https://jurnal.pustakagalerimandiri.co.id/index.php/pustakaai
DOI : https://doi.org/10.55382/jurnalpustakaai.v5i2.1109

JURNAL PUSTAKA AI
JURNAL PUSAT AKSES KAJIAN TEKNOLOGI ARTIFICIAL INTELLIGENCE
Pustaka Galeri Mandiri

Vol. 5 No. 2 (2025) 258– 264 E ISSN: 2809-4069

Klasifikasi Pengaduan Kekerasan Berbasis *Bidirectional Encoder Representations from Transformers* (BERT) pada Kementrian Pemberdayaan Perempuan dan Perlindungan Anak Kementrian (PPPA)

Muhammad Jundi[1], Rahmaddeni[2], Putri Utami[3], Satria Perdana Arifin[4], Leonardo Sinaga[5]
[123]Program Studi Teknik Informatika, Universitas Sains dan Teknologi Indonesia
[4]Program Studi Teknik Informatika, Politeknik Caltex Riau
[5]Program Studi Teknik Informatika, Universitas Sains dan Teknologi Indonesia
[1]2417052802089@usti.ac.id, [2]rahmaddeni@usti.ac.id, [3]2417052802090@usti.ac.id,[4]satria@pcr.ac.id,
[5]2417052802105@usti.ac.id

Abstract

Violence against women and children remains a critical issue in Indonesia, with thousands of reports submitted annually to protection agencies. Most of these reports are expressed in unstructured narratives, making it difficult to identify the type of violence and respond effectively. This study proposes an automated classification model based on IndoBERT to detect both the type of violence and the reporter's sentiment simultaneously. The model is trained using a multi-task learning approach with two output labels: violence category (multi-label) and sentiment (binary). The dataset consists of manually annotated Indonesian-language reports. Evaluation on a 20% test set using F1-score, precision, and recall shows promising performance, achieving an F1-score of 0.8528 for violence classification and 0.8375 for sentiment classification. Further tests on synthetic narratives confirm the model's ability to capture semantic context and emotional cues. These results demonstrate the potential of transformer-based models to support government institutions in accelerating the digital processing and response to violence reports.

Keywords: IndoBERT, classification, violence, sentiment analysis, multi-task learning

Abstrak

Kekerasan terhadap perempuan dan anak merupakan isu serius di Indonesia, dengan ribuan laporan masuk setiap tahun ke lembaga perlindungan. Sebagian besar laporan disampaikan dalam bentuk narasi bebas, yang menyulitkan proses identifikasi jenis kekerasan dan respon yang cepat. Penelitian ini mengembangkan model klasifikasi otomatis berbasis IndoBERT untuk mendeteksi jenis kekerasan dan sentimen pelapor secara simultan. Model dilatih menggunakan pendekatan multi-task learning dengan dua label keluaran: kategori kekerasan (*multi-label*) dan sentimen (biner). Dataset terdiri dari laporan berbahasa Indonesia yang telah dianotasi manual. Evaluasi dilakukan pada data uji (20%) menggunakan metrik *F1-score*, *precision*, dan *recall*. Hasil menunjukkan bahwa model mencapai *F1-score* sebesar 0,8528 untuk klasifikasi kekerasan dan 0,8375 untuk sentimen. Pengujian pada narasi fiktif juga membuktikan ketepatan model dalam menangkap konteks semantik dan ekspresi emosional pelapor. Model ini menunjukkan potensi signifikan untuk mendukung lembaga pemerintah dalam mempercepat analisis dan tindak lanjut laporan kekerasan secara digital.

Kata kunci: IndoBERT, klasifikasi, kekerasan, sentimen, *multi-task learning*

Submitted : 05-08-2025 | Reviewed : 16-08-2025 | Accepted : 31-08-2025

1. Pendahuluan

Kekerasan terhadap perempuan dan anak-anak merupakan permasalahan sosial yang serius dan kompleks di Indonesia [1]. Setiap tahun, ribuan kasus kekerasan dilaporkan kepada lembaga-lembaga terkait, mencerminkan berbagai bentuk kekerasan yang dialami oleh korban, mulai dari kekerasan fisik, psikis, seksual, hingga penelantaran [2]. Tingginya angka pengaduan ini menunjukkan perlunya penanganan yang cepat, tepat, dan berkesinambungan agar hak-hak korban dapat segera dipulihkan.

Namun demikian, proses penanganan laporan kekerasan tidak terlepas dari berbagai tantangan administratif dan teknis. Salah satu tantangan utama terletak pada proses klasifikasi isi laporan yang masih dilakukan secara manual [3]. Proses ini tidak hanya memakan waktu yang cukup lama, tetapi juga rawan terhadap ketidakkonsistenan dalam penilaian [4]. Keragaman bahasa dalam laporan, seperti penggunaan campuran bahasa formal, dialek lokal, dan ungkapan kiasan, turut menyulitkan pemahaman makna secara tepat. Sehingga dapat menghambat kecepatan dan akurasi dalam mengidentifikasi jenis kekerasan yang dilaporkan [5].

Perkembangan teknologi kecerdasan buatan (*Artificial Intelligence*/AI), khususnya di bidang pemrosesan bahasa alami atau *Natural Language Processing* (NLP), membuka peluang untuk mengatasi permasalahan tersebut. NLP memungkinkan mesin untuk memahami, menginterpretasikan, dan menghasilkan bahasa manusia secara lebih efektif. Salah satu terobosan dalam NLP adalah hadirnya model berbasis transformer seperti *Bidirectional Encoder Representations from Transformers* (BERT). BERT mampu memahami konteks kalimat secara menyeluruh dengan mempertimbangkan hubungan antar kata di kedua arah (maju dan mundur), sehingga memberikan representasi makna yang lebih akurat dibandingkan metode sebelumnya [6]. Dengan memanfaatkan BERT dan teknologi NLP lainnya, proses klasifikasi laporan pengaduan dapat dilakukan secara otomatis. Otomatisasi ini tidak hanya mempercepat proses penanganan, tetapi juga meningkatkan akurasi dan konsistensi klasifikasi jenis kekerasan. Selain itu, teknologi ini juga memungkinkan analisis sentimen pelapor, yang dapat memberikan gambaran emosional atau psikologis korban saat melaporkan kejadian, sehingga informasi yang dihasilkan lebih komprehensif dan bermanfaat dalam proses pengambilan keputusan.

Seiring pesatnya perkembangan teknologi pembelajaran mesin, berbagai riset telah mengeksplorasi otomatisasi klasifikasi teks terkait kekerasan dan ujaran bermuatan negatif. Hareem Kibriya dkk. mengembangkan sebuah model *end-to-end* berbasiskan *Deep Learning* dan *Explainable AI* (XAI) untuk deteksi dan klasifikasi *hate speech* secara otomatis. Model ini memadukan *Natural Language Processing* dan *Convolutional Neural Networks*, kemudian dilengkapi dengan mekanisme XAI guna menjelaskan keputusan model seperti menyoroti kata atau frasa yang dianggap bermuatan permusuhan, sehingga meningkatkan transparansi dan kepercayaan pengguna. Hasil uji menunjukkan peningkatan signifikan dalam akurasi deteksi serta mampu menghasilkan interpretabilitas yang jelas atas prediksi model [7].

Lebih dekat ke domain kekerasan, Alec Candidato dkk. melalui kompetisi SMM4H'22 mengembangkan *ensemble* dari berbagai model BERT/RoBERTa untuk mendeteksi laporan *intimate partner violence* di Twitter. Pendekatan bertingkat ini melampaui *baseline* hingga 13%, menunjukkan potensi model transformer dalam memahami teks kekerasan yang bersifat *self-reported*. Namun, fokus penelitian ini masih terbatas pada media sosial berbahasa Inggris dan hanya mencakup satu jenis kekerasan [8]. Penelitian terkait *violence detection* berbasis teks oleh Yang & Wang memadukan BERT dengan *fastText* untuk membangun sistem pendeteksian teks kekerasan generik. Pendekatan hibrida ini meningkatkan akurasi hingga 0,7–0,8% dibandingkan model tunggal, memperlihatkan kekuatan kombinasi transformer dan model ringan [9].

Berpindah ke Bahasa Indonesia, Farasalsabila dkk. mengembangkan sistem pendeteksi *cyberbullying* menggunakan kombinasi BERT dan Bi-LSTM. Model ini berhasil mencapai akurasi sekitar 90% dalam mengklasifikasikan berbagai variasi bahasa kasar dan pelecehan dalam teks *online*. Studi ini penting karena mengadaptasi BERT untuk konteks lokal, meskipun masih terbatas pada satu domain [10].

Terkait ujaran kebencian di sosial media Indonesia, Wijanarko dkk. mengimplementasikan model transformer seperti IndoBERT untuk mengklasifikasi ujaran kebencian berbahasa Indonesia. Sistem ini juga mendukung visualisasi melalui *dashboard* interaktif, memperlihatkan bagaimana teknologi NLP bisa diintegrasikan dengan alat pantau nyata [11].

Di sisi lain, beberapa studi juga telah mengeksplorasi pendekatan pembelajaran mesin klasik dalam analisis sentimen teks berbahasa Indonesia. Penelitian oleh Husen dkk. mengembangkan model analisis sentimen terhadap opini publik terkait Bank BSI di media sosial Twitter menggunakan algoritma *Support Vector Machine* (SVM), *Naive Bayes*, dan *Logistic Regression*. Dengan dataset berbahasa Indonesia yang berisi lebih dari 24.000 *tweet*,

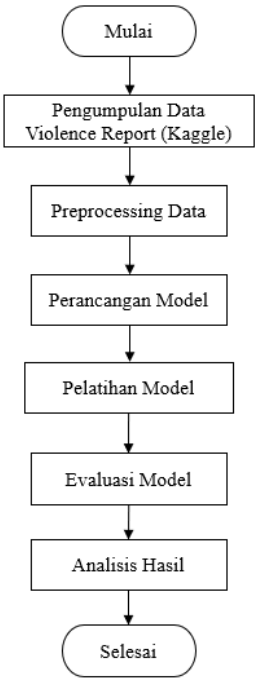
studi ini menemukan bahwa algoritma SVM menunjukkan akurasi tertinggi, yaitu sebesar 88%. Meskipun belum memanfaatkan model transformer, penelitian ini memperlihatkan bahwa pendekatan tradisional masih mampu memberikan hasil yang kompetitif. Hal ini sekaligus memperkuat urgensi untuk mengeksplorasi pendekatan yang lebih mutakhir, seperti IndoBERT, untuk menangani kompleksitas bahasa Indonesia dalam konteks yang lebih formal dan institusional [12].

Studi lain oleh Salam dkk. melakukan analisis sentimen terhadap opini masyarakat mengenai program Bantuan Langsung Tunai (BLT) BBM menggunakan algoritma *Support Vector Machine* pada data berbahasa Indonesia dari media sosial Instagram. Meskipun menggunakan jumlah data yang lebih terbatas (356 data), model berhasil mencapai akurasi sebesar 85,98% dan memperlihatkan dominasi sentimen negatif sebesar 78,61%. Penelitian ini menekankan pentingnya deteksi sentimen dalam kebijakan sosial dan menunjukkan bagaimana model pembelajaran mesin klasik dapat digunakan untuk memahami persepsi publik terhadap layanan pemerintah. Meski pendekatan ini belum memanfaatkan representasi berbasis transformer, temuan tersebut tetap relevan sebagai landasan untuk pengembangan sistem klasifikasi sentimen yang lebih kompleks dan kontekstual [13].

Berdasarkan latar belakang tersebut, penelitian ini difokuskan pada pengembangan model klasifikasi otomatis terhadap laporan kekerasan menggunakan pendekatan NLP berbasis BERT. Tujuan khusus dari penelitian ini adalah untuk mengembangkan sistem klasifikasi yang mampu mengidentifikasi jenis kekerasan secara otomatis dari teks laporan, mendeteksi sentimen atau emosi pelapor berdasarkan narasi yang disampaikan, serta meningkatkan efisiensi dan konsistensi klasifikasi dibandingkan metode manual. Berbeda dengan penelitian sebelumnya yang sebagian besar berfokus pada deteksi kekerasan dalam teks media sosial atau terbatas pada satu jenis kekerasan, studi ini menawarkan pendekatan baru dengan menerapkan model IndoBERT pada data laporan kekerasan aktual yang diterima oleh lembaga perlindungan perempuan dan anak di Indonesia. Hingga saat ini, belum ditemukan penelitian yang secara khusus menerapkan model transformer untuk menganalisis laporan kekerasan berbahasa Indonesia dalam konteks kelembagaan. Oleh karena itu, hasil dari penelitian ini diharapkan dapat berkontribusi dalam pengembangan sistem pendukung keputusan untuk lembaga perlindungan perempuan dan anak, sehingga proses advokasi dapat dilakukan secara lebih responsif dan terstandar.

2. Metode Penelitian

Penelitian ini dilakukan melalui serangkaian tahapan sistematis yang dimulai dari pengumpulan data laporan kekerasan dari Kaggle, dilanjutkan dengan proses pra-pemrosesan data untuk menyiapkan input yang bersih dan terstruktur. Selanjutnya, dilakukan perancangan model klasifikasi yang kemudian dilatih menggunakan data tersebut. Setelah model selesai dilatih, dilakukan evaluasi kinerja untuk menilai akurasi dan efektivitas model. Tahap akhir meliputi analisis hasil untuk menarik kesimpulan atas performa model yang dibangun. Alur proses lengkap ditunjukkan pada *flowchart* berikut.



Gambar 1. Diagram Alir Tahapan Penelitian

2.1. Pengumpulan Data

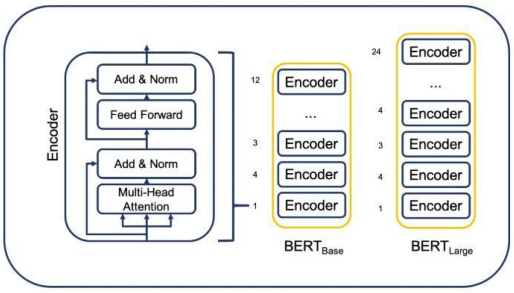
Data yang digunakan dalam penelitian ini berasal dari dataset publik berjudul *Violence Report on Women and Children* yang tersedia di *platform* Kaggle [14]. Dataset ini terdiri dari 500 entri berupa narasi laporan pengaduan dalam bahasa Indonesia, yang ini merupakan dataset dummy yang dibuat berdasarkan visualisasi data SIMFONI-PPA. Bagian "isi_laporan" pada data dibuat secara manual berdasarkan jurnal dan *google review* pada tanya jawab Kementerian Pemberdayaan Perempuan dan Perlindungan Anak di Google. Setiap entri telah diklasifikasikan ke dalam 5 jenis kekerasan (Seksual, Fisik, Psikis, Penelantaran, dan Eksploitasi) serta 2 kategori sentimen (*High* dan *Low*), sehingga memungkinkan untuk dilakukan analisis berbasis klasifikasi teks secara lebih mendalam.

2.2. Pra-pemrosesan Data

Pada tahap ini, teks laporan terlebih dahulu dibersihkan dari elemen-elemen yang tidak relevan seperti karakter khusus (misalnya simbol atau tanda baca yang tidak diperlukan), tautan yang dapat mengganggu pemaknaan konteks, serta huruf kapital yang diubah menjadi huruf kecil agar seragam. Pembersihan ini dilakukan melalui fungsi cleaning yang dirancang untuk menyaring dan menghapus komponen teks yang tidak mendukung proses analisis semantik. Setelah proses pembersihan, dilakukan tokenisasi terhadap teks menggunakan tokenizer dari model *indobenchmark/indobert-base-p1*, yaitu model BERT pra-latih untuk bahasa Indonesia. Tokenisasi ini bertujuan mengubah setiap kalimat menjadi satuan token yang dapat dikenali dan diproses oleh arsitektur model berbasis transformer. Penggunaan tokenizer BERT spesifik untuk bahasa Indonesia sangat penting karena mempertahankan nuansa sintaksis dan semantik lokal dalam representasi vektor (Purnomo et al., 2024). Selain itu, label kategorikal pada dataset, baik untuk jenis kekerasan (5 kelas) maupun untuk level sentimen (2 kelas), dikodekan ke dalam bentuk numerik menggunakan metode *Label Encoding* dengan bantuan fungsi *LabelEncoder* dari pustaka *scikit-learn*. Proses ini memungkinkan setiap label teks diubah menjadi angka unik yang dapat digunakan oleh algoritma klasifikasi dalam proses pelatihan dan evaluasi model. Seluruh tahapan pra-pemrosesan ini dirancang untuk mempersiapkan data seoptimal mungkin sehingga model dapat dilatih secara efektif dengan input yang sudah terstruktur dan bersih.

2.3. Perancangan dan Pelatihan Model

Pada tahap ini, dilakukan perancangan model klasifikasi *multi-label*, yaitu model yang memungkinkan satu entri laporan memiliki lebih dari satu jenis kekerasan secara bersamaan (misalnya kekerasan fisik dan psikis), serta diklasifikasikan ke dalam salah satu dari dua kategori sentimen. Untuk itu, digunakan arsitektur BERT yang dirancang untuk melakukan klasifikasi terhadap dua keluaran sekaligus, yakni jenis kekerasan dan label sentimen. Model yang digunakan adalah *indobenchmark* atau *indobert-base-p1*, yaitu model BERT pra-latih yang secara khusus dikembangkan untuk Bahasa Indonesia.



Gambar 2. Arsitektur BERT [15]

Untuk mendukung proses klasifikasi ganda, model ini dikembangkan dengan dua lapisan *dense* terpisah setelah representasi keluaran dari BERT, masing-masing untuk memprediksi kategori kekerasan dan sentimen.

Sebelum proses pelatihan, data dibagi menjadi dua bagian menggunakan *train-test split*, dengan proporsi 80% untuk pelatihan dan 20% untuk pengujian. Selanjutnya, data diproses menjadi Dataset khusus yang memuat token input dan *attention mask* hasil tokenisasi, serta label numerik yang telah dikodekan menggunakan *LabelEncoder*. Tokenisasi dilakukan menggunakan tokenizer BERT yang sesuai dengan model, dengan panjang maksimum input ditetapkan 128 token. Model dilatih menggunakan *DataLoader PyTorch* dengan *batch size* 8, dan seluruh proses pelatihan dilakukan dalam kerangka *multi-task learning*—yaitu satu model belajar untuk dua target secara simultan. Fungsi *loss* yang digunakan adalah gabungan dari dua *cross-entropy loss*, masing-masing untuk label kekerasan dan sentimen. Proses pelatihan dilakukan selama beberapa *epoch*, dan model dievaluasi berdasarkan metrik akurasi, presisi, *recall*, dan *F1-score* untuk masing-masing label. Strategi ini memungkinkan efisiensi dalam pelatihan serta memperbaiki generalisasi model terhadap narasi laporan yang kompleks dan beragam.

2.4. Evaluasi Model

Evaluasi model dilakukan dengan mengukur performa prediksi terhadap dua label keluaran, yaitu kategori kekerasan dan sentimen. Setelah proses pelatihan selesai, model diuji pada data uji yang telah dipisahkan sebelumnya. Evaluasi mencakup perhitungan metrik akurasi, *precision*, *recall*, dan *F1-score*, baik secara makro maupun per kelas. Penggunaan metrik ini bertujuan untuk mengukur seberapa baik model dapat mengenali masing-masing kelas, terutama dalam konteks distribusi kelas yang tidak seimbang. Evaluasi dilakukan secara terpisah untuk dua jenis label, sehingga memungkinkan identifikasi performa model pada masing-masing tugas secara lebih mendalam. Prediksi model dihasilkan melalui fungsi model.eval() dengan *no_grad* untuk menghindari perhitungan gradien, dan hasil prediksi dibandingkan dengan label asli untuk membentuk *confusion matrix* serta perhitungan metrik evaluasi. Selain itu, laporan klasifikasi (*classification report*) dari *sklearn* digunakan untuk menampilkan metrik secara lengkap per kelas. Hasil evaluasi memberikan gambaran seberapa akurat model dalam mendeteksi pola kekerasan dan menentukan kecenderungan sentimen dari teks narasi pengaduan, sekaligus menjadi dasar dalam menentukan efektivitas model yang dikembangkan.

3. Hasil dan Pembahasan

Penelitian ini berhasil mengembangkan dan melatih model klasifikasi ganda berbasis IndoBERT untuk mengidentifikasi jenis kekerasan dan tingkat sentimen dari laporan naratif kekerasan terhadap perempuan dan anak. Model dilatih selama 10 epoch menggunakan pendekatan multi-task learning. Evaluasi kinerja dilakukan terhadap data uji sebesar 20% dari total data, dan model terbaik disimpan dari *epoch* ke-6, yang ditentukan berdasarkan nilai *F1-score* tertinggi pada klasifikasi sentimen.

3.1. Performa Model

Pelatihan model multi-output berbasis IndoBERT dilakukan untuk memprediksi dua label keluaran sekaligus, yaitu kategori kekerasan dan polaritas sentimen dari laporan teks berbahasa Indonesia. Model menunjukkan performa yang tinggi dan seimbang dalam kedua tugas klasifikasi tersebut, dengan peningkatan signifikan pada fase awal pelatihan. Performa terbaik tercapai pada *epoch* ke-6, yang menjadi titik optimal berdasarkan keseluruhan metrik evaluasi.

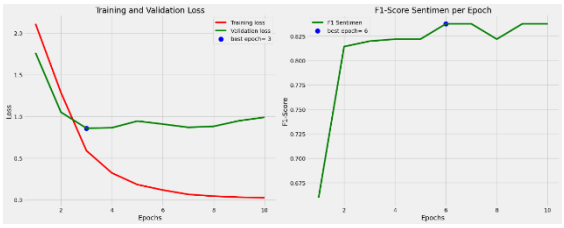
Pada tahap ini, model menghasilkan *precision* sebesar 0.8533, *recall* sebesar 0.8529, dan *F1-score* sebesar 0.8528 untuk klasifikasi kategori kekerasan. Sementara itu, pada tugas klasifikasi sentimen pelapor, model memperoleh *precision* sebesar 0.8372, *recall* sebesar 0.8382, dan *F1-score* sebesar 0.8375. Nilai metrik yang tinggi dan seimbang ini menunjukkan bahwa model memiliki kemampuan klasifikasi yang andal untuk kedua label. Kemampuan dalam membedakan berbagai jenis kekerasan secara konsisten serta menangkap ekspresi emosional pelapor dengan akurasi tinggi merupakan capaian penting. Khususnya untuk tugas klasifikasi sentimen, akurasi yang baik memiliki implikasi sosial yang signifikan karena ekspresi sentimen sering kali berkaitan dengan tingkat urgensi, kekhawatiran, dan dampak psikologis dari laporan yang diterima.

Tabel 1. Metrik Kinerja Terbaik Model

Metrik	Kategori Kekerasan	Sentimen Pelapor
Precision	0.8533	0.8372
Recall	0.8529	0.8382
F1-Score	0.8528	0.8375
Validation Loss	0.9079	—

Berdasarkan grafik *training* dan *validation loss*, *training loss* mengalami penurunan drastis dari awal pelatihan hingga akhir, menandakan bahwa model mampu mempelajari pola dari data pelatihan secara efektif. *Validation loss* mencapai nilai terendah pada epoch ketiga dan tetap relatif stabil hingga akhir pelatihan, tanpa menunjukkan peningkatan tajam yang mengindikasikan terjadinya *overfitting*. Stabilitas ini menunjukkan bahwa model tidak hanya mampu belajar dari data pelatihan, tetapi juga mempertahankan kemampuan generalisasinya terhadap data validasi.

Submitted : 05-08-2025 | Reviewed : 16-08-2025 | Accepted : 31-08-2025



(a) (b)

Gambar 3. (a) Grafik Training dan Validation Loss, (b) Grafik F1-Score Sentiment per Epoch

Gambar 3 memperlihatkan dua tren utama dari hasil pelatihan model. Gambar 3(a) menampilkan penurunan *training loss* dan *validation loss* terhadap jumlah *epoch*. Penurunan yang konsisten ini mencerminkan proses pembelajaran yang stabil. Sementara itu, Gambar 3(b) menunjukkan perkembangan nilai *F1-score* untuk klasifikasi sentimen. Terlihat adanya peningkatan tajam pada fase awal pelatihan, khususnya hingga *epoch* ke-6, kemudian mencapai titik stabil. Hal ini menunjukkan bahwa model telah mencapai konvergensi secara efisien dan pelatihan tambahan setelah titik tersebut tidak lagi memberikan manfaat yang berarti.

Secara keseluruhan, hasil pelatihan menunjukkan bahwa model *multi-output* yang dikembangkan tidak hanya efektif dalam mempelajari struktur semantik dari laporan kekerasan, tetapi juga sensitif terhadap dimensi emosional dari teks. Dengan kestabilan performa setelah *epoch* ke-6 dan tidak adanya indikasi *overfitting* yang berarti, model ini dapat dianggap cukup andal dan siap diterapkan pada sistem klasifikasi otomatis laporan kekerasan yang membutuhkan prediksi simultan terhadap tipe kekerasan dan sentimen pelapor.

3.2. Hasil Klasifikasi Laporan Baru

Sebagai simulasi praktis dan validasi terhadap generalisasi model, lima narasi laporan fiktif diuji menggunakan model yang telah dilatih. Tujuan pengujian ini adalah untuk mengevaluasi kemampuan model dalam menghadapi data nyata (*real-world inference*) yang belum pernah dilihat sebelumnya.

Tabel 2. Hasil Klasifikasi Laporan Baru

Narasi Singkat	Kategori Kekerasan	Sentimen
Ibu rumah tangga mengalami kekerasan fisik oleh suami, wajah lebam, anak-anak ketakutan	Fisik	*High*
Anak usia 10 tahun mengalami pelecehan oleh orang dewasa	Seksual	*High*
Anak 14 tahun dipaksa menikah karena utang, harus putus sekolah	Psikis	*High*
Perempuan mengalami kekerasan verbal dan tekanan mental di rumah tangga	Psikis	*Low*
Anak-anak kecil dipaksa bekerja dan tidak disekolahkan	Psikis	*High*

Model mampu mengklasifikasikan laporan-laporan tersebut dengan jenis kekerasan dan sentimen yang relevan secara semantik. Sebagai contoh:

a) Laporan mengenai kekerasan fisik oleh suami diklasifikasikan sebagai "Fisik" dan "*High*", mencerminkan trauma yang kuat dan eksplisit.
b) Kasus kekerasan verbal dan tekanan mental diklasifikasikan sebagai "Psikis" dan "*Low*", yang menunjukkan emosi lebih tertahan atau tersembunyi.
c) Laporan eksploitasi kerja anak berhasil terdeteksi sebagai bentuk kekerasan psikis, menandakan sensitivitas model terhadap konteks sosial dan hukum.

3.3. Pembahasan

Hasil evaluasi menunjukkan bahwa model IndoBERT yang dikembangkan mampu melakukan klasifikasi multi-label terhadap laporan kekerasan secara efektif. Kinerja yang seimbang antara klasifikasi jenis kekerasan dan sentimen pelapor, ditunjukkan melalui nilai *F1-score* yang tinggi, menunjukkan bahwa model tidak hanya memahami struktur linguistik Bahasa Indonesia, tetapi juga mampu menangkap nuansa semantik dari laporan naratif. Performa yang stabil setelah *epoch* ke-6 menandakan bahwa model telah mencapai konvergensi yang baik tanpa indikasi *overfitting* yang signifikan, yang memperkuat efektivitas pendekatan *multi-task learning* dalam konteks ini. Lebih lanjut, pengujian terhadap lima narasi baru memberikan gambaran bahwa model mampu menggeneralisasi terhadap data yang tidak pernah dilihat sebelumnya. Hal ini penting dalam konteks dunia nyata, di mana variasi gaya bahasa dan bentuk pelaporan sangat tinggi. Model mampu membedakan jenis kekerasan yang kompleks dan menyesuaikan penilaian sentimen sesuai dengan kekuatan emosi yang diekspresikan, seperti

membedakan sentimen "*high*" untuk kasus traumatis dan "*low*" untuk kasus verbal yang lebih tersembunyi. Hasil ini memperkuat keyakinan bahwa pendekatan berbasis transformer seperti IndoBERT cocok untuk permasalahan sosial berbasis teks yang kompleks di Indonesia.

4. Kesimpulan

Penelitian ini berhasil menjawab permasalahan klasifikasi manual yang lambat dan tidak konsisten dalam penanganan laporan kekerasan terhadap perempuan dan anak, dengan mengembangkan model otomatis berbasis IndoBERT yang mampu melakukan klasifikasi ganda terhadap jenis kekerasan dan sentimen pelapor. Hasil evaluasi menunjukkan bahwa model mampu mencapai performa yang tinggi dan seimbang, dengan nilai *F1-score* sebesar 0.8528 untuk klasifikasi kekerasan dan 0.8375 untuk klasifikasi sentimen, serta menunjukkan stabilitas pelatihan tanpa *overfitting*. Keberhasilan model dalam menggeneralisasi laporan baru yang belum pernah dilihat sebelumnya menunjukkan kemampuannya dalam memahami konteks linguistik dan emosional dari teks naratif berbahasa Indonesia. Dengan demikian, pendekatan ini tidak hanya mempercepat dan meningkatkan akurasi proses klasifikasi, tetapi juga memberikan informasi yang lebih kaya mengenai kondisi psikologis pelapor. Temuan ini mendukung pemanfaatan model transformer lokal seperti IndoBERT sebagai solusi inovatif untuk membantu lembaga perlindungan dalam pengambilan keputusan yang lebih cepat, konsisten, dan kontekstual dalam menangani laporan kekerasan.

Daftar Rujukan

[1] J. Susanto, "Data Kementrian PPPA: Kekerasan Anak Capai 28.831 Kasus Pada 2024," *NU Online*, 2024. https://nu.or.id/nasional/data-kementerian-pppa-kekerasan-anak-capai-28-831-kasus-pada-2024-npRIs?
[2] Simfoni-PPA, "Data Kasus," *Kementrian PPPA*, 2025. https://kekerasan.kemenpppa.go.id/ringkasan.
[3] E. F. Alexander and M. D. Johnson, "On categorizing intimate partner violence: A systematic review of exploratory clustering and classification studies," *J Fam Psychol*, vol. 37, no. 5, pp. 743–752, 2023, [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/36892924/.
[4] J. M. Kafka, J. P. Schleimer, O. Toomet, K. Chen, A. Ellyson, and A. Rowhani-Rahbar, "Measuring interpersonal firearm violence: natural language processing methods to address limitations in criminal charge data," *J. Am. Med. Informatics Assoc.*, vol. 31, no. 10, pp. 2374–2378, 2024, doi: https://doi.org/10.1093/jamia/ocae268.
[5] I. Astuti, "Kementerian PPPA Luncurkan Laporan Data Kekerasan," *Media Indonesia*, 2024. https://mediaindonesia.com/humaniora/692411/kementerian-pppa-luncurkan-laporan-data-kekerasan (accessed Jun. 08, 2025).
[6] L. Yang, X. Zhou, J. Fan, X. Xie, and S. Zhu, "Can bidirectional encoder become the ultimate winner for downstream applications of foundation models?," *2024 2nd Int. Conf. Found. Large Lang. Model. FLLM 2024*, pp. 526–534, 2024, doi: 10.1109/FLLM63129.2024.10852511.
[7] H. Kibriya, A. Siddiqa, W. Z. Khan, and M. K. Khan, "Towards safer online communities: Deep learning and explainable AI for hate speech detection and classification," *Comput. Electr. Eng.*, vol. 116, 2024, doi: https://doi.org/10.1016/j.compeleceng.2024.109153.
[8] A. L. Candidato, A. Gupta, X. Liu, and S. Shah, "AIR-JPMC@SMM4H'22: Classifying Self-Reported Intimate Partner Violence in Tweets with Multiple BERT-based Models," *Proc. Seventh Work. Soc. Media Min. Heal. Appl. Work. Shar. Task*, pp. 135–137, 2022, [Online]. Available: https://aclanthology.org/2022.smm4h-1.37.
[9] Y. Yang and X. Wang, "Research on Violent Text Detection System Based on BERT-fasttext Model," pp. 1–11, 2024.
[10] F. Farasalsabila, E. Utami, M. Informatika, and U. A. Yogyakarta, "Deteksi Cyberbullying Menggunakan BERT dan Bi-LSTM," vol. 17, pp. 1–6, 2024.
[11] M. I. Wijanarko, L. Susanto, P. A. Pratama, I. Idris, T. Hong, and D. Wijaya, "Monitoring Hate Speech in Indonesia : An NLP-based Classification of Social Media Texts," pp. 142–152, 2024.
[12] R. A. Husen, R. Astuti, L. Marlia, R. Rahmaddeni, and L. Efrizoni, "Husen, R, A et al., 2022," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 3, no. 2, pp. 211–218, 2023.
[13] R. R. Salam, M. F. Jamil, Y. Ibrahim, R. Rahmaddeni, S. Soni, and H. Herianto, "Analisis Sentimen Terhadap Bantuan Langsung Tunai (BLT) Bahan Bakar Minyak (BBM) Menggunakan Support Vector Machine," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 3, no. 1, pp. 27–35, 2023, doi: 10.57152/malcom.v3i1.590.
[14] T. A. Nugroho, "Violence Report on Women and Children," *Kaggle*, 2021. https://www.kaggle.com/datasets/tassamoo/violence-report-on-women-and-children.
[15] K. H. Kumar, "BERT Language Model: A Comprehensive Technical Analysis NLP and BERT," *Medium.com*, 2023. https://medium.com/@mysterious_obscure/bert-language-model-a-comprehensive-technical-analysis-bc9df450946b (accessed Jun. 09, 2025).

Submitted : 05-08-2025 | Reviewed : 16-08-2025 | Accepted : 31-08-2025

Submitted : 05-08-2025 | Reviewed : 16-08-2025 | Accepted : 31-08-2025