. Mardianta, "Implementasi Metode VADER Pada Analisis Sentimen Komentar Video Youtube Desa Wisata Bejijong," JuTI (Jurnal Teknologi Informasi), vol. 3, no. 2, pp. 91–97, 2025, doi: 10.26798/juti.v3i2.1868.
[15] N. Fauziah, M. Alkautsar, Y. Suryaman, and F. F. Roji, "Analisis Sentimen Publik Terhadap Kenaikan Tarif PPN di Indonesia dengan Pendekatan VADER," J. Akuntansi dan Keuangan, vol. 12, no. 2, pp. 228–238, 2024, doi: https://doi.org/10.29103/jak.v12i2.16796.
[16] H. Murti, R. S. Redjeki, N. Mariana, A. P. Utomo, and W. T. Handoko, "Pemodelan Sentimen Komentar YouTube Berbasis Naive Bayes," J. Inform. dan Teknol. Komput. (JITEK), vol. 6, no. 1, pp. 1–12, 2026. doi: 10.55606/jitek.v6i1.8767.
[17] D. Darwis, N. Siskawati, and Z. Abidin, "Penerapan Algoritma Naive Bayes Untuk Analisis Sentimen Review Data Twitter Bmkg Nasional," J. Tekno Kompak, vol. 15, no. 1, p. 131, 2021, doi: 10.33365/jtk.v15i1.744.
[18] D. S. Hidayat, O. Nurdiawan, F. M. Basysyar, and M. Sulaeman, "Peningkatan Model Analisis Sentimen Melalui Algoritmanaive Bayes Berdasarkan Data Komentar Youtube," J. Manaj. Inform. Dan Sist. Inf. (MISI), vol. 8, pp. 164–175, 2025, doi: 10.36595/misi.v5i2.
[19] B. Ramadhani and R. R. Suryono, "Komparasi Algoritma Naïve Bayes dan Logistic Regression Untuk Analisis Sentimen Metaverse," Jurnal Media Informatika Budidarma, vol. 8, no 2. April, pp. 714–725, 2024, doi: 10.30865/mib.v8i2.7458.
[20] F. Khoirunnisa and S. Topiq, "Analisis Sentimen Terhadap Masyarakat Pada Proses Penegak Hukum Di Indonesia Dengan Menggunakan Algoritma Naïve Bayes," J. Teknik Informatika dan Teknik Elektro Terapan (JITET), vol. 12, no. 3, pp. 2128–2139, 2024, doi: https://doi.org/10.23960/jitet.v12i3.4683.
[21] F. Septianingrum, J. H. Jaman, and U. Enri, "Analisis Sentimen Pada Isu Vaksin Covid-19 di Indonesia dengan Metode Naive Bayes Classifier," Jurnal Media Informatika Budidarma, vol. 5, pp. 1431–1437, 2021, doi: 10.30865/mib.v5i4.3260.
[22] R. Amelia, N. S. Prastiwi, and M. E. Purbaya, "Impementasi Algoritma Naive Bayes Terhadap Analisis Sentimen Opini Masyarakat Indonesia Mengenai Drama Korea Pada Twitter," Jurnal Riset Komputer (JURIKOM), vol. 9, no. 2, pp. 338–343, 2022, doi: 10.30865/jurikom.v9i2.3895.
[23] A. Karimah, G. Dwilestari, and M. Mulyawan, "Analisis Sentimen Komentar Video Mobil Listrik Di Platform Youtube Dengan Metode Naive Bayes," JATI (Jurnal Mhs. Tek. Inform., vol. 8, no. 1, pp. 767–737, 2024, doi: 10.36040/jati.v8i1.8373.