Jurnal Manajemen dan Teknologi
https://journal.stmiki.ac.id

E-ISSN: 3047-8251 | P-ISSN: 3047-826X
Vol. 3 No. 2 (2026) | November

OPEN ACCESS

https://doi.org/ 10.63447/jmt.v3i2.2196

Perbandingan IndoBERT dan mBERT Untuk Klasifikasi Toksisitas Pada Chat Roblox dan Minecraft Berbahasa Indonesia

Idris Sonny [1], Allsela Meiriza [2*], Hardini Novianti [3], Mira Afrina [4], Rizka Dhini Kurnia [5]

[1,2*,3,4,5] Program Studi Sistem Informasi, Fakultas Ilmu Komputer, Universitas Sriwijaya, Kota Palembang, Provinsi Sumatera Selatan, Indonesia.

Corresponding Email: allsela_meiriza@yahoo.co.id [1*]

Histori Artikel:
Dikirim 16 Oktober 2026; *Diterima dalam bentuk revisi* 20 Oktober 2026; *Diterima* 30 Oktober 2026; *Diterbitkan* 25 November 2026. Semua hak dilindungi oleh Lembaga Penelitian dan Pengabdian Masyarakat (LPPM) STMIK Indonesia Banda Aceh.

Abstrak

Chat toksik pada gim daring memaparkan pemain muda pada pelecehan, sementara moderasi chat berbahasa Indonesia terhambat oleh slang dan ejaan yang tidak baku. Penelitian ini bertujuan membandingkan dua encoder transformer terlatih, yaitu IndoBERT yang monolingual dan mBERT yang multibahasa, untuk klasifikasi toksisitas empat kelas (neutral, violence, racist, dan harassment) pada chat Roblox dan Minecraft berbahasa Indonesia, guna menentukan model yang lebih sesuai untuk moderasi chat. Kedua model di-fine-tune dengan pengaturan identik pada Indonesian Chat Dataset (10.702 chat) dengan tiga seed acak, dievaluasi menggunakan akurasi serta precision, recall, dan F1-score rata-rata makro, dan diuji secara statistik dengan uji McNemar. IndoBERT mencapai macro F1-score rata-rata 0,8541 ± 0,0019, lebih tinggi daripada mBERT (0,8263 ± 0,0090), dengan selisih signifikan pada dua dari tiga seed. Neutral merupakan kelas yang paling sulit dikenali kedua model, dan IndoBERT meloloskan 5,8% chat toksik sebagai neutral, dibandingkan 8,3% pada mBERT. Evaluasi tanpa 106 chat uji yang berulang di data latih, validasi atau uji tidak mengubah urutan kinerja. Pada dataset dan pengaturan pengujian ini, IndoBERT lebih unggul, dengan keterbatasan berupa label tanpa pengukuran kesepakatan antarpelabel dan data yang hanya berasal dari dua gim.

Kata Kunci: Klasifikasi Toksisitas; IndoBERT; mBERT; Chat Gim; Transformer; Bahasa Indonesia.

Abstract

Toxic chat in online games exposes young players to harassment, while moderation of Indonesian chat is hampered by slang and non-standard spelling. This study compares monolingual IndoBERT and multilingual mBERT for four-class toxicity classification (neutral, violence, racist, harassment) of Indonesian Roblox and Minecraft chat to identify the more suitable model for moderation. Both models were fine-tuned with identical settings on the Indonesian Chat Dataset (10,702 chats) using three seeds, and evaluated with accuracy, macro-averaged precision, recall, F1-score, and McNemar's test. IndoBERT reached a mean macro F1-score of 0.8541 ± 0.0019 versus 0.8263 ± 0.0090 for mBERT, significant in two of three seeds. Neutral was the hardest class, and IndoBERT let 5.8% of toxic chats pass as neutral versus 8.3% for mBERT. Excluding 106 duplicated test chats did not change the ranking. Under these settings, IndoBERT performed better, although labels lack inter-annotator agreement and data cover two games only.

Keyword: Toxicity Classification; IndoBERT; mBERT; Game Chat; Transformer; Indonesian Language.

Jurnal Manajemen dan Teknologi
https://journal.stmiki.ac.id

E-ISSN: 3047-8251 | P-ISSN: 3047-826X
Vol. 3 No. 2 (2026) | November

OPEN ACCESS

https://doi.org/ 10.63447/jmt.v3i2.2196

1. Pendahuluan

Bagi anak dan remaja, gim daring kini bukan sekadar hiburan, melainkan ruang bersosialisasi. Skalanya pun besar: Peta Ekosistem Industri Game Indonesia 2021 mencatat lebih dari 170 juta pemain gim di berbagai platform di Indonesia (Kementerian Komunikasi dan Informatika Republik Indonesia, 2022). Roblox dan Minecraft dipandang luas sebagai gim yang ramah anak dan mengandalkan konten buatan pengguna serta chat dalam gim (Prestiliano dkk., 2026). Filter chat bawaannya terutama menyasar bahasa Inggris, sehingga pesan kasar berbahasa Indonesia masih dapat sampai kepada pemain muda (Prestiliano dkk., 2026). Perilaku toksik dalam gim multipemain telah dikaji melalui tinjauan sistematis tentang definisinya (Kordyaka dkk., 2025) dan sistem intervensinya (Wijkstra dkk., 2024). Karena pelaporan manual dan penyaringan kata kunci sulit dipertahankan ketika slang terus berkembang, klasifikasi teks otomatis menjadi pelengkap yang menarik untuk moderasi chat. Kajian tentang chat gim yang toksik masih relatif terbatas dibandingkan kajian media sosial. Yang dkk. (2023) mengusulkan deteksi toksisitas waktu nyata yang memanfaatkan riwayat chat dan segmentasi pembicara, dan Hibatullah dkk. (2025) mempelajari deteksi pelecehan verbal pada gim dengan pembelajaran mesin. Untuk bahasa Indonesia, Yanti dkk. (2026) membandingkan metode pembelajaran mesin dan teknik representasi teks untuk percakapan toksik di Roblox. Prestiliano dkk. (2026) memperkenalkan Indonesian Chat Dataset berisi 10.702 chat dari pemain Roblox dan Minecraft, berlabel neutral, violence, racist, atau harassment, dan melaporkan akurasi 94,30% pada modalitas verbal dengan model BERT–BiLSTM pada validasi silang lima lipatan. Banyak penelitian toksisitas berbahasa Indonesia menyasar media sosial, tempat bahasa informal, slang, dan campur kode masih menjadi kendala utama (Ibrohim & Budi, 2023; Pamungkas dkk., 2023). Pendekatan berbasis BERT sudah dipakai untuk mendeteksi komentar toksik (Nabiilah dkk., 2023) maupun ujaran kebencian di Twitter, yaitu dengan menggabungkan IndoBERTweet, BiLSTM, dan CNN (Hakim dkk., 2024). Penelitian semacam ini juga ditopang oleh dataset beranotasi berskala besar, salah satunya IndoToxic2024 (Susanto dkk., 2024). IndoBERT juga telah dikombinasikan dengan feedforward network (Dharmawan dkk., 2023) dan DoRA, metode fine-tuning hemat parameter (Suharjanto & Sumarsono, 2025), untuk deteksi ujaran kebencian.

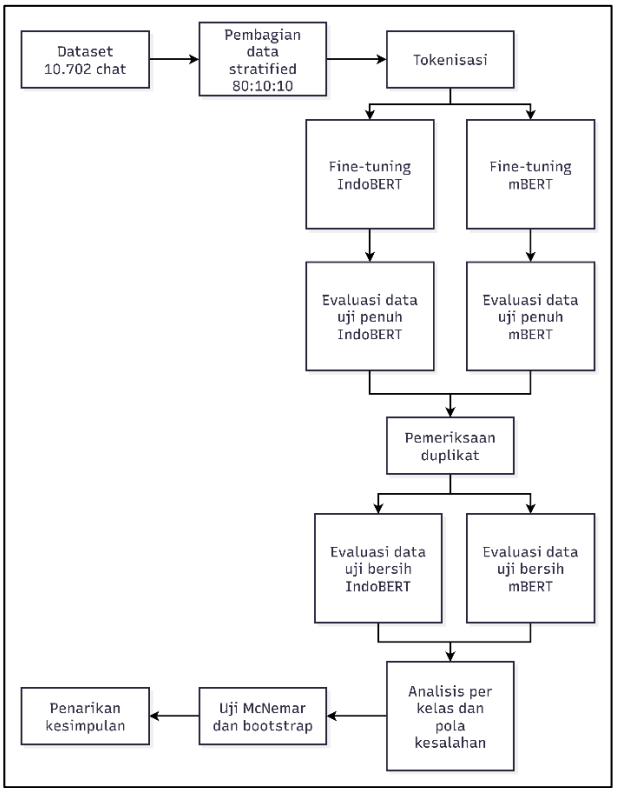
Model bahasa terlatih seperti BERT diadaptasi ke tugas hilir melalui fine-tuning (Min dkk., 2023). Untuk bahasa Indonesia, dua encoder yang umum dipakai adalah IndoBERT, yang pelatihan awalnya memakai korpus berbahasa Indonesia (Kartika dkk., 2023; Tobing dkk., 2025), dan multilingual BERT (mBERT), yang pelatihan awalnya mencakup banyak bahasa (Nabiilah dkk., 2024; Tobing dkk., 2025). Perbandingan langsung keduanya telah dilaporkan pada komentar toksik di media sosial, yang hasil terbaiknya diperoleh IndoBERT (Nabiilah dkk., 2023), dan pada deteksi hoaks politik, tempat Tobing dkk. (2025) menemukan IndoBERT lebih unggul pada seluruh metrik. Nabiilah dkk. (2024) juga mengevaluasi embedding IndoBERT dengan pengklasifikasi berbasis mBERT pada tugas multilabel. IndoBERT juga menunjukkan kinerja yang konsisten pada deteksi berita palsu dan hoaks (Isa dkk., 2022; Rahmawati dkk., 2022), sedangkan perbandingan transformer monolingual dan multibahasa untuk deteksi ujaran kebencian pada bahasa sumber daya rendah melaporkan hasil yang beragam (Ghosh & Senapati, 2025). IndoBERT dipilih karena dilatih khusus pada korpus berbahasa Indonesia sehingga diharapkan lebih peka terhadap kosakata informal, sedangkan mBERT dipilih sebagai pembanding multibahasa yang lazim dipakai dan mewakili pilihan praktis bila satu model harus melayani banyak bahasa. Keduanya berukuran base dengan arsitektur yang sama, sehingga perbedaan kinerja lebih terkait pada korpus pelatihan awal dan kosakatanya. Berdasarkan penelusuran literatur yang penulis lakukan, penelitian terdahulu pada chat gim berbahasa Indonesia belum membandingkan secara langsung encoder monolingual dan multibahasa untuk klasifikasi toksisitas multikelas. Penelitian yang memperkenalkan dataset ini menggabungkan BERT dengan BiLSTM dan membandingkannya dengan arsitektur lain, tetapi tidak melaporkan perbandingan antara encoder terlatih monolingual dan multibahasa (Prestiliano dkk., 2026), sedangkan penelitian lain membandingkan metode pembelajaran mesin dan representasi teks pada chat Roblox (Yanti dkk., 2026); keduanya tidak membandingkan IndoBERT

Jurnal Manajemen dan Teknologi
https://journal.stmiki.ac.id
E-ISSN: 3047-8251 | P-ISSN: 3047-826X
Vol. 3 No. 2 (2026) | November
OPEN ACCESS
https://doi.org/ 10.63447/jmt.v3i2.2196

dan mBERT. Perbandingan kedua model memang telah dilakukan pada komentar media sosial (Nabiilah dkk., 2023) dan hoaks politik (Tobing dkk., 2025), tetapi belum pada chat gim yang pendek (rata-rata 17,89 token) dan banyak memuat slang, sehingga belum jelas apakah keunggulan model monolingual pada domain tersebut berlaku di sini. Pemilihan encoder yang tepat menentukan seberapa andal moderasi otomatis dapat melindungi pemain muda dari chat toksik berbahasa Indonesia. Oleh karena itu, penelitian ini bertujuan membandingkan kinerja IndoBERT dan mBERT hasil fine-tuning dengan pengaturan identik pada Indonesian Chat Dataset, menguji signifikansi statistik selisihnya, serta menganalisis kinerja per kelas, pola kesalahan, dan pengaruh chat duplikat antara data latih dan data uji.

2. Metode Penelitian

Penelitian ini merupakan eksperimen kuantitatif yang membandingkan dua *encoder* transformer terlatih pada klasifikasi toksisitas empat kelas untuk chat gim berbahasa Indonesia. Seperti ditunjukkan pada Gambar 1, penelitian dimulai dari Indonesian Chat Dataset yang tersedia untuk publik melalui platform Kaggle. Data kemudian dibagi menjadi data latih, validasi, dan uji dengan rasio 80:10:10, lalu ditokenisasi dengan *tokenizer* masing-masing model. *Fine-tuning* IndoBERT dan mBERT dilakukan pada data yang sama dengan pengaturan identik, masing-masing dengan tiga *seed* acak. Setiap model dievaluasi pada data uji penuh menggunakan akurasi serta precision, recall, dan F1-score rata-rata makro, disertai analisis per kelas dan *confusion matrix*. Setelah evaluasi awal, dilakukan pemeriksaan duplikat antar subset, dan kinerja dihitung ulang pada data uji bersih yang tidak memuat chat yang berulang di data latih, validasi, atau uji. Selisih kinerja kedua model diuji secara statistik sebagai dasar penarikan kesimpulan.



Gambar 1. Tahapan penelitian

Jurnal Manajemen dan Teknologi
https://journal.stmiki.ac.id
E-ISSN: 3047-8251 | P-ISSN: 3047-826X
Vol. 3 No. 2 (2026) | November
OPEN ACCESS
https://doi.org/ 10.63447/jmt.v3i2.2196

Data yang dipakai adalah Indonesian *Chat Dataset* (Prestiliano dkk., 2026), yang dapat diunduh secara terbuka. Dataset ini terdiri atas 10.702 *chat* dengan tiga kolom, yaitu pengenal unik, teks *chat*, dan label. Menurut penyusun *dataset*, *chat* dikumpulkan dari sekitar 42 mahasiswa pemain Roblox dan Minecraft selama kurang lebih dua bulan dengan menangkap layar dan mengekstrak teksnya menggunakan *OCR*; sekitar 14.000 *chat* yang terkumpul disaring secara manual hingga menjadi 10.702. Penyusun *dataset* menyatakan bahwa pelabelan dilakukan secara manual oleh para penulis dan dinilai konsisten karena ekspresi toksik bersifat eksplisit, tetapi jumlah pelabel dan ukuran kesepakatan antarpelabel (misalnya *Cohen's kappa*) tidak dilaporkan (Prestiliano dkk., 2026). Konsistensi label dengan demikian tidak dapat diverifikasi secara kuantitatif; penelitian ini memperlakukan label yang dirilis sebagai acuan kebenaran, dan keterbatasan. Dataset memuat teks percakapan yang dapat bersifat ofensif; penelitian ini hanya memakai data yang telah dirilis publik oleh penyusunnya. Setiap *chat* termasuk dalam salah satu dari empat kelas (Prestiliano dkk., 2026). *Chat neutral* bersifat santai dan tidak bermaksud menyakiti. *Violence* mencakup makian, ancaman, hasutan untuk menyakiti, dan hinaan yang menyamakan orang dengan hewan. *Racist* mencakup pernyataan yang merendahkan orang berdasarkan ras, agama, etnis, atau kebangsaan. *Harassment* mencakup konten seksual, penghinaan fisik (*body shaming*), dan perendahan. Karena kelas *violence* mencakup makian, batasnya dengan kelas *neutral* yang bersifat santai dapat kabur. Tabel 1 menunjukkan bahwa kelas relatif seimbang, berkisar antara 23,2% dan 27,9%.

Tabel 1. Distribusi kelas pada Indonesian Chat Dataset

Kelas	Jumlah chat	Proporsi (%)
Violence	2.983	27,9
Neutral	2.729	25,5
Racist	2.506	23,4
Harassment	2.484	23,2
Total	10.702	100

2.1 Pembagian Data dan Tokenisasi

Teks telah dinormalisasi oleh penyusun dataset melalui pengubahan ke huruf kecil, penghapusan tanda baca, dan konversi angka *leetspeak* menjadi huruf (misalnya, "4nj1n9" menjadi "anjing") (Prestiliano dkk., 2026). Meskipun demikian, sebagian bentuk khas bahasa chat dan sisa artefak pengolahan masih ditemukan, yaitu penanda *user* pada 3.573 chat (33,4%), karakter baris baru dalam bentuk teks pada 832 chat (7,8%), dan urutan kode heksadesimal pada 799 chat (7,5%), serta angka yang menggantikan huruf atau menandai pengulangan (misalnya "anj11111ng" dan "deg2an"). Penelitian ini tidak melakukan pembersihan tambahan agar alur tetap selaras dengan data yang dirilis dan perbandingan kedua model tidak dipengaruhi pilihan praproses tambahan. Dataset dipartisi dengan perbandingan 80%, 10%, dan 10% untuk pelatihan, validasi, dan pengujian. Partisi dilakukan melalui dua kali pembagian bertingkat berurutan dengan *random state* 42. Nilai ini hanya mengatur pembagian data dan terpisah dari *seed* pelatihan. Pembagian ini menghasilkan 8.561, 1.070, dan 1.071 chat dengan proporsi kelas terjaga di setiap subset. Kedua model memakai pembagian yang persis sama. Setelah pembagian, setiap subset ditokenisasi dengan *tokenizer* masing-masing model, dengan panjang maksimum 128 token, *padding* hingga panjang tersebut, dan *truncation*. Chat dalam dataset tergolong pendek, dengan rata-rata 17,89 token dan median 13 pada *tokenizer* IndoBERT, dan 99,89% di antaranya muat dalam 128 token, sehingga pemotongan hampir tidak memengaruhi sampel.

2.2 Fine-Tuning Model

Penelitian ini membandingkan indobenchmark/indobert-base-p1, model yang dilatih pada korpus berbahasa Indonesia (Kartika dkk., 2023; Tobing dkk., 2025), dan bert-base-multilingual-cased (mBERT), model multibahasa yang dilatih pada banyak bahasa (Nabiilah dkk., 2024; Tobing dkk., 2025). Keduanya dimuat untuk klasifikasi sekuens, dengan kepala klasifikasi berinisialisasi acak berkeluaran empat yang diletakkan di atas representasi [CLS] dan dilatih bersama lapisan *encoder*.

Jurnal Manajemen dan Teknologi
https://journal.stmiki.ac.id
E-ISSN: 3047-8251 | P-ISSN: 3047-826X
Vol. 3 No. 2 (2026) | November
OPEN ACCESS
https://doi.org/ 10.63447/jmt.v3i2.2196

Perbedaan kedua model terletak pada korpus pelatihan awal beserta kosakata dan *tokenizer*-nya, sedangkan data, pembagian, dan pengaturan pelatihan identik. Kedua model dipilih karena berukuran sama (base, 12 lapisan) dan berarsitektur BERT, sehingga perbandingan terfokus pada pengaruh korpus pelatihan awal dan kosakata. Pelatihan menggunakan fungsi loss cross-entropy dan seluruh parameter encoder diperbarui bersama kepala klasifikasi (full fine-tuning). Pelatihan dijalankan dengan Hugging Face Trainer pada GPU Tesla T4 di Google Colab, memakai pengaturan pada Tabel 2 yang sebagian besar merupakan nilai bawaan pustaka, termasuk *optimizer* AdamW. Versi pustaka yang digunakan adalah Transformers 5.19.0, PyTorch 2.11.0+cu130, scikit-learn 1.9.1, dan datasets 5.1.0. Data validasi dievaluasi pada akhir setiap epoch, dan *checkpoint* dengan macro F1-score validasi tertinggi dimuat ulang dari penyimpanan untuk pengujian. Setiap model dilatih tiga kali dengan *seed* 1, 2, dan 3 selama empat epoch. *Seed* dikunci sebelum model dibuat sehingga inisialisasi kepala klasifikasi dan urutan data ikut terkendali, dan ketiga pelatihan bersifat independen.

Tabel 2. Konfigurasi fine-tuning yang dipakai kedua model

Pengaturan	Nilai
Learning rate	5×10^{-5}, peluruhan linear, tanpa warmup
Optimizer	AdamW, weight decay 0
Epoch	4 (2.144 langkah)
Batch size (latih / evaluasi)	16 / 32
Panjang sekuens maksimum	128
Pembagian data	80:10:10 bertingkat, random state 42
Random seed pelatihan	1, 2, 3 (tiga pelatihan independen per model)
Pemilihan model	Macro F1 validasi tertinggi, dimuat ulang sebelum pengujian

Pengaturan pada Tabel 2 tidak disesuaikan untuk salah satu model dan dipakai identik pada keduanya agar perbandingan adil. Sebagai pemeriksaan eksploratif dengan satu pelatihan per konfigurasi, variasi *learning rate* (2×10^{-5}, 3×10^{-5}, dan 5×10^{-5}) dan jumlah epoch (4 dan 6) dicoba pada IndoBERT dengan *seed* 42; macro F1-score uji berkisar 0,857–0,864. Pada mBERT, *learning rate* 2×10^{-5} menghasilkan 0,818 pada 4 epoch dan 0,842 pada 6 epoch, dibandingkan $0{,}826 \pm 0{,}009$ pada konfigurasi utama. Satu konfigurasi alternatif (*learning rate* 2×10^{-5}, *weight decay* 0,01, *warmup* 100 langkah, dan *dropout* 0,2) dicoba satu kali per model tanpa penguncian *seed* pada kepala klasifikasi, dan menghasilkan 0,856 untuk IndoBERT dan 0,801 untuk mBERT. Pelatihan dengan seed 42 tersebut dijalankan sebelum protokol final dan tidak diikutkan dalam agregat tiga seed utama (1, 2, dan 3). IndoBERT unggul pada keempat konfigurasi yang dicoba untuk kedua model (selisih 2,0–5,5 poin), sedangkan mBERT lebih sensitif terhadap pengaturan pelatihan. Pemeriksaan ini hanya bersifat eksploratif (satu pelatihan per konfigurasi, tanpa pengulangan seed), sehingga hasil utama penelitian tidak didasarkan padanya, dan pencarian *hyperparameter* yang sistematis menjadi keterbatasan.

2.3 Evaluasi

Evaluasi akhir dilakukan pada data uji yang terpisah dari proses pelatihan dan pemilihan model. Akurasi dihitung sebagai rasio prediksi yang tepat terhadap seluruh chat uji, seperti pada (1).

$$\text{Akurasi} = \frac{1}{N}\sum_{i=1}^{N} I(\hat{y}_i = y_i) \quad (1)$$

N merupakan jumlah chat uji, $\hat{y}_i$ kelas prediksi chat ke-i, y_i kelas sebenarnya, dan $I(\cdot)$ fungsi indikator yang bernilai satu bila prediksi benar dan nol bila salah. Untuk setiap kelas c, precision, recall, dan F1-score dihitung dengan (2), (3), dan (4). Pada (1), N adalah banyaknya chat uji, $\hat{y}i$ prediksi untuk chat ke-i, yi label sebenarnya, dan $I(\cdot)$ fungsi indikator yang bernilai 1 jika prediksi

Jurnal Manajemen dan Teknologi
https://journal.stmiki.ac.id
E-ISSN: 3047-8251 | P-ISSN: 3047-826X
Vol. 3 No. 2 (2026) | November
OPEN ACCESS
https://doi.org/ 10.63447/jmt.v3i2.2196

tepat dan 0 jika keliru. Metrik per kelas c, yaitu precision, recall, dan F1-score, dihitung melalui (2) sampai (4).

$$P_c = \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FP}_c} \quad (2)$$

$$R_c = \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FN}_c} \quad (3)$$

$$\mathrm{F1}_c = \frac{2P_cR_c}{P_c + R_c} \quad (4)$$

TP_c, FP_c, dan FN_c pada persamaan tersebut masing-masing adalah banyaknya *true positive*, *false positive*, dan *false negative* untuk kelas *c*, sedangkan P_c dan R_c ialah precision dan recall kelas itu. Karena sebaran keempat kelas belum sepenuhnya seimbang, skor tiap kelas dirata-ratakan secara sederhana tanpa bobot, seperti pada (5).

$$M_{\mathrm{makro}} = \frac{1}{K}\sum_{c=1}^{K} M_c \quad (5)$$

M adalah precision, recall, atau F1-score, M_c skor kelas *c*, dan $K = 4$ jumlah kelas. Di samping skor agregat, F1-score tiap kelas dan *confusion matrix* kedua model ditelaah untuk mengetahui kelas mana yang paling sering saling tertukar. Penelaahan ini memakai data uji yang sama, dan *confusion matrix* merupakan penjumlahan dari ketiga *seed* (3.213 chat per model). Untuk melihat manfaat praktisnya dalam moderasi, ketiga kelas toksik (*violence*, *racist*, dan *harassment*) digabung menjadi satu kelompok dan dikontraskan dengan *neutral*. Kelompok toksik dinilai melalui precision, recall, dan F1-score, sedangkan *neutral* dinilai melalui recall. Seluruh hasil dilaporkan sebagai rata-rata ± simpangan baku sampel dari tiga *seed* ($n = 3$); karena jumlah *seed* kecil, simpangan baku ini hanya perkiraan kasar variasi antar pelatihan. Contoh kesalahan klasifikasi ditampilkan dari sepuluh chat yang diambil secara acak dari prediksi IndoBERT (*seed* 1) dan disensor. Untuk menjajaki dugaan bahwa kosakata IndoBERT lebih cocok dengan bahasa informal, fragmentasi tokenizer diukur pada teks asli data uji tanpa pembersihan tambahan, yaitu rata-rata token per kata dengan kata dipisahkan spasi. Analisis ini bersifat eksploratif. Chat dibagi menjadi tiga kelompok sama besar menurut selisih fragmentasi mBERT terhadap IndoBERT, lalu selisih ketepatan per chat antar model (dirata-ratakan lintas tiga *seed*) dibandingkan antar kelompok dan dengan korelasi peringkat Spearman.

2.4 Pemeriksaan Duplikat dan Uji Signifikansi

Setelah evaluasi awal, dilakukan pemeriksaan duplikat. Teks chat dinormalisasi secara diperluas: huruf dijadikan kecil, penanda anonimisasi "user", sisa karakter kontrol (\n, \t) dan kode heksadesimal dihapus, angka dan tanda baca dibuang, huruf berulang lebih dari dua dipadatkan menjadi dua, dan spasi berlebih dihapus. Teks hasil normalisasi dibandingkan antar baris dan antar subset. Pemeriksaan menemukan bahwa 1.091 dari 10.702 baris (10,2%) memiliki teks yang muncul lebih dari sekali dan 26 teks muncul dengan label berbeda. Pada data uji, 92 chat memiliki teks yang sama dengan data latih, 22 dengan data validasi, dan 4 kembar dengan chat uji lain; setelah digabung (sebagian tumpang tindih), 106 chat uji disisihkan sehingga tersisa 965 chat uji bersih. Duplikat tidak dihapus dari data latih dan validasi agar alur tetap selaras dengan data yang dirilis. Sebagai analisis sensitivitas, kinerja dihitung ulang pada 965 chat uji bersih memakai prediksi yang sama tanpa pelatihan ulang, dengan metrik akurasi dan macro F1-score. Pendekatan ini menghilangkan kebocoran pada data uji, tetapi duplikat pada data latih dan validasi, termasuk teks berlabel berbeda, tetap ada dan dicatat sebagai keterbatasan. Hasil kedua model diperoleh dari chat uji yang identik, sehingga prediksinya berpasangan dan layak dibandingkan dengan uji McNemar. Untuk macro F1-score, yang dihitung atas seluruh data uji dan tidak dapat dinilai per chat, perbedaannya diestimasi dengan bootstrap. Karena tiga uji dijalankan per kumpulan data, ketahanan hasil terhadap koreksi Bonferroni ($\alpha = 0{,}05/3 \approx 0{,}017$) juga diperiksa: seed 1 dan 3 tetap signifikan, sedangkan seed 2 tidak. Perbedaan kinerja IndoBERT dan mBERT diuji secara berpasangan pada data uji yang sama, untuk setiap *seed* (IndoBERT dan mBERT dengan *seed* yang sama). Uji McNemar eksak membandingkan jumlah chat yang dijawab benar oleh satu model tetapi salah oleh model lainnya.

Jurnal Manajemen dan Teknologi
https://journal.stmiki.ac.id

E-ISSN: 3047-8251 | P-ISSN: 3047-826X
Vol. 3 No. 2 (2026) | November
OPEN ACCESS

https://doi.org/ 10.63447/jmt.v3i2.2196

Misalkan *b* adalah jumlah chat yang dijawab benar oleh IndoBERT tetapi salah oleh mBERT dan *c* kebalikannya, dengan $n = b + c$. Di bawah hipotesis nol bahwa kedua model memiliki tingkat kesalahan yang sama, nilai *p* dua sisi dihitung dengan (6).

$$p = \min\left(1, 2\sum_{k=0}^{\min(b,c)} C(n,k)\left(\frac{1}{2}\right)^{n}\right) \quad (6)$$

dengan C(*n*, *k*) koefisien binomial.

Selisih *macro* F1-score IndoBERT terhadap mBERT diperkirakan dengan *bootstrap* berpasangan dari 2.000 sampel ulang chat uji dengan pengembalian; interval kepercayaan 95% diambil dari persentil ke-2,5 dan ke-97,5 selisih tersebut, dan selisih dianggap berbeda dari nol bila interval tidak memuat nol. Bootstrap gabungan antar seed dihitung dengan mengambil sampel ulang chat uji yang sama untuk ketiga seed, lalu merata-ratakan selisih macro F1-score lintas seed; ringkasan ini hanya mencerminkan ketidakpastian akibat data uji, bukan variasi pelatihan. Taraf signifikansi yang digunakan adalah 0,05 untuk setiap *seed*. Kedua pengujian dijalankan pada data uji penuh dan data uji bersih. Pengujian yang dilakukan menilai ketidakpastian akibat ukuran data uji untuk pasangan model tertentu, sedangkan variasi yang diakibatkan pelatihan diwakili oleh tiga *seed*. Karena itu, kesimpulan didasarkan pada konsistensi arah selisih pada ketiga *seed* dan signifikansinya di tiap *seed*, bukan pada satu uji gabungan.

3. Hasil dan Pembahasan

3.1 Hasil

3.1.1 Kinerja Keseluruhan

Tabel 3 memuat kinerja kedua model pada data uji penuh (n = 1.071) sebagai rata-rata ± simpangan baku sampel dari tiga seed. IndoBERT unggul pada setiap metrik agregat, dengan selisih 2,81 poin pada akurasi dan 2,78 poin pada macro F1-score. IndoBERT juga lebih stabil antar *seed*: simpangan baku macro F1-score-nya 0,0019, dibandingkan 0,0090 pada mBERT.

Tabel 3. Kinerja pada data uji penuh (rata-rata ± simpangan baku, tiga seed, n = 1.071)

Model	Akurasi	Precision	Recall	Macro F1
IndoBERT	0,8519 ± 0,0022	0,8538 ± 0,0030	0,8554 ± 0,0016	0,8541 ± 0,0019
mBERT	0,8238 ± 0,0085	0,8262 ± 0,0098	0,8272 ± 0,0079	0,8263 ± 0,0090

Rata-rata macro F1-score validasi pada *checkpoint* terpilih adalah 0,859 untuk IndoBERT dan 0,829 untuk mBERT, sehingga skor uji sedikit di bawah skor validasi pada kedua model (0,5 dan 0,2 poin). *Checkpoint* terbaik adalah epoch ke-4 pada keenam pelatihan, artinya skor validasi tertinggi tercapai pada epoch terakhir dan anggaran empat epoch kemungkinan membatasi, terutama pada mBERT (Subbab 3.5). Signifikansi statistik selisih ini dibahas pada Subbab 3.4.

3.1.2 Kinerja per Kelas

IndoBERT mengungguli mBERT dalam F1-score, dan keunggulan itu tidak terpusat pada satu kelas saja, melainkan merata pada keempat kelas, seperti diperlihatkan pada Tabel 4. Selisih terbesar ada pada kelas *harassment* (3,8 poin), diikuti *neutral* (3,2), *racist* (2,2), dan *violence* (1,9). Selisih F1-score per kelas divisualisasikan pada Gambar 2. Kedua model paling baik mengenali *harassment* dan paling lemah pada *neutral*. IndoBERT juga lebih tinggi pada seluruh precision dan recall per kelas, tetapi selisih pada precision *violence* (0,3 poin), recall *neutral* (0,6 poin), dan precision *racist* (0,8 poin) masih berada dalam rentang simpangan baku antar *seed* sehingga tidak dapat ditafsirkan sebagai perbedaan nyata.

Jurnal Manajemen dan Teknologi
https://journal.stmiki.ac.id

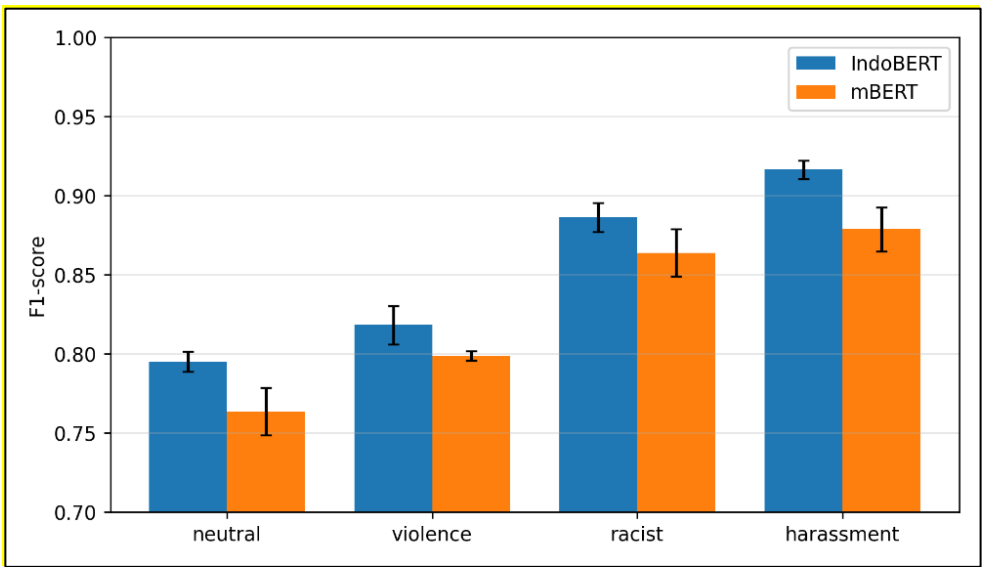
E-ISSN: 3047-8251 | P-ISSN: 3047-826X
Vol. 3 No. 2 (2026) | November

OPEN ACCESS

https://doi.org/ 10.63447/jmt.v3i2.2196

Tabel 4. Precision, recall, dan F1-score per kelas (rata-rata ± simpangan baku, tiga seed)

Kelas	Model	Precision	Recall	F1-score
Neutral	IndoBERT	0,8195 ± 0,0190	0,7729 ± 0,0229	0,7952 ± 0,0064
	mBERT	0,7609 ± 0,0260	0,7668 ± 0,0153	0,7636 ± 0,0149
Violence	IndoBERT	0,8178 ± 0,0068	0,8194 ± 0,0307	0,8183 ± 0,0121
	mBERT	0,8150 ± 0,0148	0,7837 ± 0,0196	0,7988 ± 0,0031
Racist	IndoBERT	0,8730 ± 0,0195	0,9004 ± 0,0120	0,8863 ± 0,0092
	mBERT	0,8650 ± 0,0287	0,8632 ± 0,0083	0,8639 ± 0,0150
Harassment	IndoBERT	0,9048 ± 0,0187	0,9288 ± 0,0116	0,9165 ± 0,0058
	mBERT	0,8637 ± 0,0326	0,8952 ± 0,0070	0,8788 ± 0,0138



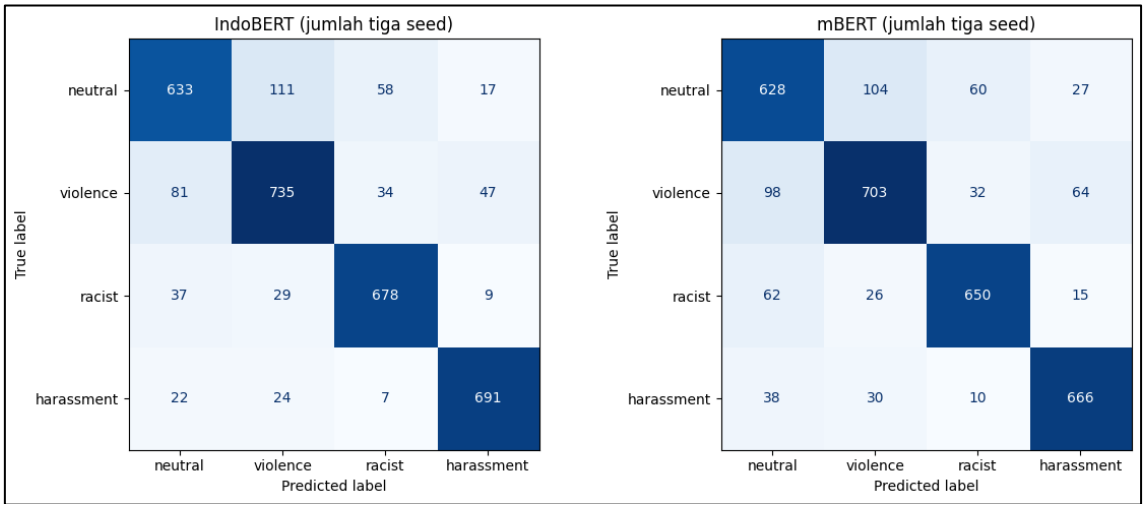
Gambar 2. F1-score per kelas (rata-rata ± simpangan baku, tiga seed; sumbu vertikal dimulai dari 0,7)

Pola precision dan recall berbeda antarkelas. Pada *neutral*, precision IndoBERT (0,820) lebih tinggi daripada recall-nya (0,773), sehingga sebagian chat *neutral* tidak dikenali meskipun chat yang diberi label *neutral* umumnya benar, sedangkan pada mBERT keduanya seimbang dan sama-sama lebih rendah (0,761 dan 0,767). Pada *violence*, precision dan recall IndoBERT seimbang (0,818 dan 0,819), sedangkan recall mBERT (0,784) lebih rendah daripada precision-nya (0,815). Pada *racist* dan *harassment*, recall IndoBERT lebih tinggi daripada precision-nya (0,900 dan 0,873 pada *racist*; 0,929 dan 0,905 pada *harassment*).

3.1.3 Analisis Kesalahan

Gambar 3 menunjukkan *confusion matrix* yang dijumlahkan dari ketiga *seed* (3.213 chat per model). IndoBERT membuat 476 kesalahan dan mBERT 566, dan pada kedua model sekitar 69% kesalahan melibatkan kelas *neutral* (326 dan 389). Kesalahan tersering pada kedua model adalah chat *neutral* yang diprediksi *violence* (111 pada IndoBERT dan 104 pada mBERT, sekitar 13% dari 819 chat *neutral*), disusul *violence* yang diprediksi *neutral* (81 dan 98).

Jurnal Manajemen dan Teknologi
https://journal.stmiki.ac.id
E-ISSN: 3047-8251 | P-ISSN: 3047-826X
Vol. 3 No. 2 (2026) | November
OPEN ACCESS
https://doi.org/ 10.63447/jmt.v3i2.2196



Gambar 3. Confusion matrix pada data uji, dijumlahkan dari tiga seed: (a) IndoBERT dan (b) mBERT

Perbedaan keduanya terutama terletak pada chat toksik yang lolos sebagai neutral: 140 pada IndoBERT dan 198 pada mBERT dari 2.394 chat toksik (5,8% dan 8,3%), dengan rincian *violence* 81 dan 98, *racist* 37 dan 62, serta *harassment* 22 dan 38. Sebaliknya, chat *neutral* yang ditandai sebagai toksik hampir sama banyak, yaitu 186 dan 191 (sekitar 23% dari 819 chat *neutral*). Bila ketiga kelas toksik digabung dan dikontraskan dengan *neutral*, kelompok toksik pada IndoBERT memperoleh precision 0,9238, recall 0,9415, dan F1-score 0,9326, sedangkan pada mBERT masing-masing 0,9200, 0,9173, dan 0,9186. Recall *neutral* berturut-turut 0,7729 dan 0,7668 untuk kedua model. Keunggulan IndoBERT dengan demikian terutama berupa berkurangnya chat toksik yang terlewat, bukan berkurangnya alarm palsu.

3.1.4 Signifikansi Statistik dan Pengaruh Data Duplikat

Sebanyak 106 chat uji yang berulang di data latih, validasi, atau uji disisihkan, sehingga evaluasi ulang dilakukan pada 965 chat uji bersih. Macro F1-score rata-rata IndoBERT turun dari 0,8541 menjadi 0,8463 dan mBERT dari 0,8263 menjadi 0,8165, sementara akurasi masing-masing turun dari 0,8519 menjadi 0,8446 dan dari 0,8238 menjadi 0,8145. Turunnya macro F1-score sebesar 0,8–1,0 poin menandakan bahwa duplikat memberi keuntungan kecil bagi kedua model, dengan mBERT sedikit lebih diuntungkan. Urutan kinerja tetap sama, dan jarak kedua model justru melebar tipis, dari 2,8 menjadi 3,0 poin. Hasil uji berpasangan untuk setiap *seed* dimuat pada Tabel 5. IndoBERT unggul pada ketiga *seed*, baik pada data uji penuh maupun bersih, dan selisihnya signifikan pada dua dari tiga *seed* (*seed* 1 dan 3; $p < 0,01$); pada *seed* 2 selisihnya positif tetapi tidak signifikan, dengan interval kepercayaan yang mencakup nol ($p = 0,0985$ pada data penuh; 0,1388 pada data bersih). Bootstrap gabungan antar *seed* memberi selisih rata-rata 0,0278 (CI 95% [0,0129; 0,0446]) pada data penuh dan 0,0298 ([0,0136; 0,0467]) pada data bersih; interval ini tidak memuat nol, tetapi hanya mencerminkan ketidakpastian akibat data uji, bukan variasi pelatihan. Sebagai informasi tambahan, pelatihan awal dengan seed 42 (di luar tiga seed utama) menghasilkan macro F1-score 0,8628 pada IndoBERT dan 0,8433 pada mBERT, selisih sekitar 2,0 poin. Selisih ini tidak signifikan pada uji McNemar ($p = 0,0686$), meskipun interval kepercayaan bootstrap-nya tidak memuat nol ([0,0007; 0,0394]). Dengan demikian, temuan bahwa IndoBERT lebih baik konsisten dalam arah pada seluruh pelatihan tetapi bukti statistiknya tidak seragam, sehingga sebaiknya dibaca sebagai kecenderungan yang kuat, bukan kepastian.

Jurnal Manajemen dan Teknologi
https://journal.stmiki.ac.id
E-ISSN: 3047-8251 | P-ISSN: 3047-826X
Vol. 3 No. 2 (2026) | November
OPEN ACCESS
https://doi.org/ 10.63447/jmt.v3i2.2196

Tabel 5. Uji berpasangan IndoBERT terhadap mBERT per seed (selisih = macro F1 IndoBERT − mBERT; data bersih: n = 965)

Seed	Data uji	Selisih macro F1	CI 95%	p McNemar
1	Penuh	+0,0345	[0,0135; 0,0571]	0,0029
2	Penuh	+0,0166	[−0,0040; 0,0357]	0,0985
3	Penuh	+0,0322	[0,0114; 0,0532]	0,0036
1	Bersih	+0,0368	[0,0136; 0,0610]	0,0025
2	Bersih	+0,0169	[−0,0042; 0,0388]	0,1388
3	Bersih	+0,0356	[0,0126; 0,0579]	0,0043
Gabungan	Penuh	+0,0278	[0,0129; 0,0446]	–
Gabungan	Bersih	+0,0298	[0,0136; 0,0467]	–

3.2 Pembahasan

Arah keunggulan IndoBERT sejalan dengan penelitian pada domain lain: Nabiilah dkk. (2023) melaporkan IndoBERT memberi hasil terbaik di antara IndoBERT, mBERT, dan IndoRoBERTa pada klasifikasi komentar toksik di media sosial, dan Tobing dkk. (2025) melaporkan IndoBERT lebih akurat daripada mBERT untuk deteksi hoaks politik. Perbandingan ini bersifat kualitatif karena tugas, jenis teks (komentar dan berita dibandingkan chat pendek), dan metrik berbeda, tetapi menunjukkan bahwa keunggulan encoder monolingual tidak terbatas pada satu domain. Penjelasan yang masuk akal adalah IndoBERT dilatih hanya pada teks berbahasa Indonesia, sedangkan mBERT membagi kosakata dan kapasitasnya ke banyak bahasa, sehingga mungkin kurang baik merepresentasikan kosakata informal, singkatan, dan slang. Selisih pada *neutral*, yang chat-nya banyak memuat bahasa sehari-hari, sejalan dengan penjelasan ini, tetapi selisih terbesar justru pada *harassment* dan selisih pada *violence* paling kecil, sehingga pola selisih antarkelas tidak sepenuhnya mendukungnya. Pemeriksaan tambahan menunjukkan bahwa mBERT memecah chat menjadi 2,13 token per kata dibandingkan 1,54 pada IndoBERT (sekitar 38% lebih banyak), meskipun chat dengan token [UNK] justru lebih banyak pada IndoBERT (9 dibandingkan 1). Selisih ketepatan per chat antar model naik dari 1,5 poin pada sepertiga chat dengan tambahan fragmentasi terendah menjadi 3,4–3,6 poin pada dua sepertiga lainnya, tetapi korelasi peringkatnya hampir nol ($\rho = 0{,}03$; $p = 0{,}41$). Hasil ini searah dengan dugaan bahwa kosakata yang lebih sesuai membantu IndoBERT, tetapi tidak membuktikannya, sehingga penjelasan tersebut tetap berupa hipotesis karena korpus pelatihan awal tidak divariasikan.

Selisihnya sekitar 2,8 poin macro F1-score dan konsisten pada ketiga *seed*, sedangkan mBERT juga lebih bervariasi antar *seed* (simpangan baku 0,0090 dibandingkan 0,0019). Pemeriksaan eksploratif pada beberapa konfigurasi (Subbab 2.3) menunjukkan bahwa IndoBERT unggul pada keempat konfigurasi yang dicoba untuk kedua model dengan selisih 2,0–5,5 poin, dan mBERT lebih sensitif terhadap pengaturan pelatihan. Karena *checkpoint* terbaik selalu epoch ke-4, sebagian selisih pada konfigurasi utama dapat berasal dari anggaran epoch yang belum cukup bagi mBERT, bukan semata kemampuan model; hal ini tidak diuji dengan epoch lebih panjang pada konfigurasi utama. Model multibahasa tetap dapat dipertimbangkan bila satu sistem harus menangani beberapa bahasa, meskipun hal itu tidak diuji di sini. Dari sudut pandang moderasi, IndoBERT meloloskan lebih sedikit chat toksik (5,8% dibandingkan 8,3%), tetapi kedua model menandai sekitar 23% chat *neutral* sebagai toksik, sehingga sistem otomatis ini tampaknya lebih cocok sebagai penyaring awal yang diikuti peninjauan manusia daripada pengambil keputusan tunggal. Sebagai ilustrasi, dari setiap 1.000 chat toksik IndoBERT meloloskan sekitar 58 dan mBERT sekitar 83, yang berarti lebih banyak konten merugikan sampai kepada pemain muda. Sebaliknya, dari setiap 1.000 chat neutral kedua model menandai sekitar 227–233 sebagai toksik, yang pada sistem otomatis penuh berisiko membungkam atau menghukum pemain secara keliru. Contoh acak chat *neutral* yang diprediksi *violence* menunjukkan bahwa sebagian memuat makian ringan dalam konteks bercanda atau mengutip sehingga batasnya dengan *violence* kabur, sedangkan sebagian lain tidak memuat petunjuk toksik dan merupakan kesalahan klasifikasi murni; sepuluh contoh ini bersifat ilustratif dan tidak mewakili seluruh kesalahan. Karena kelas *violence* mencakup makian, ambiguitas label ini dapat

E-ISSN: 3047-8251 | P-ISSN: 3047-826X
Vol. 3 No. 2 (2026) | November
OPEN ACCESS
https://doi.org/ 10.63447/jmt.v3i2.2196

menjadi salah satu penyebab sulitnya memisahkan kedua kelas. Skor penelitian ini lebih rendah daripada yang dilaporkan penyusun dataset, yaitu akurasi 94,30% pada modalitas verbal dengan model BERT–BiLSTM pada validasi silang lima lipatan (Prestiliano dkk., 2026); selisihnya sekitar 9,1 poin terhadap akurasi IndoBERT (0,8519). Kedua hasil tidak dapat dibandingkan langsung karena protokolnya berbeda (satu pembagian data di sini, lima lipatan di sana) dan arsitekturnya berbeda (kepala linear di sini, BiLSTM tambahan di sana). Dataset yang sama memuat 10,2% baris berteks duplikat; pada validasi silang, duplikat dapat jatuh di lipatan latih dan uji sekaligus dan menaikkan skor, tetapi pengaruhnya pada penelitian ini hanya 0,8–1,0 poin sehingga tidak menjelaskan selisih sekitar 9 poin itu. Mengevaluasi IndoBERT dan mBERT dengan protokol validasi silang yang sama merupakan langkah lanjutan yang jelas. Selain itu, skor 94,30% tersebut merupakan rata-rata lima lipatan yang bervariasi lebar, dari 81,83% sampai 98,79% (Prestiliano dkk., 2026). Beberapa keterbatasan perlu dicatat. Setiap model hanya dilatih tiga kali, sehingga simpangan baku bersifat kasar dan satu dari tiga uji tidak signifikan. Duplikat tidak dihapus dari data latih dan validasi (10,2% baris, dengan 26 teks berlabel berbeda), konfigurasi alternatif hanya dicoba satu kali, dan mBERT tidak dituning serta mencapai skor validasi tertinggi pada epoch terakhir. Label tidak disertai kesepakatan antaranotator yang dilaporkan, dan data hanya berasal dari dua gim, sehingga generalisasi ke platform lain belum terverifikasi. Terakhir, penjelasan tentang korpus pelatihan awal tidak diuji secara langsung.

4. Kesimpulan

Penelitian ini membandingkan IndoBERT dan mBERT untuk klasifikasi toksisitas empat kelas pada chat Roblox dan Minecraft berbahasa Indonesia dengan pengaturan *fine-tuning* yang identik dan tiga *seed*. Pada dataset dan pengaturan pengujian ini, IndoBERT cenderung lebih baik daripada mBERT untuk klasifikasi toksisitas chat gim berbahasa Indonesia: macro F1-score rata-rata 0,8541 dibandingkan 0,8263, searah pada ketiga seed dan signifikan pada dua di antaranya, dengan urutan yang sama pada data uji tanpa duplikat. IndoBERT juga meloloskan lebih sedikit chat toksik sebagai neutral (5,8% dibandingkan 8,3%), tetapi kedua model masih salah menandai sekitar 23% chat neutral sebagai toksik, sehingga paling sesuai dipakai sebagai penyaring awal dengan peninjauan manusia. Kesimpulan ini terbatas pada satu dataset dengan label tanpa pengukuran kesepakatan antarpelabel dan pengaturan pelatihan yang belum dioptimalkan secara sistematis. Penelitian selanjutnya perlu melatih ulang tanpa duplikat, menguji kesepakatan antarpelabel, mencari hyperparameter secara sistematis dengan lebih banyak seed, dan mengevaluasi gim lain.

5. Ucapan Terima Kasih

Penelitian ini tidak akan selesai tanpa bantuan baik emosional maupun material dari banyak pihak. Penulis berterima kasih kepada Prestiliano dkk. atas ketersediaan Indonesian Chat Dataset yang terbuka untuk umum. Ucapan terima kasih juga ditujukan kepada para dosen pembimbing atas arahannya, kepada keluarga atas doa dan dukungannya, serta kepada rekan-rekan seperjuangan yang selalu membantu dan menyemangati hingga naskah ini rampung. Semoga seluruh jalan kita semua selalu diiringi doa-doa ke mana pun dan di mana pun dalam pengembaraan dunia fana ini.

6. Daftar Pustaka

Dharmawan, S., Mawardi, V. C., & Perdana, N. J. (2023). Klasifikasi ujaran kebencian menggunakan metode feedforward neural network (IndoBERT). *Jurnal Ilmu Komputer dan Sistem Informasi, 11*(1). https://doi.org/10.24912/jiksi.v11i1.24066.

E-ISSN: 3047-8251 | P-ISSN: 3047-826X
Vol. 3 No. 2 (2026) | November
OPEN ACCESS
https://doi.org/ 10.63447/jmt.v3i2.2196

Ghosh, K., & Senapati, A. (2025). Hate speech detection in low-resourced Indian languages: An analysis of transformer-based monolingual and multilingual models with cross-lingual experiments. *Natural Language Processing, 31*(2), 393–414. https://doi.org/10.1017/nlp.2024.28.

Hakim, A. N., Sibaroni, Y., & Prasetyowati, S. S. (2024). Detection of hate-speech text on Indonesian Twitter social media using IndoBERTweet-BiLSTM-CNN. In *2024 12th International Conference on Information and Communication Technology (ICoICT)* (pp. 374–381). IEEE. https://doi.org/10.1109/ICoICT61617.2024.10698615.

Hibatullah, H., Ballı, T., & Yetkin, E. F. (2025). Verbal harassment detection in online games using machine learning methods. *Entertainment Computing*, Article 101009. https://doi.org/10.1016/j.entcom.2025.101009.

Ibrohim, M. O., & Budi, I. (2023). Hate speech and abusive language detection in Indonesian social media: Progress and challenges. *Heliyon, 9*(8), Article e18647. https://doi.org/10.1016/j.heliyon.2023.e18647.

Isa, S. M., Nico, G., & Permana, M. (2022). IndoBERT for Indonesian fake news detection. *ICIC Express Letters, 16*(3), 289–297. https://doi.org/10.24507/icicel.16.03.289.

Kartika, B. V., Martin, J. A., & Kusuma, G. P. (2023). Fine-tuned IndoBERT based model and data augmentation for Indonesian language paraphrase identification. *Revue d'Intelligence Artificielle, 37*(3), 733. https://doi.org/10.18280/ria.370322.

Kordyaka, B., Karaosmanoglu, S., & Laato, S. (2025). Defining toxicity in multiplayer online games: A systematic literature review. *Computers in Human Behavior Reports*, Article 100698. https://doi.org/10.1016/j.chbr.2025.100698.

Min, B., Ross, H., Sulem, E., Veyseh, A. P. B., Nguyen, T. H., Sainz, O., Agirre, E., Heintz, I., & Roth, D. (2023). Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys, 56*(2), 1–40. https://doi.org/10.1145/3605943.

Nabiilah, G. Z., Alam, I. N., Purwanto, E. S., & Hidayat, M. F. (2024). Indonesian multilabel classification using IndoBERT embedding and MBERT classification. *International Journal of Electrical and Computer Engineering, 14*(1), 1071–1078. https://doi.org/10.11591/ijece.v14i1.pp1071-1078.

Nabiilah, G. Z., Prasetyo, S. Y., Izdihar, Z. N., & Girsang, A. S. (2023). BERT base model for toxic comment analysis on Indonesian social media. *Procedia Computer Science, 216*, 714–721. https://doi.org/10.1016/j.procs.2022.12.188.

Pamungkas, E., Putri, D., & Fatmawati, A. (2023). Hate speech detection in Bahasa Indonesia: Challenges and opportunities. *International Journal of Advanced Computer Science and Applications, 14*(6), 1175–1181. https://doi.org/10.14569/IJACSA.2023.01406125.

Prestiliano, J., Azhari, A., & Nurwidyantoro, A. (2026). Multimodal deep learning violence detector for child-friendly online game. *International Journal of Serious Games, 13*(1), 41–61. https://doi.org/10.17083/9k96e890.

E-ISSN: 3047-8251 | P-ISSN: 3047-826X
Vol. 3 No. 2 (2026) | November
OPEN ACCESS
https://doi.org/ 10.63447/jmt.v3i2.2196

Rahmawati, A., Alamsyah, A., & Romadhony, A. (2022). Hoax news detection analysis using IndoBERT deep learning methodology. In *2022 10th International Conference on Information and Communication Technology (ICoICT)* (pp. 368–373). IEEE. https://doi.org/10.1109/ICoICT55009.2022.9914902.

Suharjanto, D., & Sumarsono, S. (2025). Optimalisasi DoRA untuk deteksi ujaran kebencian berbahasa Indonesia berbasis transformer. *Jurnal Nasional Teknologi dan Sistem Informasi, 11*(3), 341–349. https://doi.org/10.25077/TEKNOSI.v11i3.2025.341-349.

Susanto, L., et al. (2024). *IndoToxic2024: A demographically-enriched dataset of hate speech and toxicity types for Indonesian language*. arXiv preprint arXiv:2406.19349. https://doi.org/10.48550/arXiv.2406.19349.

Tobing, C. J. L., Wijayakusuma, I. G. N. L., & Harini, L. P. I. (2025). Perbandingan kinerja IndoBERT dan mBERT untuk deteksi berita hoaks politik dalam bahasa Indonesia. *JST (Jurnal Sains dan Teknologi), 14*(1), 114–123. https://doi.org/10.23887/jstundiksha.v14i1.92126.

Wijkstra, M., Rogers, K., Mandryk, R. L., Veltkamp, R. C., & Frommel, J. (2024). How to tame a toxic player? A systematic literature review on intervention systems for toxic behaviors in online video games. *Proceedings of the ACM on Human-Computer Interaction, 8*(CHI PLAY), 1–32. https://doi.org/10.1145/3677080.

Yang, Z., Grenon-Godbout, N., & Rabbany, R. (2023). Towards detecting contextual real-time toxicity for in-game chat. In *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 9894–9906). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.663.

Yanti, O. D., Alfariani, S., Celesta, S. N. M., Tania, K. D., & Rifai, A. (2026). Comparative analysis of the performance of machine learning methods and text embedding techniques in classifying toxic conversations in the Roblox game. *Journal of Applied Informatics and Computing, 10*(3), 2775–2789. https://doi.org/10.30871/jaic.v10i3.12646.