# PERBANDINGAN MODEL INDOBERT DAN INDONESIAN ROBERTA BASE DALAM ANALISIS SENTIMEN PENGGUNA TWITTER TERHADAP TIMNAS INDONESIA DI KUALIFIKASI PIALA DUNIA 2026

**Eben Ezer Obed Marpaung, Masmur Tarigan**
Teknik Informatika, Universitas Esa Unggul
Jalan Arjuna Utara No. 9, Kebon Jeruk, Jakarta Barat
*ebenezermarpaung@gmail.com*

**ABSTRAK**

Perkembangan teknologi pemrosesan bahasa alami (*Natural Language Processing*/NLP) memungkinkan analisis opini publik di media sosial dilakukan secara lebih efektif, termasuk terhadap topik olahraga nasional seperti performa Tim Nasional Indonesia pada Kualifikasi Piala Dunia 2026. Penelitian ini bertujuan membandingkan performa dua model *Transformer* berbahasa Indonesia, yaitu IndoBERT dan Indonesian RoBERTa Base, dalam analisis sentimen terhadap tweet pengguna Twitter mengenai Timnas Indonesia, serta menilai pengaruh *class weighting* dan *Bayesian Optimization* terhadap peningkatan performa model. Data penelitian terdiri dari 4.676 tweet yang dikumpulkan dari akun resmi @TimnasIndonesia, melalui tahap pembersihan, pelabelan otomatis menggunakan InSet Lexicon, pembagian data secara *stratified*, tokenisasi, penerapan *class weighting*, dan penyetelan *hyperparameter* dengan *Bayesian Optimization*. Hasil menunjukkan bahwa IndoBERT memiliki performa lebih unggul dengan akurasi 76,66% dan *F1-score* 71,60%, sedangkan Indonesian RoBERTa Base mencapai akurasi 74,30% dan *F1-score* 68,92%.

***Kata kunci :*** *Analisis sentimen, IndoBERT, Indonesian RoBERTa Base, Class Weighting, Bayesian Optimization*

## 1. PENDAHULUAN

Perkembangan teknologi pemrosesan bahasa alami (*Natural Language Processing*/NLP) telah memberikan peluang besar dalam menganalisis opini publik di media sosial secara cepat, dan akurat. Twitter, sebagai salah satu platform yang paling banyak digunakan di Indonesia, menjadi ruang bagi masyarakat untuk menyampaikan respons spontan terhadap berbagai peristiwa, termasuk performa Tim Nasional Indonesia pada Kualifikasi Piala Dunia 2026. Opini-opini ini penting dipelajari untuk memahami persepsi publik, baik berupa dukungan, kritik, maupun komentar netral yang muncul setelah pertandingan berlangsung.

Kemajuan model berbasis Transformer seperti BERT dan RoBERTa semakin memperkuat kemampuan analisis sentimen otomatis. Untuk konteks bahasa Indonesia, model IndoBERT dan Indonesian RoBERTa Base menjadi dua model yang populer karena keduanya dilatih menggunakan korpus besar berbahasa Indonesia. Meskipun masing-masing model telah terbukti efektif dalam berbagai penelitian bidang [1][2][3][4][5][6][7][8][9], belum banyak studi yang secara langsung membandingkan performa keduanya pada konteks opini publik terkait olahraga nasional, khususnya pada percakapan mengenai Timnas Indonesia. Selain itu, karakteristik data media sosial seperti Twitter sering kali menunjukkan ketidakseimbangan kelas sentimen, di mana sentimen negatif cenderung lebih dominan. Ketidakseimbangan ini dapat menurunkan performa model jika tidak ditangani dengan tepat. Di sisi lain, performa model Transformer juga sangat dipengaruhi oleh konfigurasi *hyperparameter* yang optimal, sehingga diperlukan metode tuning seperti *Bayesian Optimization* agar model dapat mencapai kinerja terbaiknya.

Berdasarkan kondisi tersebut, terdapat beberapa permasalahan yang perlu diteliti, yaitu belum adanya perbandingan menyeluruh antara IndoBERT dan Indonesian RoBERTa Base dalam menganalisis sentimen publik terhadap Timnas Indonesia, adanya ketidakseimbangan kelas dalam *dataset* yang berpotensi menimbulkan bias prediksi, serta perlunya proses tuning hyperparameter yang lebih sistematis agar kedua model dapat dimaksimalkan performanya. Penelitian ini kemudian bertujuan untuk mengevaluasi dan membandingkan kinerja kedua model, menganalisis pengaruh penerapan class weighting terhadap ketidakseimbangan kelas, mengoptimalkan performa model dengan penyetelan *hyperparameter* menggunakan Bayesian Optimization, serta menentukan model mana yang paling efektif digunakan untuk analisis sentimen teks berbahasa Indonesia pada media sosial.

Untuk mencapai tujuan tersebut, penelitian ini dirancang melalui beberapa langkah. Data tweet dikumpulkan dari akun resmi @TimnasIndonesia pada periode pertandingan Kualifikasi Piala Dunia 2026. Data kemudian dibersihkan melalui proses seperti penghapusan URL, *mention*, emoji, serta normalisasi teks. Pelabelan sentimen dilakukan menggunakan InSet Lexicon untuk menentukan polaritas positif, netral, atau negatif. Dataset selanjutnya dibagi secara *stratified* menjadi data latih, validasi, dan uji. Proses tokenisasi disesuaikan dengan karakteristik masing-masing model, kemudian ketidakseimbangan kelas ditangani melalui penerapan *class weighting* pada fungsi *loss*. Tahap berikutnya adalah penyetelan *hyperparameter* menggunakan metode *Bayesian Optimization* untuk menemukan kombinasi terbaik

bagi kedua model. Setelah itu, kedua model dilatih dengan konfigurasi terbaiknya dan dievaluasi menggunakan akurasi, presisi, *recall, F1-score*, serta *confusion matrix* untuk melihat distribusi kesalahan prediksi.

Melalui rangkaian proses tersebut, penelitian ini diharapkan mampu memberikan gambaran mengenai performa IndoBERT dan Indonesian RoBERTa Base dalam menganalisis sentimen publik terhadap Timnas Indonesia.

## 2. TINJAUAN PUSTAKA

### 2.1. Analisis Sentimen

Analisis sentimen adalah proses klasifikasi sentimen berdasarkan polaritas teks dalam sebuah frasa sehingga dapat ditentukan apakah sentimen tersebut positif, negatif, atau netral. Teknik ini juga digunakan untuk mengidentifikasi bagaimana sebuah sentimen diekspresikan melalui teks dan bagaimana sentimen tersebut dapat dikategorikan [10].

### 2.2. *Transformer*

Transformer adalah jaringan saraf tiruan *deep feed-forward* yang utamanya dicirikan oleh mekanisme *self-attention* , yang memungkinkannya unggul dalam pemrosesan bahasa alami (NLP) dan pemodelan data sekuensial. Komponen arsitektur utamanya mencakup *Multi-head Attention* sebagai mekanisme inti untuk memungkinkan jaringan memanfaatkan informasi kontekstual saat memprediksi token, lapisan *Token Embedding* yang berfungsi merepresentasikan setiap elemen kosakata sebagai vektor, *Positional Embedding* yang menambahkan informasi posisi token untuk memahami urutan kata, yang biasanya ditambahkan ke *token embedding* awal, *Layer Normalisation* yang secara eksplisit mengontrol mean dan varians aktivasi jaringan, dan *Unembedding* yang mengubah representasi vektor token dan konteksnya menjadi distribusi probabilitas di atas kosakata. Arsitektur Transformer yang menonjol, seperti *Encoder-Decoder* (EDT), *Encoder-only* (BERT), dan *Decoder-only* (GPT), dibentuk dari kombinasi komponen-komponen ini [11].

### 2.3. BERT dan IndoBERT

BERT (Bidirectional Encoder Representations from Transformers) diperkenalkan pada tahun 2018 oleh Jacob Devlin dan tim dari Google sebagai model representasi bahasa yang dirancang untuk melatih representasi dua arah (*bidirectional*) dari teks tanpa label dengan mengondisikan konteks kiri dan kanan di semua lapisan secara bersamaan. Secara arsitektur, model ini mengoptimalkan proses *pre-training* melalui dua mekanisme utama: *Masked Language Model* (MLM), yang memprediksi kata-kata yang disembunyikan dalam sebuah kalimat, dan *Next Sentence Prediction* (NSP), sebuah klasifikasi biner untuk memprediksi apakah dua kalimat saling berurutan atau tidak. BERT dibangun di atas arsitektur Transformer dengan menggunakan lapisan *embedding* untuk mengubah kata menjadi vektor dan memanfaatkan token khusus [CLS] di awal input untuk merepresentasikan makna keseluruhan teks [12][13].

IndoBERT merupakan model BERT yang dirancang khusus untuk bahasa Indonesia dan dilatih menggunakan lebih dari 220 juta kata dari Wikipedia Indonesia, artikel berita, dan Indonesian Web Corpus. Model ini dievaluasi menggunakan IndoLEM, sebuah benchmark komprehensif untuk mengukur kemampuan semantik dan wacana dalam NLP bahasa Indonesia [14].

### 2.4. RoBERTa dan Indonesian RoBERTa Base

RoBERTa (Robustly Optimized BERT Approach) dikembangkan oleh Facebook AI pada tahun 2019 sebagai versi yang dioptimalkan dan lebih kuat dari model BERT standar. Model ini mempertahankan arsitektur dasar Transformer milik BERT tetapi berfokus pada penyempurnaan signifikan pada prosedur pra-pelatihan dan penyesuaian hyperparameter. Perubahan arsitektural yang mendasar pada RoBERTa mencakup penggunaan Dynamic Masking, yang secara acak menghasilkan pola masker yang berbeda pada setiap epoch pelatihan, alih-alih pola masker statis, untuk meningkatkan generalisasi model. Selain itu, RoBERTa menghapus tugas pra-pelatihan Next Sentence Prediction (NSP) secara total, karena ditemukan bahwa NSP tidak memberikan kontribusi positif yang signifikan pada kinerja model. Peningkatan kinerja RoBERTa juga didukung oleh pelatihan yang menggunakan dataset yang jauh lebih besar daripada yang digunakan untuk melatih BERT standar, yang bertujuan untuk mempercepat optimasi dan menangkap representasi bahasa yang lebih kaya [12][13].

Indonesian RoBERTa Base kemudian dikembangkan sebagai model RoBERTa yang secara khusus dilatih untuk bahasa Indonesia. Model ini dibangun dari awal menggunakan data berbahasa Indonesia dari korpus OSCAR (subset unshuffled_deduplicated_id), sebuah korpus besar berbasis Common Crawl yang telah difilter menurut bahasa, yang ikembangkan dalam program JAX/Flax Community Week oleh *Hugging Face* [15].

### 2.5. Penelitian Terdahulu

Beberapa penelitian terdahulu telah memanfaatkan model *Transformer* dalam berbagai konteks analisis sentimen di Indonesia. IndoBERT telah digunakan untuk menganalisis opini mahasiswa terhadap pembelajaran daring pasca pandemi COVID-19 dan memperoleh akurasi hingga 87% [1]. Model yang sama juga dimanfaatkan untuk analisis sentimen pengguna terhadap akun resmi aplikasi DANA di Twitter dengan *F1-score* sebesar 92% [2]. Dalam konteks deteksi hoaks, IndoBERT terbukti mengungguli LSTM dengan akurasi 92,07% [3], sementara penggunaan IndoBERT Large

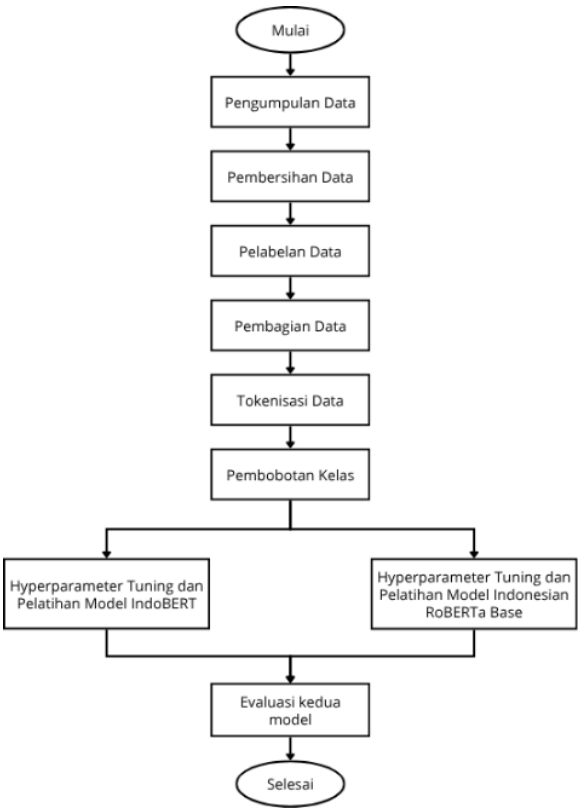
menghasilkan akurasi hingga 98% pada deteksi berita palsu [4].

Selain BERT, model RoBERTa juga banyak digunakan dalam penelitian terkini. Model RoBERTa-GRU hybrid telah diterapkan untuk menganalisis sentimen terhadap rencana pembangunan Ibu Kota Nusantara (IKN) dan mencapai *F1-score* 71,06% [5]. RoBERTa juga digunakan untuk mendeteksi ujaran kebencian pada komentar Instagram dengan akurasi rata-rata 85,09% [6], serta untuk analisis komentar YouTube terkait calon presiden dengan akurasi di atas 90% [7].

Penelitian lain membandingkan IndoBERT dan Indonesian RoBERTa Base pada analisis sentimen terhadap komentar publik mengenai kemajuan kecerdasan buatan di Indonesia dan menunjukkan bahwa IndoBERT menghasilkan akurasi lebih tinggi (83–84%) dibandingkan Indonesian RoBERTa Base [8]. Selain itu, *Bayesian Optimization* juga terbukti efektif dalam meningkatkan performa IndoBERT untuk deteksi berita palsu dengan *F1-score* mencapai 91,56% [9].

## 3. METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif untuk membandingkan kinerja dua model *Transformer* berbahasa Indonesia, yaitu IndoBERT dan Indonesian RoBERTa Base, dalam menganalisis sentimen pengguna Twitter terhadap Tim Nasional Indonesia selama Kualifikasi Piala Dunia 2026. Berikut adalah tahapan dalam penelitian ini.




Gambar 1. Bagan alur penelitian

### 3.1. Penelitian Terdahulu

Data diperoleh dari akun resmi @TimnasIndonesia menggunakan *TweetHarvest* yang terhubung dengan *Twitter API*. Tweet yang dikumpulkan adalah reply dan retweet dari thread yang diunggah setelah pertandingan Timnas Indonesia selama Kualifikasi Piala Dunia 2026 ronde ketiga, periode 5 September 2024–10 Juni 2025.

### 3.2. Pembersihan Data

Proses pembersihan data dilakukan untuk memastikan teks siap diolah oleh model. Langkah-langkahnya mencakup penghapusan URL, mention (@), tanda retweet (RT), emoji, dan simbol tidak relevan, normalisasi spasi, serta mengubah teks menjadi huruf kecil (*case folding*). Data duplikat dan kosong juga dihapus sehingga tersisa data teks bersih yang siap diberi label.

### 3.3. Pelabelan Data

Pelabelan sentimen dilakukan secara otomatis menggunakan InSet Lexicon, yaitu kamus bahasa indonesia yang berisi 3.609 kata positif dan 6.609 kata negatif dengan bobot antara -5 (sangat negatif) hingga +5 (sangat positif) [16]. Total skor setiap tweet menentukan label sentimennya:

a. Negatif jika skor < 0
b. Netral jika skor = 0
c. Positif jika skor > 0

Penentuan skor ini memungkinkan proses analisis sentimen dilakukan secara konsisten dan terstruktur, karena setiap kata dalam tweet memiliki nilai polaritas yang sudah ditetapkan. Dengan demikian, akumulasi bobot kata-kata tersebut memberikan gambaran umum mengenai kecenderungan emosi atau opini yang terkandung dalam tweet..

### 3.4. Pembagian Dataset

Dataset dibagi menjadi tiga bagian menggunakan metode *stratified split* agar proporsi tiap kelas sentimen tetap seimbang:

a. 70% data latih,
b. 10% data validasi, dan
c. 20% data uji.

Data latih digunakan untuk membangun model, data validasi untuk pemantauan performa selama pelatihan dan penyetelan hiperparameter, sedangkan data uji untuk mengevaluasi kemampuan generalisasi model.

### 3.5. Tokenisasi Data

Teks yang telah dilabeli kemudian diubah menjadi token sesuai model yang digunakan:

a. IndoBERT menggunakan algortima *WordPiece* dengan token [CLS] dan [SEP].
b. Indonesian RoBERTa Base menggunakan algoritma *Byte-Pair Encoding (BPE)* dengan token <s> dan </s>.

Tokenisasi ini mengubah teks menjadi ID numerik agar dapat diproses dalam tensor oleh model *Transformer*.

### 3.6. Pembobotan Kelas (*Class Weighting*)

Untuk mengatasi ketidakseimbangan data, diterapkan teknik *class weighting* dengan menghitung bobot berdasarkan proporsi jumlah data per kelas. Bobot kelas dihitung menggunakan fungsi *compute_class_weight()* dari pustaka *scikit-learn* dan diterapkan pada fungsi *loss* (*CrossEntropyLoss*). Bobot ini membuat kesalahan prediksi pada kelas minoritas dihitung lebih berat, sehingga model belajar secara lebih seimbang.

### 3.7. *Hyperparameter Tuning*

*Hyperparameter Tuning* dilakukan menggunakan metode *Bayesian Optimization*, yaitu salah satu metode pencarian *hyperparameter* otomatis yang diimplementasikan melalui pustaka *Optuna* [17]. Proses *tuning* dilakukan sebanyak 10 percobaan (trials) untuk setiap model guna menemukan kombinasi optimal dari *learning rate*, *batch size*, *weight decay*, dan *epoch* dari rentang tertentu. Berikut penjelasan dari tiap *hyperparameter* di-*tuning* dan rentang yang diambil pada Tabel 1.

Tabel 1. *Hyperparameter* yang dipilih untuk tuning

| ***Hyperparameter*** | **Deskripsi** | **Rentang yang Diambil** |
|---|---|---|
| *learning_rate* | Mengatur seberapa besar model memperbarui bobot saat pelatihan berlangsung. | 1e-6 – 5e-5 |
| *num_train_epochs* | Menentukan berapa kali seluruh data latih diproses selama pelatihan. | 3 – 5 |
| *batch_size* | Jumlah data yang diproses sekaligus sebelum pembaruan bobot dilakukan. | 8, 16, 32 |
| *weight_decay* | Regularisasi untuk mencegah *overfitting* dengan membatasi bobot model yang terlalu besar. | 0.0 – 0.3 |

Pemilihan kombinasi terbaik didasarkan pada nilai *F1-score* tertinggi di data validasi. Hasil optimasi ini kemudian digunakan pada tahap pelatihan kedua model.

### 3.8. Pelatihan Model

Pelatihan dilakukan menggunakan pustaka *Transformers* dari *Hugging Face* dengan kelas khusus *WeightedTrainer* untuk mendukung penerapan *class weighting*. Kedua model dilatih dengan data latih menggunakan *hyperparameter* hasil tuning dengan *Bayesian Optimization*. Evaluasi dilakukan pada setiap *epoch* menggunakan metrik *F1-score*, dan model terbaik disimpan secara otomatis pada akhir pelatihan.

### 3.9. Evaluasi Model

Evaluasi model dilakukan menggunakan data uji untuk mengukur sejauh mana model mampu melakukan klasifikasi sentimen secara akurat. Proses evaluasi diawali dengan analisis menggunakan *Confusion Matrix*, yang memberikan gambaran menyeluruh mengenai performa model dalam memprediksi setiap kelas sentimen.

*Confusion Matrix* merupakan tabel dua dimensi yang membandingkan hasil prediksi model dengan label sebenarnya pada data uji. Matriks ini menunjukkan jumlah prediksi yang benar dan salah untuk masing-masing kelas, sehingga dapat mengidentifikasi jenis kesalahan yang terjadi selama proses klasifikasi [18].

Tabel 2. Tabel *Confusion Matrix*

| **Aktual** \ **Prediksi** | **Positive** | **Negative** |
|---|---|---|
| **Positive** | True Positive (TP) | False Negative (FN) |
| **Negative** | False Positive (FP) | True Negative (TN) |

*Confusion matrix* terdiri atas empat komponen utama, yaitu True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN).

a. True Positive (TP) menunjukkan jumlah data aktual positif yang berhasil diprediksi positif.
b. True Negative (TN) adalah jumlah data aktual negatif yang berhasil diprediksi negatif.
c. False Positive (FP) merupakan data negatif yang diprediksi positif oleh model.
d. False Negative (FN) menunjukkan data positif yang diprediksi negatif oleh model.

Nilai-nilai pada *confusion matrix* digunakan untuk menghitung empat metrik utama evaluasi performa model [19], yaitu:

a. Akurasi (*Accuracy*)
Nilai akurasi merupakan persentase dari jumlah prediksi yang benar oleh model terhadap total seluruh data uji.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

b. Presisi (*Precision*)
Nilai presisi menunjukkan tingkat ketepatan antara data yang diprediksi positif oleh model dengan data yang benar-benar positif.

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

c. *Recall*
*Recall* merupakan persentase data positif yang berhasil dikenali oleh model, atau sejauh mana model mampu mendeteksi data dari kelas sebenarnya.

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

d. *F1-Score*

*F1-score* merupakan metrik evaluasi yang mengombinasikan presisi dan *recall* dalam bentuk rata-rata harmonik.

$$F1 - Score = 2 * \frac{Recall * Precision}{Recall + Precision} \quad (4)$$

Dalam penelitian ini, klasifikasi dilakukan terhadap tiga kelas sentimen, yaitu negatif, netral, dan positif. Pada kasus multi-kelas seperti ini, *confusion matrix* berbentuk 3×3, di mana setiap baris menunjukkan label aktual dan setiap kolom menunjukkan hasil prediksi. Metrik *accuracy*, *precision*, *recall*, dan *F1-score* tetap digunakan, namun masing-masing dihitung untuk setiap kelas sentimen dan kemudian dirata-ratakan agar diperoleh gambaran performa keseluruhan model secara seimbang (*macro average*).

Tabel 3. Tabel *Confusion Matrix* untuk 3 kelas

| Aktual \ Prediksi | **Positive** | **Netral** | **Negative** |
|---|---|---|---|
| **Positive** | True Positive (TP) | False Netral (FNt) | False Negative (FN) |
| **Netral** | False Positive (FP) | True Netral (TNt) | False Negative (FN) |
| **Negative** | False Positive (FP) | False Netral (FNt) | True Negative (TN) |

Hasil evaluasi kedua model dibandingkan untuk menentukan model yang memberikan performa terbaik pada analisis sentimen tweet terhadap Timnas Indonesia.

## 4. HASIL DAN PEMBAHASAN

### 4.1. Gambaran Dataset

Proses *scraping* dilakukan pada akun resmi @TimnasIndonesia menggunakan tools *TweetHarvest* yang memanfaatkan *Twitter API*. Pengambilan data dilakukan dalam waktu 1×24 jam setelah pertandingan selama periode 5 September 2024 – 10 Juni 2025, sehingga mencakup seluruh pertandingan Tim Nasional Indonesia pada Kualifikasi Piala Dunia 2026. Dari proses ini diperoleh 5.632 tweet mentah yang berisi opini dan tanggapan pengguna terhadap performa tim.

Tabel 4. Hasil scraping dari setiap tweet setelah pertandingan

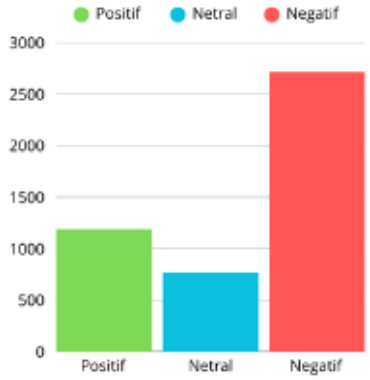
| No. | Pertandingan | Tanggal pertandingan | Jumlah Tweet |
|---|---|---|---|
| 1 | Arab Saudi vs Indonesia | 6 September 2024 | 493 |
| 2 | Indonesia vs Australia | 10 September 2024 | 605 |
| 3 | Bahrain vs Indonesia | 10 Oktober 2024 | 606 |
| 4 | China vs Indonesia | 15 Oktober 2024 | 604 |
| 5 | Indonesia vs Jepang | 15 November 2024 | 607 |
| 6 | Indonesia vs Arab Saudi | 19 November 2024 | 563 |
| 7 | Australia vs Indonesia | 20 Maret 2025 | 611 |
| 8 | Indonesia vs Bahrain | 25 Maret 2025 | 612 |
| 9 | Indonesia vs China | 5 Juni 2025 | 322 |
| 10 | Jepang vs Indonesia | 10 Juni 2025 | 609 |
| **Total** | | | 5632 |

Setelah dilakukan tahap pembersihan data, jumlah tweet yang layak digunakan menjadi 4.676 tweet. Proses pembersihan meliputi penghapusan URL, mention, hashtag, tanda "RT", emoji, angka, tanda baca, dan karakter non-alfabet, serta penerapan *case folding* dan penghapusan *null* serta duplikasi.

Tabel 5. Contoh tweet yang sudah dibersihkan

| Teks asli | Teks yang sudah dibersihkan |
|---|---|
| @TimnasIndonesia Mantap Indonesia https://t.co/ObZVE46jaC | mantap indonesia |
| @_ninuninuninu @_Frmnnnn @xymons_24 @SamFerkyFerslm @TimnasIndonesia nah udh mas Timnas fix lolos round 4. | nah udh mas timnas fix lolos round 4. |
| @TimnasIndonesia #GARUDACENGKRAMNAGA #GARUDACENGKRAMNAGA | garudacengkramnaga garudacengkramnaga |
| @TimnasIndonesia Terima kasih Timnas. INDONESIA~~ INDONESIA~ #TimnasDay | terima kasih timnas. indonesia indonesia timnasday |
| @Opungdoli_10 @TimnasIndonesia Bukan sepenuhnya salah pelatih sih Tp PSSI beg* 100% tidak ada keraguan. Mending #PSSIBubarAja | bukan sepenuhnya salah pelatih sih tp pssi beg 100 tidak ada keraguan. mending pssibubaraja |

### 4.2. Hasil Pelabelan Dataset




Gambar 2. Grafik perbandingan jumlah data antar kelas

Hasil pelabelan sentimen menggunakan InSet Lexicon menunjukkan bahwa dataset terbagi ke dalam tiga kelas sentimen, yaitu positif, netral, dan negatif. Dari total 4.676 tweet, diperoleh distribusi sebagai berikut: 1.189 tweet (25%) berkategori positif, 771 tweet (17%) berkategori netral, dan 2.716 tweet (58%) berkategori negatif.

Gambar 2 menunjukkan bahwa kelas negatif mendominasi dataset, yang menandakan kecenderungan warganet untuk lebih aktif mengekspresikan ketidakpuasan dibandingkan dukungan

### 4.3. Hasil Pembagian Dataset

Dataset yang telah dilabeli kemudian dibagi menjadi data latih (70%), data validasi (10%), dan data uji (20%) dengan metode *stratified split* agar proporsi tiap kelas tetap seimbang. Hasil pembagian ditunjukkan pada Tabel 6.

Tabel 6. Distribusi pembagian data di tiap kelas

| Kelas | Jumlah Awal | Data Latih (70%) | Data Validasi (10%) | Data Uji (20%) |
|---|---|---|---|---|
| Negatif | 2.716 | 1.901 | 272 | 543 |
| Positif | 1.189 | 832 | 119 | 238 |
| Netral | 771 | 540 | 77 | 154 |
| **Total** | **4.676** | **3.273** | **467** | **936** |

### 4.4. Hasil *Hyperparameter Tuning*

Proses *Hyperparameter Tuning* dilakukan dengan metode *Bayesian Optimization* menggunakan pustaka *Optuna* sebanyak 10 percobaan (*trials*) untuk setiap model. *Hyperparameter* yang dioptimasi meliputi *learning rate*, *batch size*, *epoch*, dan *weight decay*. Tujuan *tuning* ini adalah menemukan konfigurasi yang memberikan nilai *F1-score* tertinggi pada data validasi

### 4.5. IndoBERT

Dari 10 percobaan (*trials*) yang telah dilakukan, konfigurasi terbaik diperoleh dengan kombinasi *hyperparameter* yang didapat sebagai berikut:

a. *Learning rate*: $5.42\times10^{-6}$
b. *Epoch*: 3
c. *Batch size*: 8
d. *Weight decay*: 0.1099

Kombinasi ini menghasilkan nilai akurasi 76.66%, presisi 72.74%, *recall* 70.87%, dan *F1-score* 71.60% pada data validasi. *Hyperparameter* terbaik ini kemudian digunakan dalam pelatihan akhir model IndoBERT.

### 4.6. Indonesian RoBERTa Base

Dari 10 percobaan (trials) yang telah dilakukan, konfigurasi terbaik diperoleh dengan kombinasi *hyperparameter* yang didapat sebagai berikut:

a. *Learning rate*: $2.60\times10^{-5}$
b. *Epoch*: 3
c. *Batch size*: 32
d. *Weight decay*: 0.1574

Kombinasi tersebut menghasilkan nilai akurasi 74.30%, presisi 70.12%, *recall* 67.84%, dan *F1-score* 68.92%. *Hyperparameter* terbaik ini kemudian digunakan dalam pelatihan akhir model Indonesian RoBERTa Base.

### 4.7. Evaluasi Model

Model yang telah dituning kemudian dilatih ulang (*final training*) menggunakan data latih dan data evaluasi. Evaluasi dilakukan dengan menggunakan data uji berdasarkan empat metrik utama, yaitu akurasi, presisi, *recall*, dan *F1-score*, serta visualisasi *confusion matrix* untuk melihat distribusi kesalahan antar kelas.

### 4.8. IndoBERT

*Confusion matrix* hasil prediksi IndoBERT ditunjukkan pada tabel berikut:

Tabel 7. Hasil *Confusion Matrix* pada model IndoBERT

| Aktual \ Prediksi | Positif | Netral | Negatif |
|---|---|---|---|
| **Positif** | 168 | 21 | 49 |
| **Netral** | 22 | 99 | 33 |
| **Negatif** | 43 | 22 | 479 |

Model IndoBERT menunjukkan performa tinggi pada kelas negatif, dengan 479 prediksi benar dari 544 data. Namun, kelas netral cenderung sulit dibedakan, dengan *F1-score* hanya 0.6689. Berikut adalah hasil metrik evaluasi pada keseluruhan kelas yang ditunjukkan pada Gambar 3.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| negatif | 0.8538 | 0.8805 | 0.8670 | 544 |
| netral | 0.6972 | 0.6429 | 0.6689 | 154 |
| positif | 0.7210 | 0.7059 | 0.7134 | 238 |
| accuracy | | | 0.7970 | 936 |
| macro avg | 0.7573 | 0.7431 | 0.7498 | 936 |
| weighted avg | 0.7943 | 0.7970 | 0.7953 | 936 |

Gambar 3. Hasil metrik evaluasi model IndoBERT pada keseluruhan kelas sentimen

### 4.9. Indonesian RoBERTa Base

*Confusion matrix* untuk Indonesian RoBERTa Base adalah sebagai berikut:

Tabel 8. Hasil *Confusion Matrix* pada model Indonesian Roberta Base

| Aktual \ Prediksi | Positif | Netral | Negatif |
|---|---|---|---|
| **Positif** | 163 | 34 | 41 |
| **Netral** | 23 | 108 | 23 |
| **Negatif** | 77 | 62 | 405 |

Model Indonesian RoBERTa Base cukup baik dalam mengenali kelas netral (108 benar dari 154), tetapi lemah pada kelas negatif dan positif, di mana banyak kesalahan klasifikasi silang antar kelas.

Berikut adalah hasil metrik evaluasi pada keseluruhan kelas yang ditunjukkan pada Gambar 4.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| negatif | 0.8635 | 0.7445 | 0.7996 | 544 |
| netral | 0.5294 | 0.7013 | 0.6034 | 154 |
| positif | 0.6198 | 0.6849 | 0.6507 | 238 |
| accuracy | | | 0.7222 | 936 |
| macro avg | 0.6709 | 0.7102 | 0.6846 | 936 |
| weighted avg | 0.7466 | 0.7222 | 0.7295 | 936 |

Gambar 4. Hasil metrik evaluasi model Indonesian RoBERTa Base pada keseluruhan kelas sentimen

### 4.10. Perbandingan Hasil Evaluasi

Berdasarkan hasil evaluasi pada kedua model yang telah dilatih menggunakan kombinasi hiperparameter terbaik, diperoleh perbedaan performa yang cukup signifikan.

Model IndoBERT menunjukkan hasil lebih unggul dibandingkan Indonesian RoBERTa Base pada metrik akurasi dan *F1-score*. Kombinasi learning rate dan batch size yang lebih kecil pada IndoBERT menghasilkan pelatihan yang lebih stabil dan akurasi yang lebih tinggi.

Sementara itu, Indonesian RoBERTa Base masih memberikan hasil yang cukup baik, tetapi belum mampu melampaui IndoBERT pada kedua metrik evaluasi.

Tabel 9. Perbandingan hasil evaluasi pada kedua model

| Model | IndoBERT | Indonesian RoBERTa Base |
|---|---|---|
| *Learning rate* | $5.42\times10^{-6}$ | $2.60\times10^{-5}$ |
| *Epoch* | 3 | 3 |
| *Batch size* | 8 | 32 |
| *Weight decay* | 0.1099 | 0.1574 |
| *Accuracy* | 79.7% | 72.22% |
| *Macro Avg F1-score* | 0.7498 | 0.6846 |

## 5. KESIMPULAN DAN SARAN

Penelitian ini membandingkan kinerja IndoBERT dan Indonesian RoBERTa Base dalam klasifikasi sentimen pada 4.676 tweet tentang Tim Nasional Indonesia selama Kualifikasi Piala Dunia 2026. Setelah melalui proses pembersihan, pelabelan otomatis dengan InSet Lexicon, pemisahan data, serta penerapan *class weighting* dan *Bayesian Optimization*, hasil evaluasi menunjukkan bahwa IndoBERT unggul pada seluruh metrik dengan akurasi 79,7% dan macro *F1-score* 0,7498, dibandingkan Indonesian RoBERTa Base yang memperoleh akurasi 72,22% dan *macro F1-score* 0,6846. IndoBERT juga lebih konsisten dalam mengenali sentimen positif dan negatif, sementara RoBERTa Base sedikit lebih baik pada kelas netral. Temuan ini menegaskan bahwa IndoBERT lebih sesuai untuk analisis sentimen berbasis teks media sosial berbahasa Indonesia. Penelitian selanjutnya disarankan memperluas pencarian hiperparameter, melakukan pelabelan manual atau semi-manual, mengevaluasi model *Transformer* lain, menerapkan teknik penyeimbangan data alternatif, serta membandingkan metode *hyperparameter* tuning seperti Grid Search dan Random Search.

## DAFTAR PUSTAKA


[1] M. N. Hidayat dan R. Pramudita, "Analisis Sentimen Terhadap Pembelajaran Secara Daring Pasca Pandemi Covid-19 Menggunakan Metode IndoBERT," *Information Management for Educators and Professionals*, vol. 8, no. 2, hlm. 161–170, Des 2023, doi: 10.51211/imbi.v8i2.2719.

[2] P. J. J. S. Gautama, A. Wibowo, dan J. J. Siang, "ANALISIS SENTIMEN PENGGUNA TERHADAP AKUN X/TWITTER RESMI 'DANA' DENGAN ALGORITMA INDOBERT," *Journal of Information Technology and Computer Science (INTECOMS)*, vol. 8, no. 2, 2025, doi: 10.31539/intecoms.v8i2.14477.

[3] M. I. K. Sinapoy, Y. Sibaroni, dan S. S. Prasetyowati, "Comparison of LSTM and IndoBERT Method in Identifying Hoax on Twitter," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 7, no. 3, hlm. 657–662, 2023, doi: 10.29207/resti.v7i3.4830.

[4] M. Y. Ridho dan E. Yulianti, "From Text to Truth: Leveraging IndoBERT and Machine Learning Models for Hoax Detection in Indonesian News," *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika*, vol. 10, no. 3, hlm. 544–555, Sep 2024, doi: 10.26555/jiteki.v10i3.29450.

[5] C. A. Deagusti dan M. I. Irawan, "Analisis Sentimen terhadap Rencana Pembangunan Ibu Kota Nusantara (IKN) Berdasarkan Twitter (X) Menggunakan Metode Hybrid RoBERTa-GRU," *JURNAL SAINS DAN SENI ITS*, vol. 13, hlm. 23–30, 2024, doi: 10.12962/j23373520.v13i4.156405.

[6] A. A. Azhari, Y. Sibaroni, dan S. S. Prasetiyowati, "Detection of Indonesian Hate Speech in the Comments Column of Indone-sian Artists' Instagram Using the RoBERTa Method," *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, vol. 8, no. 3, hlm. 764–773, Sep 2023, doi: 10.29100/jipi.v8i3.3898.

[7] Y. P. Sumihar, "Sentiment Analysis of Public Opinions Regarding 'Ideas of Presidential Candidates' in YouTube Video Comments with Robustly Optimized BERT Pretraining Approach," *JUSIKOM PRIMA (Jurnal Sistem Informasi dan Ilmu Komputer Prima)*, vol. 8, no. 1, Agu 2024, doi: 10.34012/jurnalsisteminformasidanilmukomputer.v8i1.5350.

[8] N. A. R. Putri dan Ardiansyah, "Analisis Sentimen Terhadap Kemajuan Kecerdasan Buatan di Indonesia Menggunakan BERT dan

RoBERTa," *Jurnal Sains dan Informatika*, vol. 9, no. 2, hlm. 136–145, Nov 2023, doi: 10.34128/jsi.v9i2.649.

[9] A. Simanjuntak *dkk.*, "Research and Analysis of IndoBERT Hyperparameter Tuning in Fake News Detection," *Jurnal Nasional Teknik Elektro dan Teknologi Informasi*, vol. 13, no. 1, hlm. 60–67, Feb 2024, doi: 10.22146/jnteti.v13i1.8532.

[10] W. Ningsih, B. Alfianda, Rahmaddeni, dan D. Wulandari, "Perbandingan Algoritma SVM dan Naïve Bayes dalam Analisis Sentimen Twitter pada Penggunaan Mobil Listrik di Indonesia," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 4, no. 2, hlm. 556–562, Feb 2024, doi: 10.57152/malcom.v4i2.1253.

[11] M. Phuong dan M. Hutter, "Formal Algorithms for Transformers," Jul 2022, [Daring]. Tersedia pada: http://arxiv.org/abs/2207.09238

[12] U. I. Shabrina, M. I. Java, dan S. Rochimah, "OPTIMIZING SENTIMENT ANALYSIS IN EDUCATIONAL YOUTUBE VIDEOS: A COMPARATIVE STUDY OF ROBERTA AND MULTINOMIAL NAIVE BAYES," *JUTI: Jurnal Ilmiah Teknologi Informasi*, vol. 22, no. 2, hlm. 83–90, Jul 2024, doi: 10.12962/j24068535.v22i2.a1204.

[13] Moh. R. Zamroni, M. Sholihin, E. Hayati, R. A. Hamid, dan N. A. Omar, "Evaluating BiLSTM Performance with BERT, RoBERTa, and DistilBERT in Online Bullying News Detection," *JURNAL TEKNIK INFORMATIKA*, vol. 18, no. 2, hlm. 196–208, Okt 2025, doi: 10.15408/jti.v18i2.42459.

[14] F. Koto, A. Rahimi, J. H. Lau, dan T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," dalam *28th International Conference on Computational Linguistics*, 2020, hlm. 757–770. doi: 10.48550/arXiv.2011.00677.

[15] Flax Community, "Indonesian RoBERTa Base," Hugging Face. Diakses: 3 Mei 2025. [Daring]. Tersedia pada: https://huggingface.co/flax-community/indonesian-roberta-base

[16] H. Firda, P. Putra, N. R. Oktadini, P. E. Sevtiyuni, dan A. Meiriza, "Comparison of Rating-based and Inset Lexicon-based Labeling in Sentiment Analysis using SVM (Case Study: GoBiz Application Reviews on Google Play Store)," *SISTEMASI*, vol. 14, no. 2, hlm. 516, Mar 2025, doi: 10.32520/stmsi.v14i2.4795.

[17] E. Elgeldawi, A. Sayed, A. R. Galal, dan A. M. Zaki, "Hyperparameter tuning for machine learning algorithms used for arabic sentiment analysis," *Informatics*, Nov 2021, doi: 10.3390/informatics8040079.

[18] F. R. Valerian, M. Syarief, dan D. A. Fatah, "KLASIFIKASI TINGKAT OBESITAS MENGGUNAKAN METODE GBM DAN CONFUSION MATRIX," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 9, no. 2, 2025, doi: 10.36040/jati.v9i2.13062.

[19] T. Arifqi, N. Suarna, dan W. Prihartono, "PENGGUNAAN NAIVE BAYES DALAM MENGANALISIS SENTIMEN ULASAN APLIKASI MCDONALD'S DI INDONESIA," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 8, no. 2, 2024, doi: 10.36040/jati.v8i2.8740