https://doi.org/10.37817/ikraith-informatika.v10i2
P-ISSN : 2580-4316
E-ISSN : 2654-8054

KLASIFIKASI SENTIMEN ULASAN PENGGUNA HALODOC MULTI-PLATFORM MENGGUNAKAN LSTM DENGAN PELABELAN INDOBERT

Dian Dwi Pramesti[1], Nizirwan Anwar[2], Diah Aryani[3], Jefry Sunupurwa Asri[4]
[1,2] Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Esa Unggul, Jakarta, Indonesia

diandwipramesti45@student.esaunggul.ac.id, nizirwan.anwar@esaunggul.ac.id,
diah.aryani@esaunggul.ac.id jefry.sunupurwa@esaunggul.ac.id

Abstrak

Pertumbuhan layanan telemedicine di Indonesia menghasilkan volume ulasan pengguna yang besar pada toko aplikasi dan media sosial. Ulasan tersebut memuat pengalaman. kritik. serta saran yang dapat digunakan untuk mengevaluasi kualitas layanan. tetapi jumlahnya membuat analisis manual menjadi tidak efisien. Penelitian ini menerapkan Long Short-Term Memory (LSTM) untuk mengklasifikasikan sentimen ulasan Halodoc yang dihimpun dari Google Play Store dan platform X. Dataset gabungan berisi 20.535 teks. terdiri atas 10.000 ulasan Google Play Store dan 10.535 unggahan X. Tahap pengolahan meliputi case folding, cleansing, normalisasi kata, tokenisasi, dan penghapusan stopword. Label sentimen positif, netral, dan negatif dibentuk secara otomatis menggunakan IndoBERT, kemudian data dibagi menjadi 80% data latih dan 20% data uji untuk pelatihan model LSTM. Hasil pengujian menunjukkan accuracy 86,12%, precision 86,14%, recall 86,12%, dan F1-score 86,12%. Audit terhadap berkas data mentah juga menunjukkan heterogenitas panjang teks antarsumber dan keberadaan duplikasi persis yang perlu diperhatikan untuk mencegah kebocoran data. Secara umum, LSTM mampu mengklasifikasikan sentimen pada data multi-platform dengan performa yang konsisten, sedangkan validasi manual terhadap label IndoBERT dan pengendalian duplikasi menjadi langkah penting untuk memperkuat validitas hasil.

Kata kunci: *analisis sentimen, Halodoc, IndoBERT, LSTM, telemedicine.*

Abstract

The growth of telemedicine services in Indonesia has generated a large volume of user feedback on application stores and social media. These reviews contain experiences, criticisms, and suggestions that can support service evaluation, but their scale makes manual analysis inefficient. This study applies Long Short-Term Memory (LSTM) to classify the sentiment of Halodoc user reviews collected from the Google Play Store and platform X. The combined dataset contains 20,535 texts, comprising 10,000 Google Play Store reviews and 10,535 posts from X. The processing stages include case folding, cleansing, word normalisation, tokenisation, and stopword removal. Positive, neutral, and negative sentiment labels were generated automatically using IndoBERT, after which the dataset was divided into 80% training data and 20% testing data for LSTM training. The model achieved an accuracy of 86.12%, precision of 86.14%, recall of 86.12%, and F1-score of 86.12%. An audit of the raw dataset further revealed differences in text length between platforms and a substantial number of exact duplicates, which should be controlled to reduce the risk of data leakage. Overall, LSTM provides consistent performance for multi-platform sentiment classification, while human validation of IndoBERT-generated labels and duplicate-aware data splitting remain important for strengthening the validity of the findings.

Keywords: *sentiment analysis, Halodoc, IndoBERT, LSTM, telemedicine*

I. PENDAHULUAN

Telemedicine memperluas akses masyarakat terhadap konsultasi kesehatan dengan menghubungkan pasien dan tenaga kesehatan melalui teknologi informasi. Literatur menunjukkan bahwa telemedicine menawarkan kemudahan akses, efisiensi waktu, kontinuitas layanan, dan dukungan bagi wilayah yang memiliki keterbatasan fasilitas kesehatan, walaupun penerapannya tetap menghadapi hambatan teknologi, regulasi, privasi, dan kesiapan pengguna [12], [19]. Dalam konteks Indonesia, aplikasi Halodoc menjadi salah satu kanal konsultasi kesehatan daring yang menghadirkan layanan percakapan dengan dokter, pembelian obat, informasi kesehatan, serta fitur layanan pendukung lainnya.

Peningkatan pemanfaatan aplikasi digital menghasilkan jejak opini dalam jumlah besar. Di Google Play Store, pengguna memberikan ulasan langsung yang umumnya berkaitan dengan kualitas aplikasi, dokter, respons layanan, transaksi, pengiriman obat, dan kemudahan penggunaan. Di platform X, pembicaraan mengenai Halodoc muncul dalam bentuk pengalaman personal, pertanyaan, rekomendasi, keluhan, promosi, serta percakapan antarpengguna. Umpan balik semacam ini penting karena dapat mengungkap kebutuhan pengguna yang tidak selalu terlihat melalui survei formal. Pagano dan Maalej menunjukkan bahwa ulasan toko aplikasi dapat menjadi sumber informasi empiris tentang pengalaman pengguna [6], sedangkan Guzman dan Maalej memperlihatkan bahwa sentimen dapat dipetakan secara lebih terperinci terhadap fitur aplikasi [7].

Volume dan keragaman bahasa pada ulasan membuat pembacaan manual tidak lagi memadai. Analisis sentimen menyediakan mekanisme komputasional untuk mengidentifikasi orientasi opini positif, netral, atau negatif [2], [5]. Berbagai metode telah digunakan, mulai dari pendekatan leksikon dan pembelajaran mesin klasik hingga deep learning [8], [13]. Pada teks Bahasa Indonesia, tantangan tambahan muncul dalam bentuk singkatan, bahasa tidak baku, variasi ejaan, campur kode, emoji, tautan, mention, dan karakter percakapan yang berbeda antarsumber. Pendekatan stemming Bahasa Indonesia dan pengembangan sumber daya khusus seperti IndoLEM,

P-ISSN : 2580-4316
E-ISSN : 2654-8054 https://doi.org/10.37817/ikraith-informatika.v10i2

IndoBERT, dan IndoNLU menunjukkan pentingnya adaptasi model terhadap karakteristik bahasa lokal [3], [16], [17].

LSTM dirancang untuk mengatasi keterbatasan jaringan saraf berulang dalam mempertahankan informasi jangka panjang melalui mekanisme gerbang [1]. Kemampuan ini relevan untuk klasifikasi sentimen karena polaritas kalimat sering ditentukan oleh hubungan kata yang berjauhan, negasi, dan konteks urutan. Di sisi lain, arsitektur Transformer [11] dan BERT [15] menyediakan representasi kontekstual dua arah yang kuat. IndoBERT dan IndoBERTweet kemudian mengadaptasi paradigma tersebut untuk Bahasa Indonesia dan domain media sosial [16], [20]. Dalam penelitian ini, IndoBERT digunakan untuk membentuk label awal, sedangkan LSTM berperan sebagai model klasifikasi akhir.

Kajian terdahulu telah membahas klasifikasi otomatis ulasan aplikasi [9], survei analisis app store [10], analisis sentimen ulasan marketplace Indonesia menggunakan Support Vector Machine [14], serta aplikasi deep learning pada ulasan telemedicine. Mutmainah et al. menggabungkan BiLSTM dan LDA untuk analisis sentimen serta pemodelan topik aplikasi telemedicine pada Google Play [21]. Refianti et al. meneliti ulasan Halodoc di Google Play menggunakan model LSTM berbasis leksikon [22]. Penelitian lain menggunakan hibrida CNN-BiLSTM untuk mengevaluasi kepuasan pengguna Halodoc [23], SMOTE dan LSTM untuk ulasan Mobile JKN [24], serta perbandingan LSTM dan GRU pada ulasan aplikasi X di Play Store [25].

Meskipun demikian, sebagian besar penelitian terdahulu berfokus pada satu platform. Karakter Google Play Store dan X berbeda secara struktural: ulasan toko aplikasi cenderung langsung mengevaluasi layanan, sedangkan unggahan X lebih bersifat percakapan, kontekstual, dan sering mengandung tautan atau mention. Penggabungan keduanya berpotensi menghasilkan model yang lebih luas, tetapi juga menimbulkan domain shift dan risiko noise. Penelitian ini bertujuan menerapkan LSTM untuk klasifikasi tiga kelas sentimen pada data Halodoc multi-platform dengan pelabelan IndoBERT. Kontribusi penelitian adalah penyusunan dataset gabungan yang relatif seimbang antarsumber, penerapan alur IndoBERT-LSTM, dan audit karakteristik data untuk mengidentifikasi faktor yang dapat memengaruhi validitas evaluasi.

II. TINJAUAN PUSTAKA

A. Analisis Sentimen dan Ulasan Aplikasi

Analisis sentimen memetakan teks ke dalam orientasi opini tertentu. Pang et al. memperlihatkan bahwa klasifikasi sentimen memiliki karakteristik yang berbeda dari klasifikasi topik karena sinyal polaritas dapat tersebar pada banyak kata dan konstruksi bahasa [2]. Liu menjelaskan bahwa analisis opini dapat dilakukan pada tingkat dokumen, kalimat, maupun aspek [5]. Pada penelitian ini, unit analisis adalah satu ulasan atau satu unggahan, sehingga pendekatannya termasuk klasifikasi tingkat dokumen.

Ulasan aplikasi bukan sekadar penilaian numerik. Isinya dapat memuat permintaan fitur, laporan bug, pengalaman layanan, penilaian kualitas, dan informasi kontekstual lain. Klasifikasi otomatis diperlukan agar organisasi dapat memisahkan kategori umpan balik secara sistematis [9]. Survei Martin et al. menunjukkan bahwa analisis app store telah berkembang menjadi sumber bukti bagi rekayasa perangkat lunak dan pengambilan keputusan produk [10]. Oleh karena itu, klasifikasi sentimen terhadap ulasan Halodoc dapat dipandang sebagai bagian dari mekanisme intelligence untuk evaluasi layanan digital kesehatan.

B. LSTM dan Representasi Kontekstual

LSTM menggunakan sel memori dan gerbang untuk mengatur aliran informasi. Pada setiap langkah waktu, forget gate menentukan informasi lama yang dipertahankan, input gate mengendalikan informasi baru, dan output gate membentuk hidden state. Secara ringkas, operasi LSTM dapat dituliskan sebagai berikut [1]:

$$f_t = sigma(W_f [h_(t-1), x_t] + b_f) \quad (1)$$
$$i_t = sigma(W_i [h_(t-1), x_t] + b_i) \quad (2)$$
$$C\sim_t = tanh(W_C [h_(t-1), x_t] + b_C) \quad (3)$$
$$C_t = f_t * C_(t-1) + i_t * C\sim_t \quad (4)$$
$$o_t = sigma(W_o [h_(t-1), x_t] + b_o) \quad (5)$$
$$h_t = o_t * tanh(C_t) \quad (6)$$

Mekanisme tersebut memungkinkan model mempertahankan konteks yang relevan dan mengurangi masalah vanishing gradient. Pada klasifikasi sentimen, keluaran LSTM diteruskan ke lapisan klasifikasi untuk menghasilkan probabilitas kelas positif, netral, dan negatif. Perkembangan berikutnya memperkenalkan mekanisme self-attention dalam Transformer [11] serta pre-training bidirectional pada BERT [15]. Dalam penelitian ini, kedua keluarga model digunakan secara berurutan: IndoBERT sebagai pembentuk label otomatis dan LSTM sebagai classifier yang dilatih pada label tersebut.

C. NLP Bahasa Indonesia dan Pelabelan Otomatis

Teks Bahasa Indonesia pada media sosial sering berbeda dari bahasa formal. Proses normalisasi diperlukan untuk menangani variasi seperti “gak”, “nggak”, “ga”, singkatan, pengulangan karakter, serta bentuk tidak baku lainnya. Confix-stripping stemmer menjadi salah satu pendekatan awal untuk memproses morfologi Bahasa Indonesia [3]. Sumber daya yang lebih mutakhir seperti IndoLEM, IndoBERT, dan IndoNLU menyediakan benchmark dan representasi pralatih untuk beragam tugas pemahaman Bahasa Indonesia [16], [17]. Untuk data X, IndoBERTweet relevan karena kosakata dan proses pre-training-nya diarahkan pada karakter media sosial Indonesia [20].

https://doi.org/10.37817/ikraith-informatika.v10i2
P-ISSN : 2580-4316
E-ISSN : 2654-8054

Pelabelan menggunakan model pralatih dapat mempercepat penyusunan data latih, tetapi label otomatis bukan ground truth yang bebas kesalahan. Konsep weak supervision menjelaskan bahwa label yang dibentuk oleh fungsi atau model heuristik dapat digunakan untuk membangun data dalam skala besar, namun kualitas dan korelasinya perlu dikendalikan [18]. Dalam konteks penelitian ini, IndoBERT berfungsi sebagai teacher yang menghasilkan label, sedangkan LSTM mempelajari pola dari label tersebut. Konsekuensinya, performa LSTM juga mencerminkan kemampuannya meniru keputusan IndoBERT, bukan semata-mata kesesuaian dengan penilaian manusia.

Tabel 1. Perbandingan Penelitian Terkait

Penelitian	Objek	Sumber	Metode	Posisi/Karakteristik
Mutmainah et al. [21]	Aplikasi telemedicine	Google Play	BiLSTM dan LDA	Sentimen dan topik; satu platform
Refianti et al. [22]	Halodoc	Google Play	Lexicon-based LSTM	Spesifik Halodoc; satu platform
Kurniasari et al. [23]	Halodoc	Ulasan aplikasi	CNN-BiLSTM	Model hibrida untuk kepuasan
Tamami et al. [24]	Mobile JKN	Ulasan aplikasi	SMOTE-LSTM	Menangani ketidakseimbangan kelas
Wily et al. [25]	Aplikasi X	Google Play	LSTM dan GRU	Perbandingan recurrent model
Penelitian ini	Halodoc	Google Play dan X	IndoBERT-LSTM	Tiga kelas dan data multi-platform

III. METODOLOGI PENELITIAN

A. Desain Penelitian

Penelitian menggunakan pendekatan kuantitatif eksperimental untuk membangun model klasifikasi sentimen. Alur penelitian terdiri atas pengumpulan data, penggabungan dataset, audit data mentah, preprocessing, pelabelan sentimen menggunakan IndoBERT, pembagian data, pelatihan LSTM, dan evaluasi. Alur tersebut ditunjukkan pada Gambar 1. Pemisahan antara model pelabel dan model classifier dipilih agar data dalam jumlah besar dapat memperoleh label awal secara otomatis, tetapi konsekuensi validitasnya dianalisis pada bagian pembahasan.

Gambar 1. Alur penelitian IndoBERT-LSTM pada data multi-platform.

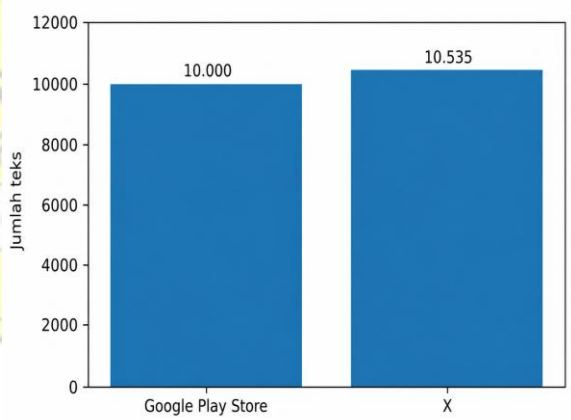
B. Dataset dan Sumber Data

Berkas Dataset_Gabungan.csv berisi 20.535 baris dan dua atribut. yaitu text dan source. Tidak terdapat nilai kosong pada kedua atribut. Sebanyak 10.000 baris berasal dari Google Play Store dan 10.535 baris berasal dari X. Komposisi tersebut setara dengan 48.70% dan 51.30% dari keseluruhan data. Dataset yang dilampirkan tidak menyertakan tanggal unggahan. rating. identitas pengguna. label sentimen. maupun keluaran prediksi. Dengan demikian. analisis deskriptif pada artikel ini dilakukan terhadap teks mentah dan sumber. sedangkan hasil klasifikasi mengacu pada ringkasan eksperimen yang tersedia.

Tabel 2. Karakteristik Dataset Mentah Berdasarkan Sumber

Sumber	Jumlah	Proporsi (%)	Rerata karakter	Median karakter	Rerata kata	Median kata
Google Play Store	10.000	48.70	54.34	29	8.27	4
X	10.535	51.30	176.59	181	26.04	27

Perbedaan panjang teks pada Tabel II menunjukkan adanya heterogenitas domain. Ulasan Google Play Store rata-rata lebih singkat dan sering berbentuk penilaian langsung, sedangkan unggahan X lebih panjang karena memuat percakapan, mention, tautan, atau konteks naratif. Distribusi jumlah data ditampilkan pada Gambar 2.



Gambar 2. Distribusi Jumlah Teks Berdasarkan Sumber.

C. Audit Kualitas Data

Audit terhadap data mentah menemukan 14.867 teks unik. Terdapat 5.668 kemunculan duplikat persis setelah kemunculan pertama atau 27.60% dari seluruh baris. Selain itu. 3.776 teks mengandung URL. 4.523 mengandung mention. dan 1.330 mengandung hashtag. Sebagian besar pola tersebut berasal dari X. Temuan ini relevan karena teks identik yang tersebar pada data latih dan data uji dapat meningkatkan nilai evaluasi secara artifisial. Oleh sebab itu. deduplikasi atau group-aware

P-ISSN : 2580-4316
E-ISSN : 2654-8054 https://doi.org/10.37817/ikraith-informatika.v10i2

split berdasarkan teks yang telah dinormalisasi direkomendasikan sebelum pembagian data.

Tabel 3. Hasil Audit Data Mentah

Indikator	Nilai	Interpretasi
Jumlah baris	20.535	Seluruh teks dari dua sumber
Teks unik	14.867	Berdasarkan kecocokan persis
Duplikat tambahan	5.668	27.60% dari dataset
Teks dengan URL	3.776	Mayoritas berasal dari X
Teks dengan mention	4.523	Perlu dihapus atau dinormalisasi
Teks dengan hashtag	1.330	Dapat dihapus atau dipertahankan sebagai token

D. Preprocessing Teks

Preprocessing bertujuan mengurangi noise dan menyamakan representasi teks antarsumber. Tahap pertama adalah case folding untuk mengubah seluruh karakter alfabet menjadi huruf kecil. Tahap kedua adalah cleansing yang menghapus atau mengganti URL, mention, simbol, karakter berulang yang tidak informatif, dan artefak HTML. Tahap ketiga adalah normalisasi kata tidak baku ke bentuk yang konsisten, misalnya penyatuan variasi negasi dan singkatan yang umum pada media sosial.

Tahap berikutnya adalah tokenisasi untuk memecah teks menjadi unit kata, kemudian stopword removal untuk menghapus kata fungsi yang dinilai tidak memberikan informasi sentimen. Penghapusan stopword harus dilakukan secara hati-hati karena kata negasi seperti “tidak”, “bukan”, “belum”, “jangan”, dan variasi informalnya dapat mengubah polaritas. Stemming dapat digunakan sebagai tahap tambahan berdasarkan karakter morfologi Bahasa Indonesia [3], tetapi abstrak eksperimen yang dilampirkan hanya menyebutkan case folding, cleansing, normalisasi, tokenisasi, dan stopword removal.

Setelah preprocessing, token diubah menjadi indeks numerik menggunakan tokenizer dan diseragamkan panjangnya melalui padding atau truncation sebelum masuk ke lapisan embedding. Karena berkas yang dilampirkan tidak memuat tokenizer, daftar kamus normalisasi, panjang sekuens maksimum, serta kosakata hasil pelatihan, parameter tersebut perlu didokumentasikan pada versi final agar penelitian dapat direplikasi.

Tabel 4. Tahap Preprocessing dan Tujuannya

Tahap	Tujuan
Case folding	Menyeragamkan kapitalisasi
Cleansing	Menghapus URL, mention, simbol, dan noise teks
Normalisasi kata	Mengubah bentuk tidak baku/singkatan menjadi bentuk konsisten
Tokenisasi	Memecah teks menjadi token
Stopword removal	Mengurangi kata fungsi yang tidak informatif
Padding/truncation	Menyeragamkan panjang masukan ke model LSTM

E. Pelabelan Sentimen Menggunakan IndoBERT

Pelabelan otomatis dilakukan menggunakan IndoBERT. Model menerima teks yang telah disiapkan dan menghasilkan probabilitas untuk tiga kelas, yaitu positif, netral, dan negatif. Kelas dengan probabilitas terbesar digunakan sebagai label. Pemanfaatan IndoBERT didasarkan pada kemampuannya menghasilkan representasi kontekstual Bahasa Indonesia [16]. Untuk unggahan X, karakter domain media sosial juga berkaitan dengan pengembangan IndoBERTweet [20], sehingga perbandingan antara IndoBERT umum dan model domain-spesifik dapat menjadi eksperimen lanjutan.

Pelabelan otomatis mempercepat anotasi 20.535 teks, tetapi perlu diperlakukan sebagai weak supervision [18]. Sampel label idealnya diverifikasi oleh dua atau lebih anotator manusia, kemudian kesepakatannya dihitung menggunakan Cohen’s kappa atau Fleiss’ kappa. Validasi ini diperlukan untuk mengetahui apakah kesalahan LSTM berasal dari model classifier atau telah diwariskan dari model pelabel.

F. Pembagian Data dan Pelatihan LSTM

Dataset dibagi menjadi 80% data latih dan 20% data uji. Dengan jumlah 20.535 baris. pembagian tersebut setara dengan sekitar 16.428 data latih dan 4.107 data uji. Pembagian harus dilakukan secara stratified berdasarkan kelas agar proporsi positif. netral. dan negatif tetap terjaga. Mengingat adanya duplikasi. pemisahan juga sebaiknya dilakukan berdasarkan kelompok teks unik agar satu isi yang sama tidak muncul pada kedua subset.

Secara konseptual, arsitektur terdiri atas lapisan embedding, satu atau lebih lapisan LSTM, regularisasi dropout, dan dense layer dengan aktivasi softmax tiga kelas. Loss yang sesuai adalah categorical cross-entropy atau sparse categorical cross-entropy, bergantung pada format label. Namun, dokumen terlampir tidak mencantumkan dimensi embedding, jumlah unit LSTM, dropout rate, optimizer, learning rate, batch size, jumlah epoch, random seed, maupun strategi early stopping. Parameter tersebut tidak direka dalam artikel ini dan perlu dilengkapi dari catatan eksperimen sebelum pengiriman ke jurnal.

G. Evaluasi Model

Evaluasi menggunakan confusion matrix dan classification report. Empat metrik utama adalah accuracy, precision, recall, dan F1-score [4]. Untuk satu kelas, rumus dasarnya adalah:

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \quad (7)$$

$$Precision = TP / (TP + FP) \quad (8)$$

https://doi.org/10.37817/ikraith-informatika.v10i2
P-ISSN : 2580-4316
E-ISSN : 2654-8054

Recall = TP / (TP + FN) (9)
*F1 = 2 * (Precision * Recall) / (Precision + Recall)* (10)

Pada masalah tiga kelas, precision, recall, dan F1 dapat dihitung dengan macro average, micro average, atau weighted average. Ringkasan hasil yang dilampirkan tidak menyebutkan skema averaging. Oleh karena itu, versi final artikel harus menuliskan skema tersebut agar interpretasi metrik konsisten, khususnya bila distribusi kelas tidak seimbang.

IV. HASIL DAN PEMBAHASAN

A. Karakteristik Dataset Multi-Platform

Komposisi data antarsumber hampir seimbang, sehingga model memperoleh eksposur yang relatif setara terhadap ulasan toko aplikasi dan percakapan media sosial. Keseimbangan pada tingkat sumber tidak otomatis berarti keseimbangan pada tingkat sentimen. Berkas data mentah tidak memuat label hasil IndoBERT, sehingga proporsi kelas positif, netral, dan negatif tidak dapat dihitung kembali dari lampiran. Distribusi kelas tersebut seharusnya dilaporkan karena berpengaruh terhadap pemilihan metrik dan strategi penanganan imbalance.

Panjang teks memperlihatkan perbedaan tajam. Rata-rata teks Google Play Store adalah 54,34 karakter atau 8,27 kata, sedangkan X mencapai 176,59 karakter atau 26,04 kata. Perbedaan ini dapat memengaruhi panjang padding, densitas informasi, dan peluang munculnya konteks non-sentimen. Ulasan sangat pendek seperti “bagus”, “mantap”, atau “sangat membantu” relatif mudah diklasifikasikan, tetapi duplikasi frasa semacam itu juga tinggi. Sebaliknya, unggahan X sering memuat percakapan panjang, kutipan, tautan, dan mention yang menambah noise.

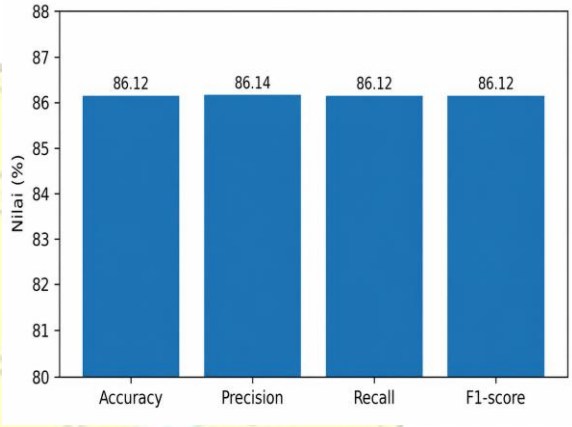
Sebanyak 27,60% baris merupakan kemunculan tambahan dari teks yang sama secara persis. Angka ini tidak selalu berarti data berkualitas buruk, karena ulasan singkat dapat memang identik secara alami. Namun, dari perspektif evaluasi pembelajaran mesin, duplikasi dapat menimbulkan data leakage. Jika frasa identik ada pada data latih dan uji, model mungkin memperoleh skor tinggi melalui memorisasi. Oleh karena itu, hasil 86,12% perlu dibaca bersama dengan prosedur deduplikasi dan pemisahan data. Tanpa informasi tersebut, skor tetap informatif sebagai hasil eksperimen, tetapi kekuatan generalisasinya belum dapat dinilai secara penuh.

B. Kinerja Model LSTM

Hasil eksperimen menunjukkan nilai accuracy 86,12%, precision 86,14%, recall 86,12%, dan F1-score 86,12%. Kedekatan keempat nilai menunjukkan bahwa performa agregat model relatif konsisten. Precision yang sedikit lebih tinggi daripada recall mengindikasikan bahwa proporsi prediksi benar di antara prediksi yang dibuat model sedikit lebih baik daripada kemampuannya menangkap seluruh contoh relevan, meskipun selisih 0,02 poin persentase terlalu kecil untuk ditafsirkan sebagai perbedaan substantif.

Tabel 5. Kinerja Agregat Model LSTM

Metrik	Nilai (%)
Accuracy	86,12
Precision	86,14
Recall	86,12
F1-score	86,12



Gambar 3. Nilai Metrik Agregat Model LSTM.

Nilai tersebut menunjukkan bahwa LSTM mampu mempelajari pola sentimen dari representasi sekuensial teks multi-platform. Temuan ini konsisten dengan penggunaan recurrent neural network pada penelitian telemedicine dan ulasan aplikasi [21], [22], [24], [25]. Walaupun demikian, perbandingan langsung dengan penelitian terdahulu tidak dilakukan karena setiap studi dapat menggunakan dataset, skema label, pembagian data, dan metrik averaging yang berbeda. Penelitian ini juga memiliki skema khusus: label target dibentuk oleh IndoBERT. Dengan demikian, accuracy 86,12% menggambarkan kesesuaian LSTM terhadap label hasil teacher model, bukan akurasi terhadap anotasi manusia yang independen.

C. Peran IndoBERT dalam Pembentukan Label

Pemanfaatan IndoBERT memberikan keuntungan praktis karena model dapat memberikan label awal secara kontekstual pada teks Bahasa Indonesia. Berbeda dengan leksikon, model BERT dapat mempertimbangkan urutan dan konteks kata [15], [16]. Hal ini penting pada kalimat seperti “dokternya bagus, tapi aplikasinya selalu gagal”, yang mengandung polaritas campuran. Namun, pengambilan satu label pada tingkat dokumen dapat menyederhanakan sentimen multi-aspek menjadi satu kelas dominan.

Hubungan teacher-student antara IndoBERT dan LSTM perlu diposisikan secara eksplisit. Jika tujuan penelitian adalah membuat model LSTM yang lebih ringan untuk meniru IndoBERT, maka evaluasi terhadap label IndoBERT relevan sebagai ukuran distillation atau surrogate modelling. Jika tujuan utamanya adalah mengukur sentimen pengguna secara

P-ISSN : 2580-4316
E-ISSN : 2654-8054 https://doi.org/10.37817/ikraith-informatika.v10i2

faktual, maka ground truth manusia tetap diperlukan. Ratner et al. menekankan bahwa weak supervision dapat mempercepat pembuatan data, tetapi kualitas sumber label harus dimodelkan dan diverifikasi [18].

Selain validasi manusia, penelitian lanjutan dapat membandingkan IndoBERT dengan IndoBERTweet untuk subset X [20]. Perbandingan tersebut dapat menguji apakah model domain-spesifik lebih baik dalam menangani slang, mention, hashtag, dan struktur percakapan. Eksperimen ablation juga dapat dilakukan dengan tiga skenario: LSTM berlabel manual, LSTM berlabel IndoBERT, dan fine-tuning IndoBERT langsung. Skenario ini akan memperjelas kontribusi setiap komponen.

D. Perbandingan dengan Penelitian Terdahulu

Penelitian ini berbeda dari studi Mutmainah et al. yang menggabungkan BiLSTM dan LDA pada ulasan Google Play [21] karena menambahkan sumber X dan menggunakan IndoBERT untuk pelabelan. Dibandingkan Refianti et al. [22], perbedaannya terletak pada penggunaan data multi-platform dan pelabelan model kontekstual, bukan hanya pendekatan leksikon. Dibandingkan model hibrida CNN-BiLSTM [23], penelitian ini mempertahankan LSTM sebagai classifier utama sehingga arsitekturnya lebih sederhana.

Studi Mobile JKN menggunakan SMOTE untuk mengatasi ketidakseimbangan kelas [24]. Pada penelitian ini, kebutuhan teknik balancing belum dapat ditentukan karena distribusi label tidak tersedia dalam dataset mentah. Studi Wily et al. membandingkan LSTM dan GRU pada ulasan aplikasi X di Play Store [25], sedangkan penelitian ini memanfaatkan percakapan pada platform X sebagai sumber tambahan. Perbedaan tersebut penting karena "ulasan aplikasi X di Play Store" dan "unggahan pada platform X tentang Halodoc" merupakan dua domain yang tidak sama.

Kontribusi utama penelitian terletak pada integrasi dua ekosistem opini. Google Play Store merepresentasikan evaluasi langsung terhadap aplikasi, sedangkan X menangkap percakapan sosial yang lebih luas. Integrasi ini dapat memperkaya representasi opini, tetapi model juga harus belajar membedakan opini pengguna, informasi netral, promosi, kutipan artikel, dan balasan layanan pelanggan. Karena itu, penyertaan source sebagai fitur tambahan atau pelatihan model domain-adversarial dapat dipertimbangkan pada penelitian berikutnya.

E. Ancaman terhadap Validitas

Tabel 6.. Ancaman Validitas Dan Mitigasi

Aspek	Risiko	Mitigasi
Construct validity	Label IndoBERT mungkin tidak sepenuhnya sama dengan penilaian manusia	Validasi sampel oleh minimal dua anotator dan laporkan kappa
Internal validity	Duplikasi dapat tersebar pada data latih dan uji	Deduplikasi atau group-aware stratified split
External validity	Data hanya berasal dari Google Play Store dan X	Tambahkan App Store, forum, atau survei pengguna
Conclusion validity	Skema averaging dan distribusi kelas tidak tersedia	Laporkan per-class metrics, macro/weighted average, dan confusion matrix
Reproducibility	Hyperparameter serta random seed tidak disertakan	Publikasikan konfigurasi, kode, dan artefak model

Ancaman yang paling menonjol adalah label otomatis dan duplikasi. Keduanya dapat menghasilkan performa yang terlihat baik tetapi tidak sepenuhnya merepresentasikan generalisasi pada opini baru. Dataset juga tidak menyertakan timestamp, sehingga perubahan kualitas layanan dari waktu ke waktu tidak dapat dianalisis. Di samping itu, tidak adanya label dan prediksi pada berkas CSV membuat confusion matrix tidak dapat direkonstruksi. Oleh karena itu, artikel ini tidak mengarang nilai per kelas dan membatasi pembahasan kuantitatif pada metrik agregat yang tersedia.

F. Implikasi Praktis

Model sentimen dapat digunakan sebagai lapisan pemantauan untuk merangkum opini pengguna secara berkala. Kelas negatif dapat menjadi sinyal awal untuk menelusuri keluhan tentang respons dokter, pembayaran, voucher, kestabilan aplikasi, pengiriman obat, atau layanan pelanggan. Kelas positif dapat mengidentifikasi atribut layanan yang dihargai pengguna, seperti kemudahan konsultasi, keramahan dokter, kecepatan respons, dan kualitas penjelasan.

Implementasi operasional sebaiknya tidak berhenti pada klasifikasi dokumen. Sentimen perlu dipadukan dengan aspect extraction agar organisasi mengetahui objek keluhan atau pujian. Pendekatan fine-grained pada ulasan aplikasi telah ditunjukkan oleh Guzman dan Maalej [7]. Dengan demikian, pipeline berikutnya dapat menghubungkan kelas sentimen dengan aspek dokter, aplikasi, transaksi, obat, pengiriman, dan dukungan pelanggan. Hasilnya akan lebih mudah diterjemahkan menjadi prioritas perbaikan layanan.

V. KESIMPULAN

Penelitian ini menerapkan LSTM untuk mengklasifikasikan sentimen ulasan Halodoc yang berasal dari Google Play Store dan platform X. Dataset gabungan memuat 20.535 teks dengan komposisi 10.000 ulasan Google Play Store dan 10.535 unggahan X. Preprocessing dilakukan melalui case folding, cleansing, normalisasi kata, tokenisasi, dan stopword removal. Label positif, netral, dan negatif dibentuk menggunakan IndoBERT, kemudian data dibagi 80:20 untuk pelatihan dan pengujian LSTM. Model menghasilkan accuracy 86,12%, precision 86,14%, recall 86,12%, dan F1-score 86,12%, yang menunjukkan performa agregat konsisten pada data

multi-platform. Audit data menunjukkan bahwa karakteristik kedua sumber berbeda dan terdapat 5.668 kemunculan duplikat tambahan. Karena itu, hasil perlu diperkuat melalui deduplikasi atau group-aware split, pelaporan distribusi kelas, skema averaging, hyperparameter, serta confusion matrix per kelas. Validasi manusia terhadap sebagian label IndoBERT juga diperlukan agar evaluasi tidak hanya mengukur kesesuaian LSTM terhadap teacher model. Dengan perbaikan tersebut, pendekatan IndoBERT-LSTM berpotensi menjadi mekanisme yang lebih kuat untuk memahami opini pengguna dan mendukung peningkatan layanan telemedicine.

KETERSEDIAAN DATA

Dataset yang dianalisis merupakan berkas Dataset_Gabungan.csv yang disediakan bersama dokumen penelitian. Berkas memuat atribut text dan source. Informasi tanggal pengambilan, rating, label IndoBERT, prediksi LSTM, dan konfigurasi pelatihan tidak terdapat dalam lampiran yang tersedia.

DAFTAR PUSTAKA

[1] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural Comput., vol. 9, no. 8, pp. 1735-1780, 1997, doi: 10.1162/neco.1997.9.8.1735.

[2] B. Pang, L. Lee, and S. Vaithyanathan, “Thumbs up? Sentiment classification using machine learning techniques,” in Proc. 2002 Conf. Empirical Methods in Natural Language Processing, 2002, pp. 79-86, doi: 10.3115/1118693.1118704.

[3] J. Asian, H. E. Williams, and S. M. M. Tahaghoghi, “Stemming Indonesian: A confix-stripping approach,” ACM Trans. Asian Lang. Inf. Process., vol. 6, no. 4, Art. no. 13, 2007, doi: 10.1145/1316457.1316459.

[4] M. Sokolova and G. Lapalme, “A systematic analysis of performance measures for classification tasks,” Inf. Process. Manage., vol. 45, no. 4, pp. 427-437, 2009, doi: 10.1016/j.ipm.2009.03.002.

[5] B. Liu, Sentiment Analysis and Opinion Mining. San Rafael, CA, USA: Morgan & Claypool Publishers, 2012, doi: 10.2200/S00416ED1V01Y201204HLT016.

[6] D. Pagano and W. Maalej, “User feedback in the AppStore: An empirical study,” in Proc. 21st IEEE Int. Requirements Engineering Conf. (RE), 2013, pp. 125-134, doi: 10.1109/RE.2013.6636712.

[7] E. Guzman and W. Maalej, “How do users like this feature? A fine grained sentiment analysis of app reviews,” in Proc. IEEE 22nd Int. Requirements Engineering Conf. (RE), Karlskrona, Sweden, 2014, pp. 153-162, doi: 10.1109/RE.2014.6912257.

[8] W. Medhat, A. Hassan, and H. Korashy, “Sentiment analysis algorithms and applications: A survey,” Ain Shams Eng. J., vol. 5, no. 4, pp. 1093-1113, 2014, doi: 10.1016/j.asej.2014.04.011.

[9] W. Maalej, Z. Kurtanovic, H. Nabil, and C. Stanik, “On the automatic classification of app reviews,” Requirements Eng., vol. 21, no. 3, pp. 311-331, 2016, doi: 10.1007/s00766-016-0251-9.

[10] W. Martin, F. Sarro, Y. Jia, Y. Zhang, and M. Harman, “A survey of app store analysis for software engineering,” IEEE Trans. Softw. Eng., vol. 43, no. 9, pp. 817-847, 2017, doi: 10.1109/TSE.2016.2630689.

[11] A. Vaswani et al., “Attention is all you need,” in Advances in Neural Information Processing Systems, vol. 30, 2017, pp. 5998-6008.

[12] C. S. Kruse, P. Karem, K. Shifflett, L. Vegi, K. Ravi, and M. Brooks, “Evaluating barriers to adopting telemedicine worldwide: A systematic review,” J. Telemed. Telecare, vol. 24, no. 1, pp. 4-12, 2018, doi: 10.1177/1357633X16674087.

[13] L. Zhang, S. Wang, and B. Liu, “Deep learning for sentiment analysis: A survey,” WIREs Data Min. Knowl. Discov., vol. 8, no. 4, Art. no. e1253, 2018, doi: 10.1002/widm.1253.

[14] A. A. Lutfi, A. E. Permanasari, and S. Fauziati, “Sentiment analysis in the sales review of Indonesian marketplace by utilizing Support Vector Machine,” J. Inf. Syst. Eng. Bus. Intell., vol. 4, no. 1, pp. 57-64, 2018, doi: 10.20473/jisebi.4.1.57-64.

[15] 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, 2019, pp. 4171-4186, doi: 10.18653/v1/N19-1423.

[16] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, “IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP,” in Proc. 28th Int. Conf. Computational Linguistics, Barcelona, Spain, 2020, pp. 757-770, doi: 10.18653/v1/2020.coling-main.66.

[17] B. Wilie 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, Suzhou, China, 2020, pp. 843-857, doi: 10.18653/v1/2020.aacl-main.85.

[18] A. Ratner, S. H. Bach, H. Ehrenberg, J. Fries, S. Wu, and C. Re, “Snorkel: Rapid training data creation with weak supervision,” VLDB J., vol. 29, pp. 709-730, 2020, doi: 10.1007/s00778-019-00552-1.

[19] A. Haleem, M. Javaid, R. P. Singh, and R. Suman, “Telemedicine for healthcare: Capabilities, features, barriers, and applications,” Sensors Int., vol. 2, Art. no. 100117, 2021, doi: 10.1016/j.sintl.2021.100117.

[20] F. Koto, J. H. Lau, and T. Baldwin, “IndoBERTweet: A pretrained language model for Indonesian Twitter with effective domain-specific vocabulary initialization,” in Proc. 2021 Conf. Empirical Methods in Natural Language Processing, Punta Cana, Dominican Republic, 2021, pp. 10660-10668, doi: 10.18653/v1/2021.emnlp-main.833.

[21] S. Mutmainah, D. H. Fudholi, and S. Hidayat, “Analisis sentimen dan pemodelan topik aplikasi telemedicine pada Google Play menggunakan BiLSTM dan LDA,” J. Media Inform. Budidarma, vol. 7, no. 1, p. 312, 2023.

[22] R. Refianti, A. B. Mutiara, and R. A. Putra, “A lexicon-based Long Short-Term Memory (LSTM) model for sentiment analysis to classify Halodoc application reviews on Google Playstore,” J. Appl. Data Sci., vol. 5, no. 1, pp. 146-157, 2024.

[23] D. Kurniasari, A. Su’admaji, F. R. Lumbanraja, and Warsono, “Evaluating user satisfaction in the Halodoc application using a hybrid CNN-BiLTSM model for sentiment analysis,” J. Tek. Inform., vol. 18, no. 2, pp. 209-225, 2025.

[24] G. Tamami, W. A. Triyanto, and S. Muzid, "Sentiment analysis Mobile JKN reviews using SMOTE based LSTM," IJCCS (Indonesian J. Comput. Cybern. Syst.), vol. 19, no. 1, p. 13, 2025.

[25] W. A. Wily, S. Anggai, and T. Tukiyat, "Analisis sentimen ulasan pengguna aplikasi media sosial X di Play Store menggunakan algoritma Long Short-Term Memory (LSTM) dan Gated Recurrent Unit (GRU)," J. SISKOM-KB (Sistem Komput. dan Kecerdasan Buatan), vol. 9, no. 1, pp. 63-72, 2025.