JITET (Jurnal Informatika dan Teknik Elektro Terapan)
Vol. 14 No. 3, pISSN: 2303-0577 eISSN: 2830-7062 http://dx.doi.org/10.23960/jitet.v14i3.10847

IMPLEMENTASI *INDOBERT* UNTUK SISTEM TRIASE CERDAS BERBASIS *NATURAL LANGUAGE PROCESSING* (NLP): STUDI KASUS PUSKESMAS KUTA UTARA

I Made Hendra Wijaya[1*], I Wayan Septa Malan Vergantana[2]

[1,2]Program Studi Sistem Informasi, Institut Teknologi dan Kesehatan Bintang Persada, Denpasar, Bali, Indonesia

Keywords:
fine-tuning;
IndoBERT;
klasifikasi teks;
Natural Language Processing;
triase.

Corespondent Email:
hendrawijaya@bintangpersada.ac.id

Abstrak. Triase merupakan tahap kritis dalam pelayanan kesehatan primer yang menentukan keselamatan pasien, namun pelaksanaannya di puskesmas masih bergantung pada penilaian klinis dan pengalaman perawat sehingga rentan terhadap subjektivitas serta variabilitas antarpetugas. Kesalahan triase berdampak serius: *under-triage* menunda penanganan pasien kritis, sedangkan *over-triage* memboroskan sumber daya yang terbatas. Penelitian ini mengembangkan sistem triase cerdas berbasis *Natural Language Processing* (NLP) melalui *fine-tuning* model *IndoBERT* (*indobert-base-p1*) untuk mengklasifikasikan keluhan pasien berbahasa Indonesia berbentuk teks bebas ke dalam tiga tingkat triase (Hijau, Kuning, Merah), dengan studi kasus di Puskesmas Kuta Utara. Model dilatih pada *dataset* sintetis seimbang berisi 10.000 keluhan dengan pembagian terstratifikasi 70/15/15 dan kombinasi teknik optimasi berupa *dynamic padding*, *cosine learning rate scheduling* dengan *warmup*, *early stopping*, *label smoothing*, serta pembobotan kelas. Model mencapai *F1-macro* 1,0000 pada *test set* (1.500 sampel), melampaui target minimum 0,85, dengan waktu pelatihan ±2 menit 51 detik dan tanpa kasus *under-triage* maupun *over-triage*. Capaian tersebut dikualifikasi secara jujur karena hanya berlaku pada data sintetis berbasis *template*, sehingga lebih tepat dibaca sebagai bukti keberhasilan rekayasa *fine-tuning* daripada ukuran akurasi klinis final. Sistem direkomendasikan diposisikan sebagai alat bantu rekomendasi awal (*assistive triage*) dengan keputusan akhir tetap pada perawat, dan penelitian lanjutan diarahkan pada uji validitas menyeluruh menggunakan data keluhan riil.

Abstract. *Triage is a critical stage in primary healthcare that determines patient safety, yet its implementation in community health centers still relies heavily on nurses' clinical judgment and experience, making it prone to subjectivity and inter-rater variability. Triage errors carry serious consequences: under-triage delays treatment for critical patients, while over-triage wastes limited resources. This study develops a Natural Language Processing (NLP)-based smart triage system by fine-tuning the IndoBERT model (indobert-base-p1) to classify Indonesian free-text patient complaints into three triage levels (Green, Yellow, Red), using Puskesmas Kuta Utara as a case study. The model was trained on a balanced synthetic dataset of 10,000 complaints with a stratified 70/15/15 split, combining dynamic padding, cosine learning rate scheduling with warmup, early stopping, label smoothing, and class weighting. The model achieved a macro F1-score of 1.0000 on the test set (1,500 samples), exceeding the 0.85 minimum target, with a training time of ±2 minutes 51 seconds and no under-triage or over-triage cases. This result is honestly qualified, since it applies only to template-based synthetic data and is better read as evidence*

of successful fine-tuning engineering rather than a final clinical accuracy measure. The system is recommended to be positioned as an assistive triage tool, with final decisions remaining with nurses, and further research is directed toward comprehensive validity testing using real complaint data.

Keywords: *classification, fine-tuning, IndoBERT, Natural Language Processing, triage*

1. PENDAHULUAN

Triase merupakan proses pemilahan pasien berdasarkan beratnya cedera atau penyakit guna menentukan jenis penanganan dan prioritas intervensi kegawatdaruratan. Di Indonesia, pelaksanaan triase pada fasilitas pelayanan kesehatan primer diatur melalui Peraturan Menteri Kesehatan Nomor 47 Tahun 2018 tentang Pelayanan Kegawatdaruratan, yang mengklasifikasikan pasien ke dalam kategori Merah (*emergency*), Kuning (*urgent*), dan Hijau (*non-urgent*) [1]. Namun, pelaksanaan triase di puskesmas masih sangat bergantung pada penilaian klinis dan pengalaman perawat, sehingga rentan terhadap subjektivitas dan variabilitas antarpetugas. Kesalahan triase membawa konsekuensi serius: *under-triage* menunda penanganan pasien kritis dan meningkatkan risiko perburukan hingga kematian, sedangkan *over-triage* memboroskan sumber daya yang terbatas [2]. Studi melaporkan tingkat *under-triage* berkisar 7–10% dan *over-triage* mencapai 52–74%, jauh melampaui ambang batas aman yang direkomendasikan.

Persoalan ini diperberat oleh keterbatasan tenaga kesehatan di Indonesia, di mana hanya sekitar 65% puskesmas yang memenuhi standar minimum ketenagaan WHO, sehingga perawat dituntut mengambil keputusan triase dalam waktu kurang dari lima menit di tengah tekanan kerja yang tinggi. Di Puskesmas Kuta Utara, yang melayani wilayah dengan mobilitas penduduk dan volume kunjungan tinggi, kebutuhan akan instrumen pendukung keputusan triase yang cepat, konsisten, dan objektif menjadi semakin mendesak. *Natural Language Processing* (NLP) telah terbukti mampu meningkatkan akurasi dan konsistensi klasifikasi triase melalui pengolahan keluhan pasien dalam bentuk teks bebas (*free-text*) dibandingkan pendekatan konvensional [3], [4]. Namun, penerapan NLP pada bahasa Indonesia menghadapi tantangan khusus sebagai bahasa dengan sumber daya komputasional terbatas (*low-resource language*), terutama karena teks medis berbahasa Indonesia memiliki ragam bahasa informal, singkatan klinis, dan istilah lokal yang khas [5].

IndoBERT, sebagai model bahasa pralatih (*pre-trained language model*) berbahasa Indonesia yang dikembangkan dalam proyek *IndoNLU* [5], menawarkan potensi besar untuk mengatasi tantangan tersebut. Kesenjangan penelitian (*gap*) yang teridentifikasi adalah belum tersedianya sistem pendukung keputusan triase berbasis model bahasa pralatih berbahasa Indonesia pada layanan kesehatan primer, mengingat sebagian besar sistem serupa dikembangkan untuk bahasa bersumber daya melimpah seperti bahasa Inggris [3], [6]. Kebaruan (*novelty*) penelitian ini terletak pada penerapan *IndoBERT* secara spesifik sebagai asisten keputusan triase perawat di puskesmas. Berdasarkan latar belakang tersebut, penelitian ini merumuskan tiga permasalahan: (1) bagaimana merancang sistem triase berbasis *IndoBERT* yang mampu mengklasifikasikan keluhan pasien berbahasa Indonesia; (2) seberapa baik akurasi model *IndoBERT* dalam mengklasifikasikan tingkat kegawatdaruratan pasien yang diukur dengan metrik *F1-macro*; dan (3) bagaimana kelayakan penerapan sistem sebagai asisten keputusan triase bagi perawat di Puskesmas Kuta Utara. Penelitian ini bertujuan merancang, membangun, dan mengevaluasi sistem tersebut, sekaligus memberikan kontribusi bagi pengembangan sumber daya NLP medis berbahasa Indonesia.

2. TINJAUAN PUSTAKA

2.1. Triase dan Variabilitas Keputusan Klinis

Triase bertujuan memastikan pasien dengan kondisi paling mengancam nyawa memperoleh penanganan lebih dahulu. Permenkes Nomor

47 Tahun 2018 menetapkan tiga kategori prioritas pada layanan kegawatdaruratan, yaitu Merah, Kuning, dan Hijau, beserta rentang waktu tanggap yang menyertainya [1]. Tinjauan sistematis Hinson dkk. [2] terhadap kinerja triase menunjukkan bahwa akurasi keputusan triase bervariasi antarpetugas maupun antarinstitusi, dengan *under-triage* dan *over-triage* sebagai dua bentuk kesalahan yang konsisten ditemukan. Temuan tersebut menegaskan bahwa instrumen bantu yang objektif dan konsisten dibutuhkan, khususnya pada fasilitas dengan keterbatasan tenaga dan beban kunjungan tinggi.

2.2. Pemanfaatan *NLP* pada Triase Layanan Kegawatdaruratan

Keluhan pasien pada tahap triase umumnya terekam sebagai teks bebas, sehingga menjadi objek yang sesuai bagi NLP. Sitthiprawiat dkk. [3] mengembangkan dan memvalidasi model triase berbasis kecerdasan artifisial untuk memprediksi luaran kritis di instalasi gawat darurat, sementara Chang dkk. [4] memadukan *machine learning* dan NLP pada data triase untuk memprediksi *clinical disposition* pasien. Tinjauan pelingkupan (*scoping review*) Liang dkk. [6] merangkum tren penerapan NLP pada triase dan mencatat bahwa mayoritas studi masih berfokus pada bahasa Inggris. Untuk konteks bahasa non-Inggris, Navarro dkk. [10] membandingkan model berbasis *transformer* (*ALBERT*) dengan pendekatan *machine learning* klasik pada catatan klinis berbahasa Spanyol dan menunjukkan bahwa model pralatih tetap unggul pada teks klinis yang tidak terstruktur [14].

2.3. Arsitektur *Transformer*, *BERT*, dan *IndoBERT*

Arsitektur *transformer* yang diperkenalkan Vaswani dkk. [8] menggantikan rekurensi dengan mekanisme *self-attention* sehingga mampu memodelkan relasi antartoken secara paralel. Devlin dkk. [7] mengembangkan *BERT* (*Bidirectional Encoder Representations from Transformers*) yang dilatih secara dwiarah melalui tugas *masked language modeling*, sehingga representasi kontekstual yang dihasilkan dapat diadaptasi ke tugas hilir hanya dengan menambahkan lapisan klasifikasi. Untuk bahasa Indonesia, Wilie dkk. [5] merilis *IndoBERT* beserta tolok ukur *IndoNLU* sebagai sumber daya rujukan pemahaman bahasa alami. Efektivitas *IndoBERT* pada domain spesifik telah dilaporkan pada klasifikasi teks kesehatan berbahasa Indonesia [11] serta pada deteksi tema ide bunuh diri pada unggahan *Twitter* berbahasa Indonesia melalui kombinasi *transfer learning* dan pengklasifikasi *Long Short-Term Memory* [12]. Pemilihan varian *indobert-base-p1* dibandingkan varian yang lebih ringan (misalnya *IndoBERT-lite*) atau model RoBERTa berbahasa Indonesia (*IndoBERT-RoBERTa*) didasarkan pada pertimbangan bahwa varian *base* menawarkan keseimbangan optimal antara kapasitas representasi bahasa dan kebutuhan komputasi, sehingga tetap dapat dijalankan secara efisien pada infrastruktur GPU kelas menengah seperti *Google Colab Pro* tanpa mengorbankan akurasi klasifikasi. Varian *lite* memang lebih ringkas dan berpotensi lebih sesuai untuk keterbatasan sumber daya komputasi di fasilitas kesehatan primer, namun umumnya disertai penurunan kapasitas representasi yang berisiko menurunkan sensitivitas terhadap nuansa bahasa informal khas keluhan pasien puskesmas, sehingga pemilihan varian *base* dipandang sebagai titik tengah yang lebih aman pada tahap pengembangan awal ini.

2.4. *Fine-tuning* dan Teknik Optimasinya

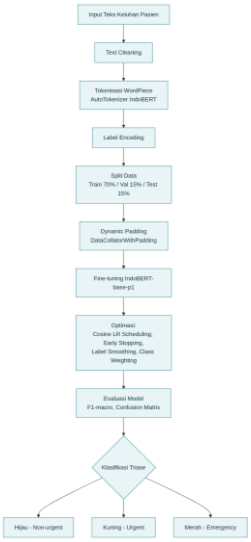
Fine-tuning adalah proses melanjutkan pelatihan model pralatih pada data tugas tertentu. Dodge dkk. [9] menunjukkan bahwa hasil *fine-tuning* sensitif terhadap inisialisasi bobot dan urutan data, serta merekomendasikan *early stopping* sebagai strategi praktis untuk menekan variansi hasil. [13] Temuan tersebut menjadi dasar pemilihan teknik optimasi pada penelitian ini, yaitu *dynamic padding*, *cosine learning rate scheduling* dengan *warmup*, *early stopping*, *label smoothing*, dan pembobotan kelas, serta pelaporan hasil yang disertai kualifikasi validitas [15].

3. METODE PENELITIAN

3.1. Rancangan Penelitian

Penelitian ini menggunakan pendekatan eksperimen kuantitatif melalui *fine-tuning* model bahasa pralatih *IndoBERT* (*indobert-base-p1*, 124.443.651 parameter) untuk tugas

klasifikasi teks tiga kelas. Penelitian dilaksanakan sebagai studi kasus di Puskesmas Kuta Utara, Bali, menggunakan infrastruktur GPU *Google Colab Pro*. Alur penelitian terdiri atas lima tahapan utama, yaitu pembuatan *dataset*, praproses data, *fine-tuning* model, evaluasi performa, dan uji validitas hasil.



Gambar 1. Diagram alir (*flowchart*) *pipeline* sistem *IndoBERT*, dari input teks keluhan hingga *output* klasifikasi triase

3.2. Dataset Penelitian

Dataset yang digunakan bersifat sintetis, dibuat menggunakan pendekatan berbasis *template* untuk merepresentasikan ragam bahasa khas puskesmas: bahasa Indonesia baku formal, bahasa informal dengan reduksi fonem ("udah" → "sudah"), singkatan klinis ("TD", "KU", "GCS"), dan istilah lokal wilayah Kuta Utara, Bali. *Dataset* terdiri atas 10.000 keluhan pasien dengan distribusi kelas seimbang (1:1:1) pada tiga label triase: Hijau (*non-urgent*, misalnya pilek dan batuk ringan), Kuning (*urgent*, misalnya nyeri abdomen sedang dan demam tinggi), serta Merah (*emergency*, misalnya sesak napas berat dan penurunan kesadaran).

3.3. Praproses dan Pembagian Data

Praproses data meliputi pembersihan teks (*text cleaning*), tokenisasi menggunakan *AutoTokenizer IndoBERT* berbasis *WordPiece* dengan *max_length* 128 token, serta pengodean label menggunakan *LabelEncoder*. *Dynamic padding* diterapkan melalui *DataCollatorWithPadding* agar setiap *batch* hanya dipadatkan hingga panjang token terpanjang dalam *batch* tersebut. Pembagian data dilakukan menggunakan *train_test_split* dengan parameter *stratify=labels* dan *random_state=42*, menghasilkan 7.000 sampel data latih (70%), 1.500 sampel data validasi (15%), dan 1.500 sampel data uji (15%).

3.4. Teknik Optimasi *Fine-tuning*

Proses *fine-tuning* menerapkan lima teknik optimasi secara bersamaan: (1) *dynamic padding* untuk efisiensi komputasi; (2) *cosine learning rate scheduling* dengan *warmup* 10% dari total langkah pelatihan dan *learning rate* puncak 2×10^{-5}; (3) *early stopping* berdasarkan *F1-macro* validasi dengan *patience* 3 *epoch*; (4) *label smoothing* dengan $\varepsilon = 0{,}1$ pada fungsi *CrossEntropyLoss*; dan (5) pembobotan kelas (*class weighting*) dengan formula sebagaimana ditunjukkan pada Persamaan (1).

$$\omega_i = N / (K \cdot n_i) \quad (1)$$

di mana N adalah total sampel, K adalah jumlah kelas, dan n_i adalah jumlah sampel pada kelas ke-i. Konfigurasi pelatihan menggunakan *batch size* 32, *mixed precision training* (FP16), *seed* 42, dan *epoch* maksimum 15.

3.5. Metrik Evaluasi dan Rancangan Uji Validitas

Evaluasi performa model menggunakan *classification_report* dan *confusion_matrix* dari *scikit-learn*, dengan *F1-macro* sebagai metrik utama karena menilai performa model secara setara pada seluruh kelas triase, termasuk kelas Merah yang secara klinis paling kritikal. Target performa minimum ditetapkan sebesar *F1-macro* ≥ 0,85 pada *test set*. Untuk menelusuri validitas hasil, penelitian ini juga merumuskan tiga hipotesis terkait potensi performa sempurna, yaitu kebocoran struktural *template* antar-*split*, separabilitas leksikal absolut antarkelas, dan absennya kasus ambigu atau *borderline* pada data sintetis.

4. HASIL DAN PEMBAHASAN

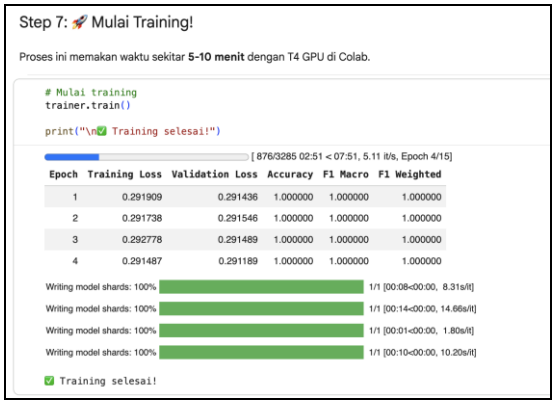
4.1. Komposisi Data dan Bobot Kelas

Pembagian data terstratifikasi menghasilkan komposisi aktual sesuai

rancangan, yaitu 7.000 sampel data latih, 1.500 sampel data validasi, dan 1.500 sampel data uji, dengan *support* 500 sampel per kelas pada *test set* yang membuktikan distribusi kelas 1:1:1 berhasil dipertahankan secara konsisten. Perhitungan *class weight* menghasilkan nilai identik 1,000 untuk ketiga kelas sebagai konsekuensi langsung dari distribusi kelas yang sudah seimbang.

4.2. Hasil Proses *Fine-tuning*

Proses *fine-tuning* berjalan dengan riwayat metrik yang mencapai *accuracy* dan *F1-macro* sempurna (1,0000) sejak *epoch* pertama dan bertahan konstan hingga pelatihan berhenti secara otomatis pada *epoch* ke-4 (dari maksimum 15 *epoch*), tepatnya setelah 876 langkah dari 3.285 langkah yang direncanakan, sebagaimana ditampilkan pada Gambar 2.



Gambar 2. Riwayat metrik proses *fine-tuning* model

Checkpoint terbaik disimpan pada *epoch* ke-1. Nilai *validation loss* konvergen pada kisaran 0,2909–0,2915, sangat dekat dengan batas bawah teoretis *cross-entropy loss* akibat *label smoothing* (H(q) ≈ 0,2909), yang menjadi bukti kuantitatif bahwa model mencapai separasi klasifikasi mendekati sempurna, bukan akurasi tinggi secara kebetulan. Keseluruhan proses pelatihan berlangsung sangat efisien, yakni sekitar 2 menit 51 detik, sebagai hasil kombinasi *dynamic padding*, *group_by_length*, dan *mixed precision training* (FP16).

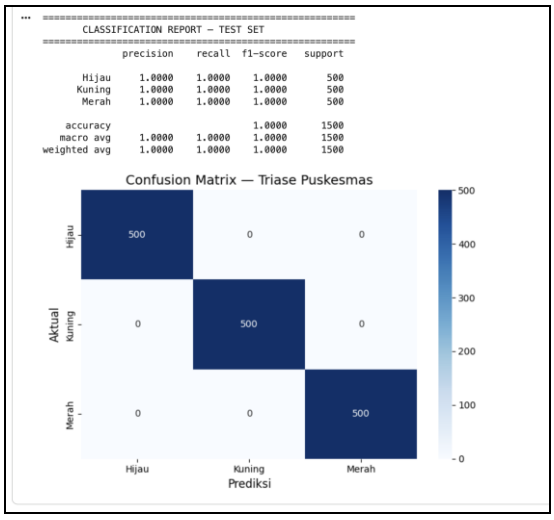
4.3. Hasil Evaluasi Model pada *Test Set*

Evaluasi model *checkpoint* terbaik pada *test set* terisolasi (1.500 sampel) menghasilkan *classification report* sebagaimana ditampilkan pada Tabel 1.

Tabel 1. *Classification report* model *IndoBERT* pada *test set*

Kelas	*Precision*	*Recall*	*F1-Score*
Hijau	1,0000	1,0000	1,0000
Kuning	1,0000	1,0000	1,0000
Merah	1,0000	1,0000	1,0000
Accuracy	–	–	1,0000
Macro avg	1,0000	1,0000	1,0000

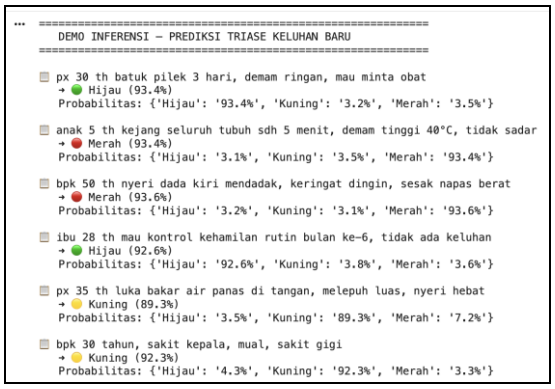
Seluruh metrik *precision*, *recall*, dan *F1-score* pada ketiga kelas triase mencapai nilai sempurna 1,0000, sehingga menghasilkan *F1-macro* sebesar 1,0000 yang jauh melampaui target performa minimum penelitian (*F1-macro* ≥ 0,85). *Confusion matrix* pada Gambar 3 menunjukkan diagonal yang sepenuhnya sempurna: seluruh 500 sampel pada masing-masing kelas diklasifikasikan dengan tepat, tanpa satu pun kasus *under-triage* (Merah salah diklasifikasikan sebagai Kuning atau Hijau) maupun *over-triage* (Hijau atau Kuning salah diklasifikasikan sebagai kelas yang lebih tinggi) pada seluruh sampel uji.



Gambar 3. *Confusion matrix* dan *classification report* pada *test set*

4.4. Kalibrasi Probabilitas dan Demo Inferensi

Analisis kalibrasi probabilitas terhadap demo inferensi enam keluhan baru menunjukkan *confidence* prediksi terkonsentrasi pada kisaran 89,3%–93,6%, konsisten dengan target *label smoothing* (≈93,3%). Hasil demo inferensi tersebut ditampilkan pada Gambar 4.



Gambar 4. Hasil uji coba prediksi pada keluhan baru

4.5. Uji Validitas dan Kualifikasi Capaian

Capaian *F1-macro* sempurna ini dikualifikasi secara jujur melalui perumusan tiga hipotesis, yaitu (H1) kebocoran struktural *template* antar-*split*, (H2) separabilitas leksikal absolut antarkelas, dan (H3) absennya kasus ambigu atau *borderline* pada data sintetis. Lima uji validitas dirancang untuk menelusuri hipotesis tersebut, yaitu deteksi kemiripan antar-*split*, *baseline* sederhana (TF-IDF dengan model klasik), generalisasi di luar *template* (*out-of-distribution*), uji *robustness* atau perturbasi, dan uji stabilitas *multi-seed*. Namun, *notebook fine-tuning* yang dieksekusi pada penelitian ini belum menyertakan modul komputasi untuk keempat uji tersebut secara lengkap, sehingga hasil kuantitatifnya belum dapat disajikan dan dicatat sebagai kebutuhan penelitian lanjutan yang krusial. Berdasarkan pola leksikal yang teramati secara kualitatif, yakni kata kunci yang tampak sangat eksklusif per kelas, terdapat indikasi awal yang kuat bahwa performa sempurna kemungkinan besar dipengaruhi oleh sifat trivial *dataset* sintetis berbasis *template*, bukan semata-mata keunggulan arsitektur *transformer* pada *IndoBERT*.

Oleh karena itu, nilai *F1-macro* 1,0000 pada penelitian ini paling tepat diinterpretasikan sebagai bukti keberhasilan proses rekayasa (*engineering*) *fine-tuning*, yaitu bahwa *pipeline*, *hyperparameter*, dan teknik optimasi berjalan dengan benar, bukan sebagai ukuran akurasi klinis final yang dapat digeneralisasi pada keluhan pasien riil. Secara arsitektural, sistem yang dirancang telah berhasil dibangun dan berfungsi sebagaimana mestinya, dibuktikan oleh keberhasilan proses pelatihan serta demo inferensi yang menghasilkan prediksi label triase beserta skor *confidence*.

4.6. Perbandingan dengan Penelitian Terdahulu dan Implikasi

Capaian penelitian ini tidak dapat diperbandingkan secara langsung dengan studi triase berbasis NLP pada data klinis riil [3], [4], [10] maupun dengan studi klasifikasi teks kesehatan berbahasa Indonesia [11], karena perbedaan mendasar pada sumber data: penelitian ini menggunakan data sintetis berbasis *template*, sedangkan studi-studi tersebut menggunakan catatan klinis atau teks pengguna yang otentik dengan derajat ambiguitas jauh lebih tinggi. Kesenjangan inilah yang menjelaskan mengapa performa sempurna pada penelitian ini tidak boleh dibaca sebagai keunggulan kompetitif terhadap studi terdahulu. Konsisten dengan rekomendasi Dodge dkk. [9] mengenai variansi hasil *fine-tuning*, klaim performa juga masih memerlukan pembuktian stabilitas *multi-seed*.

Temuan analisis kalibrasi probabilitas menunjukkan bahwa skor *confidence* semata tidak dapat dijadikan ambang tunggal yang andal untuk membedakan prediksi yang layak dipercaya dari prediksi yang memerlukan verifikasi. Implikasinya, sistem direkomendasikan diposisikan secara ketat sebagai alat bantu rekomendasi awal (*assistive triage*), bukan pengganti penilaian klinis, dengan keputusan akhir tetap berada pada perawat dan tenaga medis, terutama untuk setiap rekomendasi kelas Merah yang wajib melalui verifikasi cepat.

5. KESIMPULAN

1) Sistem triase cerdas berbasis *IndoBERT* untuk mengklasifikasikan keluhan pasien berbahasa Indonesia ke dalam tiga tingkat triase telah berhasil dirancang dan dibangun secara utuh melalui *pipeline end-to-end* yang mencakup pembuatan *dataset* sintetis, praproses dan tokenisasi berbasis *WordPiece*, *fine-tuning* model *indobert-base-p1*, serta rancangan layanan inferensi berbasis *FastAPI*.
2) Model hasil *fine-tuning* mencapai *F1-macro* sebesar 1,0000 pada *test set* terisolasi, jauh melampaui target performa minimum penelitian (*F1-macro* $\geq$ 0,85),

dengan proses pelatihan yang sangat efisien (±2 menit 51 detik).

3) Capaian tersebut wajib dikualifikasi: performa sempurna hanya berlaku pada *dataset* sintetis berbasis *template* dan belum dapat digeneralisasi sebagai representasi kemampuan model pada keluhan pasien riil, sehingga paling tepat diinterpretasikan sebagai bukti keberhasilan proses rekayasa *fine-tuning*.
4) Penelitian selanjutnya disarankan menjalankan uji validitas yang telah dirancang sebagai prioritas utama, khususnya uji *baseline* sederhana (TF-IDF dengan *Logistic Regression* dan *Multinomial Naive Bayes*) serta uji stabilitas *multi-seed* yang berbiaya komputasi rendah namun bernilai pembuktian tinggi, dilanjutkan dengan uji deteksi kemiripan antar-*split* dan penyusunan *dataset out-of-distribution* berskala memadai (100–300 keluhan) dengan label yang divalidasi tenaga kesehatan sebagai *ground truth* klinis. Lebih lanjut, apabila penelitian lanjutan hendak menggunakan data rekam medis riil pasien, pengurusan *Ethical Clearance* (Kelaikan Etik) dari komite etik penelitian kesehatan serta penerapan prosedur anonimisasi data pasien (*data de-identification*) merupakan prasyarat utama yang wajib dipenuhi sebelum data klinis riil dapat dikumpulkan, diproses, dan digunakan untuk pelatihan maupun validasi model.
5) Bagi Puskesmas Kuta Utara, sistem disarankan diposisikan secara ketat sebagai alat bantu rekomendasi awal (*assistive triage*) dengan alur kerja verifikasi berlapis, di mana setiap rekomendasi kelas Merah wajib memicu verifikasi cepat oleh perawat, serta pelibatan tenaga kesehatan dalam validasi label *dataset* sebagai prasyarat mutlak sebelum penerapan operasional di lapangan.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Institut Teknologi dan Kesehatan Bintang Persada serta kepada manajemen dan tenaga kesehatan Puskesmas Kuta Utara, Kabupaten Badung, atas dukungan, diskusi, dan masukan klinis yang diberikan selama pelaksanaan penelitian ini.

DAFTAR PUSTAKA

[1] Kemenkes RI, "Peraturan Menteri Kesehatan Republik Indonesia Nomor 47 Tahun 2018 tentang Pelayanan Kegawatdaruratan," Kementerian Kesehatan RI, Jakarta, 2018.

[2] J. S. Hinson, D. A. Martinez, S. Cabral, K. George, M. Whalen, B. Hansoti, R. H. Schmicker, H. Semple, and K. R. Maciejewski, "Triage performance in emergency medicine: a systematic review," *Annals of Emergency Medicine*, vol. 74, no. 1, pp. 140–152, 2018, doi: 10.1016/j.annemergmed.2018.09.005.

[3] P. Sitthiprawiat, B. Wittayachamnankul, W. Sirikul, and K. Laohavisudhi, "Development and internal validation of an AI-based emergency triage model for predicting critical outcomes in emergency department," *Scientific Reports*, vol. 15, p. 31212, 2025, doi: 10.1038/s41598-025-17180-1.

[4] Y.-C. Chang, Y.-C. Li, C.-T. Chen, and C.-C. Lu, "Using machine learning and natural language processing in triage for prediction of clinical disposition in the emergency department," *BMC Emergency Medicine*, vol. 24, no. 1, p. 226, 2024, doi: 10.1186/s12873-024-01152-1.

[5] B. Wilie, K. Vincentio, G. I. Winata, S. Cahyawijaya, X. Li, Z. Y. Lim, S. Soleman, et al., "IndoNLU: benchmark and resources for evaluating Indonesian natural language understanding," in *Proc. 1st Conf. Asia-Pacific Chapter of the Association for Computational Linguistics and 10th Int. Joint Conf. Natural Language Processing*, 2020.

[6] H. Liang, B. Y. Tsui, H. Ni, C. C. S. Valentini, X. Liu, and Z. Han, "Applications of natural language processing at emergency department triage: a scoping review," *Journal of the American Medical Informatics Association*, vol. 31, no. 2, pp. 351–362, 2023, doi: 10.1093/jamia/ocad212.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proc. 2019 Conf. North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser,

and I. Polosukhin, “Attention is all you need,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

[9] J. Dodge, G. Ilharco, R. Schwartz, A. Farhadi, H. Hajishirzi, and N. A. Smith, “Fine-tuning pretrained language models: weight initializations, data orders, and early stopping,” *arXiv preprint*, 2020. [Online]. Available: https://arxiv.org/abs/2002.06305

[10] F. Navarro, C. Matos, and R. Salinas, “Automated triage classification in emergency services using Spanish clinical notes: a comparative analysis between ALBERT and classical machine learning approaches,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[11] A. G. S. Moeslim, E. Firmansyah, and B. Sutara, “Perbandingan kinerja XGBoost dan IndoBERT untuk klasifikasi teks kesehatan bahasa Indonesia,” *Data Sciences Indonesia (DSI)*, vol. 5, no. 2, pp. 62–72, 2025, doi: 10.47709/dsi.v5i2.7281.

[12] M. F. R. Sujana, C. A. Goenawan, and M. Muljono, “Transfer learning using IndoBERT with long short-term memory classifier for detection of suicide ideation themed Indonesian Twitter posts,” in *2024 IEEE International Conference*, 2024, doi: 10.1109/10747816.

[13] D. Duei Putri, G. F. Nama, dan W. E. Sulistiono, “Analisis Sentimen Kinerja Dewan Perwakilan Rakyat (DPR) pada Twitter Menggunakan Metode *Naive Bayes Classifier*,” *JITET (Jurnal Informatika dan Teknik Elektro Terapan)*, vol. 10, no. 1, Jan. 2022.

[14] F. Koto, J. H. Lau, dan T. Baldwin, “IndoBERTweet: A pretrained language model for Indonesian Twitter with effective domain-specific vocabulary initialization,” dalam *Proc. 2021 Conf. Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

[15] L. Rasmy, Y. Xiang, Z. Xie, C. Tao, dan D. Zhi, “Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction,” *npj Digital Medicine*, vol. 4, art. 86, 2021.