P-ISSN : 2549-7480
E-ISSN : 2549-6301

JURNAL ILMIAH INFORMATIKA
E-ISSN : 2549-6301
ISSN : 2549-7480
Faculty of Sains and Technology, Ibrahimy University
S4-Accredited : 105/E/KPT/2022
Journal page is available to : https://journal.ibrahimy.ac.id/index.php/JIMI

ANALISIS SENTIMEN KOMENTAR YOUTUBE PERTANDINGAN PORTUGAL MELAWAN KROASIA MENGGUNAKAN TRANSFER LEARNING INDOBERT

Syahid Ahmad Yasin[1)], Josua Eriek Kurniawan[2)], Raditya Fernanda Putra[3)], Yumi Novita Dewi[4)], Waeisul Bismi[5)], Suharyanto[6)]

[1,2,3,4,5,6] *Program Studi Informatika, Fakultas Teknik dan Informatika, Universitas Bina Sarana Informatika*
Jl. Kramat Raya No.98, Senen, Jakarta Pusat, DKI Jakarta 10450

email: [1] 15241007@bsi.ac.id, [2] 15241108@bsi.ac.id, [3] 15241034@bsi.ac.id, [4] yumi.ymd@bsi.ac.id, [5] waeisul.bismi@bsi.ac.id, [6] suharyanto.shy@bsi.ac.id

ARTICLE INFO

Article History:
Received : 1 Oktober 2026
Accepted : 28 Oktober 2026
Published : 16 November 2026

Keywords:
sentiment analysis
transfer learning
IndoBERT
YouTube comments
Video Assistant Referee

IEEE style in citing this article:
S. A. Yasin, J. E. Kurniawan, R. F. Putra, Y. N. Dewi, W. Bismi, and Suharyanto, "Analisis Sentimen Komentar YouTube Pertandingan Portugal Melawan Kroasia Menggunakan Transfer Learning IndoBERT", *Jurnal Ilmiah Informatika*, vol. 11, no. 2, pp. 230-241, Des. 2026.

Corresponding Author:
Syahid Ahmad Yasin
Universitas Bina Sarana Informatika
15241007@bsi.ac.id

ABSTRACT

Many social media users become furious when refereeing decisions occur in football matches. The highlight of the Portugal vs Croatia match in July-October 2026 was uploaded to TVRI Nasional YouTube channel and received thousands of comments from users criticizing the Video Assistant Referee (VAR) decisions. Unfortunately, in the field of natural language processing (NLP), there is still little research on how to link social media users' anger to specific events that occurred on the pitch. In this research, we used 5,460 raw comments from the social media and filtered them using the Indonesian word ratio (Rid ≥ 0.40) and the football term dictionary, resulting in 3,273 clean texts (stratified 80:20 split, 262 lines for validation). We used the IndoBERT architecture, which uses the [CLS] token without dropout. On 655 test data (seed 42), the model achieved 75.42% accuracy and 72.29% Macro-F1 (ROC-AUC 87.88%; PR-AUC 76.21%) with a reading speed of 5.860 ms per text on GPU. The average of three seeds was 75.47 ± 0.23% accuracy and 72.37 ± 0.13% Macro-F1. This model outperformed four classic TF-IDF-based methods, i.e., SVM (67.63%), Logistic Regression (66.26%), Random Forest (65.34%), and Naive Bayes (61.83%) (Wilcoxon test p = 0.004; paired t-test p = 0.001). Negative-toned sentences were highly recognizable to the model with an F1-score of 82.05%.On the other hand, the model's performance on guessing the neutral class dropped to 62.19% due to the fact that social media users' question sentences often use words identical to those used in complaints. Using logistic regression, we found that VAR offside line debates increased the chance of negative comments by 3.44 times (73.91% negative) and penalty checks increased the chance by 2.25 times (64.93% negative).

Finally, we would like to note that the training corpus used in this research still employs automatic labeling with agreement with humans of only 34.67% (Kappa 0.02) and 36.75% on consensus sample (Kappa 0.05).

© 2026 Jurnal Ilmiah Informatika (Scientific Informatics Journal) with CC BY NC licence

P-ISSN : 2549-7480
S. A. Yasin dkk/ JIMI 11 (2) pp. 230-241 E-ISSN : 2549-6301

1. PENDAHULUAN

Pada umumnya, keputusan wasit di lapangan hijau memicu perdebatan panjang [1]–[3]. Ketika tayangan ulang VAR membatalkan gol atau menunjuk titik putih, penonton langsung memberikan reaksi keras di media sosial. Hal yang sama terjadi pada video cuplikan laga Portugal melawan Kroasia di kanal YouTube TVRI Nasional yang seketika dipenuhi oleh ribuan komentar dari suporter tanah air. Namun demikian, membaca teks obrolan warganet dengan program komputer tidak semudah membalik telapak tangan. Kalimat mereka penuh dengan bahasa gaul, huruf vokal sengaja diketik berulang-ulang untuk meluapkan kekesalan, singkatan muncul tanpa patokan baku, belum lagi salah ketik pada kata serapan seperti offside maupun penalti [4], [5]. Struktur kalimatnya benar-benar cair.

Pendekatan hitung kata konvensional seperti TF-IDF atau n-gram seringkali gagal menangkap maksud kalimat obrolan semacam ini. Ketika susunan kata berubah sedikit saja, bobot frekuensinya langsung meleset. Di sinilah arsitektur Transformer seperti IndoBERT lebih unggul karena mampu menimbang hubungan antarkata sesuai konteks kalimat utuh di media sosial Indonesia [6], [7]. Sebelumnya, teknik transfer learning memang pernah diuji untuk memilah emosi pengguna Twitter [8], sedangkan model regresi digunakan memetakan naik-turunnya suasana hati penonton dari menit ke menit [9]. Namun untuk urusan wasit, kajian mengenai VAR selama ini lebih sering menghitung berapa lama pertandingan terhenti di liga-liga Eropa [10], [11]. Tantangan di kolom komentar jauh lebih rumit. Penonton seringkali menyisipkan sindiran terbalik [12], [13] di tengah jumlah kelas data yang berat sebelah [14]. Itulah sebabnya, pembedahan komentar video [15] harus dipilah menurut sosok atau objek yang sedang dibicarakan [16] agar arah kritik suporter terlihat jelas dan tidak tertimbun angka rata-rata.

Ada dua hal pokok yang ingin kami jawab melalui eksperimen ini. Pertama, seberapa jauh model IndoBERT mampu melampaui empat algoritma mesin belajar klasik (SVM, Naive Bayes, Random Forest, dan Logistic Regression) ketika dipaksa membaca komentar sepak bola YouTube berbahasa Indonesia? Kedua, apakah kemunculan topik VAR tertentu benar-benar melipatgandakan peluang munculnya komentar negatif dari penonton? Untuk menjawabnya, kami membedah ketepatan tebakan IndoBERT menggunakan matriks konfusi One-vs-All sekaligus mencatat kecepatan kerjanya di kartu grafis (GPU), lalu menghitung nilai Odds Ratio pada empat kejadian lapangan: garis offside, klaim penalti, pembatalan gol, dan aturan wasit umum.

2. METODE PENELITIAN

Urutan pengerjaan yang kami lakukan tersusun secara bertahap sehingga teks sampah dapat disaring sejak dini dan data uji tidak bocor ke dalam proses pelatihan (data leakage). Seluruh tahapan pengerjaan dibagi menjadi 6 tahap berturut-turut yaitu: penarikan komentar mentah melalui YouTube Data API v3, penyaringan kata bahasa Indonesia, pembagian data 80:20 secara proporsional, pelatihan ulang (fine-tuning) IndoBERT, adu performa melawan model klasik, sampai dengan perhitungan regresi logistik untuk insiden VAR (lihat Gambar 1).

DOI : https://doi.org/10.35316/jimi.v11i2.230-241
© 2026 Jurnal Ilmiah Informatika with CC BY NC licence

P-ISSN : 2549-7480
S. A. Yasin dkk/ JIMI 11 (2) pp. 230-241 E-ISSN : 2549-6301



Gambar 1. Alur penelitian dan arsitektur *Transfer Learning* IndoBERT

Gambar 1 menunjukkan alur kerja dari tahap pengambilan komentar mentah sampai pengujian statistik atas reaksi penonton.

2.1. Akuisisi Data dan Penyortiran Leksikal Dua Sisi

Komentar-komentar tersebut dikumpulkan dari video resmi cuplikan pertandingan Portugal versus Kroasia yang diunggah ke kanal YouTube TVRI Nasional (https://youtu.be/eKLrQ2ztNhA). Data ini diperoleh dengan menggunakan YouTube Data API v3 mulai tanggal 3 Juli hingga 24 Oktober 2026, yang merupakan waktu sejak video tersebut pertama kali diterbitkan (publishedAt 2026-07-03T01:31:11Z), sehingga diperoleh 5.460 baris komentar yang belum diolah. Nama akun dan ID pengirim langsung dihapus untuk menjamin privasi penonton. Data komentar mentah yang diperoleh sebanyak ribuan baris tersebut kemudian disaring pada tahap pertama dengan menghitung persentase kata bahasa Indonesia menggunakan Persamaan (1):

$$R_{id} = \frac{N_{id}}{N_{id} + N_{asing}} \geq 0{,}40 \qquad (1)$$

Rasio Rid yang diberikan pada Persamaan (1) adalah hubungan antara jumlah kata penanda bahasa Indonesia (Nid) dan kata bahasa asing (Nasing). Komentar yang memiliki nilai Rid kurang dari 0,40 akan dibuang. Tahap kedua adalah untuk melihat apakah komentar tersebut memiliki kata kunci yang berhubungan dengan sepak bola, nama tim, atau nama pemain yang bertanding. Dengan melakukan dua lapis saringan tersebut, maka diperoleh 3.273 komentar yang sudah bersih dan siap untuk dianalisis.

2.2. Arsitektur *Transfer Learning* IndoBERT

Model pengklasifikasi kami dikembangkan dari bobot pra-latih indobenchmark/indobert-base-p1 (12 layer encoder Transformer, dimensi H = 768, 12 attention heads) dengan memodifikasi framework Shaw dkk. [8]. Kami menggunakan model dasar ini karena model ini telah terbukti efektif dalam memproses teks obrolan non-formal dalam bahasa Indonesia [6]. Setiap komentar dibatasi sampai T = 96 subkata WordPiece. Model ini di-fine-tuned dengan menggunakan optimizer AdamW dengan learning rate 3 × 10-5, weight decay 0,005, batch size 32, dan warmup 30 langkah selama 3 epoch. Untuk model baseline, empat model klasik (SVM, Logistic Regression, Random Forest, Naive Bayes) dilatih dengan menggunakan matriks TF-IDF n-gram (1,2) dengan 5.000 fitur dengan opsi sublinear_tf = True. Dalam eksperimen awal berdasarkan literatur, vektor kalimat diringkas dengan menggunakan dua layer dense network (64 dan 16 titik neuron ReLU) dengan menggunakan teknik

Global Average Pooling (GAP) dengan bobot attention mask [8] dengan rumus Persamaan (2):

$$h_{pool} = \frac{1}{\sum_t m_t} \sum_t (m_t \cdot h_t) \quad (2)$$

Berdasarkan rumus (2), nilai rata-rata dari output encoder ht pada kata yang muncul (bit mask mt = 1) diambil dan tidak memperhitungkan ruang kosong (mt = 0) sehingga dihasilkan vektor kalimat hpool dengan dimensi 768. Namun demikian, setelah dilakukan evaluasi terhadap data validasi, metode sebelumnya yang hanya mengambil vektor token [CLS] tanpa lapisan dropout tambahan menghasilkan nilai Macro-F1 yang paling baik. Dengan demikian, model akhir yang kami usulkan langsung meneruskan vektor [CLS] ke lapisan klasifikasi tiga kelas sentimen dengan aktivasi Softmax.

2.3. Skema Evaluasi Matriks Konfusi dan Pemodelan Regresi Insiden VAR

Untuk mencegah kontaminasi pada data uji selama proses pencarian parameter, data uji dijaga agar tidak tersentuh. Dari 2.618 baris data latih, 10% (262 baris) dipisahkan sebagai ruang validasi internal, sedangkan 655 baris data uji disimpan rapat sampai tahap evaluasi akhir. Sisa 2.356 baris digunakan penuh untuk melatih bobot jaringan. Agar hasilnya tidak kebetulan semata, konfigurasi model terbaik dijalankan kembali pada tiga angka acak (seed 42, 123, dan 2024) kemudian dihitung rata-rata beserta simpangan bakunya. Kami juga menguji perbedaan hasil antara teknik pooling [CLS] melawan GAP dan melihat efek penghapusan dropout. Di akhir, tabel silang 3 × 3 dipecah per kelas menjadi komponen True Positive (TPk), True Negative (TNk), False Positive (FPk), dan False Negative (FNk) untuk menghitung skor F1 melalui Persamaan (3):

$$F1_k = \frac{2 \cdot TP_k}{2 \cdot TP_k + FP_k + FN_k} \quad (3)$$

Persamaan (3) dengan nilai F1k memberikan kompromi yang optimal antara TPk (tebakan yang tepat), FPk (tebakan yang salah kelas), dan FNk (data yang tidak terjaring) [14]. Selain melakukan uji akurasi model, kami juga melakukan regresi OLS Fixed Effects serta regresi logistik untuk menentukan seberapa besar pengaruh pembicaraan VAR terhadap negativitas komentar suporter.

2.4. Pelabelan Sentimen, Pengodean Aspek, dan Pengodean Insiden VAR

Pelabelan awal pada seluruh 3.273 komentar dilakukan secara otomatis menggunakan kamus kata beraturan (rule-based lexicon). Berapa akurat label mesin ini? Kami mengambil 150 komentar secara acak (masing-masing 50 baris per kelas) untuk dinilai ulang oleh dua orang secara terpisah tanpa melihat hasil mesin. Dua penilai manusia tersebut sepakat pada 78,00% sampel dengan nilai Cohen's Kappa 0,674 (tergolong kuat/substantial). Namun ketika label otomatis mesin dibandingkan ke penilai manusia, kecocokannya jatuh ke angka 34,67% (Kappa 0,02) pada penilai tunggal, dan 36,75% (Kappa 0,05) pada 117 komentar yang disepakati kedua penilai. Fakta ini harus digarisbawahi sejak awal: dataset kami berstatus silver label, artinya IndoBERT dilatih untuk meniru logika pelabelan kamus otomatis tersebut. Adapun pengelompokan topik insiden VAR disusun menggunakan pencocokan daftar kata kunci setiap kejadian.

3. HASIL DAN PEMBAHASAN

3.1. Distribusi Sentimen Entitas dan Pemetaan Insiden VAR

Sebagian besar komentar di kolom tanggapan bernada negatif. Dari 3.273 komentar yang sudah dibersihkan, 52,82%

DOI : https://doi.org/10.35316/jimi.v11i2.230-241
© 2026 Jurnal Ilmiah Informatika with CC BY NC licence

P-ISSN : 2549-7480
S. A. Yasin dkk/ JIMI 11 (2) pp. 230-241 E-ISSN : 2549-6301

atau 1.729 komentar berisi keluhan dan kemarahan. Sementara itu, hanya 24,08% (788 komentar) yang bersentimen positif dan 23,10% (756 komentar) yang bersifat netral. Ketimpangan ini memang sering ditemukan di media sosial; biasanya yang merasa dirugikan akan lebih terdorong untuk berkomentar ketimbang yang puas. Rincian distribusi sentimen pada 655 data uji menurut subjek pembicaraan bisa dilihat pada Tabel 1.

Tabel 1. Distribusi sentimen pada aspek entitas utama (655 data uji)

Target Aspek Entitas	Negatif (%)	Netral (%)	Positif (%)	Total
Aspek Ronaldo & Portugal	86 (48,04 %)	54 (30,17 %)	39 (21,79 %)	179
Aspek Wasit & VAR	175 (68,09 %)	50 (19,46 %)	32 (12,45 %)	257
Aspek Modric & Kroasia	52 (36,88 %)	27 (19,15 %)	62 (43,97 %)	141
Lainnya / Konteks Umum Laga	33 (42,31 %)	20 (25,64 %)	25 (32,05 %)	78
Total Sampel Data Uji	346 (52,82 %)	151 (23,05 %)	158 (24,12 %)	655

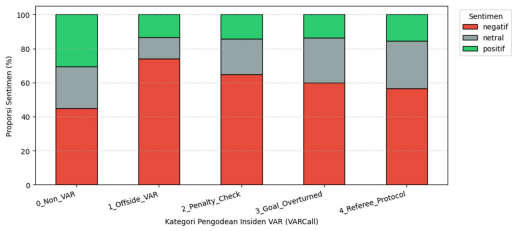
Dari data di Tabel 1, pihak yang paling banyak jadi sasaran kemarahan penonton adalah perangkat pertandingan. Pada 257 komentar uji yang menyebut wasit atau VAR, sebanyak 68,09% (175 komentar) bernada negatif dan hanya 12,45% yang memuji. Sementara itu, pembahasan tentang Luka Modric atau Kroasia justru banyak menuai apresiasi—43,97% komentar tentang mereka bernada positif; banyak warganet menilai kiprah sang kapten layak dihargai. Regresi OLS Fixed Effects yang dijalankan juga memperlihatkan pola serupa: munculnya topik VAR menurunkan skor sentimen sebesar -0,3296 poin (p = 1,57 × 10^-30), sedangkan gaya bahasa sarkasmenurunkan skor lebih dalam lagi, yakni -0,3468 poin (p = 2,61 × 10^-5). Peluang kemunculan komentar negatif pada setiap kategori insiden VAR dirangkum dalam Tabel 2.

Tabel 2. Pengodean kategori insiden VAR dan hasil estimasi regresi logistik reaksi negatif (seluruh 3.273 komentar)

Kategori Insiden VAR (VARCall)	Jumlah	Negatif (%)	Koefisien *Logit*	*Odds Ratio* (p-value)
0_Non_VAR (Komentar Umum Non-VAR)	1.960	44,95 %	-0,2229	1,000x (Baseline)
1_*Offside*_VAR (Garis *Offside* & VAR)	483	73,91 %	+1,2365	3,444x (p < 0,0001)
2_Penalty_Check (Klaim & Penalti)	211	64,93 %	+0,8090	2,246x (p < 0,0001)
3_Goal_Overturned (Gol Dianulir)	132	59,85 %	+0,5627	1,755x (p = 0,0023)
4_Referee_Protocol (Kepemimpinan Wasit)	487	56,47 %	+0,4356	1,546x (p < 0,0001)
Sarcasm_Dummy (Kontrol Sarkasme)	77	81,82 %	+1,4291	4,175x (p < 0,0001)

Dari Tabel 2 terlihat bahwa keterlibatan teknologi VAR membuat emosi penonton semakin tinggi. Komentar soal offside (1_Offside_VAR) punya peluang 3,444 kali lebih besar berakhir negatif dibandingkan obrolan pertandingan biasa (p < 0,0001). Posisi kedua ditempati oleh klaim penalti (2_Penalty_Check) dengan peluang naik 2,246 kali lipat, lalu pembatalan gol (3_Goal_Overturned; OR 1,755), dan keluhan kepemimpinan wasit (4_Referee_Protocol; OR 1,546). Yang paling menonjol adalah variabel kontrol sarkasme (Sarcasm_Dummy): jika komentar mengandung sarkasme, peluang sentimen negatif melonjak 4,175 kali lipat (p = 0,000004). Perbandingan proporsi ketiga sentimen untuk tiap kejadian bisa dilihat pada Gambar 2.

DOI : https://doi.org/10.35316/jimi.v11i2.230-241
© 2026 Jurnal Ilmiah Informatika with CC BY NC licence



Gambar 2. Proporsi sentimen suporter berdasarkan kategori pengodean insiden VAR

Dari diagram balok di Gambar 2, terlihat dua kejadian yang paling memicu kemarahan—warna merahnya paling mencolok—yaitu sengketa garis offside (73,91% negatif) dan pemeriksaan penalti (64,93% negatif). Artinya, saat wasit jeda menunggu keputusan ruang VAR, baik soal gol dianulir maupun penalti, momen ini jadi titik rawan yang kerap membuat penonton geram.

Namun demikian, perlu diingat bahwa Odds Ratio di atas tidak otomatis berarti hubungan sebab-akibat di lapangan. Faktor overlapping word juga ada, misalnya kata 'curang' dipakai buat mendeteksi topik tertentu, 94,1% (32 dari total 34 kemunculan) langsung diberi label negatif oleh aturan kamus yang dipakai.

3.2. Evaluasi Kinerja Klasifikasi IndoBERT dan Dekomposisi Matriks Konfusi

Untuk menguji performa IndoBERT dibanding model klasik, kami melakukan eksperimen dengan empat model pembanding berbasis TF-IDF (SVM, Logistic Regression, Random Forest, dan Naive Bayes) pada dataset uji yang sama sebanyak 655 baris, dan hasilnya ditampilkan pada Tabel 3.

Tabel 3. Perbandingan kinerja klasifikasi IndoBERT (*seed* 42) terhadap empat model pembanding pada data uji

Model Klasifikasi	Fitur / Arsitektur	Akurasi (%)	*Macro-F1* (%)
Multinomial Naive Bayes	TF-IDF N-Gram (1,2)	61,83	47,91
Random Forest Classifier	TF-IDF (100 Trees)	65,34	58,05
Support Vector Machine (SVM)	TF-IDF Linear Kernel	67,63	63,15
Logistic Regression	TF-IDF L2 Penalty	66,26	58,33
IndoBERT + [CLS] (Diusulkan)	*Transformer Transfer Learning*	75,42	72,29

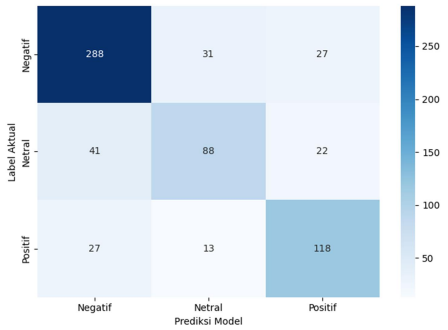
Dari Tabel 3, IndoBERT tampil jauh lebih baik dibanding keempat model lain. Dengan seed 42, model Transformer ini mencatat akurasi 75,42% dan Macro-F1 72,29%. Angka ini jauh di atas SVM yang jadi pesaing terdekat, yang hanya mencapai akurasi 67,63% dan Macro-F1 63,15%. Tiga model lain semakin tertinggal: Logistic Regression dengan akurasi 66,26% (Macro-F1 58,33%), Random Forest 65,34% (Macro-F1 58,05%), dan Multinomial Naive Bayes 61,83% (Macro-F1 47,91%). Uji statistik pada sepuluh sampel data uji menunjukkan selisih ini bukan sekadar kebetulan (Wilcoxon $p = 0{,}004$; uji t berpasangan $p = 0{,}001$). Model klasik memang kesulitan menangkap makna kata yang sering diplesetkan suporter karena TF-IDF hanya mengandalkan frekuensi kata.

P-ISSN : 2549-7480
S. A. Yasin dkk/ JIMI 11 (2) pp. 230-241 E-ISSN : 2549-6301

Tabel 4. Kinerja klasifikasi sentimen model Transfer Learning IndoBERT pada data uji

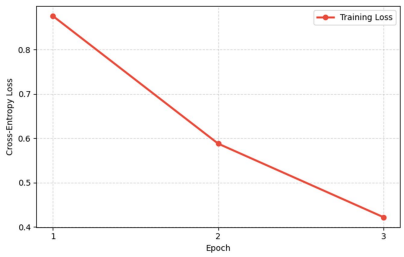
Kelas Sentimen	Presisi (%)	*Recall* (%)	F1 (%)	n / AUC
Kelas Negatif (0)	80,90	83,24	82,05	346 komentar
Kelas Netral (1)	66,67	58,28	62,19	151 komentar
Kelas Positif (2)	70,66	74,68	72,62	158 komentar
Rata-rata Makro (Macro Avg)	72,74	72,07	72,29	ROC-AUC: 87,88%
Rata-rata Terbobot (Weighted Avg)	75,15	75,42	75,20	PR-AUC: 76,21%

Jika dilihat per kelas sentimen (Tabel 4), IndoBERT paling jitu mendeteksi komentar negatif dengan F1-Score 82,05% (presisi 80,90% dan recall 83,24%). Di urutan kedua kelas positif dengan F1-Score 72,62% (presisi 70,66% dan recall 74,68%). Namun, performa menurun pada kelas netral, hanya F1-Score 62,19% (presisi 66,67% dan recall 58,28%). Hal ini terjadi karena banyak penonton yang hanya bertanya soal aturan atau jam tayang tapi bahasanya mirip protes, sehingga membingungkan pemisahan kelas. Pola kekeliruan ini terlihat jelas pada Gambar 3.



Gambar 3. Confusion matrix heatmap hasil klasifikasi sentimen model IndoBERT

Pada Gambar 3, kotak-kotak matriks konfusi memperlihatkan IndoBERT jarang sekali salah menebak komentar positif sebagai negatif atau sebaliknya. Namun, di kelas netral sering terjadi salah prediksi: dari 151 komentar netral, 63 dimasukkan ke kategori positif atau negatif. Kesalahan di kelas netral inilah yang paling banyak menyumbang dari total 161 tebakan meleset pada data uji. Sementara itu, grafik penurunan nilai loss selama pelatihan bisa dilihat di Gambar 4.



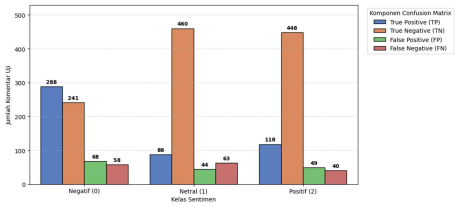
Gambar 4. Kurva penurunan training *loss* model IndoBERT terhadap *epoch*

Pada tiga epoch pelatihan di Gambar 4, nilai loss turun mulus dari 0,8759 (epoch pertama) ke 0,5880 (epoch kedua), lalu ke 0,4222 pada epoch ketiga. Walau kurva loss menurun tanpa tanda macet konvergensi, F1 di data validasi sempat fluktuatif ketika seed diubah—artinya bobot awal masih berpengaruh pada hasil akhir. Untuk melihat detail kesalahan setiap kelas, matriks 3 × 3 tadi dipecah dengan skema One-vs-All seperti di Tabel 5.

Tabel 5. Dekomposisi matriks konfusi *One-vs-All* (TP, TN, FP, FN) model IndoBERT per kelas sentimen

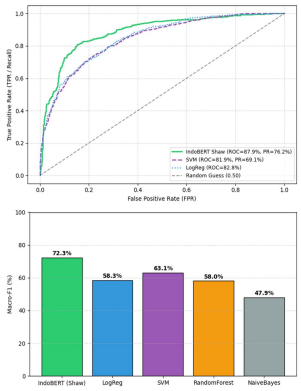
Kelas Sentimen	TP	TN	FP	FN	Total Sampel
Negatif (0)	288	241	68	58	655
Netral (1)	88	460	44	63	655
Positif (2)	118	448	49	40	655

Berdasarkan hasil One-vs-All di Tabel 5, kelas negatif menghasilkan 288 prediksi benar (TP) dan 241 benar-negatif (TN). Untuk kelas netral dan positif, TN-nya bahkan mencapai 460 dan 448 baris, dengan jumlah salah deteksi (FP) tetap di kisaran 44 dan 49 baris. Analisis tiap bagian matriks ini bisa dilihat lebih lanjut lewat diagram batang bertumpuk di Gambar 5.



Gambar 5. Distribusi komponen TP, TN, FP, dan FN per kelas sentimen pada model IndoBERT

Pada Gambar 5, terlihat bahwa prediksi benar (TP dan TN) selalu lebih banyak daripada prediksi salah untuk setiap kategori sentimen. Dari sisi hardware, file bobot model ini berukuran 475,0 MB dan butuh waktu 5,860 ms untuk memproses satu komentar dengan GPU CUDA (batch size 32, setara dengan 170,6 komentar per detik) serta memakan RAM 2.129,1 MB. Artinya, file bobot model ini cukup ringan dan bisa dipakai untuk memonitor chat live secara real time.



Gambar 6. Kurva ROC beserta skor PR-AUC (atas) dan perbandingan Macro-F1 (bawah) IndoBERT terhadap model *baseline*

Keunggulan IndoBERT dalam membedakan data juga tampak pada kurva ROC dan PR-AUC di Gambar 6 (panel atas), dengan nilai AUC 87,88% untuk ROC-AUC dan 76,21% untuk PR-AUC. Nilai ini lebih tinggi dari Logistic Regression (82,85% dan 69,89%) serta SVM (81,92% dan 69,13%). Sementara diagram batang di panel bawah menunjukkan selisih Macro-F1 yang cukup lebar, yakni 9,14 poin persentase antara IndoBERT dan SVM, membuktikan pemrosesan konteks dua arah pada IndoBERT jauh lebih efektif daripada sekadar menghitung frekuensi kata.

Apakah hasil yang diperoleh bisa konsisten jika memakai seed acak yang berbeda? Apakah benar menambah komponen yang makin kompleks selalu memperbaiki kinerja? Hasil uji dengan tiga seed dan eksperimen pengurangan komponen (ablation study) dapat dilihat pada Tabel 6.

P-ISSN : 2549-7480
S. A. Yasin dkk/ JIMI 11 (2) pp. 230-241 E-ISSN : 2549-6301

Tabel 6. Variasi antar-seed pada konfigurasi usulan dan hasil ablation study dari konfigurasi acuan

Konfigurasi	*Seed*	F1 Validasi (%)	Akurasi (%)	*Macro-F1* (%)
Usulan / Tanpa GAP (cls, dropout 0,00)	42	76,83	75,42	72,29
Usulan (cls, tanpa dropout)	123	73,16	75,73	72,51
Usulan (cls, tanpa dropout)	2024	74,44	75,27	72,30
Rata-rata ± SD (3 seed)	—	74,81 ± 1,89	75,47 ± 0,23	72,37 ± 0,13
Acuan (GAP, dropout 0,05)	42	74,70	76,64	73,93
Ablasi: Tanpa dropout (GAP, dropout 0,00)	42	76,41	76,79	73,79

Dalam pengujian model usulan dengan tiga seed yang berbeda (42, 123, dan 2024), akurasi uji model usulan hanya bervariasi antara 75,27% sampai 75,73% dengan rata-rata 75,47 ± 0,23%, Macro-F1 uji model usulan 72,37 ± 0,13%, dan rata-rata Macro-F1 validasi model usulan 74,81 ± 1,89% (per seed: 76,83%, 73,16%, 74,44%). Hasil ablation study pada seed 42 menunjukkan arsitektur sederhana yang mengambil token [CLS] secara langsung tanpa dropout memiliki skor validasi tertinggi (76,83%) daripada arsitektur acuan yang menggunakan Global Average Pooling (GAP) dengan dropout 0,05 (validasi 74,70%). Pada data uji, varian GAP acuan memiliki akurasi sedikit lebih tinggi (76,64%) dengan Macro-F1 73,93%, sedangkan varian GAP tanpa dropout memiliki akurasi sedikit lebih tinggi (76,79%) dengan Macro-F1 73,79%, namun perbedaannya tipis dan memperlihatkan bahwa penggunaan seluruh token (GAP) tidak begitu berdampak pada komentar pendek di YouTube.

3.3. Analisis Kesalahan dan Keterbatasan

Meskipun akurasi model usulan pada seed 42 sudah mencapai 75,42%, dalam pemeriksaan manual terhadap 161 komentar yang salah klasifikasi, ditemukan kelemahan klasik di mana IndoBERT sering kesulitan menangkap sindiran halus tanpa kata kasar, ataupun membaca pujian palsu yang sebenarnya bernada ejekan (Lihat Tabel 7).

Tabel 7. Contoh kesalahan prediksi berkeyakinan tinggi pada komentar implisit

Contoh Komentar Suporter	Prediksi	Aktual	Analisis Penyebab Kesalahan
Kroasia vs Wasit	Netral	Negatif	Hanya dua kata tanpa penanda polaritas; tuduhan keberpihakan wasit harus disimpulkan dari konteks laga.
MESSI GOAT *PENALDO* DI BANTU WASIT	Netral	Negatif	Huruf kapital, singkatan, dan julukan pemain mendominasi sehingga penanda negatif tidak terbaca model.
RESPECT LUCA MODRIC	Netral	Negatif	Frasa apresiatif, padahal label aktualnya negatif karena konteks mendukung Kroasia yang tersingkir.
Tambahan waktu nya sampai <MENIT>	Netral	Negatif	Token <MENIT> menjaga penanda waktu, tetapi keluhan atas molornya laga tersirat tanpa kata negatif eksplisit.

Dari Tabel 7, empat contoh data dengan confidence tertinggi di data uji, di mana kalimat yang seharusnya berlabel Negatif malah diprediksi Netral oleh IndoBERT. Pola kesalahan pada kalimat-kalimat ini serupa, yaitu kalimat pendek,

DOI : https://doi.org/10.35316/jimi.v11i2.230-241 © 2026 Jurnal Ilmiah Informatika with CC BY NC licence

penggunaan julukan ejekan ('PENALDO'), atau menyindir perpanjangan waktu wasit ('<MENIT>') tanpa kata kasar yang umum di kamus. Selain itu, data latih yang mengandung sarkasme sangat sedikit (hanya 2,52% atau 66 baris dari total 77 baris di seluruh korpus). Oleh karena itu, nilai kemiringan regresi OLS -0.3468 harus diinterpretasi dengan hati-hati sebagai indikasi awal saja.

Ada dua catatan teknis terkait pengujian yang bisa kami buka secara transparan. Pertama, ruang validasi hanya 262 baris (10% dari data latih). Oleh karena itu, nilainya sangat mudah berubah jika komposisi kalimat berubah sedikit. Hal ini mungkin menjelaskan mengapa model validasi terbaik (76,83%) justru menghasilkan Macro-F1 uji yang sedikit lebih rendah dibanding model acuan GAP (72,29% vs 73,93%). Kedua, p-value pada uji Wilcoxon (p = 0,004) dan paired t-test (p = 0,001) dihitung dari sepuluh subsampel di 655 data uji yang sama, sehingga kondisi independensi antar-pengamatan belum ideal. Sebagai pembanding yang lebih ketat, cross-validation 5-lipat pada data latih menghasilkan rata-rata Macro-F1 61,51% ± 2,27% untuk SVM dan 59,59% ± 2,23% untuk Logistic Regression.

Masalah terbesar penelitian ini adalah sumber label datanya. Dua penilai manusia yang menilai data mencapai tingkat kesepakatan tinggi, yaitu 78,00% (Cohen's Kappa 0,674). Namun, aturan kamus otomatis yang dipakai melabeli seluruh korpus hanya cocok 34,67% (Kappa 0,02) dengan penilai manusia tunggal dan 36,75% (Kappa 0,05) dengan 117 baris konsensus manusia. Artinya, semua 3.273 baris data adalah silver label—IndoBERT dan keempat model pembanding sebenarnya memprediksi pola kamus otomatis, bukan penilaian manusia secara penuh. Selain masalah label, cakupan data terbatas pada satu video pertandingan, pemenggalan kalimat pada batas 96 token, dan penghapusan angka kecuali yang merujuk menit. Untungnya, penggunaan SequenceMatcher (batas 0,80) memastikan data uji bebas duplikat: hanya ada 4 pasang kalimat mirip (8 teks atau 1,2%) dan tidak ada komentar yang benar-benar duplikat.

4. KESIMPULAN

Dari serangkaian pengujian, transfer learning IndoBERT terbukti mengungguli mesin pembelajar klasik dalam membaca komentar sepak bola YouTube berbahasa Indonesia. Dari 5.460 komentar mentah yang disaring menjadi 3.273 teks bersih, IndoBERT dengan token [CLS] tanpa dropout memperoleh rata-rata tiga seed 75,47 ± 0,23% untuk akurasi dan 72,37 ± 0,13% untuk Macro-F1, mengalahkan SVM (67,63%), Logistic Regression (66,26%), Random Forest (65,34%), dan Naive Bayes (61,83%). Pada seed 42, model ini mendapat skor ROC-AUC 87,88% dan PR-AUC 76,21% dengan waktu proses 5.860 ms per komentar di GPU. Prediksi paling akurat ada di kelas sentimen negatif (F1-score 82,05%), lalu positif (72,62%), dan netral (62,19%).

Pada pemodelan regresi logistik, perdebatan tentang keputusan wasit dan VAR muncul seiring meningkatnya komentar miring: perdebatan garis offside menaikkan peluang sentimen negatif sampai 3,44 kali lipat (73,91% negatif), lalu momen cek penalti sebesar 2,25 kali lipat (64,93% negatif). Walaupun uji statistik Wilcoxon (p = 0,004) dan paired t-test (p = 0,001) mendukung keunggulan model, semua temuan ini tetap harus dibaca sebagai hubungan kemunculan kata dari korpus berlabel otomatis (silver label; kecocokan 34,67% dengan penilai tunggal

DOI : https://doi.org/10.35316/jimi.v11i2.230-241
© 2026 Jurnal Ilmiah Informatika with CC BY NC licence

P-ISSN : 2549-7480
E-ISSN : 2549-6301

dan 36,75% dengan konsensus manusia, Kappa 0,02–0,05).

Penelitian ini memperlihatkan bahwa model bahasa lokal bisa digunakan untuk membaca respons penonton terhadap siaran olahraga digital di Indonesia. Ke depan, peneliti sebaiknya mencoba model bahasa lintas daerah untuk memahami komentar campur kode (code-switching), mengganti pelabelan otomatis dengan anotasi manual penuh terutama untuk kalimat sarkasme, dan menyandingkan waktu komentar dengan detik kejadian di video pertandingan.

5. REFERENSI

[1] A. G. Ghifari, G. Y. Ananada, dan K. Purwandari, "A Comparative Sentiment Analysis of Public Opinion on Indonesia's National Football Coach Using CRNN and SVM," Procedia Computer Science, vol. 269, hal. 1485-1493, 2025, doi: 10.1016/j.procs.2025.09.090.

[2] G. Sumantri dan B. S. H. Marwoto, "Analisis Sentimen di Twitter Terkait Tim Nasional Sepak Bola Indonesia Menggunakan Metode Support Vector Machine," Jurnal Kajian dan Terapan Matematika, vol. 10, no. 2, hal. 96-104, 2024, doi: 10.21831/jktm.v10i2.19561.

[3] D. A. Prabowo dan Sudianto, "The sentiment analysis of Indonesian football on Twitter by using K nearest neighbors and random forest," JSAI (Journal Scientific and Applied Informatics), vol. 6, no. 2, hal. 247-257, 2023, doi: 10.36085/jsai.v6i2.5337.

[4] A. Khan, D. Majumdar, dan B. Mondal, "Sentiment analysis of emoji fused reviews using machine learning and BERT," Scientific Reports, vol. 15, no. 1, Art. no. 7538, 2025, doi: 10.1038/s41598-025-92286-0.

[5] A. F. Hidayatullah, R. A. Apong, D. T. C. Lai, dan A. Qazi, "Pre-trained language model for code-mixed text in Indonesian, Javanese, and English using transformer," Social Network Analysis and Mining, vol. 15, no. 1, Art. no. 30, 2025, doi: 10.1007/s13278-025-01444-9.

[6] A. H. Nasution, A. Onan, Y. Murakami, W. Monika, dan A. Hanafiah, "Benchmarking Open-Source Large Language Models for Sentiment and Emotion Classification in Indonesian Tweets," IEEE Access, vol. 13, hal. 94009-94025, 2025, doi: 10.1109/ACCESS.2025.3574629.

[7] G. I. Winata et al., "NusaX: Multilingual Parallel Sentiment Dataset for 10 Indonesian Local Languages," dalam Proc. 17th Conf. European Chapter Assoc. Comput. Linguistics (EACL), 2023, hal. 815-834, doi: 10.18653/v1/2023.eacl-main.57.

[8] C. Shaw, P. LaCasse, dan L. Champagne, "Exploring emotion classification of indonesian tweets using large scale transfer learning via IndoBERT," Social Network Analysis and Mining, vol. 15, no. 1, Art. no. 22, 2025, doi: 10.1007/s13278-025-01439-6.

[9] C. Smith et al., "Social media perceptions of college football performance and season length," PLOS ONE, vol. 20, no. 7, Art. no. e0325840, 2025, doi: 10.1371/journal.pone.0325840.

[10] J. Walker et al., "Psychology at the screen: Investigating the current VAR protocol," PLOS ONE, vol. 21, no. 4, Art. no. e0345704, 2026, doi: 10.1371/journal.pone.0345704.

[11] M. N. Nazarudin et al., "VAR consultation patterns and their association with fouls and misconduct:

DOI : https://doi.org/10.35316/jimi.v11i2.230-241
© 2026 Jurnal Ilmiah Informatika with CC BY NC licence

An analysis of the top five European football leagues," PLOS ONE, vol. 20, no. 10, Art. no. e0334518, 2025, doi: 10.1371/journal.pone.0334518.

[12] D. Suhartono, W. Wongso, dan A. T. Handoyo, "IdSarcasm: Benchmarking and Evaluating Language Models for Indonesian Sarcasm Detection," IEEE Access, vol. 12, hal. 87323-87332, 2024, doi: 10.1109/ACCESS.2024.3416955.

[13] M. A. Rosid, D. Siahaan, dan A. Saikhu, "Sarcasm Detection in Indonesian-English Code-Mixed Text Using Multihead Attention-Based Convolutional and Bi-Directional GRU," IEEE Access, vol. 12, hal. 137063-137079, 2024, doi: 10.1109/ACCESS.2024.3436107.

[14] M. Mujahid et al., "Data oversampling and imbalanced datasets: an investigation of performance for machine learning and feature engineering," Journal of Big Data, vol. 11, Art. no. 87, 2024, doi: 10.1186/s40537-024-00943-4.

[15] M. Bal, A. G. Kara Aydemir, dan M. Coskun, "Exploring YouTube content creators' perspectives on generative AI in language learning: Insights through opinion mining and sentiment analysis," PLOS ONE, vol. 19, no. 9, Art. no. e0308096, 2024, doi: 10.1371/journal.pone.0308096.

[16] K. Aziz et al., "Unifying aspect-based sentiment analysis BERT and multi-layered graph convolutional networks for comprehensive sentiment dissection," Scientific Reports, vol. 14, no. 1, Art. no. 14646, 2024, doi: 10.1038/s41598-024-61886-7.

DOI : https://doi.org/10.35316/jimi.v11i2.230-241 © 2026 Jurnal Ilmiah Informatika with CC BY NC licence