Jurnal Sains Dan Teknologi Informasi
Vol. 5, No. 4, September 2026, Hal 189-198
ISSN 2809-610X (Media Online)
DOI 10.47065/jussi.v5i4.11226
Website https://ejurnal.seminar-id.com/index.php/jussi

Analisis Sentimen Komentar YouTube Kualifikasi Piala Dunia 2026 Ronde 4 Timnas Indonesia menggunakan Naïve Bayes

Rasyid Iskandar Prayogi*, Azhari Ali Ridha, Ade Andri Hendriadi, Nina Sulistyowati, Taufik Ridwan

Fakultas Ilmu Komputer, Sistem Informasi, Universitas Singaperbangsa Karawang, Telukjambe Timur, Jawa Barat, Indonesia
Email: 1,*2210631250028@student.unsika.ac.id, 2azhari.ali@unsika.ac.id, 3adeandri@unsika.ac.id,
4nina.sulistiyowati@unsika.ac.id, 5taufik.ridwan@cs.unsika.ac.id
(* : coressponding author)

Abstrak-Perhatian masyarakat terhadap Kualifikasi Piala Dunia 2026 Tim Nasional Indonesia terlihat dari beragam tanggapan yang disampaikan melalui komentar YouTube. Perbedaan tanggapan tersebut menunjukkan adanya sentimen yang beragam terhadap performa Tim Nasional Indonesia, sehingga diperlukan pengelompokan komentar untuk melihat kecenderungan sentimen yang muncul. Penelitian ini bertujuan menganalisis sentimen komentar YouTube terkait Tim Nasional Indonesia pada Ronde 4 Kualifikasi Piala Dunia 2026 menggunakan algoritma Multinomial Naïve Bayes. Sebanyak 17.404 komentar dikumpulkan melalui YouTube Data API v3 dari kanal Timnas Indonesia dan RCTI Sports. Data kemudian melalui tahapan preprocessing, pelabelan menggunakan InSet Lexicon yang diverifikasi oleh ahli bahasa, pembobotan menggunakan Term Frequency-Inverse Document Frequency (TF-IDF), dan klasifikasi dengan Multinomial Naïve Bayes. Pengujian dilakukan menggunakan tiga skenario pembagian data, yaitu 80:20, 75:25, dan 70:30 dengan metrik accuracy, precision, recall, dan F1-score. Hasil pengujian menunjukkan bahwa pembagian data 70:30 menghasilkan akurasi tertinggi sebesar 79,09%. Pada skenario tersebut, kelas negatif memperoleh precision sebesar 0,81, recall sebesar 0,93, dan F1-score sebesar 0,87. Kelas negatif juga menjadi sentimen yang paling dominan, sementara jumlah data pada kelas netral dan positif lebih rendah. Hasil tersebut menunjukkan bahwa Multinomial Naïve Bayes dapat digunakan untuk mengklasifikasikan komentar YouTube ke dalam sentimen negatif, netral, dan positif serta memberikan gambaran mengenai kecenderungan tanggapan masyarakat terhadap performa Tim Nasional Indonesia pada Ronde 4 Kualifikasi Piala Dunia 2026.

Kata Kunci: Analisis Sentimen; YouTube; Tim Nasional Indonesia; Naïve Bayes; TF-IDF

Abstract-*Public interest in the Indonesian National Team's 2026 World Cup Qualifiers is evident from the diverse responses expressed in YouTube comments. The diversity of these responses indicates a range of sentiments toward the Indonesian National Team's performance, making it necessary to cluster the comments to identify emerging sentiment trends. This study aims to analyze the sentiment of YouTube comments regarding the Indonesian National Team during Round 4 of the 2026 World Cup Qualifiers using the Multinomial Naïve Bayes algorithm. A total of 17,404 comments were collected via the YouTube Data API v3 from the Indonesian National Team and RCTI Sports channels. The data then underwent preprocessing, labeling using the InSet Lexicon verified by linguists, weighting using Term Frequency-Inverse Document Frequency (TF-IDF), and classification with the Multinomial Naïve Bayes algorithm. Testing was conducted using three data split scenarios 80:20, 75:25, and 70:30 with the metrics of accuracy, precision, recall, and F1-score. The test results showed that the 70:30 data split yielded the highest accuracy of 79.09%. In that scenario, the negative class achieved a precision of 0.81, a recall of 0.93, and an F1-score of 0.87. The negative class was also the most dominant sentiment, while the number of data points in the neutral and positive classes was lower. These results indicate that Multinomial Naïve Bayes can be used to classify YouTube comments.*

Keywords*: Sentiment Analysis; YouTube; Indonesian National Team; Naïve Bayes; TF-IDF*

1. PENDAHULUAN

Perkembangan teknologi informasi menghasilkan peningkatan jumlah data digital, termasuk teks yang dibuat melalui aktivitas pengguna pada beragam layanan media sosial. Data tersebut dapat merepresentasikan aktivitas dan tanggapan pengguna terhadap berbagai peristiwa. Melalui penerapan text mining serta analisis sentimen, kumpulan komentar dan unggahan dapat diproses untuk memperoleh gambaran mengenai kecenderungan opini masyarakat [1], [2]. Kondisi tersebut membuka peluang pemanfaatan data media sosial sebagai sumber informasi dalam menganalisis respons publik terhadap isu yang memiliki tingkat perhatian tinggi [3].

Sepak bola merupakan salah satu bidang yang memiliki keterlibatan masyarakat cukup besar, termasuk di Indonesia. Tingginya ketertarikan masyarakat terhadap olahraga tersebut turut meningkatkan perhatian terhadap Tim Nasional Indonesia (Timnas) dalam berbagai kompetisi internasional [4]. Perjalanan Timnas pada Kualifikasi Piala Dunia FIFA 2026 Zona Asia menjadi salah satu peristiwa yang banyak mendapat perhatian publik. Indonesia berhasil mencapai Ronde 4 untuk pertama kalinya dalam sejarah setelah melalui rangkaian pertandingan pada babak sebelumnya [5]. Pencapaian tersebut diikuti oleh berbagai tanggapan masyarakat mengenai performa tim, keputusan kepelatihan, komposisi pemain, serta kebijakan yang berkaitan dengan Tim Nasional Indonesia [6], [7], [8].

Kemajuan media digital memberikan kesempatan yang semakin besar kepada masyarakat untuk mengungkapkan berbagai tanggapan mengenai perjalanan Tim Nasional Indonesia. YouTube menjadi salah satu platform yang menyediakan konten sekaligus ruang interaksi melalui kolom komentar. Komentar yang diberikan pengguna dapat memuat dukungan, kritik, apresiasi, kekecewaan, maupun harapan terhadap suatu konten atau peristiwa yang sedang dibahas. Karakteristik tersebut menjadikan komentar YouTube sebagai sumber data tekstual yang dapat digunakan untuk memetakan opini publik terhadap suatu objek tertentu [4], [9]. Pemanfaatan

Jurnal Sains Dan Teknologi Informasi
Vol. 5, No. 4, September 2026, Hal 189-198
ISSN 2809-610X (Media Online)
DOI 10.47065/jussi.v5i4.11226
Website https://ejurnal.seminar-id.com/index.php/jussi

YouTube sebagai sumber data juga memungkinkan komentar dikaitkan secara langsung dengan video dan peristiwa yang menjadi objek pembahasan, sehingga konteks data dapat dipertahankan dalam proses analisis [6].

Permasalahan muncul ketika jumlah komentar yang besar membuat identifikasi kecenderungan opini secara manual menjadi kurang efektif. Komentar yang terkumpul memiliki variasi bahasa, penggunaan kata tidak baku, singkatan, serta bentuk ekspresi yang beragam sehingga diperlukan metode pengolahan data untuk mengelompokkan sentimen secara sistematis. Melalui analisis sentimen, opini yang terdapat dalam data teks dapat diklasifikasikan berdasarkan polaritas tertentu, yaitu positif, negatif, dan netral [10], [11]. Tahapan pengolahan teks seperti normalization, stopword removal, dan stemming juga memiliki pengaruh terhadap kualitas data sebelum digunakan untuk proses klasifikasi [12], [13], [14].

Algoritma klasifikasi merupakan salah satu metode yang dapat diterapkan untuk melakukan pengelompokan sentimen. Naïve Bayes adalah metode klasifikasi berbasis probabilitas yang menentukan kelas data dengan mempertimbangkan kemungkinan kemunculan karakteristik tertentu pada masing-masing kelas. Metode ini sering diterapkan pada klasifikasi teks karena perhitungannya cukup sederhana serta dapat diterapkan pada data dalam jumlah besar [15], [16]. Penerapannya pada analisis sentimen juga telah dilakukan pada berbagai objek, seperti ulasan pengguna, media sosial, dan komentar daring [16], [10], [17]. Naïve Bayes memiliki beberapa varian yang berbeda dalam cara memperlakukan fitur kata. Varian Multinomial memanfaatkan jumlah kemunculan kata di setiap dokumen, sedangkan varian Bernoulli hanya membedakan kata yang muncul dan yang tidak muncul. Pada klasifikasi komentar media sosial, varian Multinomial dilaporkan memberikan akurasi lebih tinggi dan waktu proses lebih singkat daripada varian Bernoulli [18], dan varian ini dipandang sebagai bentuk klasik Naïve Bayes yang sesuai untuk klasifikasi teks [17].

Term Frequency-Inverse Document Frequency (TF-IDF) dapat digunakan sebagai representasi fitur untuk mengonversi teks ke dalam bentuk numerik dengan mempertimbangkan tingkat kepentingan kata pada suatu dokumen, sehingga hasil representasi tersebut dapat dimanfaatkan untuk klasifikasi [19], [20], [17], [9]. Keunggulan pembobotan ini terletak pada kemampuannya menekan pengaruh kata yang tersebar di banyak komentar, sehingga kata yang lebih khas mendapat porsi lebih besar dalam penentuan kelas. Pada analisis sentimen dengan Naïve Bayes di Twitter, representasi TF-IDF mencapai akurasi 85% dan mengungguli Bag of Words yang hanya 64% [21], sementara pada klasifikasi ujaran kebencian, TF-IDF juga mengungguli Count Vectorizer [19]. Meski begitu, pada data komentar cyberbullying, Bag of Words justru memberi peningkatan akurasi lebih besar daripada TF-IDF [18], sehingga efektivitas pembobotan bergantung pada karakteristik data.

Sejumlah studi telah menerapkan Naïve Bayes pada permasalahan analisis sentimen di bidang olahraga maupun media digital [3], [8]. Aliyudin et al. menganalisis sentimen mengenai pembatalan Piala Dunia U-20 di Indonesia dengan memanfaatkan data YouTube serta algoritma Naïve Bayes [6]. Hasil penelitian tersebut menunjukkan bahwa komentar YouTube dapat digunakan sebagai data untuk mengidentifikasi kecenderungan sentimen masyarakat terhadap isu sepak bola. Franko et al. menerapkan Decision Tree dan Naïve Bayes untuk menganalisis sentimen mengenai naturalisasi pemain Tim Nasional Indonesia pada YouTube. Hasil pengujian menunjukkan bahwa Naïve Bayes memperoleh akurasi sebesar 85,4%, sedangkan Decision Tree menghasilkan akurasi 70%, 71,8%, dan 70,9% pada depth yang berbeda. Berdasarkan hasil tersebut, performa Naïve Bayes pada penelitian tersebut lebih tinggi dibandingkan Decision Tree [7]. Penelitian lain juga menerapkan Multinomial Naïve Bayes dengan pembobotan TF-IDF pada komentar daring untuk klasifikasi sentimen [17], [22].

Penelitian yang memiliki objek paling dekat dengan penelitian ini dilakukan oleh Muriyatmoko et al. yang menganalisis sentimen komentar YouTube mengenai performa Tim Nasional Indonesia dalam Kualifikasi Piala Dunia 2026 dengan menerapkan algoritma Naïve Bayes. Studi tersebut menggunakan komentar yang berasal dari 20 video yang mencakup Ronde 1 hingga Ronde 4 sehingga menghasilkan gambaran sentimen secara umum selama rangkaian kualifikasi [23]. Perbedaan penelitian ini terletak pada ruang lingkup data yang difokuskan pada Ronde 4 dalam Kualifikasi Piala Dunia 2026 Zona Asia. Pembatasan tersebut memungkinkan analisis diarahkan pada periode pertandingan tertentu yang memiliki dinamika tersendiri, sehingga kecenderungan sentimen masyarakat dapat diamati secara lebih spesifik dibandingkan penggabungan seluruh ronde kualifikasi.

Permasalahan yang menjadi perhatian penelitian ini adalah belum spesifiknya gambaran kecenderungan sentimen masyarakat pada komentar YouTube yang berkaitan dengan Timnas Indonesia pada Ronde 4 Kualifikasi Piala Dunia 2026. Data komentar yang tersedia dalam jumlah besar perlu diolah melalui tahapan pemrosesan teks dan klasifikasi agar dapat menghasilkan informasi mengenai distribusi sentimen secara terstruktur. Dapat menjadi dasar penerapan klasifikasi sentimen dengan algoritma Multinomial Naïve Bayes melalui representasi TF-IDF untuk mengelompokkan komentar ke dalam sentimen positif, negatif, dan netral.

State of the art penelitian ini terletak pada Ronde 4 Kualifikasi Piala Dunia 2026 Zona Asia dengan menerapkan algoritma Multinomial Naïve Bayes serta representasi TF-IDF. Fokus tersebut membedakan penelitian ini dari penelitian terdahulu yang menggunakan objek sepak bola dengan isu tertentu maupun penelitian yang mencakup keseluruhan ronde Kualifikasi Piala Dunia 2026. Pembatasan objek berdasarkan ronde juga memberikan ruang untuk melihat karakteristik opini publik pada fase tertentu secara lebih terarah.

Penelitian ini bertujuan untuk mengidentifikasi kecenderungan sentimen masyarakat terhadap Tim Nasional Indonesia berdasarkan komentar YouTube pada Ronde 4 Kualifikasi Piala Dunia 2026 Zona Asia dengan

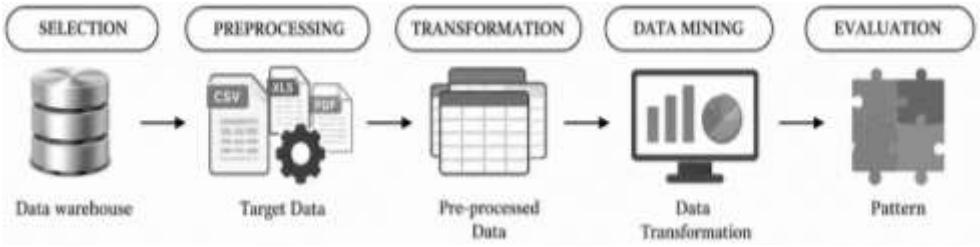
Jurnal Sains Dan Teknologi Informasi
Vol. 5, No. 4, September 2026, Hal 189-198
ISSN 2809-610X (Media Online)
DOI 10.47065/jussi.v5i4.11226
Website https://ejurnal.seminar-id.com/index.php/jussi

menerapkan algoritma Multinomial Naïve Bayes untuk mengelompokkan komentar berdasarkan sentimen positif, negatif, dan netral. Model yang dibangun kemudian dievaluasi berdasarkan metrik klasifikasi guna mengukur kemampuan algoritma dalam mengolah data komentar tersebut. Kontribusi penelitian ini adalah menyediakan gambaran berbasis data mengenai kecenderungan opini publik terhadap Tim Nasional Indonesia pada Ronde 4 serta memberikan penerapan metode analisis sentimen pada data komentar YouTube yang dapat menjadi referensi bagi penelitian sejenis di bidang analisis data olahraga dan media sosial.

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Penelitian ini menerapkan pendekatan Knowledge Discovery in Database (KDD) sebagai kerangka kerja untuk menganalisis sentimen komentar YouTube terkait Tim Nasional Indonesia pada Ronde 4 Kualifikasi Piala Dunia 2026 Zona Asia. KDD digunakan untuk mengatur proses pengolahan data secara terstruktur, dimulai dari pemilihan data hingga tahap evaluasi hasil klasifikasi[24]. Tahapan yang diterapkan meliputi data selection, data preprocessing, data transformation, data mining, dan evaluation [25]. Keseluruhan dari rancangan penelitian disajikan pada Gambar 1:



Gambar 1. Proses Pengolahan Data Menggunakan Metode KDD [26]

Tahap selection pada Gambar 1 mencakup proses pengumpulan serta penentuan komentar yang sesuai dengan objek penelitian. Data terpilih selanjutnya diproses melalui preprocessing untuk membersihkan serta menyeragamkan teks. Pada tahap transformation, data teks dikonversi ke dalam bentuk numerik menggunakan Term Frequency-Inverse Document Frequency (TF-IDF). Selanjutnya, tahap data mining menerapkan algoritma Multinomial Naïve Bayes dalam melakukan klasifikasi sentimen, sedangkan tahap evaluation berfungsi untuk mengukur kinerja model berdasarkan prediksi yang dihasilkan pada data pengujian [25].

2.2 Data Selection

Data penelitian berupa komentar YouTube berbahasa Indonesia yang berkaitan dengan Timnas Indonesia di Putaran 4 Kualifikasi Piala Dunia 2026 untuk Zona Asia. Komentar dikumpulkan menggunakan YouTube Data API v3 dari dua kanal YouTube, yaitu Timnas Indonesia dan RCTI Sports. Komentar yang diperoleh kemudian diseleksi berdasarkan kesesuaian dengan objek dan kebutuhan penelitian. Data hasil seleksi selanjutnya digunakan sebagai dataset penelitian [25]. Setelah melalui proses pengolahan teks, diperoleh sebanyak 17.404 komentar yang digunakan dalam proses analisis sentimen.

2.3 Data Preprocessing dan Pelabelan Sentimen

Komentar YouTube umumnya tersaji dalam bentuk teks yang tidak memiliki pola penulisan yang seragam. Kondisi tersebut dapat terlihat dari penggunaan kapitalisasi, istilah nonstandar, bentuk singkatan, simbol, maupun perbedaan cara penulisan. Karakteristik tersebut menyebabkan teks perlu melalui tahap preprocessing agar bentuk data menjadi lebih seragam sebelum digunakan dalam pembobotan dan proses klasifikasi [12], [27]. Tahapan preprocessing meliputi pembersihan data, case folding, tokenizing, normalisasi, stopword removal, dan stemming [25]. Pembersihan bertujuan menyingkirkan komponen yang tidak diperlukan, meliputi URL, simbol, angka, penyebutan akun, serta tagar. Case folding menyeragamkan kapitalisasi seluruh teks menjadi huruf kecil, sementara tokenizing menguraikan teks menjadi bagian-bagian berupa kata. Tahap normalisasi dilakukan dengan memetakan kata yang memiliki bentuk penulisan berbeda, termasuk istilah nonbaku, singkatan, dan kesalahan pengetikan, ke bentuk yang telah ditetapkan [12]. Selanjutnya, stopword removal menyaring kata-kata yang frekuensi penggunaannya tinggi tetapi memiliki kontribusi informasi yang rendah terhadap klasifikasi [27]. Tahap terakhir, stemming, mereduksi kata berimbuhan menjadi bentuk asalnya dengan memanfaatkan pustaka Sastrawi untuk pengolahan teks berbahasa Indonesia [28], [29].

Data komentar yang telah melalui preprocessing selanjutnya dikategorikan berdasarkan polaritas sentimennya ke dalam tiga kelompok, yakni positif, negatif, dan netral. Penentuan kategori dilakukan dengan memanfaatkan InSet Lexicon sebagai acuan untuk mengidentifikasi kecenderungan sentimen berdasarkan kata-kata yang terkandung dalam komentar (Muzaki, Ramadhani, and Kasih 2024), [30]. Label yang diperoleh

Jurnal Sains Dan Teknologi Informasi
Vol. 5, No. 4, September 2026, Hal 189-198
ISSN 2809-610X (Media Online)
DOI 10.47065/jussi.v5i4.11226
Website https://ejurnal.seminar-id.com/index.php/jussi

selanjutnya diperiksa kembali oleh ahli bahasa guna menilai keselarasan antara hasil pelabelan dan konteks yang terdapat pada masing-masing komentar.

2.4 Data Transformation

Teks yang telah melewati tahap preprocessing kemudian dikonversi menjadi representasi berbentuk angka dengan menerapkan Term Frequency-Inverse Document Frequency (TF-IDF). Pendekatan ini digunakan untuk menetapkan bobot terhadap setiap term dengan mempertimbangkan seberapa banyak term tersebut muncul dalam sebuah dokumen dan seberapa luas penyebarannya di seluruh kumpulan dokumen. Dengan demikian, term yang memiliki frekuensi tinggi pada suatu dokumen, tetapi hanya dijumpai secara terbatas pada dokumen lain, akan memiliki nilai bobot yang lebih besar. Hasil representasi tersebut kemudian dimanfaatkan sebagai fitur masukan untuk tahapan klasifikasi [31]. Perhitungan nilai TF-IDF pada dasarnya merupakan hasil pengalian antara Term Frequency (TF) dengan Inverse Document Frequency (IDF), yang dirumuskan sebagai:

$$\text{TF} = tf_{(t,d)} = tf_{(t,d)} \tag{1}$$

$$\text{IDF} = idf_{(t)} = \log\left(\frac{N}{df_{(t)}}\right) \tag{2}$$

$$\text{TF} - \text{IDF} = tf - idf_{(t,d)} = tf_{(t,d)} * idf_{(t)} \tag{3}$$

Dengan $tf_{(t,d)}$ merepresentasikan seberapa sering suatu term (t) ditemukan dalam dokumen (d), sedangkan $idf_{(t)}$ menunjukkan tingkat keunikan term dalam keseluruhan dokumen. Hasil pembobotan TF-IDF menghasilkan matriks dalam bentuk nilai numerik yang kemudian dimanfaatkan sebagai atribut masukan bagi algoritma Multinomial Naïve Bayes dalam melakukan klasifikasi.

2.5 Data Mining

Pada tahap data mining, komentar yang telah direpresentasikan dalam bentuk fitur numerik diproses dengan algoritma Multinomial Naïve Bayes untuk menentukan kategori sentimennya, yang terdiri atas positif, negatif, dan netral. Metode ini termasuk pendekatan klasifikasi berbasis probabilitas yang menerapkan Teorema Bayes dengan menganggap setiap fitur memiliki keterkaitan yang independen satu sama lain [32]. Pemilihan Multinomial Naïve Bayes didasarkan pada karakteristiknya yang sesuai untuk pengolahan dan pengelompokan data berbentuk teks. Algoritma ini juga telah digunakan dalam sejumlah penelitian analisis sentimen dengan memanfaatkan TF-IDF sebagai representasi fitur [17], [22], [30]. Perhitungan probabilitas dalam metode tersebut berlandaskan pada Teorema Bayes yang dapat dinyatakan melalui persamaan berikut:

$$P(X \mid Y) = \frac{P(Y \mid X) \cdot P(X)}{P(Y)} \tag{4}$$

Dengan P(X|Y) merupakan probabilitas hipotesis X berdasarkan data Y, P(Y|X) merupakan probabilitas data Y apabila X benar, P(X) merupakan probabilitas awal hipotesis X, dan P(Y) merupakan probabilitas data Y. Dalam penelitian ini, fitur hasil pembobotan TF-IDF digunakan sebagai masukan model. Penggunaan TF-IDF bersama Multinomial Naïve Bayes memiliki landasan empiris pada penelitian sebelumnya, antara lain Ernayanti (2023) dan Gerliandeva et al. (2024) yang menerapkan kombinasi tersebut pada klasifikasi sentimen berbasis teks [17]. Dataset kemudian dibagi menjadi data training dan testing. Data training digunakan untuk membangun model, sedangkan data testing digunakan untuk mengukur kemampuan model dalam mengklasifikasikan komentar yang belum digunakan selama proses pelatihan. Tiga skenario pembagian data digunakan dalam penelitian ini, yaitu 80:20, 75:25, dan 70:30.

2.6 Evaluation

Kinerja model klasifikasi dinilai dengan membandingkan hasil keluaran prediksi terhadap label sebenarnya menggunakan confusion matrix. Penyajian dalam bentuk matriks tersebut memungkinkan identifikasi terhadap prediksi yang tepat maupun yang keliru, sekaligus memperlihatkan distribusi kesalahan yang muncul pada setiap kategori sentimen. Informasi dari confusion matrix kemudian digunakan dalam menghitung accuracy, precision, recall, dan F1-score sebagai ukuran untuk melihat performa model dalam membedakan kelas sentimen [25], [8].

Accuracy merepresentasikan tingkat keberhasilan model dalam memberikan prediksi yang sesuai terhadap seluruh data pengujian. Precision menggambarkan tingkat ketepatan prediksi model ketika menetapkan suatu data ke kelas tertentu, sedangkan recall menunjukkan sejauh mana model mampu menemukan data yang memang berasal dari kelas tersebut [33], [14]. Adapun F1-score diperoleh dari nilai rata-rata antara precision dan recall, sehingga kedua aspek tersebut dapat dipertimbangkan secara bersamaan dalam menilai hasil klasifikasi. Rata-rata yang dimaksud adalah rata-rata harmonik, yaitu dua kali perkalian precision dan recall dibagi jumlah keduanya [25].

Jurnal Sains Dan Teknologi Informasi
Vol. 5, No. 4, September 2026, Hal 189-198
ISSN 2809-610X (Media Online)
DOI 10.47065/jussi.v5i4.11226
Website https://ejurnal.seminar-id.com/index.php/jussi

3. HASIL DAN PEMBAHASAN

3.1 Hasil Data Selection

Tahap data selection bertujuan menyaring komentar yang memiliki keterkaitan dengan fokus penelitian. Data diperoleh melalui YouTube Data API v3 dengan mengambil komentar dari video pada kanal Timnas Indonesia dan RCTI Sports yang membahas Kualifikasi Piala Dunia 2026. Dari proses pengambilan tersebut diperoleh 18.110 komentar sebagai data awal. Data yang terkumpul kemudian diperiksa berdasarkan kriteria seleksi yang telah ditentukan untuk memisahkan komentar yang tidak relevan atau tidak layak digunakan dalam proses analisis. Penyaringan dilakukan terhadap komentar yang mengandung unsur SARA, rasisme, penghinaan terhadap agama atau kelompok tertentu, pembahasan di luar objek penelitian, serta komentar yang memuat unsur vulgar.

Dari proses tersebut diperoleh 17.752 komentar yang memenuhi kriteria penelitian. Pemeriksaan berikutnya dilakukan pada keluaran preprocessing untuk memastikan data dapat diproses pada tahapan selanjutnya. Dari pemeriksaan tersebut, 348 komentar dieliminasi karena mengalami kegagalan pada proses stemming atau tidak memuat informasi yang memadai untuk kebutuhan klasifikasi. Dengan demikian, dataset yang digunakan hingga tahap analisis berjumlah 17.404 komentar. Data tersebut selanjutnya digunakan sebagai dasar untuk tahapan pelabelan sentimen, transformasi data, dan klasifikasi menggunakan Multinomial Naïve Bayes.

3.2 Hasil Preprocessing dan Pelabelan Sentimen

Data yang telah lolos tahap seleksi kemudian diproses melalui rangkaian text preprocessing yang mencakup cleaning, case folding, tokenizing, normalisasi, stopword removal, dan stemming. Rangkaian tersebut diterapkan secara bertahap dengan tujuan menyederhanakan isi teks dari komponen yang kurang diperlukan sekaligus menyelaraskan bentuk kata, sehingga data memiliki format yang lebih sesuai untuk tahap pembobotan dan klasifikasi. Proses diawali dengan cleaning untuk menyingkirkan komponen tertentu yang tidak diperlukan dalam analisis, seperti URL, angka, tanda baca, dan karakter khusus. Tahap case folding kemudian menyeragamkan penggunaan huruf dalam teks ke dalam satu bentuk. Komentar yang telah diseragamkan selanjutnya diuraikan melalui tokenizing menjadi unit-unit kata.

Normalisasi diterapkan untuk menyelaraskan kata yang ditulis secara tidak baku maupun singkatan agar mengikuti bentuk yang telah ditentukan. Tahap berikutnya adalah stopword removal, yaitu penyaringan kata yang memiliki kontribusi informasi rendah terhadap klasifikasi, kemudian stemming diterapkan dengan mereduksi kata berimbuhan ke bentuk dasarnya. Komentar yang telah selesai telah diproses kemudian dikelompokkan ke dalam tiga kelas sentimen, yaitu positif, negatif, dan netral. Pelabelan dilakukan menggunakan InSet Lexicon dan selanjutnya divalidasi oleh ahli bahasa terhadap 1.068 komentar untuk memastikan kesesuaian label dengan konteks komentar. Dari 18.110 komentar yang diperoleh pada tahap pengumpulan data, setelah melalui proses seleksi dan preprocessing diperoleh 17.404 komentar yang digunakan sebagai dataset akhir yang sudah dibersihkan.

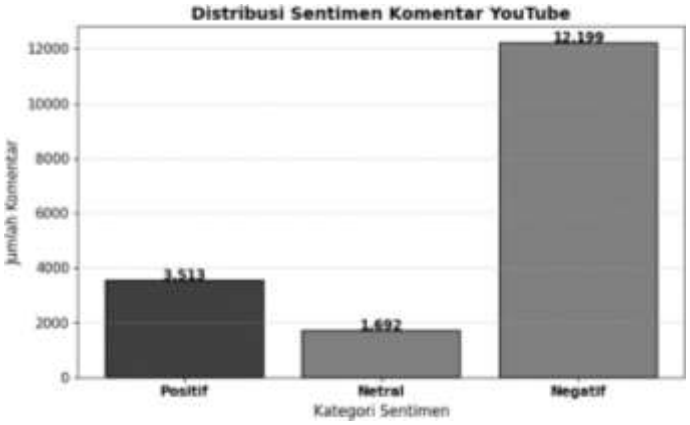
Sebagai gambaran awal mengenai karakteristik teks dalam dataset, komentar YouTube yang telah melalui proses pembersihan kemudian divisualisasikan menggunakan Wordcloud. Visualisasi tersebut digunakan untuk melihat kata-kata yang memiliki frekuensi kemunculan tinggi dalam kumpulan komentar.



Gambar 2. Wordcloud Komentar YouTube

Gambar 2 memperlihatkan distribusi kata yang muncul dalam komentar YouTube setelah melalui proses pembersihan dan penyeragaman teks. Perbedaan ukuran kata menunjukkan tingkat frekuensi kemunculannya, sehingga kata dengan ukuran lebih besar memiliki frekuensi yang lebih tinggi dibandingkan kata lainnya. Visualisasi ini memberikan gambaran awal mengenai kata-kata yang dominan dalam dataset. Hasil pelabelan sentimen pada komentar YouTube kemudian divisualisasikan untuk melihat distribusi masing-masing kelas sentimen.

Jurnal Sains Dan Teknologi Informasi
Vol. 5, No. 4, September 2026, Hal 189-198
ISSN 2809-610X (Media Online)
DOI 10.47065/jussi.v5i4.11226
Website https://ejurnal.seminar-id.com/index.php/jussi



Gambar 3. Distribusi Kelas Sentimen

Gambar 3 memperlihatkan bahwa kelas negatif terdiri atas 12.199 komentar, sedangkan kelas positif berjumlah 3.513 komentar dan kelas netral sebanyak 1.692 komentar. Jumlah tersebut menunjukkan bahwa komentar dengan sentimen negatif lebih banyak ditemukan dibandingkan komentar pada dua kategori lainnya.

3.3 Hasil Data Transformation

Teks yang telah selesai melalui tahap preprocessing dan pelabelan selanjutnya diubah ke dalam representasi numerik dengan menerapkan Term Frequency-Inverse Document Frequency (TF-IDF). Pengubahan bentuk data ini diperlukan karena algoritma klasifikasi melakukan proses pembelajaran berdasarkan fitur yang dapat direpresentasikan dalam nilai numerik dan diolah secara matematis. Pemberian bobot TF-IDF dilakukan dengan mempertimbangkan frekuensi suatu term pada sebuah dokumen serta persebarannya di dalam kumpulan dokumen.

Term yang kemunculannya relatif tinggi pada satu dokumen, tetapi tidak banyak dijumpai pada dokumen lainnya, akan memperoleh bobot yang lebih besar. Sebaliknya, term yang tersebar pada banyak dokumen cenderung menghasilkan bobot yang lebih kecil karena daya pembeda terhadap karakteristik dokumen menjadi rendah. Proses tersebut menghasilkan matriks fitur TF-IDF yang kemudian diteruskan ke tahap data mining. Setiap komentar yang semula berupa rangkaian teks akhirnya direpresentasikan dalam bentuk nilai numerik sehingga dapat diterima dan diolah oleh algoritma Multinomial Naïve Bayes. Hasil transformasi TF-IDF ditampilkan melalui 20 kata dengan bobot TF-IDF tertinggi sebagai gambaran fitur yang terbentuk.

```
20 Kata Dengan Bobot TF-IDF Tertinggi
main         0.045856
indonesia    0.034743
latih        0.033982
kluivert     0.033215
patrick      0.032365
yong         0.030475
tae          0.030349
shin         0.030318
tim          0.030102
wajah        0.029446
nasional     0.027091
menang       0.022915
mata         0.022080
kalah        0.021610
piala        0.020685
gembira      0.020663
air          0.020518
erick        0.020346
thohir       0.020256
urai         0.020200
dtype: float64
```

Gambar 4. Representasi Matriks TF-IDF

Gambar 4 memperlihatkan kata “main” memiliki bobot TF-IDF tertinggi sebesar 0,045856, diikuti oleh “indonesia” sebesar 0,034743 dan “latih” sebesar 0,033982. Kata-kata tersebut memiliki bobot yang lebih tinggi

Jurnal Sains Dan Teknologi Informasi
Vol. 5, No. 4, September 2026, Hal 189-198
ISSN 2809-610X (Media Online)
DOI 10.47065/jussi.v5i4.11226
Website https://ejurnal.seminar-id.com/index.php/jussi

dibandingkan kata lainnya dalam daftar, sehingga menjadi bagian dari fitur yang terbentuk melalui proses transformasi TF-IDF.

3.4 Hasil Klasifikasi

Pada tahap data mining, model dibangun menggunakan algoritma Multinomial Naïve Bayes untuk menentukan kategori sentimen pada setiap komentar. Klasifikasi dilakukan terhadap tiga kategori, yakni positif, negatif, dan netral. Kinerja model kemudian diuji melalui tiga skenario proporsi data training dan testing, yaitu 80:20, 75:25, serta 70:30. Untuk membandingkan hasil dari ketiga skenario pembagian data tersebut, nilai accuracy yang diperoleh disajikan pada Tabel 1.

Tabel 1. Perbandingan Hasil Klasifikasi

Split Ratio	Data Training	Data Testing	Accuracy
80:20	13.923	3.481	79,00%
75:25	13.053	4.351	78,81%
70:30	12.182	5.222	79,09%

Ketiga skenario menghasilkan nilai accuracy yang relatif berdekatan. Skenario 80:20 memperoleh accuracy sebesar 79,00%, sedangkan skenario 75:25 menghasilkan 78,81%. Sementara itu, pembagian data 70:30 menghasilkan nilai accuracy sebesar 79,09%, sehingga menjadi skenario dengan nilai accuracy tertinggi dalam penelitian. Perbandingan pada Tabel 1 menunjukkan bahwa selisih accuracy antar-skenario relatif kecil. Skenario 70:30 kemudian digunakan sebagai dasar evaluasi lanjutan karena menghasilkan accuracy sebesar 79,09%. Selanjutnya, hasil klasifikasi divisualisasikan berdasarkan masing-masing kategori sentimen untuk melihat karakteristik kata yang muncul pada komentar.



Gambar 5. Wordcloud Berdasarkan Tiga Kategori Sentimen

Visualisasi pada Gambar 5 dari ketiga kategori tersebut memperlihatkan adanya perbedaan kata yang dominan pada masing-masing sentimen. Perbedaan karakteristik tersebut dapat digunakan sebagai pendukung dalam menginterpretasikan hasil pengelompokan komentar ke dalam kelas positif, netral, dan negatif.

3.5 Evaluasi Model

Penilaian terhadap model dilakukan untuk melihat sejauh mana Multinomial Naïve Bayes mampu membedakan setiap kategori sentimen yang digunakan dalam penelitian. Pengukuran kinerja mencakup precision, recall, F1-score, accuracy, serta confusion matrix. Dari ketiga skenario pembagian data yang diuji, proporsi 70:30 memperoleh nilai accuracy paling tinggi sehingga pembahasan performa model selanjutnya mengacu pada hasil pengujian dengan skenario tersebut. Hasil pengujian disajikan melalui classification report pada Tabel 2.

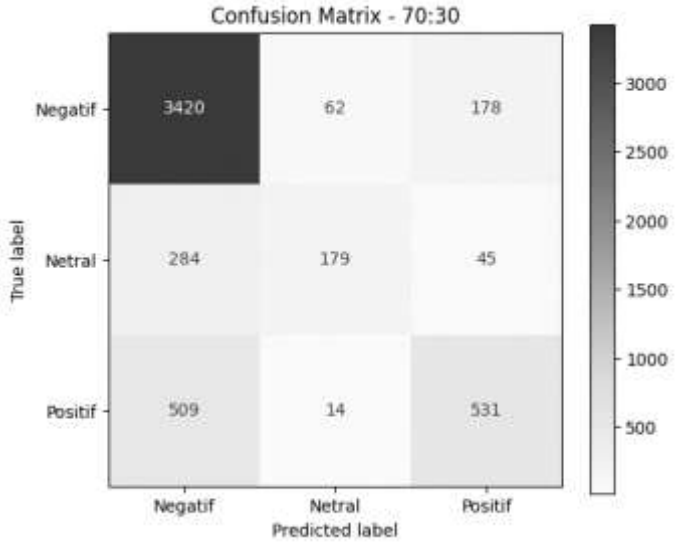
Tabel 2. Classification Report Skenario 70:30

Kelas	Precision	Recall	F1-Score	Support
Negatif	0,81	0,93	0,87	3.660

Jurnal Sains Dan Teknologi Informasi
Vol. 5, No. 4, September 2026, Hal 189-198
ISSN 2809-610X (Media Online)
DOI 10.47065/jussi.v5i4.11226
Website https://ejurnal.seminar-id.com/index.php/jussi

Kelas	Precision	Recall	F1-Score	Support
Netral	0,70	0,35	0,47	508
Positif	0,70	0,50	0,59	1.054
Accuracy 79,09%				

Tabel 2 memperlihatkan bahwa kelas negatif memperoleh precision sebesar 0,81, recall sebesar 0,93, dan F1-score sebesar 0,87. Sementara itu, kelas netral memiliki F1-score sebesar 0,47 dan kelas positif sebesar 0,59. Perbedaan nilai tersebut berkaitan dengan jumlah data pada masing-masing kelas, di mana kelas negatif memiliki jumlah data yang lebih besar dibandingkan kelas netral dan positif. Untuk melihat hasil prediksi model pada setiap kelas secara lebih rinci, distribusi prediksi divisualisasikan menggunakan confusion matrix.



Gambar 6. Confusion Matrix Skenario 70:30

Gambar 6 memperlihatkan perbandingan antara label aktual dan hasil prediksi model pada masing-masing kategori sentimen. Confusion matrix menunjukkan jumlah data pada setiap kombinasi kelas aktual dan kelas hasil prediksi, sehingga dapat terlihat penyebaran hasil prediksi pada kelas positif, netral, dan negatif.

3.6 Pembahasan

Hasil penelitian menunjukkan bahwa Multinomial Naïve Bayes menghasilkan nilai accuracy yang relatif stabil pada ketiga skenario pembagian data. Perbedaan nilai accuracy antar-skenario tergolong kecil, dengan nilai tertinggi sebesar 79,09% pada pembagian data 70:30. Hasil tersebut menunjukkan bahwa perubahan proporsi data training dan testing memberikan perbedaan yang relatif terbatas terhadap hasil klasifikasi yang diperoleh. Berdasarkan Tabel 2, variasi hasil evaluasi pada masing-masing kelas berkaitan dengan komposisi sentimen dalam dataset. Dari total 17.404 komentar, sebanyak 12.199 komentar atau 70,09% termasuk dalam kelas negatif. Jumlah tersebut menjadikan kelas negatif sebagai kategori dengan representasi terbesar dalam dataset. Kondisi tersebut memberikan lebih banyak contoh bagi model untuk mempelajari karakteristik kata dan pola yang berkaitan dengan sentimen negatif. Hal ini tercermin dari nilai recall sebesar 0,93 dan F1-score sebesar 0,87 pada kelas negatif. Sementara itu, kelas netral dan positif memiliki jumlah data yang lebih terbatas sehingga karakteristik kedua kelas tersebut direpresentasikan oleh jumlah contoh yang lebih sedikit dalam proses pembelajaran.

Gambaran hasil prediksi tersebut dapat dilihat pada confusion matrix pada Gambar 6. Model berhasil mengklasifikasikan 3.420 komentar negatif, 179 komentar netral, dan 531 komentar positif sesuai dengan kelas aktualnya. Selain itu, terdapat 284 komentar netral dan 509 komentar positif yang memperoleh prediksi sebagai kelas negatif. Pola tersebut menunjukkan bahwa dominasi jumlah data pada kelas negatif turut membentuk kecenderungan hasil prediksi, karena model memperoleh lebih banyak representasi untuk mempelajari karakteristik kelas tersebut selama proses pelatihan.

Selain distribusi kelas, karakteristik bahasa pada komentar YouTube juga menjadi bagian yang perlu diperhatikan dalam memahami hasil klasifikasi. Data komentar yang digunakan memiliki bentuk bahasa yang beragam, seperti penggunaan kata tidak baku, singkatan, typo, serta ungkapan yang bergantung pada konteks pembicaraan. Variasi tersebut dapat membuat suatu kata memiliki makna yang berbeda ketika digunakan dalam konteks tertentu. Dalam penelitian ini, TF-IDF merepresentasikan teks berdasarkan bobot kemunculan term, sehingga informasi kontekstual yang tidak secara langsung tercermin melalui bobot kata menjadi salah satu aspek yang perlu dipertimbangkan dalam interpretasi hasil klasifikasi.

Jurnal Sains Dan Teknologi Informasi
Vol. 5, No. 4, September 2026, Hal 189-198
ISSN 2809-610X (Media Online)
DOI 10.47065/jussi.v5i4.11226
Website https://ejurnal.seminar-id.com/index.php/jussi

Secara keseluruhan, penerapan KDD, TF-IDF, dan Multinomial Naïve Bayes menghasilkan proses klasifikasi yang dapat mengelompokkan komentar YouTube ke dalam tiga kategori sentimen, yaitu positif, negatif, dan netral. Nilai accuracy sebesar 79,09% pada skenario 70:30 menunjukkan hasil klasifikasi yang diperoleh dari dataset dengan karakteristik bahasa informal dan distribusi kelas yang tidak merata. Oleh karena itu, interpretasi hasil penelitian tidak hanya mempertimbangkan nilai accuracy, tetapi juga memperhatikan distribusi data serta karakteristik hasil pada masing-masing kelas sentimen.

Untuk pengembangan penelitian selanjutnya, beberapa pendekatan dapat dipertimbangkan untuk mengoptimalkan klasifikasi pada dataset dengan distribusi kelas yang tidak merata. Alternatif yang dapat digunakan antara lain Synthetic Minority Over-sampling Technique (SMOTE), Cost-Sensitive Learning, serta Complement Naïve Bayes (CNB). SMOTE dapat digunakan untuk meningkatkan representasi kelas dengan jumlah data yang lebih sedikit, sedangkan Cost-Sensitive Learning dapat memberikan bobot yang berbeda pada setiap kelas selama proses pembelajaran. Sementara itu, Complement Naïve Bayes merupakan salah satu varian Naïve Bayes yang dapat diterapkan pada klasifikasi teks dan menjadi alternatif untuk menguji pendekatan klasifikasi pada kondisi distribusi kelas yang tidak seimbang. Penggunaan beberapa pendekatan tersebut dapat menjadi arah pengembangan penelitian berikutnya untuk melihat pengaruh strategi penanganan distribusi kelas dan penggunaan varian algoritma terhadap hasil analisis sentimen.

4. KESIMPULAN

Penelitian analisis sentimen komentar YouTube pada Kualifikasi Piala Dunia 2026 Tim Nasional Indonesia dengan pendekatan Knowledge Discovery in Database, pembobotan TF-IDF, dan algoritma Multinomial Naïve Bayes menghasilkan klasifikasi ke dalam tiga kategori sentimen, yaitu positif, netral, dan negatif. Dari proses pengolahan data diperoleh 17.404 komentar yang dapat digunakan dalam proses analisis, dengan pengujian model dilakukan melalui tiga skenario pembagian data, yaitu 80:20, 75:25, dan 70:30. Hasil pengujian menunjukkan bahwa skenario 70:30 memberikan accuracy sebesar 79,09% dan menjadi hasil tertinggi dibandingkan skenario lainnya, sedangkan skenario 80:20 memperoleh accuracy sebesar 79,00% dan skenario 75:25 sebesar 78,81%. Pada skenario pembagian data 70:30, kelas negatif memperoleh precision sebesar 0,81, recall sebesar 0,93, dan F1-score sebesar 0,87, sedangkan kelas netral memperoleh F1-score sebesar 0,47 dan kelas positif sebesar 0,59. Berdasarkan hasil pelabelan sentimen, komentar negatif menjadi kategori dengan jumlah terbanyak, yaitu 12.199 komentar, diikuti komentar positif sebanyak 3.513 komentar dan komentar netral sebanyak 1.692 komentar. Komposisi tersebut menggambarkan distribusi sentimen yang tidak merata pada data komentar yang dianalisis. Secara keseluruhan, penerapan metode tersebut dapat digunakan untuk mengidentifikasi kecenderungan sentimen pada komentar YouTube terkait Tim Nasional Indonesia dan memberikan gambaran mengenai karakteristik opini yang muncul dalam dataset penelitian.

REFERENCES

[1] F. Lubis, K. B. Ayuningtyas, S. Rahmadani, and Z. Purba, "Penggunaan Metode Text Mining Untuk Mengekstrak Informasi Penting," *Motiv. Pendidik. dan Bhs.*, vol. 1, no. 4, 2023.

[2] S. F. Handayani, R. W. Pratiwi, and D. Intan, "Analisis Sentimen pada Data Ulasan Twitter dengan Long-Short Term Memory," vol. 7, no. 1, pp. 39–46, 2022, doi: 10.31544/jtera.v7.i1.2022.39-46.

[3] A. Muzaki, R. A. Ramadhani, and P. Kasih, "Analisis Sentimen Kemenangan Timnas U-23 Menggunakan Naive Bayes," vol. 8, pp. 902–911, 2024.

[4] E. A. Zuraida, A. R. Berto, and A. W. Utoyo, "Pemetaan Youtube sebagai Dasar Pengembangan Strategi Komunikasi Timnas U-23," vol. 6, no. 6, pp. 2765–2774, 2024.

[5] Y. C. A. Sanjaya, "Cetak Sejarah, Indonesia Lolos Putaran Keempat Kualifikasi Piala Dunia 2026," Kompas.com. Accessed: Sep. 30, 2026. [Online]. Available: https://www.kompas.com/tren/read/2025/06/06/013000965/cetak-sejarah-indonesia-lolos-putaran-keempat-kualifikasi-piala-dunia-2026?page=all

[6] A. Aliyudin, U. Siliwangi, and K. Tasikmalaya, "Analisis Sentimen Terhadap Piala Dunia U-20 Yang Batal Diselenggarakan Di Indonesia Pada Media Youtube Dengan Menggunakan Metode Naïve Bayes" vol. 8, no. 2, pp. 71–78, 2023.

[7] B. Franko, N. Wilyanto, H. Irsyad, U. Multi, and D. Palembang, "Analisis Sentimen Terhadap Naturalisasi Pemain pada Youtube Menggunakan Decision Tree dan Naive Bayes Sentiment Analysis of Player Naturalization on Youtube Using Decision Trees and Naive Bayes," vol. 03, no. September, pp. 8–16, 2024, doi: 10.57203/session.v3i1.2024.8-16.

[8] V. P. Prasetyo and K. Cimahi, "Platform Mengenai Pro Kontra Terhadap Naturalisasi Indonesia," vol. 13, no. 3.

[9] D. F. Zhafira, B. Rahayudi, U. Brawijaya, and P. Korespondensi, "Analisis Sentimen Kebijakan Kampus Merdeka Menggunakan Naive Bayes Dan Pembobotan TF-IDF Sentiment Anlaysis Of Kampus Merdeka Policy Using Naive Bayes And TF-IDF Term Weighting Based On Youtube," vol. 2, no. 1, pp. 55–63, 2021.

[10] R. Hidayat, M. Fikry, F. Yanto, and E. P. Cynthia, "Penerapan Naïve Bayes Classifier dalam Klasifikasi Sentimen Publik di Twitter terhadap Puan Maharani," vol. 6, pp. 100–108, 2024.

[11] A. P. Widyassari, D. Salsabilla, and M. A. Amrozi, "Analisis sentimen publik di twitter terhadap pelantikan presiden Prabowo menggunakan algoritma Naïve Bayes Analysis of public sentiment on twitter towards president Prabowo ' s

Jurnal Sains Dan Teknologi Informasi
Vol. 5, No. 4, September 2026, Hal 189-198
ISSN 2809-610X (Media Online)
DOI 10.47065/jussi.v5i4.11226
Website https://ejurnal.seminar-id.com/index.php/jussi

inauguration using the Naïve Bayes algorithm," vol. 10, no. 1, pp. 13–24, 2025.

[12] R. L. Nurdiansyah and K. E. Dewi, "Pengaruh Information Gain Dan Normalisasi Kata Pada Analisis Sentimen Berbasis Aspek Komputa : Jurnal Ilmiah Komputer dan Informatika," *KOMPUTA J. Ilm. Komput. dan Inform.*, vol. 12, no. 2, 2023.

[13] M. U. Albab, Y. K. P, and M. N. Fawaiq, "Optimization of the Stemming Technique on Text preprocessing President 3 Periods Topic," vol. 20, no. 2, pp. 1–10, 2023.

[14] B. Hakim, "Analisa Sentimen Data Text Preprocessing Pada Data Mining Dengan Menggunakan Machine Learning Data Text Pre-Processing Sentiment Analysis In Data Mining Using Machine Learning School of Computer Science and Technology , Harbin Institute of Technology," vol. 4, no. 2, pp. 16–22, 2021.

[15] R. R. Sani, Y. A. Pratiwi, S. Winarno, and E. D. Udayanti, "Analisis Perbandingan Algoritma Naive Bayes Classifier dan Support Vector Machine untuk Klasifikasi Hoax pada Berita Online Indonesia," vol. 13, no. 2, pp. 85–98, 2022.

[16] T. Ernayanti, "Penggunaan seleksi fitur chi-square dan algoritma multinomial naïve bayes untuk analisis sentimen pelangggan tokopedia 1,2,3," vol. 11, no. 1, pp. 562–571, 2023, doi: 10.14710/j.gauss.11.4.562-571.

[17] A. Gerliandeva, Y. H. Chrisnanto, and H. Ashaury, "Optimasi Klasifikasi Sentimen pada Komentar Online menggunakan Multinomial Naïve Bayes dan Ekstraksi Fitur TF-IDF serta N-grams Optimization of Sentiment Classification on Online Comments using Multinomial Naïve Bayes and TF-IDF Feature Extraction and N-grams," no. X, pp. 259–272, 2024, doi: 10.56873/jpkm.v9i2.5585.

[18] W. Mega, P. Dhuhita, F. Zone, P. S. Informatika, F. I. Komputer, and U. A. Yogyakarta, "Perbandingan Kinerja Algoritma Naïve Bayes Model Multinomial dan Bernoulli pada Pengklasifikasian Komentar Cyberbullying Performance Comparison of Naive Bayes Multinomial and Bernoulli Models in Cyberbullying Comment Classification," vol. 12, no. 148, pp. 109–117, 2023, doi: 10.34010/komputika.v12i2.9767.

[19] K. M. Suryaningrum, "Comparison of the TF-IDF Method with the Count Vectorizer to Classify Hate Speech," vol. 5, no. 2, pp. 79–83, 2023, doi: 10.21512/emacsjournal.v5i2.9978.

[20] I. Widaningrum, D. Mustikasari, R. Arifin, S. L. Tsaqila, and D. Fatmawati, "Algoritma Term Frequency – Inverse Document Frequency Clustering Untuk Menentukan Kategori Dokumen," 2023.

[21] R. Suryanti and P. T. Prasetyaningrum, "Perbandingan Metode TF-IDF dan Bag of Words dalam Analisis Sentimen Diet Kopi Americano di Media Sosial Twitter Menggunakan Naïve Bayes," vol. 7, no. 1, 2025, doi: 10.47065/bits.v7i1.7244.

[22] I. Imelda and A. R. Kurnianto, "Naïve Bayes and TF-IDF for Sentiment Analysis of the Covid-19 Booster Vaccine," vol. 5, no. 158, pp. 1–2, 2026.

[23] D. Muriyatmoko, A. Musthafa, and Y. Al Banna, "Sentiment Analysis of YouTube Comments on Indonesia ' s Performance in 2026 World Cup Qualifiers Using Naive Bayes," no. November 2025, 2026.

[24] A. A. Baskara, N. M. Piranti, and M. F. Romdendine, "," vol. 9, no. 3, pp. 4886–4895, 2025. Framework Data Mining: Sebuah Survei

[25] A. H. Nurdy and A. Rahim, "Analisis Sentimen Ulasan Game Stumble Guys Pada Playstore Menggunakan Algoritma Naïve Bayes Sentiment Analysis of Stumble Guys Game Reviews on Playstore Using the Naïve Bayes Algorithm," vol. 13, no. November, pp. 388–395, 2024, doi: 10.34148/teknika.v13i3.993.

[26] A. Fahlevi, "Proses Data Mining KDD," Binus University, School of Information System. Accessed: Dec. 11, 2025. [Online]. Available: https://sis.binus.ac.id/2021/09/30/proses-data-mining-kdd/

[27] S. J. Angelina, A. Bijaksana, P. Negara, and H. Muhardi, "Analisis Pengaruh Penerapan Stopword Removal Pada Performa Klasifikasi Sentimen Tweet Bahasa Indonesia Analyzing The Impact Of Applying Stopword Removal On Indonesian Tweet Sentiment Classification," vol. 02, no. 1, pp. 165–173, 2023, doi: 10.26418/juara.v2i1.69680.

[28] A. Nafi, A. Tri, J. Harjanta, B. A. Herlambang, and S. Fahmi, "Analisis Sentimen Review Pelanggan Lazada dengan Sastrawi Stemmer dan SVM-PSO untuk Memahami Respon Pengguna," no. 204, pp. 330–339, 2024.

[29] M. Rosidin, M. F. Gustafi, and S. A. Pratiwi, "Jurnal Mantik Optimizing nazief adriani ' s stemmer algorithm in detecting indonesian word errors using sastrawi," vol. 7, no. 3, 2023.

[30] R. B. Aprianto, A. Eviyanti, H. Setiawan, I. Ratna, and I. Astutik, "Dedi Mulyadi Dengan Metode Machine Learning (Studi Kasus: Kenakalan Remaja)," vol. 10, no. 2, pp. 2997–3004, 2026.

[31] A. H. Dani, E. Y. Puspaningrum, R. Mumpuni, U. Pembangunan, N. Veteran, and J. Timur, "Studi Performa TF-IDF dan Word2Vec Pada Analisis Sentimen Cyberbullying," no. 2, 2024.

[32] N. A. Sivi, I. Mualim, and M. T. Kussofyan, "Penerapan Algoritma Naïve Bayes untuk Analisis Sentimen Ulasan Produk," 2024.

[33] T. Adrian and N. Suarna, "Implementasi Data Mining Untuk Mengklasifikasi Hasil Kelulusan Madrasah Menggunakan Algoritma Naïve Bayes," vol. 8, no. 1, pp. 1000–1004, 2024.