Jurnal Teknologi Informasi dan Multimedia
p-ISSN : 2715-2529
e-ISSN : 2684-9151
https://journal.sekawan-org.id/index.php/jtim

Implementasi Ensemble Voting Classifier untuk Analisis Sentimen Publik terhadap Kontroversi Pertandingan Indonesia vs Bahrain pada Platform Youtube

Dlovan Ferdiansyah [1,*], Susanto [1], dan **Basworo Ardi Pramono [1]**

[1] Program Studi Teknik Informatika, Universitas Semarang, Indonesia.

* Korespondensi: dlovanferdiansyah10@gmail.com

Sitasi: D. Ferdiansyah, S. Susanto, and B. A. Pramono, "Implementasi Ensemble Voting Classifier untuk Analisis Sentimen Publik terhadap Kontroversi Pertandingan Indonesia vs Bahrain pada Platform Youtube", *Jurnal Teknologi Informasi Dan Multimedia*, vol. 8, no. 3, pp.. 465-483, 2026. https://doi.org/10.35746/jtim.v8i3.1029

Diterima: 08-05-2026
Direvisi: 09-06-2026
Disetujui: 17-06-2026

Abstract: The qualifying match between the Indonesian National Team and Bahrain sparked diverse public responses that escalated into controversy, particularly in digital spaces such as YouTube comment sections. This phenomenon prompted research to analyze public sentiment using a machine learning approach to understand trends and polarization of public opinion. The research data was initially collected as many as 1,000 raw comments through the YouTube scraping process, which then went through a cleaning stage resulting in 984 valid comments. The research data was obtained through scraping YouTube comments, The valid data is then labeled using a combination of lexicon (au-to-suggest) and manual validation approaches into three categories, namely positive, negative and neutral. The preprocessing stage focused on normalizing non-standard language (slang) and handling negations to maintain contextual meaning. Next, feature extraction was performed using Feature Union, which combines word- and character-based TF-IDF, as well as numeric features such as text length and punctuation proportion. To address data imbalance, the SMOTE method was applied to improve minority class representation. The model used was an Ensemble Voting Classifier with a soft voting approach, which combines a calibrated Support Vector Machine, Logistic Regression, and Random Forest. Model optimization was performed using GridSearchCV to obtain the best parameters. The evaluation results showed that the model performed well with an accuracy of 89.34%, a precision of 89.17%, a recall of 89.34%, and an F1-score of 89.18%. Furthermore, the application of SMOTE and negation handling has been shown to help reduce bias toward the majority class. The application of the SMOTE method has been shown to significantly improve model performance compared to the baseline model without oversampling, which only achieved an accuracy of 88.83%. Overall, this ensemble approach with multidimensional feature engineering is effective in producing an accurate sentiment analysis model for evaluating public response to sporting events.

Keywords: Sentiment Analysis; YouTube Comments; Machine Learning; Ensemble Voting Classifier

Abstrak: Pertandingan kualifikasi antara Timnas Indonesia dan Bahrain memunculkan beragam respons publik yang berkembang menjadi kontroversi, terutama di ruang digital seperti kolom komentar YouTube. Fenomena ini mendorong dilakukannya penelitian untuk menganalisis sentimen publik menggunakan pendekatan machine learning guna memahami kecenderungan dan polarisasi opini masyarakat. Data penelitian pada awalnya dikumpulkan sebanyak 1.000 komentar mentah melalui proses *scraping* YouTube, yang kemudian melalui tahap pembersihan menghasilkan 984 komentar valid. Data penelitian diperoleh melalui proses scraping komentar YouTube, data valid tersebut selanjutnya dilakukan pelabelan dengan pendekatan kombinasi

leksikon (auto-suggest) dan validasi manual ke dalam tiga kategori, yaitu positif, negatif, dan netral. Tahap praproses difokuskan pada normalisasi bahasa tidak baku (slang) serta penanganan kata negasi agar makna kontekstual tetap terjaga. Selanjutnya, ekstraksi fitur dilakukan menggunakan Feature Union yang menggabungkan TF-IDF berbasis kata dan karakter, serta fitur numerik seperti panjang teks dan proporsi tanda baca. Untuk mengatasi ketidakseimbangan data, diterapkan metode SMOTE guna meningkatkan representasi kelas minoritas. Model yang digunakan adalah Ensemble Voting Classifier dengan pendekatan soft voting, yang mengombinasikan Support Vector Machine terkalibrasi, Logistic Regression, dan Random Forest. Optimasi model dilakukan menggunakan GridSearchCV untuk memperoleh parameter terbaik. Hasil evaluasi menunjukkan bahwa model mampu memberikan kinerja yang baik dengan akurasi sebesar 89.34%, precision 89.17%, recall 89.34%, dan F1-score 89.18%. Selain itu, penerapan SMOTE dan penanganan negasi terbukti membantu mengurangi bias terhadap kelas mayoritas. Penerapan metode SMOTE ini terbukti secara signifikan meningkatkan performa model dibandingkan dengan model baseline tanpa *oversampling* yang hanya menghasilkan akurasi sebesar 88,83%. Secara keseluruhan, pendekatan ensemble dengan rekayasa fitur multidimensi ini efektif dalam menghasilkan model analisis sentimen yang akurat untuk mengevaluasi respons publik terhadap peristiwa olahraga.

Kata kunci: Analisis Sentimen; Komentar Youtube; Machine Learning; Ensemble Voting Classifier

1. Pendahuluan

Perkembangan teknologi informasi yang semakin pesat telah mendorong media sosial, khususnya YouTube, menjadi ruang publik utama bagi masyarakat dalam menyampaikan opini dan mengekspresikan diri secara interaktif. Dalam konteks olahraga nasional, isu yang berkaitan dengan Tim Nasional (Timnas) Sepak Bola Indonesia hampir selalu memicu diskusi luas dan perdebatan di ranah digital. Oleh karena itu, pemetaan sentimen publik dalam skala besar menjadi hal yang penting untuk memahami dan mengevaluasi respons sosial masyarakat terhadap berbagai dinamika tersebut [1]. Salah satu peristiwa yang menarik perhatian adalah pertandingan kualifikasi antara Timnas Indonesia dan Bahrain, yang menimbulkan kontroversi serta memicu beragam reaksi emosional yang terekam dalam kolom komentar YouTube.

Dalam kajian akademis, analisis sentimen telah banyak dimanfaatkan untuk menggali informasi dari data teks media sosial. Sejumlah penelitian sebelumnya mencoba memetakan opini publik di bidang olahraga dengan menggunakan algoritma machine learning tunggal. Prasetyo dan Dahlan [2], misalnya, menggunakan metode Naïve Bayes untuk menganalisis sentimen terkait isu naturalisasi pemain Timnas Indonesia dan memperoleh akurasi sebesar 72,55%. Keterbatasan Naïve Bayes pada data teks YouTube terletak pada asumsi independensi probabilitas antar kata, sehingga algoritma ini sering gagal menangkap makna berurutan atau konteks negasi (misalnya: "tidak adil") yang sangat krusial dalam klasifikasi sentimen. Sementara itu, Artamevia dan Wibowo [3] menerapkan algoritma K-Nearest Neighbour (KNN) untuk mengklasifikasikan kontroversi pergantian pelatih Timnas, namun hanya mencapai akurasi 68%. Performa KNN yang rendah ini dipengaruhi oleh ketidakmampuannya beradaptasi dengan data teks berdimensi tinggi (*sparse matrix* dari pembobotan fitur) serta sifatnya yang sangat rentan (bias) terhadap ketidakseimbangan distribusi data kelas mayoritas.Di sisi lain, Permana [4] membandingkan Naïve Bayes dan Support Vector Machine (SVM) pada komentar YouTube di ranah esports, dengan hasil SVM sedikit lebih unggul melalui akurasi 73,97%.

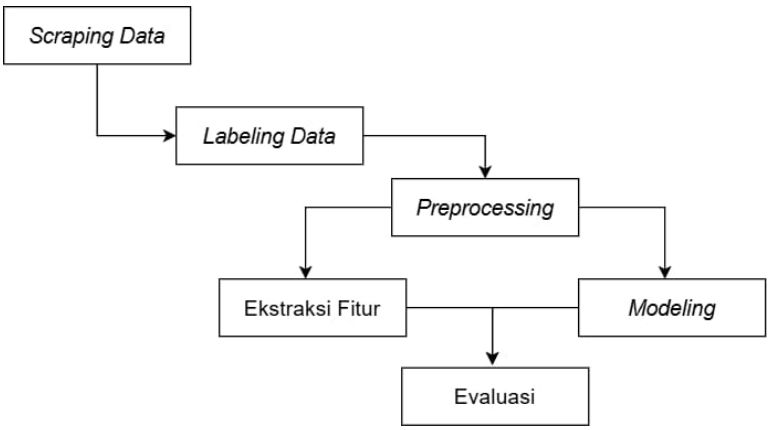
Berdasarkan temuan dari berbagai penelitian tersebut, terdapat celah penelitian (research gap) yang jelas yaitu penggunaan model klasifikasi tunggal belum mampu secara optimal menangani kompleksitas data teks pada komentar YouTube. Keterbatasan terse-

but terlihat dalam kemampuan model untuk mengatasi penggunaan bahasa yang tidak baku, memahami konteks negasi yang bersifat implisit, serta menangani masalah ketidakseimbangan distribusi kelas sentimen secara menyeluruh. Oleh karena itu, penelitian ini berfokus pada pengembangan suatu arsitektur analitik yang lebih tangguh dan efektif untuk mengatasi kelemahan metode tunggal, terutama dalam menghadapi noise linguistik dan kondisi data yang tidak seimbang (imbalanced) pada kolom komentar YouTube.

Sebagai solusi dan **kebaruan (*novelty*)**, penelitian ini mengusulkan pendekatan yang lebih komprehensif melalui penerapan *Ensemble Voting Classifier*. Model ini mengombinasikan SVM terkalibrasi, *Logistic Regression*, dan *Random Forest* dengan metode *soft voting*. Berbeda dengan penelitian terdahulu [2], [3], integrasi *ensemble* ini yang juga dipadukan dengan metode SMOTE untuk mengatasi ketidakseimbangan data dan *Feature Union* (berbasis TF-IDF kata dan karakter) untuk mempertahankan konteks bahasa belum pernah diterapkan dalam konteks analisis sentimen opini olahraga Timnas Indonesia di platform YouTube.

2. Bahan dan Metode

Penelitian ini dilakukan melalui tahapan-tahapan yang disusun secara sistematis dan berurutan, sebagaimana ditampilkan pada Gambar 1. Setiap tahapan dirancang dengan tujuan untuk memastikan proses pengembangan model machine learning berlangsung secara terstruktur, sehingga model yang dihasilkan mampu mencapai kinerja yang optimal serta selaras dengan tujuan penelitian.



Gambar 1. Alur Penelitian

Alur penelitian ini mencakup beberapa tahapan utama. Sebagaimana diilustrasikan pada Gambar 1, dimulai dari tahap pengambilan data melalui proses scraping untuk memperoleh korpus opini publik secara langsung dari berbagai platform digital. Data mentah yang diperoleh kemudian diproses pada tahap pelabelan guna mengklasifikasikan teks berdasarkan polaritas sentimen sebagai acuan ground truth. Selanjutnya, data tersebut melalui tahap preprocessing untuk membersihkan noise linguistik serta menyeragamkan struktur bahasa agar lebih siap diolah oleh sistem komputasi. Hasil dari proses ini kemudian digunakan dalam tahap ekstraksi fitur yang bertujuan mengubah data teks menjadi representasi numerik dalam bentuk vektor. Representasi tersebut selanjutnya dimanfaatkan pada tahap pemodelan untuk membangun dan melatih algoritma machine learning. Keseluruhan proses ini diakhiri dengan tahap evaluasi, di mana performa model diukur secara objektif berdasarkan metrik seperti akurasi, presisi, dan keandalan guna menghasilkan kesimpulan yang valid [5].

2.1 Scraping Data

Penelitian ini dimulai dengan pengumpulan data opini publik dari YouTube yang dapat diakses melalui https://www.youtube.com/watch?v=h2Rs6u41nIg. melalui teknik web scraping otomatis dengan metode batching untuk menghindari pembatasan system. Data difokuskan pada video terkait pertandingan Indonesia melawan Bahrain, dengan pengambilan komentar berdasarkan popularitas dan kebaruan guna memperoleh representasi yang lebih seimbang. Selanjutnya, data dibersihkan dari duplikasi dan disimpan dalam format CSV sebagai korpus mentah yang siap dianalisis.

Tabel 1. Contoh data mentah

Username	Teks Komentar	Likes
@RCTIMEGAENTERTAINMENT	Kesal Sedih kecewa jadi satu!!! Gimana pertandingan ini menurut kalian??	4.800
@Wanderlust58	Indonesia should have win the match. Big respect from London❤️❤️❤️	1.300
@Gua_1212	Demi Allah, Dzat yang maha Esa dan Kuasa, berikanlah Keadilanmu atas Kedzaliman dan Kecurangan yang terjadi dari Pertandingan ini	453
@SaifulBahri-e1l	Kesini gara gara kangen coach shin yg bgtu semangat ngasi instruksi gk jayak5 kang Patrick yg cuma plonga plongo	11

Sebagaimana terlihat dari Tabel 1, dari proses *scraping* tersebut, berhasil dikumpulkan sebanyak 1.000 baris data komentar mentah (*raw data*). Setiap baris data memuat berbagai atribut metadata, termasuk nama pengguna (*username*), teks komentar (*comment*), jumlah suka (*likes*).

2.2 Labeling Data

Proses kedua merupakan anotasi sentimen pada data mentah, dalam penelitian ini dilakukan dengan pendekatan semi-otomatis berbasis leksikon untuk mencapai keseimbangan antara efisiensi dan tingkat akurasi [6]. Pada tahap awal, sistem menghitung frekuensi kemunculan kata-kata yang mengandung muatan positif, negatif, dan netral berdasarkan kamus leksikon kustom yang dibuat secara mandiri khusus untuk domain olahraga sepak bola (*custom domain-specific lexicon*).

Penggunaan kamus buatan sendiri ini dipilih karena kamus standar yang tervalidasi seperti InSet atau SentiWordNet Indonesia umumnya berbasis bahasa formal dan belum mencakup variasi bahasa informal, istilah *slang* media sosial, serta ekspresi kontekstual khusus yang muncul dalam kontroversi pertandingan. Kamus kustom ini disusun dengan memasukkan kata kunci spesifik yang relevan dengan objek penelitian, seperti istilah *slang* dan frasa kritik ('wasit tidak adil', 'mafia', 'botak', 'curang', 'ancur', 'noob') untuk kelas negatif, serta kata kunci ekspresi dukungan ('garuda', 'mantap', 'respect', 'indonesia bisa', 'joss') untuk kelas positif.

Berdasarkan perhitungan tersebut, sistem kemudian menentukan kelas sentimen sekaligus menghasilkan tingkat keyakinan prediksi dalam bentuk confidence score. Untuk mengurangi potensi bias dari proses pelabelan otomatis, data dengan nilai confidence rendah (≤ 0,5) dipisahkan dan selanjutnya melalui proses validasi manual oleh peneliti atau pakar melalui pendekatan *human-in-the-loop*. Dengan mekanisme ini, kualitas label sebagai *ground truth* dapat lebih terjamin sebelum digunakan dalam proses pelatihan model.

2.3 Preprocessing

Setelah melalui proses pelabelan, korpus data kemudian diproses lebih lanjut pada tahap prapemrosesan teks dengan tujuan menghilangkan *noise* sekaligus menstandarkan bentuk bahasa yang umumnya tidak baku pada media social [7]. Tahap ini mencakup

beberapa langkah, seperti *case folding* untuk menyeragamkan huruf, penghapusan elemen metadata seperti URL, *mention, hashtag*, dan karakter non-ASCII, serta normalisasi kata tidak baku menjadi bentuk baku menggunakan kamus khusus. Salah satu aspek kebaruan dalam penelitian ini terletak pada penerapan strategi penanganan negasi, di mana kata negasi seperti "tidak" digabungkan secara langsung dengan kata sifat yang mengikutinya menggunakan tanda underscore, misalnya "tidak_adil". Pendekatan ini bertujuan untuk menjaga makna kontekstual agar tetap utuh dan tidak mengalami distorsi saat memasuki tahap *stemming* dengan bantuan *library Sastrawi* maupun proses penghapusan *stopword*.

2.4 Ekstraksi Fitur

Teks yang telah melalui tahap pembersihan selanjutnya diubah ke dalam bentuk representasi numerik agar dapat diproses oleh model komputasi berbasis matematis. Transformasi ini dilakukan dengan memanfaatkan arsitektur *Feature Union* yang mengintegrasikan beberapa jenis fitur secara parallel [8]. Pada dimensi pertama, digunakan pembobotan TF-IDF berbasis kata untuk mengidentifikasi istilah yang memiliki tingkat kepentingan tinggi dalam dokumen. Dimensi kedua melengkapi representasi tersebut melalui TF-IDF berbasis karakter dalam bentuk *n-gram*, yang berfungsi untuk menangkap pola morfologis seperti imbuhan maupun variasi penulisan yang tidak baku [9]. Selain itu, dimensi ketiga mencakup ekstraksi fitur leksikal tambahan, seperti proporsi penggunaan huruf kapital, jumlah tanda seru, panjang teks, serta skor berbasis leksikon dasar. Seluruh fitur tambahan ini kemudian dinormalisasi menggunakan *MinMaxScaler* agar berada dalam rentang nilai yang seragam sebelum digunakan dalam proses pemodelan.

Tabel 2. Rincian Ekstraksi Fitur

Jenis Pipeline	Algoritma	Deskripsi dan Parameter	Tujuan Ekstraksi
Fitur Word	*Term Frequency-Inverse Document Frequency* (TF-IDF) Tingkat Kata	Menggunakan analyzer="word" dengan pembobotan skala logaritmik (sublinear_tf=True) dan batas maksimal 10.000 fitur konseptual teratas. Diekstrak dari teks bersih (clean).	Menangkap bobot signifikansi kata tunggal yang sering muncul dalam sebuah komentar namun jarang muncul di komentar lain (keunikan term).
Fitur Char	TF-IDF Tingkat Karakter (*N-Gram*)	Menggunakan analyzer="char" dengan konfigurasi *n-gram* rentang 2 hingga 4 karakter, serta batas maksimal 5.000 fitur. Diekstrak dari teks bersih (clean).	Menangkap pola morfologis pembentukan kata, seperti imbuhan, singkatan tak baku, maupun kesalahan ketik (*typo*) ejaan yang lolos dari normalisasi.
Fitur Numerik	*Custom Text Feature Extractor* & *MinMaxScale*	Mengekstraksi 10 variabel metadata spesifik dari teks mentah (raw). Hasilnya kemudian diskalakan ke rentang 0-1 menggunakan *MinMaxScaler* agar homogen dengan nilai TF-IDF.	Merekam indikator emosi atau karakteristik linguistik eksplisit audiens melalui panjang kalimat, tanda baca, dan skor kamus leksikal dasar.

Integrasi tiga dimensi ekstraksi fitur melalui arsitektur Feature Union seperti yang terlihat pada table 2, menghasilkan representasi vektor berdimensi tinggi yang komprehensif. Pendekatan multidimensi ini mampu mengatasi keterbatasan metode konvensional seperti bag-of-words yang hanya berfokus pada frekuensi kata. Dengan menggabungkan n-gram berbasis karakter serta fitur struktural dan emosional misalnya

penggunaan huruf kapital dan tanda baca model dapat menangkap nuansa ekspresi, intensitas emosi, serta variasi bahasa informal yang umum ditemukan di YouTube. Representasi fitur yang dihasilkan kemudian digunakan sebagai variabel prediktor (X). Mengingat distribusi kelas yang cenderung tidak seimbang pada data opini publik, matriks fitur tersebut diseimbangkan menggunakan metode SMOTE sebelum dilatih dengan arsitektur Ensemble Voting Classifier untuk menghasilkan prediksi yang lebih optimal dan minim bias.

2.5 Modeling

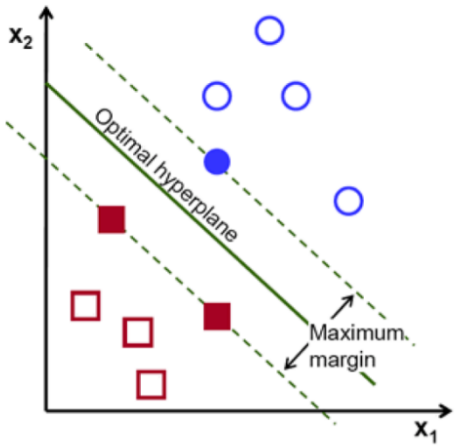
Tahap ini menjadi bagian utama dalam pengembangan sistem cerdas yang diusulkan. Sebelum proses pelatihan dilakukan, dataset terlebih dahulu dibagi menjadi data latih sebesar 80% dan data uji sebesar 20%. Untuk mengatasi permasalahan ketidakseimbangan kelas yang umum ditemukan pada data opini, data latih kemudian diseimbangkan melalui teknik augmentasi sintetis menggunakan metode *Synthetic Minority Over-sampling Technique* (SMOTE) [10]. Sebelum dilakukan augmentasi, distribusi data latih didominasi oleh kelas netral (sekitar 450 sampel). Melalui teknik ini, sampel sintetis dibuat pada kelas minoritas (positif dan negatif) hingga menyamai jumlah sampel kelas mayoritas (netral). Hasilnya, jumlah total data latih setelah augmentasi meningkat menjadi 1.350 sampel, di mana masing-masing kelas sentimen memiliki distribusi yang seimbang yaitu sebanyak 450 sampel. Model yang dikembangkan menggunakan pendekatan Ensemble Voting Classifier dengan mekanisme *soft voting*, yang menggabungkan probabilitas prediksi dari tiga algoritma dasar, yaitu Support Vector Machine (SVM) diimplementasikan menggunakan LinearSVC dengan parameter bobot kelas seimbang (*class_weight='balanced'*) untuk menangani masalah ketidakseimbangan data latih, serta batas iterasi maksimum (max_iter) ditetapkan sebesar 5.000 untuk memastikan konvergensi model pada ruang fitur berdimensi tinggi. Logistic Regression dikonfigurasi dengan pembobotan kelas seimbang (*class_weight='balanced'*) untuk meminimalkan bias terhadap kelas mayoritas, serta batas iterasi maksimum (*max_iter*) sebesar 1.000 guna mengoptimalkan konvergensi pencarian bobot koefisien pada matriks teks yang bersifat *sparse*.Selanjutnya Random Forest dikonfigurasi sebagai model berbasis *bagging* dengan jumlah pohon keputusan sebanyak 100 unit (n_estimators=100) untuk menangkap pola linguistik non-linier yang kompleks, serta dilengkapi dengan penyesuaian bobot kelas secara dinamis menggunakan parameter *class_weight='balanced'*. Untuk menjaga konsistensi, objektivitas, serta keterulangan eksperimen (*reproducibility*), seluruh algoritma dasar di atas dikunci menggunakan parameter pengacak yang sama, yaitu random_state=42. Konfigurasi awal ini berfungsi sebagai *baseline* kokoh sebelum sistem melakukan pencarian otomatis terhadap kombinasi parameter regularisasi (C) dan batas fitur maksimum (*max_features*) terbaik melalui metode *GridSearchCV*.

2.5.1 Ensemble Voting Classifier

Penelitian ini menawarkan kebaruan dalam perancangan arsitektur prediktif dengan tidak hanya mengandalkan satu model klasifikasi, tetapi menggunakan pendekatan *Ensemble Voting Classifier*. Metode ini dikembangkan untuk meningkatkan akurasi sekaligus menjaga stabilitas kemampuan generalisasi model melalui penggabungan prediksi dari beberapa algoritma dasar. Pendekatan yang digunakan adalah *soft voting*, di mana keputusan akhir ditentukan berdasarkan rata-rata probabilitas kelas yang dihasilkan oleh masing-masing model. Arsitektur ensemble ini dibentuk oleh tiga algoritma utama yang bekerja secara bersama dalam menghasilkan prediksi.

a. Support Vector Machine (SVM)

Dalam konteks klasifikasi teks berbasis TF-IDF, SVM ber-kernel linier banyak digunakan karena efisien secara komputasi serta mampu menangani data berdimensi besar yang bersifat sparse dengan baik [12].

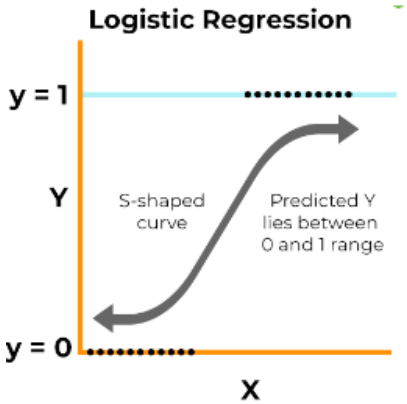


Gambar 2. *Support Vector Machine*

Pada Gambar 2, merupakan algoritma pembelajaran mesin yang bekerja dengan menentukan hyperplane optimal untuk memisahkan data dalam ruang fitur berdimensi tinggi dengan margin maksimum. Diimplementasikan menggunakan kernel *LinearSVC* yang dikenal sangat efisien dalam menangani data teks berdimensi tinggi. Untuk memastikan keluaran SVM kompatibel dengan mekanisme *soft voting* (yang membutuhkan nilai probabilitas), *CalibratedClassifierCV* diterapkan untuk mengonversi jarak *margin hyperplane* SVM menjadi estimasi probabilitas menggunakan kalibrasi fungsi logistik bersilang. Di dalamnya, model inti menggunakan LinearSVC dengan parameter C=1.0 dan dual=False untuk optimalisasi pada dataset teks yang memiliki jumlah fitur lebih besar dari jumlah sampel. Kalibrasi dilakukan melalui metode validasi silang internal (cv=3) guna mentransformasikan skor jarak *hyperplane* menjadi estimasi probabilitas yang konsisten. Selain itu, parameter class_weight='balanced' diterapkan untuk memberikan penalti yang lebih tinggi pada kesalahan klasifikasi di kelas minoritas.

b. Logistic Regression (LR)

Logistic Regression merupakan salah satu algoritma klasifikasi yang paling sederhana namun tetap efektif dalam memodelkan hubungan antara sekumpulan fitur dengan probabilitas suatu kejadian [13].

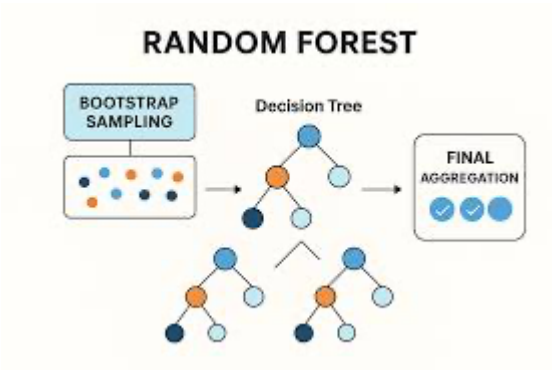


Gambar 3. *Logistic Regression*

Pada Gambar 3, algorima ini bekerja dengan memetakan nilai fitur masukan menjadi nilai probabilitas, sehingga sangat sesuai untuk menangani klasifikasi multikelas pada analisis sentimen komentar YouTube. Salah satu keunggulan Logistic Regression adalah kemampuannya menghasilkan prediksi dalam bentuk distribusi probabilitas untuk setiap kelas (positif, negatif, dan netral), bukan sekadar label kaku. Hal ini memungkinkan interpretasi yang lebih mendalam terhadap tingkat keyakinan model dalam menentukan polaritas opini publik terkait kontroversi pertandingan Indonesia vs Bahrain. Menggunakan mode multi_class='multinomial' dengan *solver* lbfgs untuk menangani klasifikasi tiga kelas (positif, negatif, netral) secara simultan. Parameter max_iter ditetapkan pada angka 1000 untuk menjamin konvergensi algoritma pada ruang fitur yang kompleks hasil ekstraksi TF-IDF. Seperti halnya SVM, model ini juga menggunakan class_weight='balanced' guna memastikan model tetap sensitif terhadap distribusi data yang tidak seimbang pasca proses SMOTE.

c. *Random Forest* (RF)

Random Forest merupakan algoritma pembelajaran mesin yang dikembangkan untuk menyelesaikan permasalahan klasifikasi maupun regresi dengan memanfaatkan sekumpulan pohon keputusan sebagai dasar prediksi [14]. Dalam proses kerjanya, setiap pohon keputusan menghasilkan prediksi yang kemudian digabungkan melalui mekanisme mayoritas suara untuk menentukan hasil akhir. Pendekatan ini memungkinkan Random Forest mampu mengenali pola nonlinier yang kompleks pada data teks secara lebih efektif, sekaligus mengurangi varians model dan meminimalkan risiko overfitting yang umumnya terjadi pada penggunaan pohon keputusan tunggal.



Gambar 4. *Random Forest*

Komponen *Random Forest* dikonfigurasi sebagai model berbasis *bagging* dengan jumlah pohon keputusan sebanyak 100 unit (*n_estimators*=100). Setiap pohon dibangun secara acak untuk mendeteksi korelasi non-linier antara fitur numerik (seperti rasio huruf kapital) dan teks. Penggunaan n_jobs=-1 memungkinkan proses pelatihan dilakukan secara paralel pada seluruh *core* prosesor untuk efisiensi waktu komputasi. Strategi penyeimbangan kelas juga diintegrasikan melalui parameter *class_weight='balanced_subsample'* yang menyesuaikan bobot secara dinamis pada setiap *bootstrap* sampel.

Seluruh konfigurasi model tersebut kemudian diintegrasikan ke dalam arsitektur VotingClassifier, di mana keputusan akhir ditentukan berdasarkan rata-rata probabilitas tertinggi dari setiap algoritma. Pendekatan ini memungkinkan model tetap mempertahankan stabilitas performa ketika salah satu pengklasifikasi mengalami ketidakpastian dalam mengenali pola teks tertentu, seperti komentar yang bersifat sarkastik, karena prediksi dari model lain dapat saling melengkapi. Selanjutnya,

arsitektur ini dioptimalkan menggunakan GridSearchCV untuk memperoleh kombinasi parameter regularisasi dan kompleksitas fitur yang paling optimal.

2.5.2 Optimalisasi *Hyperparameter* (*Hyperparameter Tuning*)

Performa arsitektur *Ensemble Voting Classifier* sangat dipengaruhi oleh konfigurasi parameter pada setiap algoritma dasar serta proses ekstraksi fitur yang digunakan. Untuk memperoleh kinerja yang optimal sekaligus mengurangi risiko overfitting, penelitian ini menerapkan penalaan hyperparameter secara otomatis menggunakan metode *GridSearchCV*. Teknik ini bekerja dengan menelusuri berbagai kombinasi hyperparameter yang telah ditentukan sebelumnya secara iteratif guna menemukan konfigurasi model yang paling efektif.

Dalam penerapannya, proses pencarian parameter difokuskan pada beberapa komponen penting dalam pipeline, seperti penentuan jumlah fitur maksimum (*max_features*) pada ekstraksi Word TF-IDF dan Character TF-IDF, serta pengaturan nilai parameter regularisasi (*C*) pada algoritma Logistic Regression. Untuk menjaga objektivitas dan kestabilan model selama proses optimasi, setiap kombinasi parameter divalidasi menggunakan metode *Stratified K-Fold Cross Validation* dengan tiga lipatan ($k = 3$). Pendekatan stratified digunakan agar proporsi distribusi kelas sentimen tetap seimbang pada setiap pembagian data validasi. Seluruh kombinasi model kemudian dievaluasi berdasarkan nilai *weighted F1-Score*, dan konfigurasi yang memperoleh skor validasi tertinggi dipilih sebagai parameter akhir dalam sistem prediksi.

2.6 Evaluasi

Dalam penelitian ini, evaluasi kinerja model dilakukan menggunakan sejumlah metrik standar yang umum digunakan dalam machine learning. Metrik-metrik tersebut berfungsi sebagai indikator untuk menilai tingkat ketepatan prediksi model, yang selanjutnya akan dijelaskan secara lebih rinci pada bagian berikutnya [15].

Sebelum membahas rumus perhitungan, terlebih dahulu perlu dipahami empat komponen utama dalam confusion matrix pada tugas klasifikasi. True Positive (TP) mengacu pada jumlah data yang sebenarnya bernilai positif dan berhasil diprediksi positif oleh model. True Negative (TN) menunjukkan jumlah data yang memang bernilai negatif dan juga diprediksi negatif dengan tepat. Sementara itu, False Positive (FP) merupakan data yang sebenarnya negatif namun secara keliru diprediksi sebagai positif, sedangkan False Negative (FN) adalah data yang sebenarnya positif tetapi salah diklasifikasikan sebagai negatif oleh model.

2.6.1 Akurasi

Akurasi adalah tingkat kedekatan nilai yang diprediksi oleh model dengan nilai yang sebenarnya. Metrik ini mengukur seberapa banyak prediksi yang benar secara keseluruhan (baik positif maupun negatif) dibandingkan dengan total seluruh data [15]. Rumus akurasi dinyatakan sebagai:

$$Akurasi = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

2.6.2 Presisi

Presisi adalah ukuran seberapa baik model dapat memberikan jawaban yang benar kepada pengguna saat model tersebut memprediksi kelas positif. Metrik ini sangat berguna untuk melihat seberapa banyak prediksi positif yang *benar-benar* positif (meminimalisasi *False Positive*) [15]. Rumus presisi dinyatakan sebagai:

$$Presisi = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

2.6.3 Recall

Recall adalah tingkat keberhasilan sebuah model dalam menemukan atau mengenali kembali semua contoh yang relevan di dalam data. Metrik ini fokus pada kemampuan model meminimalisasi *False Negative* (memastikan tidak ada data positif yang terlewat) [15]. Rumus recall dinyatakan sebagai:

$$Recall = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

2.6.4 F1-Score

F1-Score merupakan hasil dari evaluasi yang membandingkan (mengambil rata-rata harmonis) antara nilai *Precision* dengan nilai Recall. Metrik ini sangat penting digunakan jika data Anda tidak seimbang (*imbalanced data*), karena menggabungkan keseimbangan antara presisi dan recall [15]. Rumus F1-Score dinyatakan sebagai:

$$F1 - Score = 2\text{x}\frac{\text{Presisi x Recall}}{\text{Presisi} + \text{Recall}}$$

3. Hasil

Bagian ini menyajikan hasil eksperimen yang dilakukan secara sistematis mengikuti tahapan metodologi yang telah dirancang. Pembahasan dimulai dari hasil scraping data hingga evaluasi kinerja akhir menggunakan metrik standar.

3.1. Scraping Data

Dalam penelitian ini, proses pengumpulan data dilakukan tanpa menggunakan API resmi YouTube, melainkan melalui teknik web scraping menggunakan library Python youtube-comment-downloader. Pendekatan ini dipilih karena mampu mengekstraksi data berformat JSON langsung dari antarmuka frontend YouTube tanpa memerlukan API Key, serta lebih fleksibel dalam menghindari pembatasan kuota pengambilan data yang umumnya diterapkan pada API resmi.

```
============================================================
  SCRAPING KOMENTAR YOUTUBE — BATCH MODE
============================================================

============================================================
[Video 1/1] https://www.youtube.com/watch?v=h2Rs6u41nIg
============================================================
  Target    : 1000 komentar
  Batch size: 200 | Jeda: 3s

  Mode [Populer]
    Batch #1 (Populer) — target batch: 200 | terkumpul: 0 ... ✅ +200 | Total: 200
    Batch #2 (Populer) — target batch: 200 | terkumpul: 200 ... ✅ +200 | Total: 400
    Batch #3 (Populer) — target batch: 200 | terkumpul: 400 ... ✅ +200 | Total: 600
    Batch #4 (Populer) — target batch: 200 | terkumpul: 600 ... ✅ +200 | Total: 800
    Batch #5 (Populer) — target batch: 200 | terkumpul: 800 ... ✅ +200 | Total: 1000

  Video selesai: 1000 komentar | Grand total: 1000

============================================================
TOTAL KOMENTAR: 1000
✅ Disimpan ke: dataset_komentar_raw.csv

STATISTIK:
  Total komentar  : 1000
  Komentar utama  : 64
  Balasan (reply) : 936
```

Gambar 5. Hasil *Scarping Data*

Berdasarkan pada Gambar 5, proses pengambilan data dilakukan secara bertahap dan berurutan dengan menetapkan kapasitas sebanyak 200 komentar pada setiap batch,

disertai jeda waktu selama 3 detik antar proses pengambilan. Pendekatan heuristik tersebut menunjukkan kinerja yang efektif dalam mendukung proses ekstraksi data. Melalui mekanisme ini, sistem berhasil memenuhi target pengumpulan data dengan memperoleh sebanyak 1.000 baris komentar mentah (raw data) dari tautan video yang menjadi objek penelitian. Seluruh data yang berhasil dikumpulkan, meliputi atribut pengguna, isi komentar, serta jumlah likes, kemudian diekspor dan disimpan dalam format Comma Separated Values (CSV) dengan nama berkas dataset_komentar_raw.csv. Dataset mentah tersebut selanjutnya digunakan sebagai input pada tahap pelabelan sentimen (sentiment labeling).

3.2. Hasil Pelabelan Data

Setelah proses pengumpulan data selesai dilakukan, dataset mentah yang terdiri atas 1.000 komentar memasuki tahap anotasi sentimen guna menentukan kelas ground truth pada setiap data, yaitu positif, negatif, atau netral. Tahapan pelabelan ini menggunakan pendekatan hybrid dengan mengombinasikan metode pelabelan otomatis berbasis leksikon (*lexicon-based auto-suggest*) dan proses validasi manual oleh peneliti (*human-in-the-loop*).

Pada tahap awal, sistem melakukan identifikasi serta pencocokan kata dasar dari setiap komentar dengan kamus leksikon bahasa Indonesia yang telah disiapkan sebelumnya. Hasil pencocokan tersebut kemudian digunakan untuk menghitung frekuensi kemunculan kata-kata tertentu sebagai dasar dalam menentukan kecenderungan sentimen secara otomatis. Melalui mekanisme ini, algoritma mampu menghasilkan distribusi awal sentimen publik terhadap data komentar yang dianalisis. Untuk memastikan kualitas dan reliabilitas data latih, setiap hasil pelabelan otomatis juga dievaluasi secara ketat menggunakan nilai probabilitas keyakinan (confidence score) sebelum memasuki tahap validasi lanjutan.

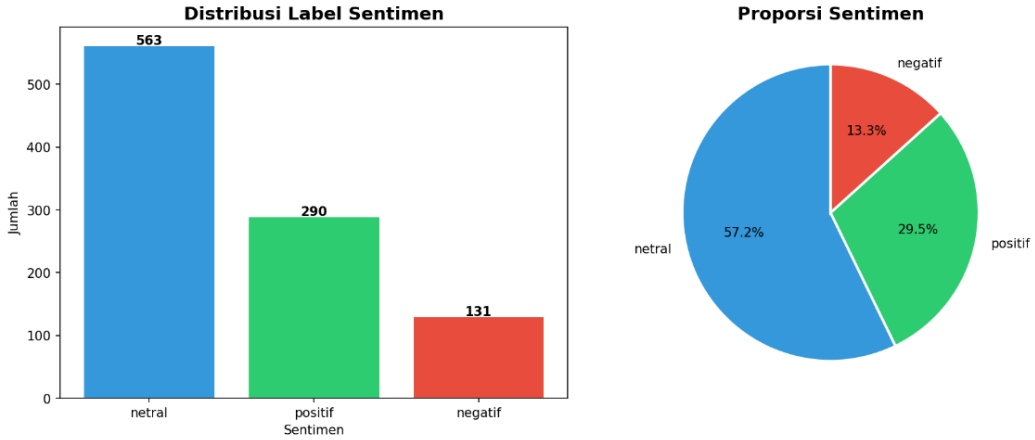
Tabel 3. Distribusi Final Label Sentimen Komentar YouTube

Kelas Sentimen	Deskripsi Dominan	Jumlah Data	Presentase
Negatif	Berisi kritik tajam, kekecewaan, kemarahan, dan penggunaan *slang* kasar terkait wasit atau hasil pertandingan.	131	13.3%
Netral	Berisi pertanyaan, diskusi taktis tanpa tendensi emosional, atau informasi di luar konteks pertandingan.	563	57.2%
Positif	Berisi dukungan, apresiasi, pujian, dan ungkapan kebanggaan terhadap performa Timnas Indonesia.	290	29.5%
Total		**984**	**100%**

Tabel 3 menyajikan distribusi final label sentimen komentar publik terhadap kontroversi pertandingan Indonesia vs Bahrain di platform YouTube setelah melalui proses pelabelan otomatis dan validasi manual. Distribusi ini menggambarkan kecenderungan opini publik yang terbentuk berdasarkan analisis terhadap 984 komentar yang valid dan berhasil diklasifikasikan ke dalam tiga kategori sentimen, yaitu negatif, netral, dan positif. Setiap kategori merepresentasikan karakteristik respons audiens yang berbeda terhadap isu pertandingan yang dianalisis.

Berdasarkan hasil distribusi tersebut, sentimen netral menjadi kategori yang paling dominan dengan jumlah 563 komentar atau 57,2% dari keseluruhan data. Dominasi ini menunjukkan bahwa sebagian besar pengguna memberikan komentar yang bersifat informatif, pertanyaan, maupun diskusi umum tanpa kecenderungan emosi yang kuat.

Sementara itu, sentimen positif tercatat sebanyak 290 komentar atau 29,5%, yang umumnya berisi dukungan, apresiasi, dan ungkapan kebanggaan terhadap performa Timnas Indonesia. Di sisi lain, sentimen negatif berjumlah 131 komentar atau 13,3%, yang mayoritas memuat kritik, kekecewaan, kemarahan, serta penggunaan bahasa kasar terkait keputusan pertandingan maupun kinerja pihak tertentu. Distribusi akhir ini selanjutnya digunakan sebagai dasar dalam proses pemodelan dan evaluasi performa metode *Ensemble Voting Classifier* pada penelitian ini.



Gambar 6. Plot Distribusi Label

Gambar 6 menunjukkan implementasi sistem pelabelan sentimen otomatis pada komentar publik terkait kontroversi pertandingan Indonesia vs Bahrain di platform YouTube menggunakan pendekatan berbasis leksikon (*lexicon-based labeling*). Proses dimulai dengan tahap *preprocessing* berupa normalisasi teks, penghapusan URL, *mention*, *hashtag*, dan karakter non-alfabetik sebelum sistem mengklasifikasikan komentar ke dalam kategori positif, negatif, atau netral berdasarkan kamus sentimen bahasa Indonesia. Selain menentukan label sentimen, sistem juga menghitung tingkat keyakinan (*confidence score*) untuk setiap komentar.

Hasil pengujian menunjukkan bahwa sebanyak 984 komentar berhasil diproses, dengan distribusi sentimen terdiri atas 563 komentar netral (57,2%), 290 komentar positif (29,5%), dan 131 komentar negatif (13,1%). Dominasi sentimen netral menunjukkan bahwa sebagian besar pengguna memberikan komentar yang bersifat informatif dan diskusi umum terkait pertandingan. Sistem juga mengidentifikasi 563 komentar dengan nilai *confidence* ≤ 0,5 yang memerlukan validasi manual akibat tingginya penggunaan bahasa slang, campuran bahasa asing, dan ekspresi sarkastik pada media sosial YouTube. Oleh karena itu, komentar dengan tingkat keyakinan rendah dipisahkan ke dalam berkas *dataset_review_manual.csv* untuk dilakukan koreksi manual sebelum digunakan pada tahap pemodelan *Ensemble Voting Classifier*.

3.3. Hasil Prapemrosesan Teks

Korpus komentar yang telah dilengkapi dengan label *ground truth* kemudian memasuki tahap prapemrosesan teks komputasional sebagai bagian penting sebelum proses pemodelan dilakukan. Karakteristik bahasa pada media sosial YouTube cenderung tidak terstruktur karena pengguna sering menggunakan karakter non-alfabetik, tautan (URL), bahasa gaul (slang), serta berbagai variasi penulisan yang tidak baku. Kondisi tersebut menyebabkan proses preprocessing menjadi sangat penting untuk menstandarkan representasi fitur teks agar lebih konsisten dan mudah dianalisis oleh sistem.

Pada tahap ini, sistem menjalankan serangkaian proses pembersihan data secara bertahap dan berurutan. Proses tersebut mencakup penghapusan tautan (URL) dan sebutan (*mention*), konversi seluruh teks menjadi huruf kecil (*case folding*), serta eliminasi karakter khusus dan tanda baca dengan pengecualian pada karakter garis bawah. Setelah itu, dilakukan normalisasi kata tidak baku dengan memetakan berbagai istilah slang, seperti "bgt" atau "yg", ke dalam bentuk bahasa Indonesia yang lebih standar, yaitu "sangat" dan "yang", menggunakan kamus kustom yang telah disiapkan sebelumnya.

Salah satu capaian paling penting dalam tahapan prapemrosesan ini adalah keberhasilan penerapan algoritma *negation handling*. Sistem mampu mengidentifikasi kata-kata negasi, seperti "tidak", "bukan", dan "belum", kemudian menggabungkannya secara leksikal dengan kata sifat yang mengikutinya menggunakan karakter *underscore*. Sebagai contoh, frasa "tidak adil" diubah menjadi "tidak_adil". Pendekatan tersebut terbukti efektif dalam mempertahankan makna sentimen yang terkandung di dalam kalimat, sehingga nuansa negatif maupun positif tidak hilang ketika data memasuki tahap penghapusan kata hubung (*stopword removal*).

Tabel 4. Hasil Preprocessing Teks

Teks Mentah	Teks Bersih	Keterangan
Wah wasitnya curang bgt sih!!! 😡 merugikan indonesia	wah wasit curang sangat rugi indonesia	Hapus emoji, URL, tanda baca; normalisasi "bgt" -> "sangat"; *stopword removal*.
Permainan timnas hari ini tidak bagus, sy kecewa sm pelatihnya.	main timnas hari tidak_bagus saya kecewa sama latih	Penanganan negasi ("tidak bagus" -> "tidak_bagus"); normalisasi "sy" dan "sm".
@AFC_Official INI BUKAN SEPAK BOLA, INI MAFIA!	ini bukan_sepak bola ini mafia	*Case folding*; hapus *mention* @; Penanganan negasi ("bukan sepak" -> "bukan_sepak").

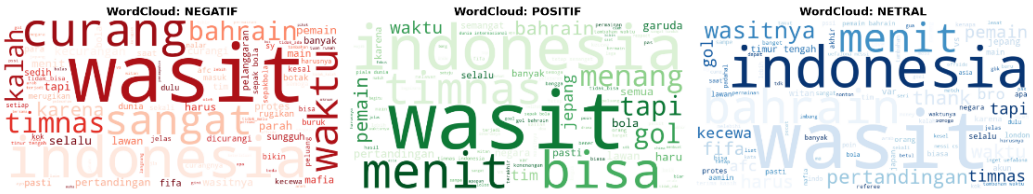
Tabel 4 mendemonstrasikan perbandingan antara teks mentah awal dan hasil akhir teks bersih (*clean text*) yang telah melewati keseluruhan *pipeline preprocessing*.

3.4. Hasil Pemodelan dan Optimalisasi Hyperparameter

Tahap pemodelan dimulai dengan membagi 1.000 data teks yang telah melalui proses pembersihan ke dalam dua kelompok, yaitu data pelatihan sebesar 80% atau 800 data dan data pengujian sebesar 20% atau 200 data menggunakan metode stratified split. Pembagian ini dilakukan untuk menjaga proporsi distribusi kelas sentimen agar tetap konsisten pada kedua kelompok data. Karena distribusi kelas pada dataset awal cenderung tidak seimbang, algoritma SMOTE diterapkan secara khusus pada data pelatihan guna menghasilkan sampel sintetis pada kelas minoritas hingga tercapai keseimbangan proporsi antar ketiga kategori sentimen.

Setelah proses penyeimbangan data selesai dilakukan, model Ensemble Voting Classifier dilatih menggunakan kombinasi fitur multidimensi. Untuk memperoleh performa yang optimal, sistem menerapkan proses optimasi hyperparameter menggunakan metode GridSearchCV dengan skema validasi silang tiga lipatan (3-*Fold Cross Validation*). Melalui proses pencarian iteratif terhadap berbagai kombinasi parameter, sistem berhasil menemukan konfigurasi komputasi yang paling efektif. Konfigurasi terbaik tersebut menggunakan batas maksimum 5.000 fitur konseptual pada representasi Word TF-IDF, 3.000 fitur morfologis pada Character TF-IDF, serta nilai parameter regularisasi (C) sebesar 10 pada algoritma Logistic Regression. Selama fase pelatihan, konfigurasi ini menunjukkan performa yang paling stabil dan robust dengan rata-rata nilai Cross-Validation F1-

Score mencapai 0,8817 atau 88,17%. Oleh karena itu, konfigurasi tersebut kemudian ditetapkan sebagai model final untuk tahap evaluasi lebih lanjut.



Gambar 7. *Word cloud*

Visualisasi WordCloud pada Gambar 7 menampilkan representasi leksikal dari kumpulan komentar YouTube yang telah melalui tahapan preprocessing teks komputasional. Visualisasi ini digunakan untuk menggambarkan kecenderungan topik pembahasan serta intensitas emosional audiens digital pada masing-masing kategori sentimen secara lebih intuitif.

Pada kategori sentimen negatif, korpus komentar didominasi oleh kemunculan kata-kata seperti *"wasit"*, *"curang"*, *"waktu"*, dan *"bahrain"*. Ukuran kata yang paling menonjol pada istilah *"curang"* dan *"wasit"* menunjukkan bahwa ketidakpuasan publik banyak diarahkan pada keputusan pengadil pertandingan dan durasi waktu permainan yang dianggap kontroversial. Temuan ini juga sejalan dengan hasil pemodelan yang menunjukkan bahwa fitur-fitur tersebut memiliki kontribusi bobot yang tinggi dalam proses klasifikasi sentimen.

Sementara itu, pada sentimen positif, kata-kata dominan yang muncul antara lain *"indonesia"*, *"bisa"*, *"menang"*, dan *"timnas"*. Kemunculan kata-kata tersebut merefleksikan bentuk dukungan, apresiasi, dan optimisme audiens terhadap performa tim nasional Indonesia. Menariknya, istilah *"wasit"* juga tetap muncul pada kelompok ini, namun dalam konteks yang lebih menekankan apresiasi terhadap perjuangan tim di tengah dinamika pertandingan. Hal ini selaras dengan hasil evaluasi model yang menunjukkan performa terbaik pada kelas positif dengan nilai *F1-score* sebesar 0,91.

Pada kategori sentimen netral, distribusi kata lebih banyak berfokus pada informasi teknis pertandingan, seperti *"menit"*, *"bahrain"*, *"indonesia"*, dan *"gol"*. Dominasi kata *"menit"* menunjukkan bahwa sebagian besar komentar dalam kategori ini cenderung bersifat informatif, berupa pelaporan jalannya pertandingan atau diskusi kronologi permainan tanpa disertai emosi yang kuat.

Secara keseluruhan, visualisasi *WordCloud* ini menunjukkan bahwa proses *preprocessing* berhasil menstandarkan berbagai bentuk bahasa tidak baku dan slang menjadi representasi linguistik yang lebih konsisten. Dengan demikian, fitur-fitur teks yang dihasilkan memiliki relevansi semantik yang lebih baik dan mampu mendukung kinerja *Ensemble Voting Classifier* dalam memetakan polaritas opini publik terhadap kontroversi olahraga di ruang digital secara lebih akurat.

3.5. Hasil Evaluasi Kinerja Model

Model *ensemble* yang telah dioptimasi kemudian diuji menggunakan set data pengujian yang belum pernah dilihat sebelumnya oleh sistem (sebanyak 20% dari total korpus). Pengujian ini bertujuan untuk mengukur kemampuan generalisasi model terhadap opini publik secara *real-time*. Hasil kalkulasi performa model secara komprehensif disajikan pada Tabel 5.

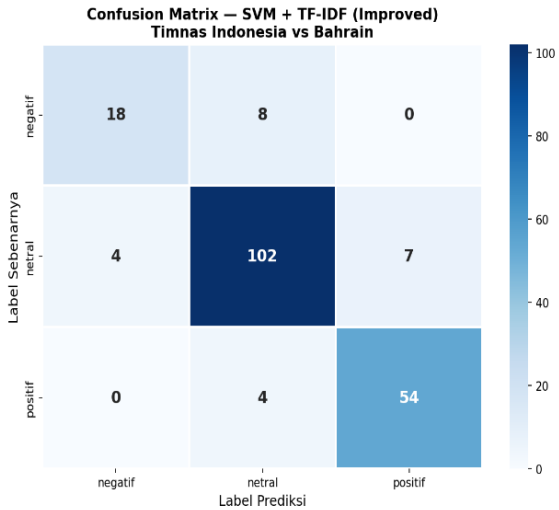
Tabel 5. Confusion Matriks

Kelas Sentimen	Presisi	Recall	F1-Score
Negatif	0.82	0.69	0.75
Netral	0.90	0.92	0.91
Positif	0.92	0.93	0.92
Akurasi	0.89	0.89	0.89

Tabel 5 memperlihatkan hasil evaluasi performa model Ensemble Voting Classifier pada tahap pengujian akhir menggunakan data testing yang tidak terlibat dalam proses pelatihan model. Tahap evaluasi ini bertujuan untuk mengetahui sejauh mana model mampu melakukan klasifikasi sentimen publik terhadap kontroversi pertandingan Indonesia vs Bahrain secara tepat dan konsisten. Berdasarkan hasil pengujian yang diperoleh, model mencatat nilai accuracy sebesar 89,34%, yang menunjukkan bahwa mayoritas data uji berhasil diprediksi sesuai dengan label sentimen yang sebenarnya. Selain itu, model juga menghasilkan nilai precision sebesar 0,8817, recall sebesar 0,8934, serta F1-score sebesar 0,8918. Perolehan nilai tersebut menunjukkan bahwa model memiliki performa klasifikasi yang cukup stabil dan seimbang pada seluruh kategori sentimen.

Jika ditinjau lebih lanjut berdasarkan masing-masing kelas sentimen, kategori positif menunjukkan performa terbaik dengan nilai F1-score sebesar 0,91. Tingginya performa tersebut didukung oleh nilai recall sebesar 0,93 yang menandakan bahwa model mampu mengenali komentar bernada positif dengan sangat baik. Pada kategori netral, model juga memperlihatkan hasil yang optimal dengan nilai F1-score sebesar 0,91, sehingga dapat disimpulkan bahwa model cukup efektif dalam mengidentifikasi komentar yang bersifat informatif maupun tidak mengandung kecenderungan emosi tertentu. Sementara itu, kategori negatif memperoleh nilai F1-score sebesar 0,75 dengan recall sebesar 0,69. Rendahnya recall ini mengindikasikan bahwa model masih mengalami kesulitan dalam menangkap seluruh sampel negatif secara komprehensif, sehingga menghasilkan tingkat *False Negative* yang cukup tinggi (banyak komentar negatif yang gagal terdeteksi dan keliru diklasifikasikan ke kelas lain, terutama netral). Hal tersebut disebabkan oleh karakteristik komentar negatif di media sosial yang umumnya mengandung unsur ambigu, sarkasme, serta penggunaan bahasa slang yang beragam dan kompleks.

Untuk melihat persebaran prediksi dan menganalisis secara detail letak kesalahan klasifikasi (*misclassification*) antar kelas, kinerja model divisualisasikan menggunakan *Confusion Matrix* berskala 3x3, sebagaimana disajikan pada Gambar 7.



Gambar 8. *Confusion Matrix 3x3*

Berdasarkan Gambar 8, angka yang berada pada matriks diagonal merepresentasikan jumlah prediksi yang benar (*True Positives*), yaitu 18 untuk kelas negatif, 102 untuk kelas netral, dan 54 untuk kelas positif.

Analisis terhadap matriks tersebut menunjukkan bahwa *misclassification* paling banyak terjadi pada persinggungan antara kelas negatif dan netral. Terdapat 8 komentar yang sebenarnya bernilai negatif namun keliru diprediksi sebagai netral oleh sistem. Selain itu, terdapat 7 komentar netral yang diprediksi sebagai positif, dan 4 komentar netral yang diprediksi sebagai negatif. Kesalahan klasifikasi pada kelas negatif menjadi netral ini memperkuat analisis sebelumnya bahwa karakteristik komentar negatif di media sosial seperti penggunaan sarkasme, ambiguitas, dan bahasa kiasan seringkali tidak memiliki bobot fitur leksikal negatif yang cukup eksplisit. Akibatnya, model cenderung mengklasifikasikannya ke dalam kelas mayoritas atau kelas dengan muatan emosi yang paling bersinggungan, yaitu netral.

Secara umum, hasil evaluasi tersebut membuktikan bahwa metode Ensemble Voting Classifier memiliki kemampuan yang baik dan cukup robust dalam melakukan klasifikasi sentimen terhadap opini publik di platform YouTube. Tingginya nilai evaluasi yang diperoleh menunjukkan bahwa penerapan tahapan preprocessing, teknik penyeimbangan data menggunakan SMOTE, serta ekstraksi fitur berbasis TF-IDF mampu meningkatkan kemampuan generalisasi model dalam mengenali pola sentimen pada data komentar baru.

3.6. Perbandingan Model

Untuk memvalidasi signifikansi dari teknik penanganan data yang tidak seimbang (*imbalanced data*), penelitian ini melakukan studi ablasi (*ablation study*) dengan membandingkan kinerja model *Ensemble* dalam dua skenario: menggunakan distribusi data asli (tanpa *oversampling*) dan menggunakan teknik augmentasi SMOTE. Hasil komparasi performa dari kedua skenario tersebut disajikan pada Tabel 6.

Tabel 6. Perbandingan Model SMOTE dan Tanpa SMOTE

Model	Akurasi	Presisi	Recall	F1-Score
Tanpa SMOTE	88.83	88.67	88.63	88.67
Model Usulan SMOTE	**89.34**	**89.17**	**89.34**	**89.18**

Pada Tabel 6 penerapan algoritma SMOTE terbukti berhasil menjadi solusi yang efektif untuk mengatasi bias tersebut. Dengan mensintesis data buatan pada kelas minoritas (positif dan negatif) secara adaptif di sekitar batas keputusan (*decision boundary*), SMOTE membantu menyeimbangkan distribusi ruang fitur tanpa merusak konteks linguistik asli. Hasilnya, implementasi SMOTE sukses mendongkrak performa puncak model, di mana tingkat akurasi meningkat secara signifikan menjadi **89,34%** dan *F1-Score* naik menjadi **89,18%**.

Perbandingan yang disajikan pada tabel di bawah memperlihatkan peningkatan performa klasifikasi yang cukup signifikan pada model yang dikembangkan dalam penelitian ini dibandingkan dengan beberapa penelitian terdahulu di bidang yang sama.

Tabel 7. Perbandingan Model

Peneliti	Model/Algoritma	Konteks	Akurasi
Prasetyo dan Dahlan	Naïve Bayes	Analisis sentimen isu naturalisasi pemain Timnas Indonesia	72,55%
Artamevia dan Wibowo	K-Nearest Neighbour (KNN)	Klasifikasi sentimen kontroversi pergantian pelatih Timnas Indonesia	68%

Peneliti	Model/Algoritma	Konteks	Akurasi
Permana	Support Vector Machine (SVM)	Analisis sentimen komentar YouTube di ranah esports	73,97%
Peneliti (Model Usulan)	Ensemble Voting Classifier	Analisis sentimen kontroversi pertandingan Indonesia vs Bahrain di YouTube	89,34%

Pada Tabel 7 melalui pendekatan yang diusulkan dalam penelitian in, integrasi arsitektur *Ensemble Voting Classifier* dengan mekanisme *soft voting* terbukti mampu mengagregasi keunggulan dari masing-masing algoritma dasar. Ditambah dengan implementasi rekayasa fitur multidimensi menggunakan *Feature Union* serta penanganan data tidak seimbang menggunakan SMOTE , model usulan ini berhasil mencatatkan performa tertinggi dengan tingkat akurasi mencapai 89,34%. Hasil ini membuktikan secara empiris bahwa kombinasi metode yang dirancang mampu menjadi solusi yang lebih kokoh dan efektif untuk melakukan pemetaan opini publik pada platform digital.

4. Pembahasan

Hasil penelitian ini menunjukkan bahwa penerapan metode Ensemble Voting Classifier dengan pendekatan soft voting mampu menghasilkan performa klasifikasi sentimen yang kuat dalam menghadapi kompleksitas opini publik. Model berhasil memperoleh tingkat akurasi sebesar 89,34%, yang menunjukkan peningkatan performa dibandingkan beberapa penelitian sebelumnya yang hanya menggunakan model tunggal, seperti Naïve Bayes dengan akurasi 72,55% maupun KNN sebesar 68% pada kajian sentimen di bidang olahraga. Pencapaian akurasi tersebut didorong oleh kontribusi sinergis dari masing-masing komponen di dalam sistem. Ekstraksi fitur menggunakan *Feature Union* terbukti sangat berkontribusi; di mana *Word* TF-IDF mengisolasi kata kunci sentimen, *Char* TF-IDF menangkap variasi kesalahan ketik emosional (seperti "curanggg"), dan fitur numerik merekam intensitas ekspresi audiens. Tingginya nilai *F1-score* pada kategori sentimen positif sebesar 0,92 serta kategori netral sebesar 0,91 mengindikasikan bahwa kombinasi fitur multidimensi ini sangat efektif dalam mengenali pola dukungan maupun komentar informatif yang muncul pada platform YouTube.

Di sisi lain, performa pada kategori sentimen negatif masih menunjukkan keterbatasan, terutama pada nilai recall yang hanya mencapai 0,69. Kondisi tersebut memperlihatkan bahwa komentar bernada kritik, kemarahan, atau sindiran di media sosial cenderung sulit diidentifikasi secara optimal oleh model berbasis statistik. Hal ini disebabkan oleh banyaknya penggunaan sarkasme, bahasa kiasan, serta variasi ekspresi informal yang memiliki makna kontekstual kompleks. Meskipun demikian, penerapan teknik penanganan negasi (*negation handling*), seperti penggabungan frasa "tidak_adil", telah membantu model dalam mempertahankan konteks makna kalimat sehingga proses klasifikasi menjadi lebih baik. Tantangan penelitian selanjutnya terletak pada kemampuan model dalam memahami konteks linguistik yang lebih mendalam, khususnya terkait deteksi sarkasme dan makna implisit dalam komentar pengguna.

Selain rekayasa fitur, penggunaan metode SMOTE (dengan variasi adaptif) dalam proses penyeimbangan data terbukti memberikan kontribusi paling penting dalam meningkatkan performa model. Melalui studi ablasi, teknik ini terbukti mendongkrak akurasi dari 88,83% (model dasar tanpa penyeimbangan) menjadi 89,34%. Teknik ini berhasil mengurangi bias klasifikasi akibat ketidakseimbangan distribusi data secara efektif pada batas keputusan (*decision boundary*), terutama karena jumlah komentar negatif hanya mencakup sekitar 13,3% dari keseluruhan korpus data.

Dengan demikian, kombinasi antara *preprocessing* yang optimal, representasi fitur multidimensi, augmentasi adaptif SMOTE, serta *Ensemble Voting Classifier* mampu

menghasilkan model klasifikasi sentimen yang lebih stabil. Lebih jauh, alur komputasi (*pipeline*) yang dibangun memiliki sifat independen (*domain-agnostic*), yang berarti model ini memiliki **kemampuan generalisasi yang baik** untuk diimplementasikan pada analisis opini publik di peristiwa olahraga lain. Proses transferabilitas model ke cabang olahraga berbeda dapat dilakukan secara robust, dengan syarat dilakukan penyesuaian (substitusi) pada entri kata kunci di dalam kamus leksikon kustom agar selaras dengan terminologi target.

5. Kesimpulan

Berdasarkan hasil analisis dan pembahasan yang telah dilakukan, penelitian ini berhasil menjawab permasalahan terkait keterbatasan metode klasifikasi tunggal dalam memproses *noise* linguistik dan ketidakseimbangan data opini publik di media sosial. Penerapan arsitektur *Ensemble Voting Classifier* yang dipadukan dengan rekayasa fitur multidimensi (*Feature Union*) dan metode penanganan data tidak seimbang (*oversampling*) variasi SMOTE secara adaptif terbukti mampu menghasilkan model klasifikasi sentimen yang jauh lebih tangguh dan terarah.

Secara kuantitatif, penelitian ini berhasil mencatatkan tingkat akurasi puncak sebesar **89,34%**. Hasil ini secara signifikan mengungguli performa model *baseline* tanpa penyeimbangan data (88,83%) maupun metode klasifikasi tunggal pada penelitian-penelitian terdahulu. Ekstraksi fitur melalui *Feature Union* (kombinasi *Word* TF-IDF, *Char* TF-IDF, dan metadata numerik) terbukti efektif merepresentasikan kompleksitas bahasa informal YouTube, sementara penerapan metode penyeimbangan berbasis SMOTE secara adaptif sukses meminimalkan bias klasifikasi pada kelas minoritas tanpa merusak struktur batas keputusan (*decision boundary*). Selain dari sisi komputasional, penelitian ini juga mengonfirmasi bahwa pola komunikasi publik terkait kontroversi olahraga (Indonesia vs Bahrain) lebih didominasi oleh diskusi informatif (netral) sebesar 57,2%, dibandingkan sekadar ekspresi emosional negatif.

Meskipun demikian, penelitian ini masih menemukan tantangan dalam proses identifikasi komentar negatif, terutama yang mengandung unsur sarkasme, bahasa kiasan, maupun ekspresi informal yang kompleks. Oleh karena itu, pengembangan penelitian di masa mendatang perlu diarahkan pada penerapan teknik pemrosesan bahasa alami yang lebih lanjut berbasis *Deep Learning* (seperti *Transformer* atau *Large Language Models*) guna menangkap relasi semantik dan konteks kalimat secara lebih komprehensif. Dengan demikian, sistem analisis dan pemantauan opini publik dapat menghasilkan klasifikasi sentimen yang lebih akurat dan adaptif terhadap karakteristik bahasa di media sosial.

Referensi

[1] B. Ariyadi Jasno, A. Ariful Fathoni, D. Dharma Putra, M. Zidane Hasan, F. Amsury, and H. Supendar, "Analisis Sentimen Komentar Berpotensi Toxic Pada Media Sosial Tiktok Menggunakan Metode Decision Tree," HOAQ (High Education of Organization Archive Quality) : Jurnal Teknologi Informasi, vol. 16, pp. 193-201, 2025. https://doi.org/10.52972/hoaq.vol16no2.p193-201.

[2] V. P. Prasetyo, "Analisis Sentimen Platform Youtube Mengenai Pro Kontra Terhadap Naturalisasi Pemain Sepakbola Timnas Indonesia Menggunakan Metode Naïve Bayes," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 13, no. 3, pp. ,2225-2232, Jul. 2025, https://doi.org/10.23960/jitet.v13i3.7235.

[3] A. M. Artamevia and J. S. Wibowo, "Analisis Sentimen Terhadap Kontroversi Pergantian Pelatih Timnas Indonesia Menggunakan Metode KNN (K-Nearest Neighbour)," vol. 9, no. 5, pp. , 8945-, 8952, 2025, https://doi.org/10.36040/jati.v9i5.15228.

[4] D. T. Permana, Y. B. Pratama, Z. Wahyuzi, E. Altiarika, and A. Pramudyantoro, "Perbandingan Performa Algoritma Naive Bayes Dan Svm Untuk Analisis Sentimen Komentar Youtube Terhadap Industri Esports Di Indonesia," vol. 2, no. 6, pp. , 1391-1399, Nov. 2025, https://ejurnal.kampusakademik.co.id/index.php/jinu/article/view/6753.

[5] A. Pradana and S. Susanto, "Implementasi Model Machine Learning untuk Deteksi Phishing dengan Pendekatan Ekstraksi Fitur yang Dioptimalkan," *Jurnal Teknologi Informasi dan Multimedia*, vol. 8, no. 1, pp. 27–40, Jan. 2026, https://doi.org/10.35746/jtim.v8i1.881.

[6] F. A. Hakim, I. W. D. Prastya, J. R. Budiani, "Sentiment Analysis of the Free Nutritious Meal Program (MBG) on Social Media X (Twitter) Using K-Nearest Neighbor and Artificial Neural Network," Journal of Applied Informatics and Computing, vol. 10, no. 1, pp. 865-876, 2026. https://doi.org/10.30871/jaic.v10i1.12205.

[7] M. I. Raif, N. N. Hidayati, and T. Matulatan, "Otomatisasi Pendeteksi Kata Baku Dan Tidak Baku Pada Data Twitter Berbasis KBBI," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 11, no. 2, pp. 337–348, Apr. 2024, https://doi.org/10.25126/jtiik.20241127404.

[8] A. Jafar and M. Lee, "High Accuracy COVID-19 Prediction Using Optimized Union Ensemble Feature Selection Approach," *IEEE Access*, vol. 12, pp. 122942–122958, 2024, https://doi.org/10.1109/ACCESS.2024.3424231.

[9] F. D. Adhiatma and A. Qoiriah, "Penerapan Metode TF-IDF dan Deep Neural Network untuk Analisa Sentimen pada Data Ulasan Hotel," *Journal of Informatics and Computer Science,* vol. 4, no. 2, pp. 183-193, 2022, https://doi.org/10.26740/jinacs.v4n02.p183-193.

[10] R. Ridwan, E. H. Hermaliani, and M. Ernawati, "Penerapan Metode SMOTE Untuk Mengatasi Imbalanced Data Pada," Computer Science (CO-SCIENCE), vo. 4, no. 1, pp. 80-88, 2024. https://doi.org/10.31294/coscience.v4i1.2990.

[11] G. N. Ahmad, H. Fatima, S. Ullah, A. S. Saidi, and I. Imdadullah, "Efficient Medical Diagnosis of Human Heart Diseases Using Machine Learning Techniques with and Without GridSearchCV," *IEEE Access*, vol. 10, pp. 80151–80173, 2022, https://doi.org/10.1109/ACCESS.2022.3165792.

[12] T. R. Alfiansyah, A. A. Hidayat, A. H. Pratama, A. A. Mahenda, M. Rafly, and F. N. Hasan, "Analisis Sentimen Kebijakan Penempatan Dana 200 Triliun Bank BUMN Menggunakan Algoritma Support Vector Machine," *Journal of Computing and Informatics Research*, vol. 5, no. 1, pp. 410-420, 2025, https://journal.fkpt.org/index.php/comforch/article/view/2329.

[13] Z. Z. H. Al Abrori and E. R. Subhiyakto, "Analisis Komparatif Akurasi Prediksi Kanker Payudara Menggunakan Algoritma Random Forest dan Logistic Regression," *Jurnal Algoritma*, vol. 22, no. 1, pp. 300–311, May 2025, https://doi.org/10.33364/algoritma/v.22-1.2164.

[14] R. Saputra and E. Hartati, "Deteksi Website Phising Menggunakan Algoritma Random Forest Dengan Optimalisasi Gridsearch," UTIM (Jurnal Teknik Informatika Musirawas), vo. 10, no. 1, pp. 55-67, Jul. 2025. Accessed: Dec. 10, 2025. https://jurnal.univbinainsan.ac.id/index.php/jutim/article/view/2674

[15] F. A. Indriyani, A. Fauzi, and S. Faisal, "Analisis sentimen aplikasi tiktok menggunakan algoritma naïve bayes dan support vector machine," *TEKNOSAINS : Jurnal Sains, Teknologi dan Informatika*, vol. 10, no. 2, pp. 176–184, Jul. 2023, https://doi.org/10.37373/tekno.v10i2.419.