Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 7, No. 4, August 2026, Page 160-169
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v7i4.9330

Deteksi Cyberbullying pada Komentar Instagram Berbahasa Indonesia Menggunakan Fine-Tuning IndoBERT dengan Analisis Kesalahan Model

Dewi Irma Afriayanti, Budi Tjahjono*

Ilmu Komputer, Teknik Informatika, Universitas Esa Unggul, Jakarta, Indonesia
Email: [1]dewiirmaafriayanti@student.esaunggul.ac.id, [2,*]budi.tjahjono@esaunggul.ac.id
Email Penulis Korespondensi: budi.tjahjono@esaunggul.ac.id
Submitted: 01/02/2026; Accepted: 29/08/2026; Published: 31/08/2026

Abstrak–Cyberbullying pada media sosial merupakan permasalahan yang sulit dideteksi secara otomatis karena komentar dapat menggunakan bahasa informal, sarkasme, body shaming tersirat, dan makna yang bergantung pada konteks. Penelitian ini bertujuan mengevaluasi kemampuan IndoBERT dalam mendeteksi cyberbullying pada komentar Instagram berbahasa Indonesia sekaligus menganalisis pola kesalahan klasifikasinya. Dataset terdiri dari 1.050 komentar yang dikumpulkan dari unggahan Instagram publik dan dilabeli secara manual ke dalam dua kelas, yaitu cyberbullying dan non-cyberbullying. Model IndoBERT-base-uncased dilatih menggunakan pendekatan fine-tuning dan dievaluasi dengan confusion matrix, accuracy, precision, recall, dan F1-score. Hasil pengujian menunjukkan accuracy sebesar 84,76% dan F1-score sebesar 0,8462. Pada kelas cyberbullying, precision mencapai 0,9294, sedangkan recall sebesar 0,7524, yang menunjukkan bahwa sebagian komentar perundungan masih terlewat sebagai false negative. Analisis kesalahan menunjukkan bahwa kasus sulit terutama berkaitan dengan sarkasme, body shaming yang tidak eksplisit, bahasa informal, dan komentar ambigu. Kontribusi penelitian ini adalah menyediakan evaluasi empiris IndoBERT pada komentar Instagram berbahasa Indonesia, memperlihatkan research gap antara pendekatan machine learning terdahulu dan kebutuhan pemahaman konteks bahasa Indonesia, serta memberikan analisis kesalahan yang dapat menjadi dasar pengembangan sistem moderasi konten berbasis AI. Temuan ini menunjukkan bahwa IndoBERT memiliki potensi yang baik, tetapi peningkatan dataset, pemodelan konteks percakapan, dan penanganan bahasa pragmatik masih diperlukan.

Kata Kunci: Cyberbullying; Instagram; IndoBERT; NLP; Klasifikasi Teks; Transformer

Abstract–Cyberbullying on social media is difficult to detect automatically because comments may contain informal language, sarcasm, implicit body shaming, and meanings that depend on context. This study evaluates the ability of IndoBERT to detect cyberbullying in Indonesian-language Instagram comments and analyzes the model's classification errors. The dataset consists of 1,050 comments collected from public Instagram posts and manually labeled into two classes: cyberbullying and non-cyberbullying. The IndoBERT-base-uncased model was fine-tuned and evaluated using a confusion matrix, accuracy, precision, recall, and F1-score. The experimental results show an accuracy of 84.76% and an F1-score of 0.8462. For the cyberbullying class, precision reached 0.9294, while recall was 0.7524, indicating that a portion of bullying comments were still missed as false negatives. Error analysis shows that difficult cases were mainly associated with sarcasm, implicit body shaming, informal language, and ambiguous comments. The contributions of this study are an empirical evaluation of IndoBERT on Indonesian Instagram comments, a clear identification of the research gap between conventional machine-learning approaches and the need for contextual Indonesian-language modeling, and an error analysis that provides practical evidence for improving AI-based content moderation. The findings indicate that IndoBERT has promising performance, but larger datasets, conversational context, and pragmatic language modeling are still required.

Keywords: Cyberbullying; Instagram; IndoBERT; NLP; Text Classification; Transformer

1. PENDAHULUAN

Perkembangan media sosial telah meningkatkan intensitas interaksi digital dan memperluas ruang bagi masyarakat untuk berbagi informasi, pengalaman, serta pendapat secara cepat. Di sisi lain, kemudahan berkomunikasi juga meningkatkan peluang munculnya perilaku negatif, salah satunya cyberbullying. Cyberbullying merupakan bentuk perundungan yang dilakukan melalui media digital dan dapat menimbulkan dampak psikologis yang serius bagi korban. Literatur terbaru menunjukkan bahwa cyberbullying berkaitan dengan berbagai masalah psikososial, terutama pada kelompok remaja dan pengguna media sosial dengan intensitas interaksi tinggi [1], [2], [3], [4], [5]. Dalam media sosial, cyberbullying tidak selalu disampaikan melalui kata-kata kasar secara langsung, tetapi dapat muncul sebagai sindiran, sarkasme, body shaming tersirat, bahasa tidak baku, singkatan, atau ungkapan yang maknanya bergantung pada konteks. Karena itu, deteksi otomatis cyberbullying tidak cukup hanya mengenali kata, tetapi juga membutuhkan pemahaman hubungan semantik dan konteks [6].

Kebutuhan terhadap sistem deteksi otomatis semakin penting karena volume komentar media sosial sangat besar dan berubah secara dinamis. Moderasi manual membutuhkan waktu dan sumber daya yang tinggi sehingga Natural Language Processing (NLP) dan machine learning banyak digunakan untuk membantu identifikasi konten berbahaya. Model konvensional dapat memanfaatkan TF-IDF, n-gram, dan fitur leksikal untuk membedakan cyberbullying dan non-cyberbullying. Namun, pendekatan tersebut cenderung bergantung pada pola kata yang tampak sehingga dapat mengalami kesulitan ketika komentar bersifat metaforis, sarkastik, ambigu, atau memiliki makna yang tidak dapat ditentukan hanya dari kata-kata secara terpisah. Perkembangan

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 7, No. 4, August 2026, Page 160-169
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v7i4.9330

deep learning meningkatkan kemampuan ekstraksi pola bahasa, sedangkan Transformer menawarkan representasi kontekstual dengan mempertimbangkan hubungan antarkata secara lebih menyeluruh [7], [8], [9], [10], [11], [12], [13].

Sejumlah penelitian terbaru menunjukkan perkembangan metode deteksi cyberbullying sekaligus memperlihatkan keterbatasan yang masih perlu ditangani. El Koshiry et al. (2024) membandingkan algoritma machine learning dan deep learning dengan embedding GloVe serta focal loss. Hasilnya menunjukkan performa tinggi, tetapi recall pada beberapa model masih relatif lebih rendah sehingga sebagian kasus cyberbullying belum teridentifikasi [7]. Perera dan Fernando (2024) menggunakan supervised machine learning pada data Twitter dengan Support Vector Machine, Naive Bayes, dan Logistic Regression serta fitur NLP seperti sentiment analysis, n-gram, TF-IDF, dan deteksi kata profan [14]. Pendekatan tersebut efektif untuk fitur tekstual eksplisit, tetapi belum menggunakan model bahasa Transformer untuk membentuk representasi kontekstual. Selain itu, perbedaan platform dan bahasa membuat hasil tersebut belum dapat langsung menggambarkan karakteristik komentar Instagram berbahasa Indonesia.

Kulkarni et al. (2024) melakukan analisis berbagai pendekatan machine learning untuk deteksi cyberbullying dan menunjukkan bahwa pemilihan model memengaruhi efektivitas klasifikasi, sementara teks ambigu tetap menjadi tantangan [15]. Sayed et al. (2025) menggunakan 39.870 posting dan komentar Twitter serta membagi cyberbullying ke dalam kategori agama, usia, gender, etnis, dan non-cyberbullying. Lima classifier machine learning dibandingkan dan Random Forest memperoleh akurasi tertinggi sebesar 94% [16]. Studi tersebut unggul dari sisi jumlah data dan keragaman kategori, tetapi masih menggunakan pendekatan machine learning konvensional dan data berbahasa Inggris. Joshi et al. (2025) juga menggunakan NLP dan TF-IDF dengan linear SVC serta stochastic gradient descent pada data Twitter dan menunjukkan bahwa metode supervised machine learning dapat digunakan secara efisien [17]. Namun, pendekatan berbasis fitur tersebut masih memiliki keterbatasan dalam memahami hubungan semantik dan konteks implisit.

Di sisi Transformer, Bhagyalekshmi, Harikrishnan, dan Salaji (2025) menerapkan fine-tuning BERT untuk deteksi cyberbullying dan memberikan bukti bahwa model Transformer dapat digunakan untuk memahami konteks teks secara lebih baik [18]. Meskipun demikian, penelitian tersebut menggunakan konteks dan dataset yang berbeda dari penelitian ini. Jika dibandingkan secara keseluruhan, penelitian terdahulu memperlihatkan tiga pola utama: machine learning konvensional masih dominan pada beberapa dataset cyberbullying; deep learning non-Transformer mampu mengekstraksi pola tekstual tetapi menghadapi keterbatasan konteks; sedangkan BERT menunjukkan potensi yang lebih baik namun penerapannya pada komentar Instagram berbahasa Indonesia masih relatif terbatas.

Berdasarkan penelitian-penelitian tersebut, research gap penelitian ini dapat dirumuskan secara jelas. Pertama, sebagian studi menggunakan Twitter atau sumber media sosial berbahasa Inggris, sedangkan kajian yang berfokus pada komentar Instagram berbahasa Indonesia masih terbatas [14], [15], [16], [17]. Perbedaan platform penting karena komentar Instagram cenderung pendek, informal, menggunakan slang, variasi ejaan, singkatan, dan ekspresi yang sangat bergantung pada konteks unggahan [3]. Kedua, pendekatan machine learning konvensional umumnya mengandalkan TF-IDF, n-gram, atau fitur kata sehingga belum sepenuhnya menangkap hubungan semantik dua arah [14], [16], [17]. Ketiga, penelitian BERT menunjukkan potensi Transformer, tetapi belum secara khusus menjawab kebutuhan model bahasa Indonesia pada komentar Instagram [18]. Keempat, penelitian terdahulu lebih sering menekankan metrik agregat, sedangkan analisis penyebab false negative pada sarkasme, body shaming tersirat, bahasa informal, dan komentar ambigu masih perlu diperdalam. Gap tersebut menjadi dasar penelitian ini untuk mengevaluasi IndoBERT sekaligus melakukan analisis kesalahan. Penelitian ini menggunakan IndoBERT-base-uncased melalui fine-tuning pada 1.050 komentar Instagram berbahasa Indonesia yang dikumpulkan dari unggahan publik dan dilabeli secara manual ke dalam kelas cyberbullying dan non-cyberbullying. Pemilihan IndoBERT didasarkan pada kesesuaiannya dengan bahasa Indonesia dan kemampuannya membentuk contextual embedding melalui arsitektur Transformer. Evaluasi dilakukan menggunakan confusion matrix, accuracy, precision, recall, dan F1-score, dilengkapi analisis kesalahan untuk mengidentifikasi komentar yang paling sulit diklasifikasikan.

Kontribusi penelitian ini terdiri atas tiga aspek. Pertama, penelitian menyediakan evaluasi empiris fine-tuning IndoBERT pada dataset komentar Instagram berbahasa Indonesia yang dilabeli secara manual. Kedua, penelitian tidak hanya mengukur performa agregat, tetapi menganalisis false positive dan false negative, khususnya pada sarkasme, body shaming tersirat, bahasa informal, dan komentar ambigu. Ketiga, temuan penelitian memberikan implikasi praktis bagi pengembangan sistem moderasi konten berbasis AI serta menunjukkan kebutuhan peningkatan dataset, variasi bahasa, hyperparameter tuning, dan pemanfaatan konteks percakapan. Dengan demikian, penelitian ini memberikan kontribusi teknis dan empiris terhadap pengembangan deteksi cyberbullying yang lebih sesuai dengan karakteristik bahasa Indonesia di media sosial.

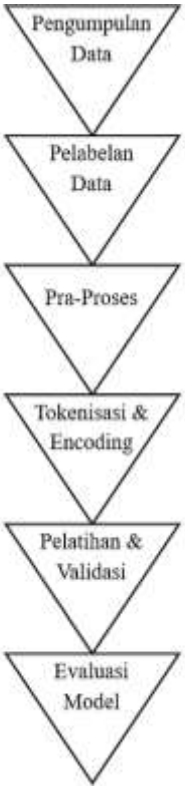
2. METODOLOGI PENELITIAN

Penelitian ini menggunakan pendekatan eksperimen dengan tujuan untuk mengevaluasi kinerja model Transformer berbasis BERT dalam mendeteksi cyberbullying pada komentar Instagram berbahasa Indonesia.

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 7, No. 4, August 2026, Page 160-169
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v7i4.9330

Metodologi penelitian disusun secara sistematis agar proses penelitian dapat direplikasi oleh peneliti lain. Tahapan penelitian meliputi pengumpulan data, pelabelan data, prapemrosesan teks, persiapan dan fine-tuning model, proses pelatihan dan validasi, serta evaluasi kinerja model.

2.1 Tahapan Penelitian



Gambar 1. Tahapan Penelitian

Gambar 1 menunjukkan tahapan penelitian yang dilakukan dalam pengembangan sistem deteksi cyberbullying. Penelitian diawali dengan pengumpulan data komentar Instagram yang relevan, kemudian dilanjutkan dengan proses pelabelan data ke dalam dua kelas, yaitu cyberbullying dan non-cyberbullying. Data yang telah dilabeli selanjutnya melalui tahap prapemrosesan teks untuk mengurangi noise dan menormalkan format data. Setelah itu, teks diproses menggunakan tokenizer BERT untuk menghasilkan representasi token yang sesuai dengan arsitektur model. Tahap berikutnya adalah proses fine-tuning model BERT menggunakan data latih, yang dilanjutkan dengan pelatihan dan validasi model. Tahap akhir penelitian adalah evaluasi performa model menggunakan metrik klasifikasi untuk mengukur efektivitas sistem deteksi yang dikembangkan.

a. Pengumpulan Data dan Deskripsi Dataset
   Dataset yang digunakan dalam penelitian ini terdiri dari 1.050 komentar Instagram berbahasa Indonesia. Data dikumpulkan melalui proses scraping pada unggahan Instagram publik yang memiliki tingkat interaksi tinggi. Pemilihan unggahan dengan interaksi tinggi bertujuan untuk memperoleh variasi komentar yang lebih beragam, termasuk komentar positif, netral, maupun komentar yang mengandung unsur perundungan. Setiap komentar kemudian diberi label secara manual ke dalam dua kelas, yaitu cyberbullying dan non-cyberbullying. Pelabelan dilakukan berdasarkan definisi cyberbullying yang mencakup tindakan penghinaan, pelecehan, ujaran kebencian, serta ungkapan negatif yang menyerang individu atau kelompok tertentu. Proses pelabelan manual dilakukan untuk memastikan akurasi label dan meningkatkan reliabilitas dataset. Dataset selanjutnya dibagi menjadi dua bagian, yaitu data latih dan data uji. Sebanyak 80% data (840 komentar) digunakan sebagai data latih, sedangkan 20% data (210 komentar) digunakan sebagai data uji. Pembagian ini bertujuan untuk menguji kemampuan generalisasi model terhadap data yang tidak digunakan selama proses pelatihan.
b. Prapemrosesan Data
   Sebelum digunakan dalam proses pemodelan, data teks melalui tahap prapemrosesan untuk mengurangi noise tanpa menghilangkan makna utama teks. Tahapan prapemrosesan meliputi case folding dengan mengubah seluruh teks menjadi huruf kecil serta penghapusan karakter non-alfanumerik seperti tanda baca dan simbol khusus. Tahapan ini bertujuan untuk menormalkan teks dan mengurangi variasi yang tidak diperlukan. Tokenisasi dilakukan menggunakan tokenizer BERT yang memecah teks menjadi unit subword sesuai dengan arsitektur BERT. Panjang maksimum sekuens ditetapkan sebesar 128 token untuk menjaga keseimbangan antara kelengkapan konteks dan efisiensi komputasi. Proses padding dan truncation diterapkan

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 7, No. 4, August 2026, Page 160-169
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v7i4.9330

agar seluruh input memiliki panjang yang seragam. Proses preprocessing pada penelitian ini tetap mempertahankan struktur kalimat utama, meskipun dilakukan pembersihan terhadap noise seperti URL dan karakter non-relevan. Hal ini dilakukan dengan mempertimbangkan bahwa model BERT mampu menangkap konteks melalui tokenisasi subword. Namun, keterbatasan preprocessing ini diakui dapat memengaruhi representasi ekspresi emosional seperti tanda baca.

c. Persiapan Model dan Fine-Tuning
Model yang digunakan dalam penelitian ini adalah IndoBERT-base-uncased [9], [10], yaitu model BERT yang telah dilatih sebelumnya menggunakan korpus bahasa Indonesia. Pemilihan model ini didasarkan pada kemampuannya dalam memahami struktur dan konteks bahasa Indonesia secara lebih baik dibandingkan model BERT umum. Proses pemodelan dilakukan dengan pendekatan fine-tuning, yaitu menyesuaikan bobot model BERT yang telah dilatih sebelumnya menggunakan dataset cyberbullying Instagram. Pada tahap ini, sebuah lapisan klasifikasi ditambahkan di atas encoder BERT untuk melakukan klasifikasi biner terhadap komentar. Pemilihan hyperparameter seperti learning rate 2e-5 dan jumlah epoch didasarkan pada praktik umum fine-tuning model BERT. Namun, penelitian ini tidak melakukan eksplorasi hyperparameter tuning secara ekstensif, yang menjadi salah satu keterbatasan penelitian dan dapat memengaruhi performa model pada dataset berukuran kecil.

d. Proses Pelatihan dan Validasi
Pelatihan model dilakukan menggunakan optimizer AdamW dengan nilai learning rate sebesar 5×10^{-5} dan ukuran batch sebesar 16. Model dilatih selama 3 epoch. Pemilihan jumlah epoch didasarkan pada rekomendasi penelitian sebelumnya yang menyatakan bahwa model BERT umumnya mencapai performa optimal pada rentang 2–4 epoch, khususnya untuk dataset berukuran kecil hingga menengah. Selama proses pelatihan, nilai training loss dan validation loss dipantau untuk mengevaluasi stabilitas model dan menghindari overfitting.

e. Evaluasi Model
Evaluasi kinerja model dilakukan menggunakan data uji dengan beberapa metrik klasifikasi, yaitu confusion matrix, precision, recall, F1-score, dan akurasi. Confusion matrix digunakan untuk menggambarkan distribusi prediksi benar dan salah pada setiap kelas. Precision dan recall digunakan untuk mengukur ketepatan dan kelengkapan prediksi model, sedangkan F1-score digunakan sebagai metrik keseimbangan antara precision dan recall. Akurasi digunakan untuk memberikan gambaran umum mengenai kinerja model dalam mengklasifikasikan komentar secara keseluruhan. Penggunaan kombinasi metrik evaluasi tersebut bertujuan untuk memberikan penilaian yang komprehensif terhadap performa model BERT dalam mendeteksi cyberbullying pada komentar Instagram berbahasa Indonesia.

2.2 Spesifikasi Dataset

Tabel 1. Spesifikasi Dataset

Komponen	Deskripsi
Sumber Data	Komentar Instagram
Bahasa	Bahasa Indonesia
Jumlah Data	1.050 komentar
Kelas	Cyberbullying, Non-Cyberbullying
Teknik Pengumpulan	Web scraping
Metode Pelabelan	Manual labeling
Data Latih	840 komentar (80%)
Data Uji	210 komentar (20%)

Tabel 1 menyajikan spesifikasi dataset yang digunakan dalam penelitian ini. Dataset terdiri dari 1.050 komentar Instagram berbahasa Indonesia yang dikumpulkan melalui teknik web scraping pada unggahan publik. Setiap komentar diberi label secara manual ke dalam dua kelas untuk memastikan akurasi data. Dataset kemudian dibagi menjadi data latih dan data uji dengan rasio 80:20 guna mengevaluasi kemampuan generalisasi model terhadap data yang tidak digunakan selama pelatihan.

2.3 Konfigurasi Model & Pelatihan

Tabel 2. Konfigurasi Model & Pelatihan

Parameter	Nilai
Model	IndoBERT-base-uncased
Arsitektur	Transformer
Panjang Maks. Token	128
Batch Size	16
Epoch	3
Optimizer	AdamW
Learning Rate	5×10^{-5}

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 7, No. 4, August 2026, Page 160-169
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v7i4.9330

Parameter	Nilai
Fungsi Loss	Cross-Entropy Loss
Platform	Google Colab

Tabel 2 menunjukkan konfigurasi model dan parameter pelatihan yang digunakan dalam penelitian ini. Model IndoBERT-base-uncased dipilih karena telah dilatih pada korpus bahasa Indonesia. Panjang maksimum token ditetapkan sebesar 128 untuk menjaga keseimbangan antara konteks teks dan efisiensi komputasi. Proses pelatihan dilakukan selama 3 epoch menggunakan optimizer AdamW dengan learning rate sebesar 5×10^{-5}. Pemilihan parameter ini didasarkan pada rekomendasi penelitian sebelumnya yang menyatakan bahwa model BERT umumnya mencapai performa optimal pada rentang epoch yang relatif kecil, khususnya pada dataset berukuran menengah.

3. HASIL DAN PEMBAHASAN

Penelitian ini menggunakan pendekatan kuantitatif dengan metode eksperimen, yang bertujuan untuk menguji kinerja model deep learning berbasis Transformer dalam mendeteksi cyberbullying pada komentar Instagram. Metode eksperimen dipilih karena memungkinkan pengukuran performa model secara objektif berdasarkan metrik evaluasi yang terstandar. Permasalahan dalam penelitian ini dirumuskan sebagai klasifikasi biner, yaitu mengelompokkan komentar ke dalam dua kelas: cyberbullying dan non-cyberbullying. Evaluasi model dilakukan menggunakan confusion matrix serta metrik akurasi, precision, recall, dan F1-score untuk menilai kemampuan model dalam mengenali pola perundungan secara kontekstual.

Data yang digunakan dalam penelitian ini berupa 1.050 komentar Instagram berbahasa Indonesia yang diambil dari unggahan publik dengan tingkat interaksi tinggi. Setiap komentar diberi label secara manual ke dalam dua kelas, yaitu cyberbullying dan non-cyberbullying. Proses pelabelan manual dilakukan untuk memastikan keakuratan label serta mempertimbangkan konteks bahasa informal yang umum digunakan pada media sosial. Dataset kemudian dibagi menjadi data latih dan data uji dengan perbandingan 80% data latih dan 20% data uji guna menguji kemampuan generalisasi model.

Hasil evaluasi menunjukkan bahwa model IndoBERT mencapai akurasi sebesar 84,76% dengan F1-score sebesar 0,8462. Meskipun nilai precision cukup tinggi, nilai recall untuk kelas cyberbullying hanya sebesar 0,75. Nilai recall tersebut menunjukkan bahwa model masih gagal mendeteksi sekitar 25% komentar cyberbullying, yang merupakan keterbatasan signifikan dalam konteks sistem moderasi konten. Hal ini mengindikasikan bahwa model cenderung menghasilkan false negative pada komentar yang mengandung makna implisit.

Berdasarkan analisis kesalahan, model mengalami kesulitan dalam mendeteksi komentar yang bersifat sarkastik, menggunakan bahasa tidak baku, serta mengandung body shaming secara tersirat. Hal ini menunjukkan bahwa meskipun model berbasis Transformer mampu memahami konteks, pemahaman terhadap aspek pragmatik bahasa masih menjadi tantangan.

Selain itu, ukuran dataset yang relatif terbatas juga mempengaruhi kemampuan generalisasi model. Dengan jumlah data yang kecil, model berpotensi belum mampu menangkap variasi bahasa yang kompleks dalam media sosial. Perbandingan dengan penelitian terdahulu tidak dilakukan secara langsung pada nilai performa, mengingat perbedaan dataset, bahasa, dan domain yang digunakan.

3.1 Distribusi Data Uji

Sebelum melakukan evaluasi model, distribusi kelas pada data uji dianalisis untuk memastikan keseimbangan data. Dari total 210 data uji, masing-masing kelas terdiri dari 105 komentar cyberbullying dan 105 komentar non-cyberbullying. Distribusi yang seimbang ini penting untuk mencegah bias model terhadap salah satu kelas dan memastikan bahwa nilai akurasi dan metrik lainnya merepresentasikan kinerja model secara adil.

Tabel 3. Distribusi Data Uji

Kelas	Jumlah Data
Cyberbullying	105
Non-Cyberbullying	105
Total	210

Distribusi data yang seimbang memungkinkan evaluasi model dilakukan secara objektif tanpa dipengaruhi oleh dominasi salah satu kelas seperti yang divisualisasikan oleh tabel 3.

3.2 Hasil Pelatihan Model

Model IndoBERT pada penelitian ini dilatih menggunakan teknik fine-tuning dengan dataset komentar Instagram berbahasa Indonesia yang telah dilabeli menjadi dua kelas, yaitu cyberbullying dan non-cyberbullying. Proses pelatihan dilakukan menggunakan pembagian data sebesar 80% untuk data latih dan 20% untuk data uji. Pelatihan model dilakukan dalam beberapa epoch untuk mengamati perubahan performa.

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 7, No. 4, August 2026, Page 160-169
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v7i4.9330

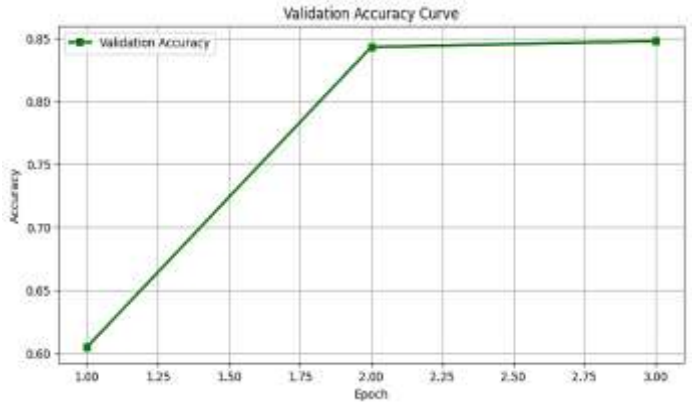
Berdasarkan hasil eksperimen, model menunjukkan peningkatan performa hingga epoch ke-4, yang kemudian dipilih sebagai model terbaik.

Tabel 3. Distribusi Data Uji

Epoch	Accuracy	Precision	Recall	F1-score
3	83,90%	0,84	0,83	0,83
4	84,76%	0,86	0,75	0,8462

Berdasarkan Tabel 3, terlihat bahwa peningkatan epoch dari 3 ke 4 memberikan peningkatan pada nilai accuracy dan F1-score. Namun, terjadi penurunan pada nilai recall. Hal ini menunjukkan adanya trade-off antara precision dan recall, di mana model menjadi lebih selektif dalam mengklasifikasikan komentar sebagai cyberbullying. Akibatnya, jumlah false positive berkurang, tetapi jumlah false negative meningkat.

Gambar 2. Training Loss



Gambar 3. Validation Accuracy

Grafik pelatihan pada gambar 2 dan gambar 3 menunjukkan bahwa selisih antara training loss dan validation loss relatif kecil pada epoch terakhir. Hal ini mengindikasikan bahwa model tidak hanya belajar dengan baik pada data latih, tetapi juga mampu melakukan generalisasi terhadap data uji. Dengan jumlah epoch yang terbatas, model berhasil mencapai konvergensi tanpa menunjukkan gejala overfitting yang berlebihan, sehingga pemilihan 3 epoch dapat dianggap optimal untuk dataset yang digunakan.

3.3 Evaluasi Menggunakan Confusion Matrix

Evaluasi performa model dilakukan menggunakan confusion matrix untuk melihat distribusi prediksi benar dan salah pada masing-masing kelas.

Tabel 4. Confusion Matrix

Actual \ Predicted	Non-Cyberbullying	Cyberbullying
Non-Cyberbullying	99	6
Cyberbullying	26	79

Berdasarkan Tabel 4, nilai True Positive (TP) sebesar 79 menunjukkan jumlah komentar cyberbullying yang berhasil dideteksi dengan benar, sedangkan True Negative (TN) sebesar 99 menunjukkan komentar non-cyberbullying yang terklasifikasi dengan tepat. Namun demikian, terdapat False Negative (FN) sebesar 26 yang menunjukkan bahwa sejumlah komentar cyberbullying tidak terdeteksi oleh model. Selain itu, terdapat False

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 7, No. 4, August 2026, Page 160-169
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v7i4.9330

Positive (FP) sebesar 6, yaitu komentar non-cyberbullying yang salah diklasifikasikan sebagai cyberbullying. Tingginya nilai FN dibanding FP menunjukkan bahwa model cenderung lebih konservatif dalam mendeteksi cyberbullying, sehingga berpotensi melewatkan beberapa kasus perundungan yang bersifat implisit.

3.4 Classification Report

Evaluasi kuantitatif lanjutan dilakukan menggunakan metrik precision, recall, F1-score, dan akurasi.

Tabel 5. Classification Report

Kelas	Precision	Recall	F1-score	Support
Non-Cyberbullying	0,792	0,9429	0,8609	105
Cyberbullying	0,9294	0,7524	0,8316	105
Accuracy			0,8476	210
Macro Avg	0,8607	0,8476	0,8462	210
Weighted Avg	0,8607	0,8476	0,8462	210

Berdasarkan Tabel 5, model IndoBERT menunjukkan performa yang cukup baik dengan nilai akurasi sebesar 84,76% dan F1-score sebesar 0,8462. Nilai precision yang relatif tinggi menunjukkan bahwa model mampu mengklasifikasikan komentar cyberbullying dengan tingkat ketepatan yang baik. Namun demikian, nilai recall sebesar 0,75 menunjukkan bahwa model masih belum mampu mendeteksi seluruh komentar cyberbullying secara optimal. Hal ini mengindikasikan bahwa sekitar 25% komentar cyberbullying tidak teridentifikasi oleh model (false negative), yang menjadi perhatian penting dalam konteks moderasi konten.

3.5 Analisis Kesalahan (Error Analysis)

Analisis kesalahan dilakukan untuk memahami karakteristik data yang sulit diklasifikasikan oleh model. Berdasarkan hasil pengujian, kesalahan klasifikasi sebagian besar terjadi pada:

a. Komentar sarkastik, yang secara literal tampak positif namun memiliki makna negatif.
b. Body shaming tersirat, yang tidak secara eksplisit menggunakan kata kasar.
c. Bahasa informal dan slang, yang memiliki variasi tinggi dan tidak selalu terdapat dalam data pelatihan.

Dari perspektif teknis, keterbatasan ini disebabkan oleh representasi embedding pada model yang belum sepenuhnya mampu menangkap konteks pragmatik dan nuansa emosional dalam bahasa informal. Meskipun IndoBERT memiliki kemampuan contextual embedding, pemahaman terhadap makna implisit masih menjadi tantangan.

	text	true_label	pred_label
58	penyihir secantik ini gue rela dikutuk jadi to...	0	1
61	hanya bisa melihat hujatan klean kasian si lis...	0	1
124	heran gua mah kalau gasuka gausah di follow da...	0	1
127	padahal gw g mau benci ama lu bro tpi ada aja ...	0	1
151	bully an di ig lebih ngeri daripada di fb ajg ...	0	1
193	username aurel itu cantik tau pernah liat asli...	0	1

Gambar 4. False Positive

	text	true_label	pred_label
6	sampah masyarakat harus di musnahkan segera	1	0
9	username lah goblok amat gak semua netizen yg ...	1	0
21	jijik saya liatnyakaya ulet bulu kluget2 gituh...	1	0
23	susah kalo artis kampung	1	0
27	sakit jiwa ini upil gajah kasihan	1	0
30	masih jauh kalo ngarep bisa ikutan kontes miss...	1	0
41	norak bener kegaet penjahat baru rasa	1	0
42	kayak banci dh mirip nenek2 pny cucu mw di op...	1	0
43	intinya kalau kesel dengan att nya gausah ke a...	1	0
46	lututnya masih hitam gosok maning biar tambah ...	1	0

Gambar 5. False Negative

Gambar 4 dan 5 menampilkan contoh komentar yang mengalami kesalahan klasifikasi oleh model, yang terdiri dari false positive dan false negative. Kesalahan false positive terjadi ketika komentar non-cyberbullying diprediksi sebagai cyberbullying, umumnya disebabkan oleh penggunaan kata bernada negatif namun tidak ditujukan untuk menyerang individu tertentu. Sebaliknya, false negative terjadi ketika komentar cyberbullying

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 7, No. 4, August 2026, Page 160-169
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v7i4.9330

tidak terdeteksi oleh model, terutama pada komentar yang bersifat sarkastik, implisit, atau menggunakan bahasa slang yang kompleks. Analisis kesalahan ini menunjukkan bahwa meskipun model BERT mampu memahami konteks secara umum, masih terdapat tantangan dalam mengenali ekspresi perundungan yang bersifat tidak langsung. Hasil ini menjadi dasar untuk pengembangan model di masa mendatang, seperti penambahan data latih atau integrasi fitur pragmatik.

3.6 Perbandingan Manual vs Prediksi Model

Untuk melengkapi evaluasi kuantitatif, dilakukan analisis kualitatif dengan membandingkan hasil prediksi model dan penilaian manual pada beberapa sampel komentar. Pada sebagian besar kasus, model mampu mengklasifikasikan komentar secara konsisten dengan penilaian manual, terutama pada komentar yang mengandung ujaran perundungan secara eksplisit. Namun, perbedaan muncul pada komentar yang bersifat ambigu atau mengandung campuran pujian dan sindiran. Pada kasus tersebut, model terkadang memberikan prediksi yang berbeda dari penilaian manual. Hal ini menunjukkan bahwa pemahaman konteks sosial dan pragmatik bahasa masih menjadi tantangan dalam sistem deteksi otomatis.

	text	true_label	pred_label
75	jijik astagah sok bat cakep njs	1	1
56	karena muka lo pasaran mel di pasar sapi juga ...	1	1
104	kok mbak meldi sekarang tambah cantik ya rahas...	0	0
193	username aurel itu cantik tau pernah liat asli...	0	1
145	lo sintingnya udah tingkat dewa	1	1
126	jari jari nya gede gede banget apalagi pas ngi...	1	0
81	smoga mas agus menjadi pemimpin di masa akan d...	0	0
124	heran gua mah kalau gasuka gausah di follow da...	0	1
156	lelaki yg mendapatkan km adalah lelaki yg beru...	0	0
2	cantik pintar bertalenta ortu mana yg gak kepe...	0	0

Gambar 6. Perbandingan Manual vs Prediksi Model

Gambar 6 menunjukkan perbandingan antara hasil pelabelan manual dengan prediksi yang dihasilkan oleh model BERT pada beberapa contoh komentar. Pada sebagian besar data, prediksi model sesuai dengan label manual, yang menandakan bahwa model mampu mengenali pola cyberbullying secara konsisten. Namun, terdapat beberapa perbedaan prediksi yang umumnya muncul pada komentar ambigu, bercanda, atau mengandung pujian yang disertai kata informal. Perbandingan ini memperlihatkan bahwa model BERT telah mampu mendekati penilaian manusia dalam mengklasifikasikan komentar cyberbullying, meskipun masih memerlukan peningkatan untuk menangani konteks sosial dan pragmatik yang lebih kompleks.

3.7 Perbandingan dengan Penelitian Terdahulu

Tabel 6. Perbandingan Penelitian

Peneliti	Metode	Dataset	Temuan Utama	Keterbatasan
Joshi et al. (2025) [17]	SVM	Twitter (EN)	Model cepat dan efisien untuk klasifikasi teks	Tidak mampu memahami konteks kalimat secara mendalam
Kulkarni et al. (2024) [15]	LSTM	Sosial media (EN)	Mampu menangkap urutan kata dalam teks	Kurang efektif pada teks ambigu dan sarkasme
El Koshiry et al. (2024) [7]	CNN + GloVe	Review teks (EN)	Baik dalam ekstraksi fitur lokal	Tidak memahami konteks global secara optimal
Penelitian ini (2025)	BERT Transformer	Instagram (ID)	Memahami konteks dua arah (bidirectional context)	Recall masih terbatas pada teks implisit dan sarkastik

Tabel 6 menyajikan perbandingan antara penelitian terdahulu dan penelitian ini berdasarkan metode yang digunakan, jenis dataset, serta karakteristik hasil yang diperoleh [13]. Perbandingan ini tidak dilakukan berdasarkan nilai performa secara langsung, melainkan difokuskan pada pendekatan metodologi dan kemampuan model dalam memahami konteks teks. Metode konvensional seperti SVM cenderung lebih cepat dan efisien, namun memiliki keterbatasan dalam memahami konteks semantik. Model berbasis deep learning seperti LSTM dan CNN menunjukkan peningkatan kemampuan dalam menangkap pola teks, namun masih memiliki kelemahan dalam memahami makna implisit dan sarkasme [9], [10], [13]. Sementara itu, model Transformer seperti IndoBERT memiliki keunggulan dalam memahami konteks dua arah, sehingga lebih efektif dalam menangani teks berbahasa Indonesia yang bersifat kontekstual. Namun, berdasarkan hasil penelitian ini, performa model masih dipengaruhi oleh karakteristik data, terutama pada komentar yang mengandung makna implisit, sarkasme, dan bahasa informal. Perbandingan performa secara langsung tidak dilakukan karena

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 7, No. 4, August 2026, Page 160-169
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v7i4.9330

perbedaan dataset, bahasa, dan domain pada masing-masing penelitian dapat menyebabkan bias dalam evaluasi model [18], [19], [20].

3.8 Implikasi dan Diskusi Hasil Penelitian

Hasil penelitian menunjukkan bahwa model IndoBERT yang di-fine-tune mampu mencapai akurasi 84,76% dan F1-score 0,8462 dalam mendeteksi cyberbullying pada komentar Instagram berbahasa Indonesia. Kinerja ini konsisten dengan temuan mutakhir bahwa model berbasis Transformer efektif untuk klasifikasi teks karena mampu menangkap konteks dua arah dan relasi semantik antar token secara lebih komprehensif dibanding pendekatan konvensional [8], [9], [10], [20].

Meskipun demikian, nilai recall sebesar 0,75 mengindikasikan bahwa sekitar 25% komentar cyberbullying belum terdeteksi (false negative). Fenomena ini sejalan dengan studi terbaru pada deteksi ujaran kebencian/cyberbullying yang melaporkan bahwa model masih kesulitan pada teks dengan makna implisit, sarkasme, dan konteks sosial yang kompleks. Dalam konteks media sosial, bentuk bahasa tidak literal tersebut sering muncul sehingga menurunkan sensitivitas model [16], [21], [22].

Dari sisi teknis, kesalahan ini berkaitan dengan keterbatasan representasi embedding dalam menangkap aspek pragmatik dan nuansa emosional. Meskipun Transformer menghasilkan contextual embeddings, pemahaman terhadap makna tersirat sering memerlukan sinyal tambahan di luar teks (misalnya konteks percakapan atau pengetahuan dunia), yang belum dimodelkan pada penelitian ini [3]. Selain itu, ukuran dataset yang relatif terbatas (1.050 data) turut memengaruhi kemampuan generalisasi model. Literatur terbaru menegaskan bahwa performa model pre-trained sangat dipengaruhi oleh ukuran dan keragaman data fine-tuning; pada data kecil, model cenderung overfit dan kurang stabil pada pola bahasa yang jarang muncul [10].

Karakteristik bahasa Indonesia di media sosial seperti penggunaan slang, singkatan, dan variasi ejaan—juga meningkatkan kompleksitas klasifikasi. Variasi ini memperlebar distribusi linguistik sehingga menyulitkan model dalam membentuk representasi yang konsisten, meskipun telah menggunakan model pra-latih berbahasa Indonesia. Dengan demikian, IndoBERT menunjukkan potensi yang baik untuk deteksi cyberbullying, namun peningkatan masih diperlukan, khususnya pada pemodelan konteks pragmatik, penanganan bahasa informal, serta perluasan dataset. Implikasi praktisnya, sistem moderasi konten berbasis AI sebaiknya dikombinasikan dengan strategi peningkatan data dan/atau pemodelan konteks percakapan untuk meningkatkan sensitivitas terhadap kasus implisit.

4. KESIMPULAN

Penelitian ini mengevaluasi kemampuan IndoBERT-base-uncased melalui fine-tuning untuk mendeteksi cyberbullying pada 1.050 komentar Instagram berbahasa Indonesia. Berdasarkan pengujian pada 210 data uji, model memperoleh accuracy sebesar 84,76% dan F1-score sebesar 0,8462. Confusion matrix menunjukkan 79 True Positive, 99 True Negative, 6 False Positive, dan 26 False Negative. Nilai precision kelas cyberbullying sebesar 0,9294 menunjukkan bahwa komentar yang diprediksi sebagai cyberbullying umumnya tepat, tetapi recall sebesar 0,7524 menunjukkan bahwa masih terdapat sekitar seperempat kasus cyberbullying yang belum berhasil dikenali. Analisis kesalahan memperlihatkan bahwa kelemahan utama model muncul pada komentar yang mengandung sarkasme, body shaming tersirat, slang, variasi bahasa informal, dan makna ambigu. Kontribusi penelitian ini terletak pada tiga aspek. Pertama, penelitian memberikan bukti empiris mengenai penerapan fine-tuning IndoBERT pada komentar Instagram berbahasa Indonesia yang dilabeli secara manual. Kedua, penelitian memperluas evaluasi performa dengan analisis false positive dan false negative sehingga karakteristik kesalahan model dapat diidentifikasi, bukan hanya dilaporkan dalam bentuk nilai accuracy atau F1-score. Ketiga, hasil penelitian memberikan dasar praktis bagi pengembangan sistem moderasi konten berbasis AI pada konteks bahasa Indonesia. Temuan ini juga memperjelas bahwa penggunaan Transformer tidak secara otomatis menyelesaikan persoalan cyberbullying yang bersifat pragmatik dan implisit. Oleh karena itu, penelitian selanjutnya disarankan memperbesar dan memperkaya dataset, melakukan hyperparameter tuning secara sistematis, memasukkan konteks percakapan atau informasi unggahan, serta mengembangkan pendekatan yang lebih sensitif terhadap sarkasme dan bahasa informal. Dengan langkah tersebut, sensitivitas model terhadap false negative dapat ditingkatkan tanpa mengorbankan ketepatan klasifikasi secara keseluruhan.

REFERENCES

[1] F. Ö. Öztürk, M. Tamaddon, and A. Tezel, "Cyberbullying, psychosocial problems and affecting factors among adolescents," *Arch. Psychiatr. Nurs.*, vol. 54, pp. 12–17, 2025, doi: https://doi.org/10.1016/j.apnu.2024.12.001.

[2] G. Gohal *et al.*, "Prevalence and related risks of cyberbullying and its effects on adolescent," *BMC Psychiatry*, vol. 23, no. 1, p. 39, 2023, doi: 10.1186/s12888-023-04542-0.

[3] P. Yi and A. Zubiaga, "Session-based cyberbullying detection in social media: A survey," *Online Soc. Netw. Media*, vol. 36, p. 100250, 2023, doi: https://doi.org/10.1016/j.osnem.2023.100250.

ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 7, No. 4, August 2026, Page 160-169
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v7i4.9330

[4] B. Akdeniz and A. Doğan, "Cyberbullying: Definition, Prevalence, Effects, Risk and Protective Factors," *Psikiyatride Güncel Yaklaşımlar*, vol. 16, no. 3, pp. 425–438, 2024, doi: 10.18863/pgy.1325195.
[5] Ditch the Label, "Cyberbullying: The impact, prevalence, and ways to address it - 2024 report." Accessed: Jun. 13, 2025. [Online]. Available: https://www.ditchthelabel.org/cyberbullying-report-2024
[6] M. S. Jahan and M. Oussalah, "A systematic review of hate speech automatic detection using natural language processing," *Neurocomputing*, vol. 546, p. 126232, 2023, doi: https://doi.org/10.1016/j.neucom.2023.126232.
[7] A. M. El Koshiry, E. H. I. Eliwa, T. Abd El-Hafeez, and M. Khairy, "Detecting cyberbullying using deep learning techniques: a pre-trained glove and focal loss technique," *PeerJ Comput. Sci.*, vol. 10, p. e1961, Mar. 2024, doi: 10.7717/peerj-cs.1961.
[8] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep Learning–based Text Classification: A Comprehensive Review," *ACM Comput. Surv.*, vol. 54, no. 3, Apr. 2021, doi: 10.1145/3439726.
[9] A. Vaswani *et al.*, "Attention Is All You Need," 2023.
[10] H. Wang, J. Li, H. Wu, E. Hovy, and Y. Sun, "Pre-Trained Language Models and Their Applications," *Engineering*, vol. 25, pp. 51–65, 2023, doi: https://doi.org/10.1016/j.eng.2022.04.024.
[11] B. Ogunleye and B. Dharmaraj, "The Use of a Large Language Model for Cyberbullying Detection," *Analytics*, vol. 2, no. 3, pp. 694–707, Sep. 2023, doi: 10.3390/analytics2030038.
[12] M. K. Ojha, N. M. Patil, and M. Joshi, "A Comparative Study of Classification Models for Cyberbullying Detection," in *2024 International Conference on Inventive Computation Technologies (ICICT)*, 2024, pp. 1531–1535. doi: 10.1109/ICICT60155.2024.10544792.
[13] H. Pen, N. Anne Huiying TEO, Z. Wang, N. Anne Huiying, and N. Teo, "Comparative analysis of hate speech detection: Traditional vs. Comparative analysis of hate speech detection: Traditional vs. deep learning approaches deep learning approaches Part of the Artificial Intelligence and Robotics Commons, and the Databases and Information Systems Commons Citation Citation Comparative Analysis of Hate Speech Detection: Traditional vs. Deep Learning Approaches," 2024. [Online]. Available: https://ink.library.smu.edu.sg/sis_research
[14] A. Perera and P. Fernando, "Cyberbullying Detection System on Social Media Using Supervised Machine Learning," *Procedia Comput. Sci.*, vol. 239, pp. 506–516, 2024, doi: https://doi.org/10.1016/j.procs.2024.06.200.
[15] R. Kulkarni, S. Chakrabarti, S. Salunke, T. Wagh, and A. Thool, "A Comprehensive Analysis of Cyberbullying Detection Using Various Machine Learning Approaches," 2024, pp. 15–25. doi: 10.1007/978-981-97-6678-9_2.
[16] F. R. Sayed, E. H. Elnashar, and F. A. Omara, "Cyberbullying detection in social media using natural language processing," Jun. 01, 2025, *Elsevier B.V.* doi: 10.1016/j.sciaf.2025.e02713.
[17] B. Joshi, B. K. Joshi, S. Pant, A. Kumar, and H. K. Sharma, "An Efficient Method for Detecting Cyberbullying Using Supervised Machine Learning Techniques," in *Procedia Computer Science*, Elsevier B.V., 2025, pp. 1254–1261. doi: 10.1016/j.procs.2025.04.359.
[18] B. S, H. D, and S. S, *Cyberbullying Detection Using Fine-Tuned BERT*. 2025. doi: 10.1109/ICTEST64710.2025.11042412.
[19] J. Peng and K. Han, "Survey of Pre-trained Models for Natural Language Processing," in *2021 International Conference on Electronic Communications, Internet of Things and Big Data (ICEIB)*, 2021, pp. 277–280. doi: 10.1109/ICEIB53692.2021.9686420.
[20] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *CoRR*, vol. abs/1810.04805, 2018, [Online]. Available: http://arxiv.org/abs/1810.04805
[21] P. K. Nag, A. Bhagat, R. Vishnu Priya, and D. K. Khare, "Emotional Intelligence Through Artificial Intelligence: NLP and Deep Learning in the Analysis of Healthcare Texts," in *2023 International Conference on Artificial Intelligence for Innovations in Healthcare Industries (ICAIIHI)*, 2023, pp. 1–7. doi: 10.1109/ICAIIHI57871.2023.10489117.
[22] A. Mansoori *et al.*, "Detection of Sarcasm in News Headlines Using NLP and Machine Learning," 2025, pp. 503–517. doi: 10.1007/978-3-031-89175-5_31.