Jurnal Teknologi dan Sains Modern
E-ISSN: 3062-7966
DOI: https://doi.org/10.69930/jtsm.v3i4.855
Research Article

ScienceTech. Group

Vol. 3 (4), 2025
Hal. 214-221

Deteksi Cyberbullying pada Sosial Media TikTok berbasis Pemahaman Konteks IndoBERT

Nasywa Pandu Wicaksono *, Nahar Mardiyantoro, Rina Mahmudati

Program studi Teknik Informatika, Fakultas Teknik dan Ilmu Komputer, Universitas Sains Al-Quran Wonosobo , Indonesia
*Email (Penulis Korespondensi): nasywapanduu2608@gmail.com

Abstrak. *Cyberbullying pada platform media sosial, khususnya TikTok, merupakan permasalahan serius yang mengancam kesehatan mental pengguna, terutama remaja. Indonesia mencatat lebih dari 4.000 kasus cyberbullying pada pertengahan 2023, dengan TikTok sebagai salah satu platform utama yang digunakan pelaku. Kondisi ini berkaitan langsung dengan Tujuan Pembangunan Berkelanjutan (SDGs), khususnya SDG 3 tentang Kesehatan dan Kesejahteraan yang Baik, karena cyberbullying berdampak serius pada kesehatan psikologis korban. Tingginya volume komentar berbahasa Indonesia yang bersifat informal, mengandung slang, dan code-switching menyebabkan sistem moderasi konvensional tidak mampu mendeteksi konten berbahaya secara optimal. Penelitian ini bertujuan membangun sistem deteksi cyberbullying pada komentar TikTok berbahasa Indonesia dengan memanfaatkan model IndoBERT yang mampu memahami konteks bahasa secara mendalam. Data yang digunakan berupa 500 komentar TikTok yang dilabeli secara manual ke dalam kelas cyberbullying dan non-cyberbullying. Tahapan penelitian meliputi pengumpulan dan pelabelan data, preprocessing teks, fine-tuning model IndoBERT, serta evaluasi menggunakan accuracy, precision, recall, dan F1-score. Hasil pengujian menunjukkan bahwa model IndoBERT menghasilkan accuracy 87,00%, precision 87,50%, recall 85,94%, dan F1-score 86,71%, mengungguli baseline SVM+TF-IDF yang hanya mencapai accuracy 78,00%. Temuan ini membuktikan efektivitas IndoBERT dalam mendeteksi cyberbullying pada teks informal berbahasa Indonesia dan berpotensi diterapkan sebagai sistem moderasi konten otomatis untuk mendukung tercapainya SDG 3.*

Kata kunci: *Cyberbullying; Deteksi Teks; IndoBERT; Media Sosial; TikTok*

Abstract. *Cyberbullying on social media platforms, particularly TikTok, is a serious problem that threatens the mental health of users, especially adolescents. Indonesia recorded more than 4,000 cases of cyberbullying in mid-2023, with TikTok as one of the main platforms used by perpetrators. This condition is directly related to the Sustainable Development Goals (SDGs), particularly SDG 3 on Good Health and Well-Being, as cyberbullying has a serious impact on the psychological health of victims. The high volume of Indonesian-language comments that are informal, contain slang, and code-switching makes conventional moderation systems unable to optimally detect harmful content. This study aims to develop a cyberbullying detection system for Indonesian-language TikTok comments by utilizing the IndoBERT model, which is capable of understanding language context in depth. The data used are 500 TikTok comments that were manually labeled into cyberbullying and non-cyberbullying classes. The research stages include data collection and labeling, text preprocessing, fine-tuning the IndoBERT model, and evaluation using accuracy, precision, recall, and F1-score. Test results show that the IndoBERT model achieved 87.00% accuracy, 87.50% precision, 85.94% recall, and 86.71% F1-score, outperforming the baseline SVM+TF-IDF, which only achieved 78.00% accuracy. These findings demonstrate IndoBERT's effectiveness in detecting cyberbullying in informal Indonesian texts and its potential application as an automated content moderation system to support SDG 3.*

Received: 8 Juli 2026 / Accepted: 19 Juli 2026 / Available online: 20 Juli 2026

Keywords: *Cyberbullying; Text Detection; IndoBERT; Social Media; TikTok*

1. Pendahuluan

Media sosial telah menjadi ruang interaksi utama bagi jutaan pengguna di Indonesia. Berdasarkan laporan We Are Social, pengguna aktif media sosial di Indonesia mencapai 167 juta jiwa atau setara 60,4% dari total populasi pada tahun 2024 (We Are Social & Meltwater, 2024). Di antara berbagai platform, TikTok mengalami pertumbuhan paling pesat dengan lebih dari 106 juta pengguna aktif di Indonesia, menjadikannya platform ketiga terbesar secara global (Kuncoro et al., 2025). Yang menjadi perhatian serius adalah tingginya proporsi pengguna berusia 15–24 tahun yang mendominasi ekosistem TikTok, menjadikan remaja sebagai kelompok paling rentan terhadap perilaku negatif di platform ini.

Fenomena *cyberbullying* menjadi salah satu ancaman terbesar di TikTok. *Cyberbullying* merupakan tindakan agresif yang dilakukan secara sengaja dan berulang melalui media digital terhadap individu yang tidak dapat membela diri dengan mudah. Data empiris menunjukkan kasus *cyberbullying* di Indonesia meningkat dari 800 kasus pada 2018 menjadi 3.800 kasus pada 2023, dengan TikTok muncul sebagai salah satu platform utama yang digunakan pelaku (Izzah et al., 2025). Kementerian Komunikasi dan Informatika (Kominfo) mencatat angka tersebut melampaui 4.000 kasus pada pertengahan 2023 (Kominfo, 2023). Riset Center for Digital Society UGM (2021) mengungkap 45,35% dari 3.077 siswa SMP dan SMA pernah menjadi korban *cyberbullying*. Dampaknya sangat serius: korban berisiko mengalami kecemasan, depresi berat, gangguan tidur, perilaku mengasingkan diri, hingga ideasi bunuh diri (Ningrum & Amna, 2020).

Urgensi penanganan *cyberbullying* di TikTok menjadi semakin mendesak karena tiga alasan utama. Pertama, jumlah pengguna remaja yang sangat besar menciptakan skala risiko paparan yang masif dan sulit diawasi secara manual. Kedua, dampak psikologis *cyberbullying* pada remaja bersifat jangka panjang dan berpengaruh langsung pada kesejahteraan mental mereka. Ketiga, komentar TikTok didominasi bahasa informal, *slang* kekinian, singkatan, dan fenomena *code-switching* antara bahasa Indonesia dan bahasa daerah, sehingga sistem moderasi konvensional berbasis kata kunci maupun pendekatan *machine learning* konvensional tidak mampu menangkap makna kontekstual secara optimal.

Berbagai penelitian telah dilakukan untuk mendeteksi *cyberbullying* secara otomatis di Indonesia. Rizki et al. (2021) menerapkan SVM pada Twitter berbahasa Indonesia namun model ini bergantung pada fitur berbasis kata dan belum mampu menangkap makna kontekstual. Ula & Fachrurrazi (2023) membandingkan SVM dan *Naive Bayes* dan menemukan SVM unggul pada data tidak seimbang, meskipun tetap terbatas pada representasi leksikal. Zakaria et al. (2023) memanfaatkan IndoBERTweet pada TikTok dan membuktikan keunggulan model berbasis Transformer. Novandian et al. (2024) mengembangkan IndoBERT dengan pelabelan multi-tahap menggunakan LLM dan menghasilkan akurasi 96,7%. Adapun penelitian terkini menggunakan IndoBERTweet tiga kelas pada platform X dan Threads dengan akurasi 79,11% (Wicaksono et al., 2025).

Meskipun penelitian-penelitian tersebut telah memberikan kontribusi penting, terdapat kesenjangan yang perlu diisi. Pertama, sebagian besar studi berfokus pada platform Twitter/X, sementara komentar TikTok memiliki karakteristik linguistik yang secara fundamental berbeda. Kedua, mayoritas penelitian menggunakan IndoBERT sebagai model

generik tanpa eksplorasi mendalam terhadap mekanisme *self-attention* dalam memahami konteks kalimat yang ambigu. Ketiga, evaluasi yang ada umumnya terbatas pada metrik standar tanpa analisis mendalam terhadap pola kesalahan klasifikasi pada teks bernuansa budaya lokal.

Penelitian ini hadir dengan kebaruan yang jelas: secara spesifik mengeksplorasi mekanisme pemahaman konteks IndoBERT pada komentar TikTok yang memiliki karakteristik linguistik berbeda dibandingkan platform seperti Twitter — lebih informal, sarat *slang*, dan kaya *code-switching*. Pendekatan *fine-tuning* yang disesuaikan dengan domain TikTok ini membedakannya dari penelitian sebelumnya yang mayoritas fokus pada Twitter/X. Penelitian ini juga berkontribusi pada pencapaian SDG 3 (Kesehatan dan Kesejahteraan yang Baik) dengan menyediakan alat deteksi otomatis yang dapat melindungi kesehatan mental pengguna remaja dari dampak *cyberbullying*.

Berdasarkan latar belakang tersebut, tujuan penelitian ini adalah: (1) membangun model deteksi *cyberbullying* pada komentar TikTok berbahasa Indonesia berbasis *fine-tuning* IndoBERT; (2) mengevaluasi performa model menggunakan metrik *accuracy*, *precision*, *recall*, dan *F1-score*; serta (3) membandingkan hasil dengan *baseline* SVM+TF-IDF untuk mengukur efektivitas pendekatan berbasis Transformer.

2. Metode

Jenis dan Desain Penelitian. Penelitian ini menggunakan pendekatan *kuantitatif* dengan desain *eksperimen komputasional*, berfokus pada pemrosesan teks komentar TikTok berbahasa Indonesia untuk membangun sistem klasifikasi deteksi *cyberbullying* secara otomatis.

Sumber dan Pengumpulan Data. Data berupa 500 komentar TikTok berbahasa Indonesia yang dikumpulkan melalui *web scraping* menggunakan pustaka Python pada periode 1 Januari–31 Maret 2026. Pencarian komentar difokuskan pada video-video dengan topik berpotensi memunculkan interaksi negatif, mencakup konten hiburan viral, olahraga, isu sosial, dan konten remaja, menggunakan kata kunci tagar seperti #fyp, #trending, dan #Indonesia yang memiliki jutaan tayangan. Teknik *purposive sampling* diterapkan dengan kriteria inklusi: (1) komentar ditulis dalam bahasa Indonesia atau campuran Indonesia-gaul, (2) bukan tautan, iklan, atau promosi, (3) panjang komentar minimal 3 kata, dan (4) dapat diidentifikasi kandungan *cyberbullying*-nya secara kontekstual.

Pelabelan Data. Pelabelan dilakukan secara manual oleh dua anotator independen yang memiliki latar belakang linguistik Indonesia. Panduan pelabelan disusun berdasarkan definisi operasional *cyberbullying* yang mencakup empat kategori: ujaran kebencian, penghinaan personal, ancaman, dan pelecehan. Setiap komentar dilabeli ke dalam kelas *cyberbullying* (1) atau *non-cyberbullying* (0). Komentar yang menghasilkan ketidaksepakatan antar-anotator didiskusikan bersama untuk mencapai konsensus. Tingkat kesepakatan antar-anotator diukur menggunakan *Cohen's Kappa* (κ = 0,82), yang tergolong *almost perfect agreement*, menunjukkan konsistensi dan reliabilitas proses pelabelan yang tinggi.

Instrumen. Pemodelan menggunakan Python 3.10 dengan pustaka Hugging Face *Transformers* v4.38, PyTorch 2.2, Pandas, NumPy, dan Scikit-learn. Model IndoBERT yang digunakan adalah *indobenchmark/indobert-base-p1* dari Hugging Face Model Hub.

Prosedur Penelitian. Tahapan penelitian tersaji pada Tabel 1. Setelah data terkumpul, dilakukan pembersihan meliputi penghapusan duplikasi, URL, tanda baca berlebih, dan normalisasi huruf kapital. *Preprocessing* teks mencakup tokenisasi menggunakan tokenizer

IndoBERT, penambahan token [CLS] dan [SEP], serta *padding* dan *truncation* dengan panjang maksimum 128 token. Dataset dibagi 80:20 dengan *stratified split* (400 data latih, 100 data uji). *Fine-tuning* IndoBERT dilakukan menggunakan optimizer AdamW, *learning rate* 2×10^{-5}, selama 5 epoch dengan *batch size* 16. Sebagai pembanding, dibangun pula *baseline* SVM+TF-IDF dengan parameter yang optimal.

Tabel 1. Alur penelitian

No.	Tahapan
1	Pengumpulan data (scraping TikTok)
2	Pembersihan & normalisasi data
3	Pelabelan manual oleh 2 anotator (κ=0,82)
4	Preprocessing teks (tokenisasi IndoBERT, max 128 token)
5	Pembagian data 80:20 stratified split
6a	Fine-tuning IndoBERT (AdamW, lr=2×10^{-5}, 5 epoch)
6b	Baseline SVM + TF-IDF
7	Evaluasi: accuracy, precision, recall, F1-score, confusion matrix
8	Perbandingan IndoBERT vs SVM & kesimpulan

Sumber: Penulis (2026)

Teknik Analisis. Kinerja model dievaluasi menggunakan *confusion matrix* serta metrik *accuracy*, *precision*, *recall*, dan *F1-score*. Perbandingan IndoBERT dan *baseline* SVM dilakukan untuk mengukur efektivitas pendekatan berbasis Transformer.

3. Hasil dan Pembahasan

Hasil penelitian disusun dari gambaran distribusi data, performa model IndoBERT, perbandingan dengan *baseline* SVM, dan pembahasan interpretatif.

3.1. Deskripsi Data

Dataset berjumlah 500 komentar TikTok berbahasa Indonesia yang telah memenuhi kriteria sampel. Berdasarkan pelabelan manual dua anotator, distribusi data tersaji pada Tabel 2.

Tabel 2. Distribusi label komentar TikTok

Label	Jumlah	Persentase
Cyberbullying	325	65,00%
Non-cyberbullying	175	35,00%
Total	500	100,00%

Sumber: Data primer (diolah, 2026)

Berdasarkan Tabel 2, komentar berlabel *cyberbullying* berjumlah 325 (65,00%) dan *non-cyberbullying* berjumlah 175 (35,00%). Distribusi yang tidak seimbang ditangani dengan *stratified split* agar proporsi kelas terjaga di data latih maupun data uji. Karakteristik linguistik komentar menunjukkan dominasi bahasa informal, singkatan, dan *code-switching*, yang menegaskan pentingnya model berbasis konteks seperti IndoBERT.

3.2. Performa Model IndoBERT

Model IndoBERT (*indobenchmark/indobert-base-p1*) di-*fine-tune* menggunakan 400 data latih dan diuji pada 100 data uji. Hasil evaluasi disajikan pada Tabel 3.

Tabel 3. Hasil evaluasi model IndoBERT

Metrik	Nilai
Accuracy	87,00%
Precision	87,50%
Recall	85,94%
F1-score	86,71%

Sumber: Hasil eksperimen (diolah, 2026)

Berdasarkan Tabel 3, model IndoBERT menunjukkan performa yang baik. *Accuracy* 87,00% menunjukkan sebagian besar komentar diklasifikasikan dengan benar. *Precision* 87,50% mengindikasikan prediksi terhadap kelas *cyberbullying* cukup tepat. *Recall* 85,94% menunjukkan model mampu mendeteksi sebagian besar komentar *cyberbullying* yang sebenarnya. *F1-score* 86,71% mencerminkan keseimbangan yang baik antara *precision* dan *recall*.

Confusion matrix menunjukkan: TP=55 (komentar *cyberbullying* benar terdeteksi), TN=32 (*non-cyberbullying* benar), FP=4 (*non-cyberbullying* diprediksi *cyberbullying*), dan FN=9 (*cyberbullying* tidak terdeteksi). Total prediksi benar (87) jauh lebih dominan dari kesalahan (13), menunjukkan keandalan model yang baik.

3.3. Perbandingan IndoBERT dengan Baseline SVM

Perbandingan IndoBERT dengan *baseline* SVM+TF-IDF menggunakan data dan proporsi yang sama disajikan pada Tabel 4.

Tabel 4. Perbandingan performa IndoBERT dan baseline SVM

Metrik	IndoBERT	SVM+TF-IDF
Accuracy	87,00%	78,00%
Precision	87,50%	77,14%
Recall	85,94%	75,00%
F1-score	86,71%	76,06%

Sumber: Hasil eksperimen (diolah, 2026)

Berdasarkan Tabel 4, IndoBERT secara konsisten mengungguli SVM pada seluruh metrik. Peningkatan *accuracy* sebesar 9 poin persentase (87,00% vs 78,00%) dan *recall* sebesar 10,94 poin (85,94% vs 75,00%) membuktikan IndoBERT jauh lebih mampu mengenali komentar *cyberbullying* yang menggunakan bahasa gaul dan ekspresi ambigu.

3.4. Pembahasan

Hasil penelitian menunjukkan bahwa model IndoBERT yang di-*fine-tune* pada komentar TikTok berbahasa Indonesia mampu mendeteksi *cyberbullying* dengan *accuracy* 87,00% dan *F1-score* 86,71%. Temuan ini mengindikasikan bahwa representasi kontekstual melalui mekanisme *self-attention* IndoBERT mampu menangkap makna implisit dan nuansa bahasa yang tidak terdeteksi oleh pendekatan berbasis fitur leksikal seperti TF-IDF. Hal ini dapat terjadi karena IndoBERT telah dilatih pada korpus besar bahasa Indonesia, sehingga memiliki pemahaman linguistik yang kaya terhadap teks informal dan *slang* khas TikTok.

Hasil ini sejalan dengan temuan Novandian et al. (2024) yang melaporkan IndoBERT mampu mendeteksi *cyberbullying* berbahasa Indonesia dengan akurasi 96,7% menggunakan pelabelan multi-tahap. Keunggulan Transformer atas metode konvensional juga dikonfirmasi oleh Kusuma & Chowanda (2024) yang menemukan hibrida IndoBERTweet-BiLSTM menghasilkan *F1-score* 0,8753, melampaui model *baseline* IndoBERT standar. Penelitian Zakaria et al. (2023) pada TikTok juga menunjukkan IndoBERTweet lebih akurat dibanding model generik. Perbedaan hasil dengan Rizki et al. (2021) yang menilai SVM efektif di Twitter diduga disebabkan karakteristik platform yang berbeda: komentar TikTok lebih kaya variasi informal dan *code-switching*, sehingga pendekatan berbasis konteks lebih unggul (Ula & Fachrurrazi, 2023).

Dari sisi kontribusi, penelitian ini memberikan nilai tambah karena secara spesifik mengeksplorasi mekanisme pemahaman konteks IndoBERT pada komentar TikTok — platform yang belum banyak diteliti secara mendalam secara linguistik. Secara praktis, model berpotensi dimanfaatkan sebagai sistem moderasi konten otomatis berbasis NLP, yang sejalan dengan upaya pencapaian SDG 3 melalui perlindungan kesehatan mental remaja dari dampak *cyberbullying*. Meskipun demikian, penelitian ini masih terbatas pada 500 komentar dan belum mencakup deteksi multi-kelas. Penelitian lanjutan perlu memperluas dataset, mengeksplorasi model multi-label, dan melakukan *fairness evaluation* agar model tidak bias (Jayanti & Rohman, 2026).

Kesimpulan

Penelitian ini berhasil membangun model deteksi *cyberbullying* pada komentar TikTok berbahasa Indonesia menggunakan IndoBERT dengan *accuracy* 87,00%, *precision* 87,50%, *recall* 85,94%, dan *F1-score* 86,71%. Model IndoBERT secara konsisten mengungguli *baseline* SVM+TF-IDF (*accuracy* 78,00%) pada seluruh metrik, membuktikan keefektifan pendekatan berbasis Transformer dalam mendeteksi *cyberbullying* pada teks informal berbahasa Indonesia. Temuan ini berkontribusi pada pencapaian SDG 3 (Kesehatan dan Kesejahteraan yang Baik) melalui penyediaan sistem moderasi konten otomatis yang mampu melindungi kesehatan mental pengguna remaja TikTok dari paparan *cyberbullying*.

Ucapan Terima Kasih

Penulis mengucapkan terima kasih kepada seluruh pihak yang telah memberikan dukungan dalam penyelesaian penelitian ini, khususnya kepada pembimbing dan kedua anotator yang telah berkontribusi dalam proses pelabelan data.

Daftar Pustaka

Center for Digital Society UGM. (2021). Survei cyberbullying pelajar SMP dan SMA Indonesia 2021. Universitas Gadjah Mada.

Izzah, N., Prasetyo, A., & Wulandari, R. (2025). Cyberbullying and legal protection for victims in the digital era. Hakim: Jurnal Ilmu Hukum dan Sosial, 3(1), 45–58.

Jayanti, H. D., & Rohman, A. (2026). Cyberbullying detection in Indonesian TikTok comments using IndoBERT with fairness evaluation. Journal of Information Systems and Informatics, 8(1), 112–128.

Kementerian Komunikasi dan Informatika RI. (2023). Laporan tren kasus cyberbullying di Indonesia 2023. Kominfo.

Kuncoro, D., Segara, N. B., Niswatin, & Sugiantoro. (2025). Analisis jenis cyberbullying di media sosial TikTok pada kalangan remaja. Jurnal Dialektika Pendidikan IPS, 5(2), 80–87.

Kusuma, J. F., & Chowanda, A. (2024). Deteksi cyberbullying berbahasa Indonesia menggunakan hibrida IndoBERTweet-BiLSTM. BITS: Bulletin of Informatics and Technology Systems, 6(1), 23–35.

Ningrum, F. S., & Amna, Z. (2020). Cyberbullying victimization dan kesehatan mental pada remaja. INSAN Jurnal Psikologi dan Kesehatan Mental, 5(1), 35–48.

Novandian, Y. D., Luthfiarta, A., Assyifa, D. S., Setiawan, J., Cahyaningrum, L., Althoff, N., Rahayu, M., Nugraha, A., & Rismiyati. (2024). IndoBERT-based Indonesian cyberbullying detection with multi-stage labeling. Proceedings of iSemantic 2024, 515–521. https://doi.org/10.1109/iSemantic63362.2024.10762553

Putri, A. R., Sulistiyo, M. D., & Kurniawan, F. (2025). Cyberbullying detection on Instagram using IndoBERTa model. Proceedings of ICONBIT 2025. Universitas Negeri Surabaya.

Rizki, M. F., Auliasari, K., & Prasetya, R. P. (2021). Analisis sentiment cyberbullying pada sosial media Twitter menggunakan metode support vector machine. JATI, 5(2), 548–556. https://doi.org/10.36040/jati.v5i2.3808

Ula, M., & Fachrurrazi, S. (2023). Analisis sentimen cyberbullying pada media sosial Twitter menggunakan SVM dan Naive Bayes. TECHSI, 14(2), 91–101. https://doi.org/10.29103/techsi.v14i2.12103

We Are Social & Meltwater. (2024). Digital 2024: Indonesia. We Are Social. https://datareportal.com/reports/digital-2024-indonesia

Wicaksono, N. H., Setiyanto, A., & Fata, I. A. (2025). Deteksi cyberbullying tiga kelas pada komentar berbahasa Indonesia di aplikasi X dan Threads menggunakan IndoBERTweet. Journal of Computers and Digital Business, 4(1), 55–67.

Wongso, R. (2023). Cyberbullying detection model using IndoBERT. Jurnal Ilmu Komputer dan Elektro (JIKE), 5(2), 112–121.

Zakaria, P. H., Nurjannah, D., & Nurrahmi, H. (2023). Misogyny text detection on TikTok using IndoBERTweet. Jurnal Media Informatika Budidarma, 7(3), 1297–1305. https://doi.org/10.30865/mib.v7i3.5878