Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 223–232
DOI : 10.47065/bulletinds.v5i3.10129
https://ejurnal.seminar-id.com/index.php/bulletinds

Analisis Sentimen Komentar YouTube Terkait Indonesia Tidak Lolos Kualifikasi Piala Dunia Menggunakan Support Vector Machine

Idris Syahrudin[1,*], Sufajar Butsianto[2], Andini Putri Riandani[3]

[1] Fakultas Teknik, Program Studi Teknik Informatika, Universitas Pelita Bangsa, Kota Bekasi, Indonesia
Jl. Inspeksi Kalimalang Tegal Danas, Cikarang Pusat, Kab. Bekasi - Jawa Barat, 17530., Indonesia
[2] Fakultas Teknologi, Program Studi Teknologi Hasil Pertanian, Univeristas Pelita Bangsa, Kota Bekasi, Indonesia
Jl. Inspeksi Kalimalang Tegal Danas, Cikarang Pusat, Kab. Bekasi - Jawa Barat, 17530., Indonesia
Email: [1,*]idrissyahrudin15552@gmail.com, [2]sufajar@pelitabangsa.ac.id, [3]andiniriandani@pelitabangsa.ac.id
Email Penulis Korespondensi: idrissyahrudin15552@gmail.com

Abstrak−Penelitian ini bertujuan untuk menganalisis sentimen opini publik pada media sosial YouTube terkait Tim Nasional Indonesia tidak lolos dalam kualifikasi Piala Dunia. Fokus utama penelitian ini adalah mengklasifikasikan komentar pengguna ke dalam tiga kategori, yaitu positif, negatif, dan netral, untuk memahami pola reaksi masyarakat terhadap performa tim nasional yang di mana komentar-komentar yang muncul dapat dianalisis untuk memahami kecenderungan sentimen dan pola interaksi pengguna. Metodologi yang digunakan mengikuti kerangka Knowledge Discovery in Databases. Data sebanyak 15.169 komentar dikumpulkan melalui proses crawling menggunakan YouTube Data API v3. Data tersebut kemudian melalui tahap pra-pemrosesan yang intensif, meliputi case folding, cleaning, tokenisasi, normalisasi kata, dan stopword removal, sehingga menghasilkan 7.584 data yang siap untuk diklasifikasikan. Algoritma Support Vector Machine dengan kernel linear diterapkan pada dataset yang dibagi dengan rasio 80:20 antara data latih dan data uji. Hasil penelitian menunjukkan bahwa mayoritas opini masyarakat didominasi oleh sentimen negatif (44,03%) dibandingkan sentimen netral (31,34%) dan positif (24,62%). Pengujian model SVM menghasilkan tingkat akurasi sebesar 80,43% dengan nilai weighted F1-score sebesar 0,81. Hal ini menunjukkan bahwa metode SVM sangat efektif dalam menangani data teks yang tidak terstruktur dan noisy komentar media sosial YouTube. Penelitian ini membuktikan keandalan algoritma SVM dalam analisis sentimen bersekala besar. Penelitian ini memberikan kontribusi dalam pemanfaatan komentar YouTube sebagai sumber data untuk memetakan opini publik terhadap isu olahraga nasional secara otomatis menggunakan Support Vector Machine. Selain itu, penelitian ini memperluas penerapan analisis sentimen berbahasa Indonesia pada data komentar yang bersifat informal dan noisy, sehingga dapat menjadi referensi bagi penelitian sejenis pada konteks media sosial dan isu publik lainnya.

Kata Kunci: Analisis Sentimen; YouTube; Support Vector Machine; Machine Learning; Timnas Indonesia

Abstract−This study aims to analyze public opinion sentiment on YouTube regarding the Indonesian National Team's failure to qualify for the World Cup. The main focus of this study is to classify user comments into three categories: positive, negative, and neutral, to understand the pattern of public reaction to the national team's performance. The comments can then be analyzed to understand sentiment tendencies and user interaction patterns. The methodology used follows the Knowledge Discovery in Databases framework. Data of 15,169 comments were collected through a crawling process using the YouTube Data API v3. The data then underwent an intensive pre-processing stage, including case folding, cleaning, tokenization, word normalization, and stopword removal, resulting in 7,584 data points ready for classification. A Support Vector Machine algorithm with a linear kernel was applied to the dataset, which was divided with an 80:20 ratio between training and test data. The results showed that the majority of public opinion was dominated by negative sentiment (44.03%) compared to neutral sentiment (31.34%) and positive sentiment (24.62%). The SVM model test yielded an accuracy rate of 80.43% with a weighted F1-score of 0.81. This indicates that the SVM method is highly effective in handling unstructured and noisy text data from YouTube social media comments. This study proves the reliability of the SVM algorithm in large-scale sentiment analysis. This study contributes to the use of YouTube comments as a data source to automatically map public opinion on national sports issues using Support Vector Machines. In addition, this study expands the application of Indonesian-language sentiment analysis to informal and noisy comment data, thus serving as a reference for similar research in the context of social media and other public issues.

Keywords: Sentiment Analysis; YouTube; Support Vector Machine; Machine Learning; Indonesian National Team

1. PENDAHULUAN

Kemajuan teknologi informasi dan media sosial telah mengubah cara masyarakat memperoleh informasi, menyampaikan pendapat, dan mengekspresikan sikap terhadap suatu peristiwa. Salah satu platform yang paling banyak digunakan untuk interaksi publik adalah YouTube, karena menyediakan ruang komentar yang memungkinkan pengguna memberikan tanggapan secara langsung terhadap video yang ditonton. Dalam konteks analisis sentimen, komentar YouTube menjadi sumber data yang penting karena merepresentasikan opini publik secara alami, spontan, dan sering kali dipengaruhi oleh bahasa informal serta variasi penulisan yang beragam [1],[2],[3].

Berbagai penelitian terdahulu telah dilakukan untuk menganalisis sentimen pada media sosial dengan objek dan metode yang beragam. Salah satunya yaitu meneliti sentimen Tim Nasional Indonesia pada ajang Piala Dunia U-17 dengan algoritma Naive Bayes dan menunjukkan bahwa metode tersebut mampu mengklasifikasikan opini pengguna media sosial dengan cukup baik.[4]. Juga penelitan terkait membandingkan Naive Bayes dan Support Vector Machine pada analisis sentimen RUU Kesehatan di Twitter, namun fokus penelitian tersebut masih terbatas pada platform Twitter dan tidak membahas karakteristik komentar YouTube [5]. Sementara itu ada beberapa yang meneliti atau menerapkan Support Vector Machine pada analisis sentimen komentar YouTube pada channel beauty vlogger dan memperoleh hasil yang baik, tetapi objek penelitiannya bukan pada isu olahraga nasional sehingga karakteristik opini publik yang dianalisis masih berbeda [6]. Dan juga penelitian yang mengkaji komentar YouTube terkait pembangunan IKN, namun topik

© 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 223
This work is licensed under a Creative Commons Attribution 4.0 International License.

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 223−232
DOI : 10.47065/bulletinds.v5i3.10129
https://ejurnal.seminar-id.com/index.php/bulletinds

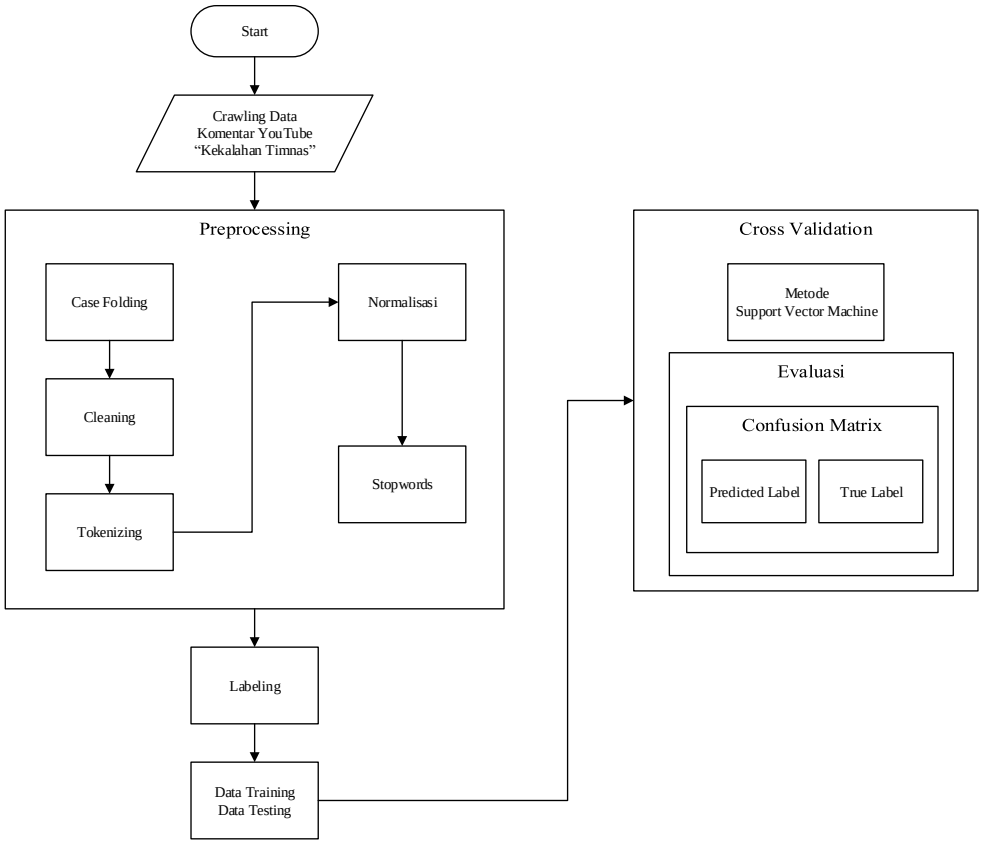
tersebut lebih berorientasi pada isu pembangunan nasional dan belum mengarah pada reaksi publik terhadap kegagalan Tim Nasional Indonesia [7].

Meskipun penelitian mengenai analisis sentimen di media sosial sudah cukup banyak, masih terdapat celah penelitian yang penting. Pertama, sebagian besar penelitian sebelumnya meneliti topik yang berbeda, sehingga belum menggambarkan dinamika opini publik pada isu kekalahan Tim Nasional Indonesia di kualifikasi Piala Dunia. Kedua, penelitian yang berfokus pada komentar YouTube masih jarang mengangkat isu olahraga nasional dengan karakter bahasa Indonesia yang cenderung informal, noisy, dan mengandung variasi slang yang tinggi. Ketiga, meskipun SVM telah terbukti efektif dalam beberapa studi sentimen berbahasa Indonesia, penerapannya pada komentar YouTube terkait kekalahan Tim Nasional Indonesia pada kualifikasi Piala Dunia 2026 belum banyak dieksplorasi secara spesifik. Dengan demikian, research gap pada penelitian ini terletak pada belum tersedianya kajian yang secara khusus menganalisis sentimen komentar YouTube terkait kekalahan Tim Nasional Indonesia pada kualifikasi Piala Dunia 2026 menggunakan Support Vector Machine untuk menangani data teks yang tidak terstruktur dan berpotensi tidak seimbang [8].

Penelitian ini bertujuan untuk menganalisis sentimen komentar pengguna YouTube terkait kekalahan Tim Nasional Indonesia pada kualifikasi Piala Dunia menggunakan metode Support Vector Machine. Penelitian ini diharapkan dapat memberikan gambaran yang lebih objektif mengenai kecenderungan opini publik serta menunjukkan efektivitas SVM dalam mengklasifikasikan komentar berbahasa Indonesia pada media sosial YouTube. Oleh karena itu, penelitian ini tidak hanya berfokus pada klasifikasi sentimen, tetapi juga pada pengisian kekosongan kajian mengenai analisis opini publik pada isu olahraga nasional berbasis komentar YouTube berbahasa Indonesia.

2. METODOLOGI PENELITIAN

Penelitian ini bersifat kuantitatif dengan fokus pada validitas dan reliabilitas model SVM terhadap data teks noisy khas komentar media sosial Indonesia, sehingga dihasilkan insight mengenai kecenderungan sentimen dominan (kekecewaan negatif vs dukungan tetap positif) kekalahan kualifikasi Timnas Indonesia yang mendukung analisis dinamika opini publik olahraga nasional seperti pada Gambar 1.



Gambar 1. Tahapan Penelitian

Alur dari metode Support Vector Machine yaitu proses analisis sentimen komentar YouTube terkait Kekalahan Timnas. Proses dimulai sebagai berikut;

1. Start
   Tahap awal sebagai titik dimulainya seluruh proses analisis, menandakan sistem siap menjalankan pipeline analisis sentimen.
2. Crawling Data

© 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 224
This work is licensed under a Creative Commons Attribution 4.0 International License.

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 223−232
DOI : 10.47065/bulletinds.v5i3.10129
https://ejurnal.seminar-id.com/index.php/bulletinds

Crawling data merupakan proses mengumpulkan data dari pangkalan data berbasis web dengan menggunakan teknik crawling dan pembersihan data yang mendukung pengolahan data berbasis sistem. Sistem crawling data melakukan crawling melalui berbagai situs web, menghasilkan berkas data yang siap digunakan. proses pengambilan data secara otomatis dari platform YouTube, Data berupa komentar pengguna yang mengandung opini, emosi, atau reaksi terhadap kekalahan timnas, Output tahap ini adalah dataset mentah (raw data) yang belum terstruktur dan masih mengandung noise.

3. Preprocessing
   a) Case Folding
      Pada tahap awal preprocessing teks, case folding mengubah semua huruf dalam dokumen teks menjadi huruf kecil yang sama. Ini membantu menyederhanakan dan memperbaiki konsistensi struktur teks sebelum analisis lebih lanjut [9]. Mengubah seluruh huruf menjadi huruf kecil, Bertujuan menghindari duplikasi kata (misalnya "Timnas" dan "timnas" dianggap sama).
   b) Cleaning
      Preprocessing yang dikenal sebagai pembersihan bertujuan untuk membersihkan data mentah dengan mengidentifikasi dan mengoreksi duplikasi, kesalahan, dan ketidakkonsistenan, data yang lebih bersih, konsisten, dan siap untuk analisis atau pemodelan lebih lanjut. Menghapus elemen yang tidak relevan seperti, tanda baca, angka, URL/link, emoji atau simbol tertentu, bertujuan mengurangi noise dalam data.
   c) Tokenizing
      Dalam proses preprocessing teks, tokenisasi memecah kalimat atau rangkaian kata menjadi bagian lebih kecil yang disebut token, biasanya kata tunggal. Ini menghasilkan daftar kata yang dapat diolah untuk klasifikasi, analisis sentimen, atau pemodelan bahasa natural (NLP) [10] .Memecah kalimat menjadi unit kata (token), "timnas bermain buruk" ["timnas", "bermain", "buruk"].
   d) Normalisasi
      Normalisasi, dalam konteks teks preprocessing, adalah proses mengubah kata-kata tak baku (seperti singkatan, slang, atau penulisan yang tidak konsisten) menjadi bentuk baku atau standar bahasa yang dapat dikenali dan diakui oleh sistem analisis. Ini membuat teks lebih terstruktur dan lebih mudah diolah untuk tugas proses pengolahan bahasa natural (NLP) seperti klasifikasi sentimen atau analisis topik [11]. Mengubah kata tidak baku/slang menjadi bentuk baku, "gk", "ga", "nggak" → "tidak". Penting untuk konsistensi makna dalam analisis.
   e) Stopwords Removal
      Pada tahap preprocessing teks, proses yang dikenal sebagai stopword removal menghilangkan kata-kata yang umum dan sering muncul tetapi tidak memberikan kontribusi signifikan terhadap keseluruhan dokumen (seperti kata penghubung atau artikel), sehingga hanya kata-kata yang lebih informatif yang disimpan untuk analisis lebih lanjut [12]. Menghapus kata-kata umum yang tidak memiliki makna signifikan, "yang", "dan", "di", "ke". Membantu fokus pada kata yang memiliki nilai sentimen.
4. Labeling
   Proses pemberian label pada setiap data komentar, label biasanya terdiri dari Positif, Negatif ,Netral label dapat dilakukan secara manual atau semi-otomatis.Tahap ini penting karena SVM merupakan metode supervised learning.
5. Data Training dan Data Testing
   Dua komponen terpisah dari dataset yang digunakan untuk berbagai tujuan dalam proses pembelajaran mesin adalah pelatihan data. Pelatihan data mengacu pada kumpulan data yang digunakan untuk melatih model, misalnya untuk menemukan pola atau aturan klasifikasi (seperti bobot, hyperplane, atau parameter lainnya) melalui pembelajaran dari label kelas yang sudah diberikan [13]. Dataset dibagi menjadi dua bagian, Data Training digunakan untuk melatih model agar mengenali pola sedangkan Data Testing digunakan untuk menguji kemampuan model pada data baru. Pembagian umum misalnya 80:20 atau 70:30.
6. Cross Validation
   Teknik validasi dengan membagi data menjadi beberapa bagian (fold), model dilatih dan diuji secara bergantian pada setiap fold, untuk menghindari overfitting mendapatkan performa model yang lebih stabil dan objektif. Misalnya, dalam proses cross-fold validation, model dilatih sebanyak beberapa kali dengan kombinasi data yang berbeda sehingga setiap bagian data sempat menjadi data uji. Tujuan dari proses ini adalah untuk mendapatkan estimasi kinerja yang lebih konsisten dan mengevaluasi seberapa baik model dapat menggeneralisasi pada data yang tidak pernah digunakan selama waktu pelatihan [14].
7. Support Vector Machine
   Support Vector Machine adalah salah satu algoritma machine learning yang banyak digunakan dalam analisis sentimen karena kemampuannya dalam mengklasifikasikan data dengan margin pemisah maksimal. Dalam konteks analisis sentimen komentar YouTube, SVM efektif dalam menangani data teks yang tidak terstruktur dan menghadapi permasalahan imbalance data dengan bantuan teknik seperti SMOTE untuk data balancing dan TF-IDF untuk ekstraksi fitur. Teknik yang digunakan antara lain klasifikasi, clustering, prediksi, dan asosiasi, yang bertujuan untuk mengidentifikasi hubungan tersembunyi di dalam data dan mendukung pengambilan keputusan yang berbasis informasi [15].
8. Evaluasi

© 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 225
This work is licensed under a Creative Commons Attribution 4.0 International License.

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 223–232
DOI : 10.47065/bulletinds.v5i3.10129
https://ejurnal.seminar-id.com/index.php/bulletinds

Digunakan untuk mengukur performa model klasifikasi, confusion matrix membandingkan Predicted Label dan True Label untuk menghitung metrik evaluasi Akurasi, Presisi, Recall F1-Score

2.1 YouTube

Youtube Merupakan platform video terbesar yang memungkinkan pengguna mengunggah, menonton, dan berinteraksi melalui berbagai video. Fitur utama yang mendukung interaksi pengguna adalah kolom komentar, di mana pengguna dapat memberikan kritik, dukungan, dan opini mereka tentang video yang telah ditonton. Komentar ini memiliki banyak data tekstual, yang dapat digunakan untuk mempelajari perasaan publik terhadap berbagai topik tertentu. YouTube adalah media sosial yang sangat efektif dalam membentuk opini publik dan memungkinkan komunikasi dua arah antara penonton dan creator [6]. Komentar-komentar dalam YouTube sering kali menggunakan bahasa informal serta variasi slang, sehingga menjadi tantangan tersendiri dalam analisis sentimen. Metode Support Vector Machine (SVM) terbukti efektif dalam mengklasifikasikan sentimen komentar dengan tingkat akurasi yang tinggi, apalagi jika didukung dengan teknik normalisasi bahasa dan pengolahan data. Dengan menggunakan metode ini, sangat bermanfaat untuk menyaring sentimen netral, positif, dan negatif dari komentar pengguna namun, identifikasi sentimen negatif memerlukan penyempurnaan tambahan [16]. Interaksi di kolom komentar tidak hanya memengaruhi persepsi penonton YouTube, tetapi juga tingkat keterlibatan penonton dengan video. Semakin banyak interaksi antara pengguna dan kreator melalui reaksi dan komentar, semakin banyak penonton yang terlibat dengan video. Hal ini menunjukkan bahwa analisis komentar YouTube dapat menjadi alat penting untuk melacak opini publik secara real-time, membantu pemerintah mengantisipasi dan menanggapi perubahan sosial [7].

2.2 Pengumpulan Data

Pengumpulan data dilakukan melalui teknik web crawling menggunakan YouTube Data API v3 yang disediakan oleh Google, memastikan akses legal dan efisien terhadap komentar publik tanpa melanggar Terms of Service YouTube. Proses ini menghasilkan dataset mentah dalam format CSV yang siap untuk tahap preprocessing selanjutnya. Sangat penting bagi pengembang untuk memiliki dokumentasi API yang baik untuk membantu mereka memahami cara API bekerja, fitur apa yang dapat diakses, dan metode autentikasi. Ini membuat adopsi API lebih cepat dan integrasi lebih mudah [17].

1. Persiapan API
   a. API Key: Diperoleh dari Google Cloud Console dengan mengaktifkan YouTube Data API v3.
   b. Library Python: googleapiclient.discovery untuk membangun service client (build('youtube', 'v3', developerKey=api_key)).

   Pada Tabel 1 merupakan sebuah parameter utama dari parameter API.

Tabel 1. Jenis jenis database

Parameter	Nilai	Fungsi
part	snippet	Mengambil metadata komentar
videoId	EtBfy-ch7cM	ID video Timnas target
maxResults	20000	Maksimal komentar per request
textFormat	plainText	Format teks bersih tanpa HTML

2.2 Text Preprocessing

Text Preprocessing merupakan tahap penting dalam analisis sentimen komentar YouTube yang berperan untuk membersihkan dan menyiapkan data teks mentah agar dapat diproses oleh algoritma SVM secara efektif. Proses ini meliputi penghapusan tanda baca, tokenisasi, penghilangan stop words, serta normalisasi teks seperti atau lemmatization. Tahapan ini sangat diperlukan mengingat karakter komentar YouTube sering menggunakan bahasa informal, singkatan, dan emotikon yang jika tidak diolah dengan baik dapat menurunkan akurasi model klasifikasi sentiment [18].

2.2 Pelabelan Data Sentimen

Pelabelan merupakan tahap krusial untuk mengubah teks komentar YouTube yang telah dipra-proses menjadi dataset supervised learning dengan tiga kelas sentimen Positif, Negatif, dan Netral. Pendekatan ini melibatkan penentuan kriteria yang jelas untuk memisahkan kelompok, sering kali bersifat hierarkis dari kategori umum ke spesifik [19]. Teknik ini diterapkan luas dalam berbagai bidang untuk mengubah data acak menjadi pengetahuan yang dapat diakses dan dimanfaatkan secara efektif [20]. Selain itu, penggunaan YouTube Data API v3 memungkinkan proses pengambilan data dilakukan secara terstruktur dan konsisten. Setiap komentar yang diperoleh dapat disimpan bersama informasi pendukung seperti jumlah suka, waktu publikasi, dan identitas pengguna yang telah dianonimkan. Ketersediaan metadata tersebut dapat membantu proses analisis lanjutan apabila penelitian dikembangkan pada aspek perilaku pengguna atau pola interaksi komentar di masa mendatang. Proses ini menghasilkan file hasil labelling yang berisi kolom stopword_removal sebagai fitur teks dan kolom Sentimen sebagai target label untuk pelatihan model SVM. Seperti pada Tabel 2.

© 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 226
This work is licensed under a Creative Commons Attribution 4.0 International License.

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 223–232
DOI : 10.47065/bulletinds.v5i3.10129
https://ejurnal.seminar-id.com/index.php/bulletinds

Tabel 2. Pelabelan Data Sentimen

Index	Stopword Removal	Score	Sentiment
7	suarakan abang erick thohir kejahatan mengatur komposisi pemain gonta ganti formasi sesuai ya mengakibatkan pecatnya sty menuruti keinginannyadan menghancurkan mimpi pecinta sepak bola bersalah	0	Netral
8	pemerintah menindas olahraga indonesia masyarakat olahraga indonesia diam lawan diam tolol	-2	Negatif
10	point bentuk tanggung et pakai diganti pelatih kompeten salahkan masyarakat kecewa disalahin ekspektasi masyarakat trust the proses sty prosesnya hentikan jalan for better good	2	Positif

2.4 Confuison Matrix

Confusion matrix adalah tabel yang digunakan untuk mengevaluasi performa model klasifikasi dengan membandingkan hasil prediksi model terhadap label aktual, terdiri dari empat elemen utama: True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN). Confusion matrix merupakan tabel yang merangkum prediksi model dengan membandingkan data aktual dan prediksi, di mana diagonal menunjukkan prediksi benar (TP dan TN) sementara off-diagonal menunjukkan kesalahan klasifikasi (FP dan FN). Dalam penelitian ini, penggunaan confusion matrix tidak hanya bertujuan memperoleh nilai akurasi, tetapi juga untuk mengidentifikasi pola kesalahan klasifikasi antar kelas sentimen. Analisis terhadap distribusi false positive dan false negative membantu memahami apakah model cenderung mengelompokkan komentar netral ke dalam kelas negatif atau positif, yang umum terjadi pada data komentar media sosial yang mengandung ambiguitas bahasa, ironi, dan variasi penulisan informal Selain itu, confusion matrix memberikan gambaran yang lebih rinci mengenai kemampuan model dalam mengenali karakteristik setiap kelas sentimen secara individual. Penggunaan confusion matrix dalam penelitian ini juga memberikan informasi yang lebih mendalam dibandingkan hanya menggunakan nilai akurasi. Melalui matriks tersebut dapat diketahui kelas sentimen mana yang paling mudah dikenali maupun yang masih sering mengalami kesalahan klasifikasi. Informasi ini menjadi dasar penting dalam mengevaluasi efektivitas model dan menentukan strategi peningkatan performa pada penelitian selanjutnya. Seperti pada Tabel 3 di bawah ini.

Tabel 3. Confusion Matrix

Metrik	Rumus	Interpretasi
Accuracy	(TP + TN) / Total	Proporsi prediksi benar secara keseluruhan
Precision	TP / (TP + FP)	Dari prediksi kelas tertentu, berapa persen yang benar-benar termasuk kelas tersebut
Recall	TP / (TP + FN)	Dari seluruh data yang sebenarnya termasuk kelas tersebut, berapa persen berhasil terdeteksi
F1-Score	2 × (Precision × Recall) / (Precision + Recall)	Harmonic mean yang menyeimbangkan nilai precision dan recall

3. HASIL DAN PEMBAHASAN

3.1 Hasil Pengumpulan Data

Pengumpulan data dilakukan menggunakan YouTube Data API melalui proses web crawling pada platform Google Colaboratory. Data yang dikumpulkan berupa komentar pengguna pada video YouTube yang membahas isu terkait Tim Nasional Indonesia Tidak Lolos Kualifikasi Piala Dunia.

Tabel 4. Hasil Crawling Data

Kolom	Non-Null Count	Tipe Data
author	15169 non-null	object
comment	15169 non-null	object
likes	15169 non-null	int64
published_at	15169 non-null	object
Total	15169 baris	4 kolom

Pada Tabel 4 tersebut menunjukkan proses pengambilan data komentar YouTube menggunakan bahasa pemrograman Python dan library pandas. Pada bagian awal program, ditentukan ID video YouTube yang akan diambil komentarnya yaitu EtBfy-ch7cM. Selanjutnya, variabel total_comments diatur sebesar 25000 sebagai target jumlah

© 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 227
This work is licensed under a Creative Commons Attribution 4.0 International License.

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 223−232
DOI : 10.47065/bulletinds.v5i3.10129
https://ejurnal.seminar-id.com/index.php/bulletinds

komentar yang ingin dikumpulkan. Hasil crawling kemudian disimpan ke dalam file CSV dengan nama INDgagallolosPialaDunia.csv.

Proses pengambilan komentar dilakukan menggunakan fungsi scrape_comments_from_videos(video_ids, total_comments, output_filename). Setelah data berhasil dikumpulkan, data dimasukkan ke dalam DataFrame menggunakan library pandas melalui perintah pd.DataFrame(comments) agar lebih mudah dianalisis dan diproses pada tahap selanjutnya.

3.2 Hasil Pra-Pemrosesan Data

Tahap pra-pemrosesan data dilakukan untuk membersihkan dan menyiapkan data komentar agar dapat diproses secara optimal oleh algoritma klasifikasi. Proses ini diawali dengan case folding, yaitu mengubah seluruh huruf pada teks menjadi huruf kecil untuk menyeragamkan bentuk penulisan kata. Selanjutnya dilakukan cleaning untuk menghapus berbagai karakter yang tidak relevan seperti tanda baca, angka, simbol, URL, maupun karakter khusus lainnya yang dapat mengganggu proses analisis. Setelah itu, dilakukan tokenizing yang bertujuan memecah kalimat menjadi kumpulan kata atau token sehingga lebih mudah diproses oleh sistem.

Tahap berikutnya adalah normalisasi kata, yaitu mengubah kata-kata tidak baku, singkatan, atau bahasa slang menjadi bentuk baku sesuai kaidah bahasa Indonesia. Kemudian dilakukan stopword removal untuk menghilangkan kata-kata umum yang tidak memiliki kontribusi signifikan terhadap pembentukan sentimen, seperti kata penghubung dan kata depan. Setelah seluruh tahapan pra-pemrosesan selesai dilakukan, diperoleh sebanyak 7.584 komentar yang layak digunakan sebagai dataset pada proses klasifikasi sentimen menggunakan metode Support Vector Machine.

Tabel 5. Jumlah Data Setelah Pra-Pemrosesan

Keterangan	Jumlah Data
Data awal hasil crawling	15.169
Data akhir siap klasifikasi	7.584

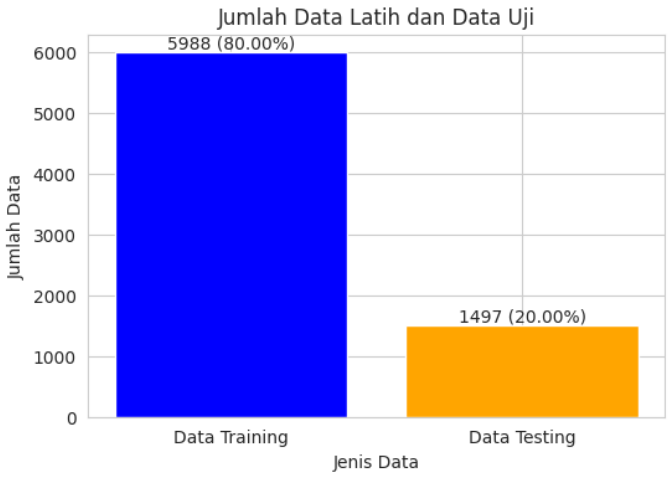
Pada Tabel 5 pengurangan jumlah data dari 15.169 komentar menjadi 7.584 komentar menunjukkan bahwa tahap pra-pemrosesan berhasil menghilangkan komentar duplikat, kosong, sangat pendek, atau mengandung noise yang tinggi. Walaupun jumlah data berkurang cukup signifikan, kualitas representasi teks meningkat sehingga fitur TF-IDF yang dihasilkan menjadi lebih relevan untuk proses klasifikasi SVM. Hasil pra-pemrosesan yang baik juga berpengaruh terhadap efisiensi komputasi pada tahap pelatihan model. Berkurangnya jumlah fitur yang tidak relevan menyebabkan proses ekstraksi fitur TF-IDF dan pembentukan model SVM dapat dilakukan dengan lebih cepat tanpa mengurangi informasi penting yang terkandung dalam data. Kualitas hasil pra-pemrosesan memiliki pengaruh yang sangat besar terhadap keberhasilan proses klasifikasi. Semakin baik proses pembersihan dan normalisasi data dilakukan, maka semakin representatif pula fitur yang dihasilkan untuk menggambarkan sentimen pada setiap komentar.

Tahap pra-pemrosesan tidak hanya berfungsi untuk membersihkan data, tetapi juga meningkatkan kualitas representasi teks yang akan digunakan dalam proses ekstraksi fitur. Pada data komentar YouTube, sering ditemukan penggunaan bahasa tidak baku, singkatan, kata serapan, hingga kombinasi huruf dan angka yang dapat mengganggu proses klasifikasi apabila tidak ditangani dengan baik. Melalui tahapan case folding, cleaning, tokenizing, normalisasi, dan stopword removal, struktur data menjadi lebih seragam sehingga memudahkan algoritma dalam mengenali pola sentimen yang terkandung dalam komentar. Selain itu, proses ini juga membantu mengurangi dimensi data yang tidak relevan sehingga model dapat bekerja lebih efisien. Dengan demikian, hasil pra-pemrosesan yang baik menjadi salah satu faktor utama yang mendukung keberhasilan klasifikasi sentimen menggunakan metode Support Vector Machine.

Berkurangnya jumlah data yang tidak relevan menyebabkan kebutuhan memori dan waktu komputasi menjadi lebih rendah dibandingkan apabila seluruh data mentah digunakan secara langsung. Kondisi tersebut sangat penting terutama ketika jumlah komentar yang dianalisis mencapai ribuan data seperti pada penelitian ini. Oleh karena itu, keberhasilan tahap pra-pemrosesan dapat dianggap sebagai salah satu faktor utama yang mendukung tercapainya performa klasifikasi yang baik pada model Support Vector Machine yang digunakan.Hasil Pembagian Data Training Dan Testing Dataset kemudian dibagi menjadi dua bagian menggunakan metode train-test split dengan rasio 80:20

© 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 228
This work is licensed under a Creative Commons Attribution 4.0 International License.

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 223–232
DOI : 10.47065/bulletinds.v5i3.10129
https://ejurnal.seminar-id.com/index.php/bulletinds



Gambar 3. Pembagian Data Training dan Testing

Pada Gambar 3 tersebut menampilkan visualisasi pembagian dataset ke dalam dua bagian utama, yaitu data latih (training data) dan data uji (testing data) pada proses klasifikasi sentimen. Pembagian dataset dilakukan menggunakan metode train-test split dengan rasio 80:20, yang merupakan salah satu metode umum dalam machine learning untuk melatih dan mengevaluasi performa model klasifikasi.

Berdasarkan grafik, sebanyak 5.988 data atau sekitar 80% dari total dataset digunakan sebagai data training. Data training berfungsi untuk melatih model agar dapat mempelajari pola, karakteristik, serta hubungan antar fitur pada data komentar yang telah diproses sebelumnya. Pada tahap ini, algoritma klasifikasi mempelajari berbagai pola sentimen seperti positif, negatif, maupun netral dari data yang tersedia sehingga model dapat melakukan prediksi dengan lebih baik.

Sementara itu, sebanyak 1.497 data atau 20% dari total dataset digunakan sebagai data testing. Data testing digunakan untuk menguji kemampuan model setelah proses pelatihan selesai. Pengujian dilakukan menggunakan data yang belum pernah dilihat oleh model sebelumnya sehingga hasil evaluasi dapat menunjukkan tingkat akurasi model secara objektif dalam melakukan klasifikasi sentimen komentar YouTube.

Pada grafik terlihat bahwa batang berwarna biru merepresentasikan data training dengan jumlah yang lebih besar dibandingkan batang berwarna oranye yang merepresentasikan data testing. Sumbu horizontal menunjukkan jenis data, sedangkan sumbu vertikal menunjukkan jumlah data yang digunakan. Pembagian dataset dengan rasio 80:20 dipilih karena dianggap mampu memberikan keseimbangan antara proses pelatihan model dan evaluasi performa, sehingga model dapat menghasilkan prediksi yang lebih optimal dan mengurangi risiko overfitting.

3.3 Hasil Pelabelan Data Sentimen

Proses pelabelan sentimen dilakukan terhadap seluruh data komentar yang telah melewati tahap pra-pemrosesan. Setiap komentar diklasifikasikan ke dalam salah satu dari tiga kategori sentimen, yaitu positif, negatif, dan netral berdasarkan kandungan opini yang terdapat pada teks komentar. Sentimen positif merepresentasikan komentar yang mengandung dukungan, apresiasi, optimisme, atau kepercayaan terhadap Tim Nasional Indonesia. Sentimen negatif menunjukkan adanya kritik, kekecewaan, ketidakpuasan, maupun ekspresi emosional yang bernada kontra terhadap hasil pertandingan dan performa tim. Sementara itu, sentimen netral mencakup komentar yang bersifat informatif, pertanyaan, atau pernyataan yang tidak menunjukkan kecenderungan emosi tertentu.

Tahap pelabelan merupakan salah satu proses yang sangat penting dalam analisis sentimen karena kualitas label akan berpengaruh langsung terhadap performa model klasifikasi yang dibangun. Dataset yang memiliki label yang konsisten dan representatif memungkinkan algoritma Support Vector Machine mempelajari pola bahasa yang membedakan masing-masing kategori sentimen secara lebih optimal. Sebaliknya, kesalahan pelabelan dapat menyebabkan model mengalami kesulitan dalam mengenali karakteristik setiap kelas sehingga berdampak pada penurunan akurasi klasifikasi. Distribusi hasil pelabelan juga dapat digunakan sebagai indikator awal tingkat keseimbangan data pada setiap kelas sentimen. Informasi tersebut penting karena perbedaan jumlah data yang terlalu besar antar kelas berpotensi memengaruhi kemampuan model dalam mempelajari pola sentimen secara merata. Dengan mengetahui distribusi kelas sejak awal, proses pelatihan model dapat dilakukan dengan lebih terarah dan menghasilkan performa yang lebih stabil.

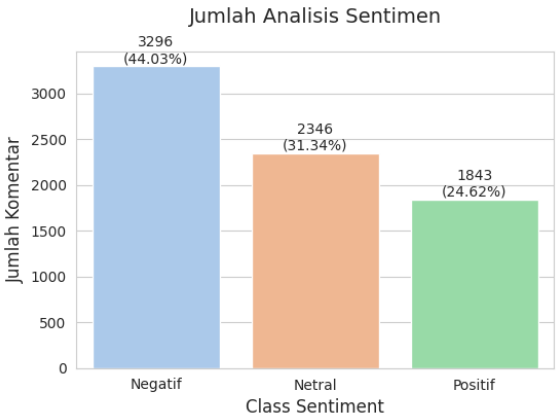
Distribusi hasil pelabelan sentimen memberikan informasi awal mengenai persepsi masyarakat terhadap isu yang sedang dibahas. Dominasi salah satu kelas sentimen dapat mencerminkan kecenderungan opini publik terhadap suatu peristiwa tertentu. Dalam konteks penelitian ini, proses pelabelan tidak hanya berfungsi sebagai target klasifikasi bagi algoritma SVM, tetapi juga menjadi dasar untuk memahami pola komunikasi pengguna pada media sosial. Informasi tersebut penting karena dapat digunakan sebagai bahan evaluasi terhadap respons masyarakat serta sebagai sumber masukan bagi pihak-pihak yang berkepentingan dalam dunia olahraga nasional. Oleh sebab itu, kualitas dan konsistensi

© 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 229
This work is licensed under a Creative Commons Attribution 4.0 International License.

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 223−232
DOI : 10.47065/bulletinds.v5i3.10129
https://ejurnal.seminar-id.com/index.php/bulletinds

pelabelan menjadi aspek yang harus diperhatikan agar hasil analisis yang diperoleh dapat menggambarkan kondisi sebenarnya dari opini publik yang diamati.

Selain itu, hasil pelabelan juga memberikan gambaran awal mengenai kecenderungan opini publik terhadap topik yang diteliti. Seperti pada Gambar 4.



Gambar 4. Pelabelan Data Sentimen

Grafik pada Gambar 4 tersebut merupakan visualisasi hasil analisis sentimen yang membagi data komentar ke dalam tiga kategori utama: Negatif, Netral, dan Positif. Berdasarkan sebaran datanya, terlihat jelas bahwa sentimen negatif mendominasi dengan jumlah mencapai 3.296 komentar atau sekitar 44,03%. Hal ini mengindikasikan bahwa mayoritas respon atau opini dalam dataset tersebut cenderung berisi kritik, ketidakpuasan, atau pernyataan yang bersifat kontra.

Di posisi kedua, terdapat sentimen netral dengan jumlah 2.346 komentar (31,34%). Kategori ini menunjukkan adanya porsi yang cukup besar dari audiens yang memberikan respon objektif, sekadar bertanya, atau memberikan pernyataan yang tidak memihak. Sementara itu, sentimen positif merupakan kategori dengan jumlah terendah, yakni hanya 1.843 komentar atau 24,62% dari total keseluruhan data.

Jika dilihat secara kolektif, perbandingan antara sentimen negatif dan positif menunjukkan selisih yang cukup signifikan, di mana jumlah komentar negatif hampir dua kali lipat dari komentar positif. Secara keseluruhan, total data yang diolah adalah sebanyak 7.485 komentar, dan visualisasi ini secara efektif merangkum bahwa tren opini publik terhadap subjek yang diteliti cenderung mengarah pada persepsi yang kurang baik atau kritis.

3.4 Hasil Klasifikasi Metode SVM

Hasil Kalsifikasi dilakukan menggunakan metrik accuracy, precision, recall, dan F1-score.

Tabel 6. Hasil Klasifikasi SVM

Kelas Sentimen	Precision	Recall	F1-Score	Support
Negatif	0.89	0.85	0.87	659
Netral	0.70	0.78	0.74	469
Positif	0.80	0.75	0.77	369
Macro Avg	0.80	0.79	0.79	1497
Weighted Avg	0.81	0.80	0.81	1497
Accuracy	-	-	80.43%	1497

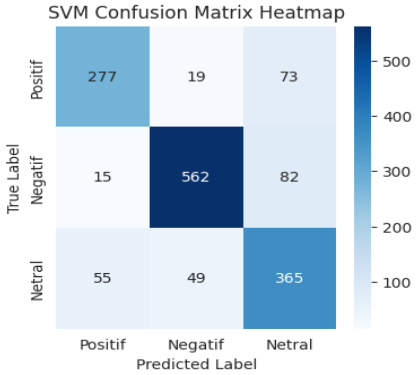
Pada Tabel 6 model SVM memperoleh tingkat akurasi sebesar 80.43%, yang menunjukkan bahwa sebagian besar data testing berhasil diklasifikasikan dengan benar. Nilai weighted F1-score sebesar 0,81 menunjukkan bahwa model memiliki performa yang cukup baik, terutama pada kelas dengan jumlah data lebih banyak seperti sentimen negatif. Sementara itu, nilai macro average sebesar 0,79 yang sedikit lebih rendah mengindikasikan adanya perbedaan performa antar kelas akibat ketidakseimbangan jumlah data. Secara umum, hasil klasifikasi menunjukkan bahwa algoritma Support Vector Machine mampu membentuk batas pemisah yang baik antar kelas sentimen meskipun data komentar memiliki karakteristik yang beragam. Performa yang diperoleh membuktikan bahwa pendekatan berbasis machine learning masih menjadi solusi yang efektif untuk mengolah data teks berbahasa Indonesia pada platform media sosial yang cenderung tidak terstruktur. Temuan ini juga memperkuat hasil penelitian sebelumnya yang menyatakan bahwa SVM merupakan salah satu algoritma yang efektif untuk klasifikasi teks berbahasa Indonesia.

© 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 230
This work is licensed under a Creative Commons Attribution 4.0 International License.

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 223–232
DOI : 10.47065/bulletinds.v5i3.10129
https://ejurnal.seminar-id.com/index.php/bulletinds

3.5 Evaluasi Model Confusion Matrix

Pada tahap ini confusion matrix merupakan salah satu metode evaluasi yang digunakan untuk mengetahui kinerja model klasifikasi dengan membandingkan antara label aktual dan label prediksi.



Gambar 5. Heatmap Confuison Matrix Model SVM

Pada Gambar 5 model SVM menunjukkan kinerja yang cukup baik karena sebagian besar data berhasil diklasifikasikan dengan benar (terlihat pada nilai diagonal utama). Kelas negatif memiliki prediksi benar tertinggi yaitu 562 data, diikuti kelas netral sebanyak 365 data, dan kelas positif sebanyak 277 data. Namun, masih terdapat kesalahan klasifikasi, terutama antara kelas positif dan netral serta antara negatif dan netral. Hal ini menunjukkan bahwa model masih mengalami kebingungan dalam membedakan sentimen yang memiliki karakteristik mirip, khususnya pada kelas netral. Secara keseluruhan, model sudah cukup baik dalam mengenali pola sentimen, tetapi masih perlu peningkatan untuk mengurangi kesalahan antar kelas.

3.6 Pembahasan

Pembahasan penelitian ini menunjukkan bahwa sentimen komentar YouTube terkait kekalahan Tim Nasional Indonesia pada kualifikasi Piala Dunia didominasi oleh sentimen negatif, diikuti netral dan positif, yang menandakan bahwa isu olahraga nasional memicu respons emosional publik yang kuat temuan ini sejalan dengan penelitian sebelumnya pada konteks Timnas Indonesia di platform YouTube maupun X, seperti studi pada Piala Dunia U-17 [4], penelitian pada AFC U23 Asian Cup [1]. pada komentar YouTube beauty vlogger [6], yang sama-sama menunjukkan bahwa pendekatan klasifikasi sentimen dapat mengidentifikasi pola opini publik secara efektif, meskipun objek dan platform berbeda. Namun, dibandingkan penelitian terdahulu, penelitian ini menegaskan bahwa komentar YouTube pada isu kekalahan Timnas cenderung lebih noisy, informal, dan tidak seimbang sehingga menjadi tantangan tersendiri bagi klasifikasi, tetapi SVM tetap mampu memberikan kinerja yang cukup baik dengan akurasi 80,43% dan weighted F1-score 0,81 hasil ini memperkuat temuan studi lain yang menyatakan bahwa SVM efektif untuk teks berbahasa Indonesia pada media sosial, termasuk pada komentar yang kompleks dan bernada politis atau emosional.

4. KESIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan, proses analisis sentimen komentar YouTube terkait kekalahan Tim Nasional Indonesia pada kualifikasi Piala Dunia dapat dilakukan menggunakan metode Support Vector Machine (SVM). Tahapan penelitian dimulai dari pengumpulan data komentar YouTube, kemudian dilanjutkan dengan proses pra-pemrosesan teks seperti cleaning, case folding, tokenisasi, normalisasi kata, dan stopword removal. Setelah data diproses, dilakukan ekstraksi fitur dan pembentukan model klasifikasi menggunakan algoritma SVM untuk mengenali pola sentimen pada komentar pengguna.Hasil penelitian menunjukkan bahwa metode Support Vector Machine mampu mengklasifikasikan komentar ke dalam tiga kategori sentimen, yaitu positif, negatif, dan netral. Berdasarkan hasil klasifikasi, sentimen negatif menjadi kategori yang paling dominan dibandingkan sentimen positif dan netral. Hal tersebut menunjukkan bahwa sebagian besar pengguna YouTube memberikan respons negatif terhadap kekalahan Tim Nasional Indonesia pada pertandingan kualifikasi Piala Dunia.Berdasarkan hasil evaluasi model menggunakan confusion matrix, diperoleh tingkat akurasi sebesar 80,43% dengan nilai weighted F1-score sebesar 0,81. Hasil tersebut menunjukkan bahwa model Support Vector Machine memiliki performa yang baik dalam mengidentifikasi dan mengklasifikasikan sentimen komentar YouTube secara otomatis. Dengan demikian, metode SVM dinilai efektif digunakan untuk analisis sentimen berbasis teks pada media sosial. Kontribusi utama penelitian ini adalah menyediakan gambaran empiris mengenai dominasi sentimen publik pada isu kekalahan Tim Nasional Indonesia di YouTube serta menunjukkan bahwa Support Vector Machine mampu digunakan secara efektif untuk mengklasifikasikan komentar berbahasa Indonesia yang informal dan beragam. Hasil ini dapat menjadi dasar pengembangan penelitian lanjutan pada analisis sentimen media sosial dengan objek, platform, atau algoritma yang berbeda.

© 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 231
This work is licensed under a Creative Commons Attribution 4.0 International License.

ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 223−232
DOI : 10.47065/bulletinds.v5i3.10129
https://ejurnal.seminar-id.com/index.php/bulletinds

REFERENCES

[1] D. Pangestu, M. Malik, and M. R. Pribadi, "Analisis Sentimen Hasil Pertandingan Sepakbola Timnas Indonesia di Piala Asia U-23 pada Platform Youtube menggunakan Algoritma Suport Vector Machine (SVM)," *Appl. Inf. Technol. Comput. Sci.*, vol. 3, no. 1, pp. 38–48, 2024.

[2] A. S. Aiman and K. M. Lhaksmana, "Topic Classification of Quranic Verses in English Translation," *J. REST*, vol. 5, no. 158, pp. 803–809, 2026.

[3] J. Maulani and M. Sari, "Komparasi Metode K-Nearest Neighbor (Knn) Dengan Support Vector Machine (Svm) Terhadap Tingkat Akurasi Klasifikasi Kualitas Air," *Smart Comp Jurnalnya Orang Pint. Komput.*, vol. 12, no. 2, pp. 430–435, 2023, doi: 10.30591/smartcomp.v12i2.4205.

[4] T. Juniardi and C. A. Sugianto, "Analisis Sentimen Tim Nasional Sepak Bola Indonesia Di Turnamen Piala Dunia U-17 Indonesia Pada Twitter (X) Menggunakan Algortima Naïve Bayes," *J. Inform. dan Tek. Elektro Terap.*, vol. 12, no. 3S1, 2024, doi: 10.23960/jitet.v12i3s1.5188.

[5] Tetrian Widyanto, Ina Ristiana, and Arief Wibowo, "Komparasi Naïve Bayes dan SVM Analisis Sentimen RUU Kesehatan di Twitter," *Jurasik (Jurnal Ris. Sist. Inf. dan Tek. Inform.*, vol. 3, no. 2, pp. 300–309, Dec. 2023.

[6] S. Chairani Siregar, R. T. Adek, and Z. Fitri, "Sentiment Analysis of Comments on Youtube Channel Beauty Vlogger in Indonesian Language Using Support Vector Machine Method," *J. Akuntansi, Audit dan Sist. Inf. Akunt.*, vol. 2, no. 3, pp. 1–6, 2024, doi: 10.29103/icomden.v2.xxxx.

[7] I. Sa'diyah, A. T. Aviolla Terza, C. C. B. Lima, M. R. M. Ariefean, and I. Athallah, "Sentiment Analysis of Netizen's Comments on YouTube about IKN (Capital City) Development in Indonesia," *GHANCARAN J. Pendidik. Bhs. dan Sastra Indones.*, vol. 6, no. 2, Jan. 2025, doi: 10.19105/ghancaran.v6i2.15432.

[8] B. S. Pringgodani and A. Supriyanto, "Sentiment Analysis of YouTube Comments on Free Lunch Program Using Machine Learning," *bit-Tech*, vol. 8, no. 2, pp. 1367–1375, 2025, doi: 10.32877/bt.v8i2.2908.

[9] D. Rifaldi, Abdul Fadlil, and Herman, "Teknik Preprocessing Pada Text Mining Menggunakan Data Tweet 'Mental Health,'" *Decod. J. Pendidik. Teknol. Inf.*, vol. 3, no. 2, pp. 161–171, Apr. 2023, doi: 10.51454/decode.v3i2.131.

[10] Putri Nur Apriliyanti, Moh. Dasuki, and M. Rahman, "Klasifikasi Sentimen Positif dan Negatif Ulasan Aplikasi GetContact Dengan Algoritma Naïve Bayes," *JUSTIFY J. Sist. Inf. Ibrahimy*, vol. 4, no. 2, pp. 123–129, Jan. 2026, doi: 10.35316/justify.v4i2.9133.

[11] P. M. S. Ardinata, A. A. J. Permana, I. N. S. W. Wijaya, F. Teknik, and D. Kejuruan, "IDENTIFIKASI DAN NORMALISASI TEKS SLANG DENGAN FASTTEXT PADA TWITTER DALAM BAHASA INDONESIA," *J. Pendidik. Teknol. dan Kejuru.*, vol. 21, no. 1, 2024.

[12] S. Jessica Angelina, A. Bijaksana Putra Negara, H. Muhardi, J. H. Nawawi, and K. Barat, "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, 2023, doi: 10.26418/juara.v2i1.69680.

[13] Baiq Nurul Azmi, Arief Hermawan, and Donny Avianto, "Analisis Pengaruh Komposisi Data Training dan Data Testing pada Penggunaan PCA dan Algoritma Decision Tree untuk Klasifikasi Penderita Penyakit Liver," *JTIM J. Teknol. Inf. dan Multimed.*, vol. 4, no. 4, pp. 281–290, Feb. 2023, doi: 10.35746/jtim.v4i4.298.

[14] R. N. Irawan, K. M. Hindrayani, and M. Idhom, "Penerapan Cross Validation sebagai Analisis Sentimen Pelayanan Publik Kereta Api Lokal Daop 8 Menggunakan Metode Multinomial Naïve Bayes," *G-Tech J. Teknol. Terap.*, vol. 8, no. 2, pp. 954–963, Apr. 2024, doi: 10.33379/gtech.v8i2.4117.

[15] A. A. N. Mostafa and H. E. A. Mahmoud, "Review of Data Mining Concept and its Techniques," *Int. J. Acad. Res. Bus. Soc. Sci.*, vol. 12, no. 6, Jun. 2022, doi: 10.6007/ijarbss/v12-i6/13135.

[16] A. Nur, A. Saputra, R. E. Saputro, and S. Saputra, "Enhancing Sentiment Analysis Accuracy Using SVM and Slang Word Normalization on YouTube Comments," *J. dan Penelit. Tek. Inform.*, vol. 9, no. 2, 2025, doi: 10.33395/sinkron.v9i2.14513.

[17] C. R. Park, H. Heo, C. H. Suh, and W. H. Shim, "Uncover This Tech Term: Application Programming Interface for Large Language Models," Aug. 2025, *Korean Radiological Society*. doi: 10.3348/kjr.2025.0360.

[18] J. Homepage, R. Rahman Salam, M. Fajri Jamil, and Y. Ibrahim, "MALCOM: Indonesian Journal of Machine Learning and Computer Science Sentiment Analysis of Cash Direct Assistance Distribution for Fuel Oil Using Support Vector Machine Analisis Sentimen Terhadap Bantuan Langsung Tunai (BLT) Bahan Bakar Minyak (BBM) Menggunakan Support Vector Machine," vol. 3, pp. 27–35, 2023.

[19] A. F. A. H. Alnuaimi and T. H. K. Albaldawi, "An overview of machine learning classification techniques," in *BIO Web of Conferences*, EDP Sciences, Apr. 2024. doi: 10.1051/bioconf/20249700133.

[20] S. Susandri, S. Defit, and M. Tajuddin, "SENTIMENT LABELING AND TEXT CLASSIFICATION MACHINE LEARNING FOR WHATSAPP GROUP," *JITK (Jurnal Ilmu Pengetah. dan Teknol. Komputer)*, vol. 9, no. 1, pp. 119–125, Aug. 2023, doi: 10.33480/jitk.v9i1.4201.

© 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 232
This work is licensed under a Creative Commons Attribution 4.0 International License.