Journal of Applied Informatics and Computing (JAIC)
Vol.10, No.1, February 2026, pp. 1093~1100
e-ISSN: 2548-6861

Comparative Analysis of BERT and LSTM Models for Sentiment Classification of Mobile Game User Reviews

Toto Indriyatmoko [1]*, Majid Rahardi [2]*, Hastari Utama [3]*, Arvin Claudy Frobenius [4]*
* Fakultas Ilmu Komputer, Universitas Amikom Yogyakarta
toto.indriyatmoko@amikom.ac.id [1], majid@amikom.ac.id [2], utama@amikom.ac.id [3] arvinclaudy@amikom.ac.id [4]

Article Info

Article history:

Received 2026-01-01
Revised 2026-01-26
Accepted 2026-02-11

Keyword:

Sentiment Classification,
LSTM,
BERT Multilanguage,
Direct Ads,
Mobile Game Reviews.

ABSTRACT

Sentiment classification of user reviews for mobile games that rely on direct advertising (direct ads) is crucial for understanding player perceptions and improving user experience. This study aims to compare the performance of two deep learning architectures, Long Short-Term Memory (LSTM) and multilingual Bidirectional Encoder Representations from Transformers (BERT) in classifying sentiment in reviews into three categories, positive, negative, and neutral. The dataset used consists of reviews from games employing direct ads, which underwent rule-based labeling and text preprocessing. The LSTM model was built from scratch using a custom embedding layer, while the multilingual BERT model was fine-tuned using a transfer learning approach. Evaluation was conducted based on accuracy, precision, recall, and F1-score metrics. Experimental results show that multilingual BERT achieves superior validation loss compared to LSTM (0.37 vs. 0.44). BERT also outperforms LSTM significantly in terms of F1-score and its ability to understand multilingual linguistic context. However, LSTM demonstrates advantages in computational efficiency and training speed. These findings offer practical recommendations for developers in selecting an appropriate sentiment analysis model based on accuracy requirements and resource availability.

I. PENDAHULUAN

Perkembangan industri game mobile yang pesat dalam dekade terakhir mendorong penggunaan model monetisasi berbasis iklan langsung (*direct ads*) seperti game Royal Match yang dapat diunduh pada *play store* [1], di mana pengalaman bermain sering kali diselingi tayangan iklan sebagai sumber pendapatan pengembang. Game dengan jenis iklan seperti ini sangat bergantung pada kepuasan dan retensi pengguna, sehingga menurut Adigüzel (2024) pemahaman terhadap sentimen pemain melalui ulasan menjadi krusial untuk pengambilan keputusan strategis bagi pengembang [2].

Sistem analisis sentimen otomatis yang memungkinkan pengembang mengevaluasi persepsi pengguna secara *real-time*, mengidentifikasi keluhan, serta meningkatkan kualitas dan pengalaman bermain tentu dapat mempermudah pengembang mendapatkan masukan yang dibutuhkan untuk game yang dikembangkan. Terkait hal tersebut, dibutuhkan klasifikasi sentimen teks ulasan dalam pemrosesan bahasa alami (*Natural Language Processing/NLP*) dalam konteks multilingual seperti game yang populer di berbagai negara, termasuk Indonesia [3].

Dua pilihan pendekatan yang digunakan untuk klasifikasi sentimen ini adalah Long Short-Term Memory (LSTM) dan Bidirectional Encoder Representations from Transformers (BERT). LSTM dirancang khusus untuk memodelkan urutan data (*sequential data*), seperti kalimat dalam teks. Dengan mekanisme gating (*input gate, forget gate, output gate*), LSTM dapat secara selektif menyimpan informasi penting dan mengabaikan yang tidak relevan. Menurut Mahadevaswam (2023) kelebihan ini sangat berguna dalam analisis sentimen, di mana kata-kata awal dalam kalimat misalnya "Awalnya bagus, tapi..." bisa berdampak besar pada makna keseluruhan [4]. Arsitektur RNN klasik sering gagal belajar dari urutan panjang karena gradien yang mengalir mundur selama pelatihan menjadi sangat kecil (*vanishing gradient*). Tulisan Al-Selwi (2023) menyebutkan bahwa LSTM mampu mengatasi masalah tersebut melalui jalur

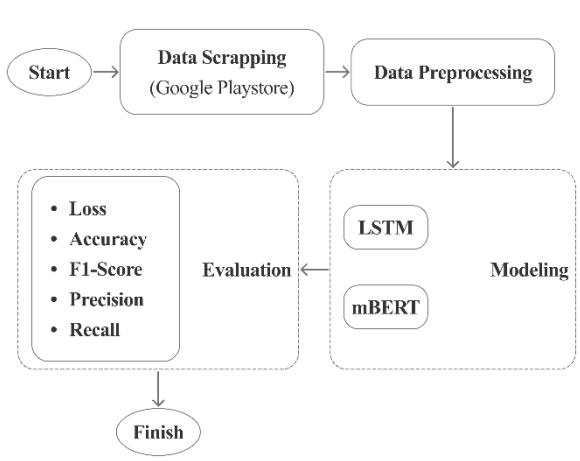
memori jangka panjang (*cell state*) yang memungkinkan gradien mengalir tanpa tereduksi drastis, sehingga model tetap mampu belajar dari dependensi jangka menengah [5]. Menurut Khan (2024), LSTM sebagai arsitektur berbasis *recurrent neural network* (RNN) mampu menangkap dependensi urutan dalam teks namun masih terbatas dalam memahami konteks jangka panjang [6].

Sebaliknya, BERT Multilingual memanfaatkan mekanisme attention dan dilatih secara pre-trained pada korpus multibahasa, sehingga lebih unggul dalam menangkap nuansa semantik dan konteks linguistik yang kompleks [7]. Shobana (2023) menjelaskan bahwa arsitektur *Transformer* dengan mekanisme *self-attention* yang terdapat pada BERT memungkinkan setiap kata dalam kalimat diproses dalam konteks seluruh kalimat secara simultan, baik dari kiri maupun kanan [8]. Hal ini sangat penting dalam analisis sentimen, karena makna suatu kata bisa berubah drastis tergantung pada konteks. Contohnya pada komentar "Game ini keren banget, tapi iklannya bikin kesel." BERT dapat memahami bahwa "keren" bersifat positif dan "kesel" negatif yang mana kalimat ini mengandung dua polaritas berbeda sehingga lebih akurat dibandingkan denngan model sekuensial seperti LSTM. BERT Multilingual telah dilatih dengan 104 bahasa termasuk Bahasa Indonesia tanpa pembatasan eksplisit antar bahasa sehingga dapat menangani teks dalam satu bahasa meski tidak pernah melihat label pelatihan dalam bahasa tersebut (*zero-shot cross-lingual transfer*) [9]. BERT juga mampu memahami kata-kata yang muncul di beberapa bahasa (misalnya "game", "fun", "lag") dalam konteks yang sesuai. Namun dengan banyaknya parameter dan kompleksitas BERT Multilingual (mBERT) tersebut mengakibatkan tingginya kebutuhan sumber daya untuk pelatihan dan inferensi yang efisien.

Berdasarkan temuan diatas, penelitian ini dilakukan untuk membandingkan kinerja model LSTM dan BERT Multilingual dalam mengklasifikasikan sentimen ulasan pengguna dari game yang menggunakan *direct ads* ke dalam tiga kategori, yaitu positif, netral, dan negatif. Evaluasi dilakukan berdasarkan metrik akurasi, presisi, recall, dan F1-score, serta dianalisis dari aspek performa dan efisiensi komputasi. Temuan dari penelitian ini diharapkan dapat memberikan rekomendasi praktis bagi pengembang game dalam memilih arsitektur model yang sesuai dengan kebutuhan akurasi dan ketersediaan sumber daya teknis. Selain itu, studi ini juga memberikan kontribusi terhadap penerapan metode NLP berbasis deep learning dalam konteks bahasa non-Inggris, khususnya ulasan game berbahasa Indonesia dan campuran.

II. METODE

Penelitian ini menggunakan pendekatan kuantitatif eksperimen, yaitu metode penelitian yang bertujuan untuk menguji hipotesis atau menjawab pertanyaan penelitian melalui manipulasi variabel dan pengukuran hasil secara objektif, terukur, dan dapat direplikasi [10]. Dalam konteks ini, variabel independen adalah arsitektur model *deep learning* Long Short-Term Memory (LSTM) dan BERT sedangkan variabel dependen adalah kinerja model dalam mengklasifikasikan sentimen ulasan pengguna, yang diukur melalui metrik kuantitatif seperti akurasi, presisi, recall, dan F1-score.



Gambar 1. Alur Penelitian

A. Pengumpulan Data

Penelitian ini bertujuan memahami persepsi dan sentimen pengguna terhadap game Android yang menggunakan model monetisasi *direct ads* iklan yang muncul otomatis selama atau setelah bermain, sering tanpa interaksi pengguna, dan kerap menjadi sumber keluhan di ulasan Google Play Store.

Dari game berjudul Royal Match mendapatkan dataset sebanyak 222.299 ulasan pengguna dari Google Play Store. Sebagai langkah awal dalam pipeline pemrosesan data, penelitian ini menggunakan pustaka Python Pandas karena kemampuannya dalam menangani data terstruktur berbentuk tabel (*DataFrame*), mendukung pemfilteran, transformasi, dan agregasi untuk pembersihan dan pelabelan data, serta kompatibel dengan berbagai format file terutama CSV, yang merupakan format utama dataset ini [11]. Dataset dalam format CSV ini berisi kolom seperti *reviewId*, *userName*, *content*, *score*, *thumbsUpCount*, *reviewCreatedVersion*, dan at. Data dimuat ke dalam *DataFrame* menggunakan *pd.read_csv(),* lalu dilakukan pemeriksaan awal meliputi dimensi data, tipe data tiap kolom, dan contoh entri untuk memastikan integritas dan kesiapan data bagi proses selanjutnya [12].

Karena dataset berasal dari sumber publik (*Google Play Store*), terdapat kemungkinan adanya entri yang tidak relevan, duplikat, atau tidak lengkap. Oleh karena itu, tahap pembersihan dilakukan dengan tiga langkah utama [13]. Langkah pertama menghapus ulasan dengan entri kosong. Kolom *content* yang berisi teks ulasan, menjadi fokus utama analisis sentimen. Entri kosong atau *NaN* dihapus dengan *.dropna(subset=['content'])* agar hanya teks yang dapat diproses yang digunakan. Berikutnya, kolom *content*

dikonversi ke tipe *string* dengan *.astype(str)* agar pemrosesan teks seperti pencarian pola atau ekspresi reguler berjalan lancar tanpa kesalahan tipe data. Langkah terakhir, ulasan yang hanya berisi spasi atau karakter kosong dihapus dengan menggunakan *df['content'].str.strip() != ''*, sehingga hanya teks bermakna yang dipertahankan (*Whitespace-Only*).

Setelah tahap ini, dataset yang tersisa terdiri dari ulasan yang valid secara kontekstual dan siap untuk dianalisis lebih lanjut dengan Pelabelan Sentimen Berbasis Aturan (*Rule-Based Sentiment Labeling*). Tahapan ini menganalisis reaksi pengguna terhadap iklan *direct ads* dengan menggabungkan rating bintang dan konteks teks ulasan [14]. Untuk itu, dibuat fungsi aturan khusus *label_sentiment(score, text)* yang memadukan kedua aspek tersebut. Fungsi ini menerapkan ulasan dengan skor 1 atau 2 diklasifikasikan sebagai negatif, karena skor rendah umumnya mencerminkan ketidakpuasan. Ulasan dengan skor 4 atau 5 biasanya mencerminkan pengalaman positif, namun jika teks ulasan mengandung frasa-frasa kunci yang mengindikasikan keluhan terhadap iklan seperti "iklan mulu", "keluar sendiri", "ganggu main", "lemot karena iklan", "banyak iklan", atau "nggak bisa main tenang" maka ulasan tersebut tetap dikategorikan sebagai negatif, meskipun skornya tinggi. Hal ini mengakomodasi fenomena umum di mana pengguna memberikan rating tinggi karena menyukai *gameplay*, tetapi tetap mengekspresikan ketidakpuasan terhadap sistem iklan. Ulasan dengan skor 3 (netral) diklasifikasikan sebagai netral, kecuali jika teks ulasannya mengandung keluhan yang eksplisit dan spesifik terhadap iklan atau kinerja aplikasi. Dalam kasus tersebut, ulasan tersebut dinaikkan ke kategori negatif untuk menangkap nuansa ketidakpuasan yang tidak tercermin dalam skor.

Pendekatan ini dipilih untuk memastikan konsistensi penuh dalam pelabelan, menghindari subjektivitas antar anotator, dan merefleksikan fenomena nyata di mana pengguna sering memberikan rating tinggi meskipun menyampaikan ketidakpuasan terhadap iklan. Karena seluruh proses dilakukan secara otomatis oleh sistem, maka tidak diperlukan *inter-annotator agreement* (IAA) atau tim anotator.

Setelah fungsi pelabelan selesai, label sentimen diterapkan ke seluruh dataset menggunakan metode ***.apply()*** pada *DataFrame Pandas*, dengan parameter ***axis=1*** untuk memproses setiap baris secara individual. Hasil akhir berupa *DataFrame* yang diperkaya dengan kolom baru bernama "sentimen", yang berisi nilai positif, netral, atau negatif untuk setiap ulasan. Sebagai validasi awal, distribusi sentimen dianalisis dengan menghitung jumlah dan persentase ulasan positif, netral, dan negatif. Misalnya, dominasi ulasan negatif terkait iklan memberikan gambaran dampak iklan *direct ads* terhadap pengalaman pengguna. Ringkasan statistik ini disimpan dalam file *sentimen_summary.txt* yang berisi total ulasan, rincian tiap kategori sentimen, dan catatan singkat tentang logika pelabelan.

Dari total 22.288 ulasan yang valid setelah pembersihan awal, terdapat 17.757 ulasan positif (79,7%), 3.187 negatif (14,3%), dan hanya 1.344 yang dilabeli sebagai netral (6,0%). Dominasi kelas positif mencerminkan kecenderungan pengguna untuk memberikan rating tinggi meskipun menyertakan keluhan terhadap iklan. Namun, ketimpangan ini berpotensi menyebabkan bias evaluasi model, di mana performa tinggi bisa jadi hanya mencerminkan kemampuan model mengenali kelas mayoritas, bukan generalisasi yang sebenarnya. Oleh karena itu, dataset diseimbangkan menjadi 20.000 sampel per kelas melalui kombinasi *oversampling* (untuk kelas minoritas) dan *undersampling* (untuk kelas mayoritas), sehingga total menjadi 60.000 ulasan dengan distribusi sempurna (1:1:1). Pendekatan ini memastikan bahwa metrik evaluasi (akurasi, F1-score, dll.) benar-benar mencerminkan kemampuan model dalam membedakan ketiga kelas secara adil.

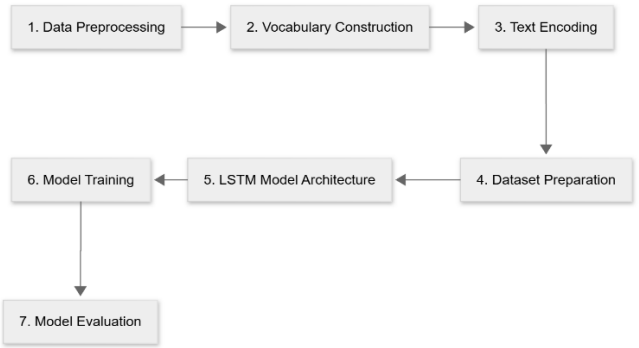
Secara linguistik, dataset yang diperoleh bersifat multilingual sehingga mencerminkan basis pengguna global Royal Match. Analisis awal menunjukkan komposisi bahasa Indonesia: ~70% ("tolong tambahkan bahasa indonesia", "susah banget dapat bintang"), bahasa Inggris ~25% ("*good game but too many ads*", "*love it!*") dan campuran sebanyak ~5% ("game nya bagus tapi iklannya *annoying* banget").

Fenomena multilingual ini menjadikan tugas klasifikasi sentimen lebih menantang, karena model harus mampu menangani variasi linguistik, termasuk ekspresi dan struktur kalimat campuran. Karakteristik ini juga menjadi dasar pemilihan arsitektur BERT Multilingual, yang dirancang khusus untuk memahami konteks lintas bahasa.

Dengan rangkaian langkah di atas, penelitian ini memastikan bahwa data yang digunakan tidak hanya valid dan bersih, tetapi juga relevan secara kontekstual dengan tujuan untuk memahami bagaimana model iklan direct ads dalam game Android memengaruhi persepsi dan kepuasan pengguna, sebagaimana tercermin dalam ulasan mereka di platform digital.

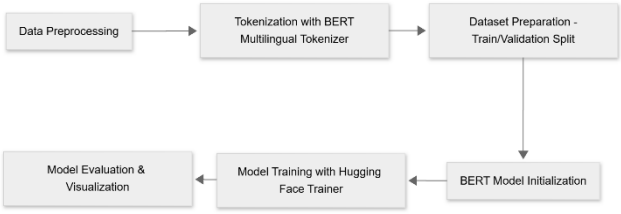
B. Teknik Analisis Data

Dalam penelitian ini, dua pendekatan *deep learning Long Short-Term Memory* (LSTM) dan *Bidirectional Encoder Representations from Transformers* (BERT) Multilanguage digunakan untuk mengklasifikasikan sentimen ulasan pengguna dari game mobile yang menggunakan sistem monetisasi berbasis iklan langsung (*direct ads*). Pendekatan LSTM berfokus pada model berbasis *Recurrent Neural Network* (RNN), yang mampu menangkap dependensi urutan kata dalam teks ulasan. Langkah-langkah teknis yang diterapkan dapat dilihat pada Gambar 2.



Gambar 2. Alur pengembangan model LSTM untuk pemrosesan teks

Proses pengembangan model LSTM dimulai dari pra-pemrosesan data, di mana teks mentah dibersihkan dan dinormalisasi melalui langkah-langkah seperti *lowercasing*, penghapusan *stopwords*, dan tokenisasi [15]. Selanjutnya, konstruksi kosakata dilakukan untuk membangun kamus indeks unik dari seluruh token yang muncul yang kemudian digunakan dalam tahap pengkodean teks untuk mengkonversi urutan kata menjadi representasi numerik. Setelah itu, data yang telah dikodekan disusun dan dipartisi dalam *dataset preparation*, termasuk penyesuaian panjang urutan (*padding* atau *truncating*) serta pembagian menjadi *data training*, validasi, dan/atau pengujian. Model LSTM kemudian dibangun dalam tahap arsitektur model, yang mencakup penentuan jumlah lapisan, ukuran *hidden state*, serta lapisan pendukung seperti *embedding* dan *dense layer*. Proses pelatihan model dilakukan secara iteratif dengan optimisasi berbasis gradien untuk meminimalkan *data lost* [16], sementara performa dipantau melalui validasi untuk menghindari *overfitting* [17]. Terakhir, model dievaluasi dalam tahap evaluasi model menggunakan metrik kinerja yang relevan terhadap data uji independen guna mengukur kemampuan generalisasi dan kesiapan implementasi. Secara garis besar, model LSTM yang digunakan terdiri dari dua lapisan rekursif dengan dimensi *hidden state* sebesar 128. *Embedding layer* memiliki ukuran 128 dimensi, dan *dropout* sebesar 0.3 diterapkan setelah representasi LSTM sebelum klasifikasi akhir. Model dilatih selama 50 *epoch* dengan *optimizer AdamW (learning rate = 2×10^{-3}, weight decay = 0.01)*, serta *gradient clipping* pada norma gradien maksimum 1.0. *Learning rate* dikurangi separuh setiap 10 *epoch* menggunakan *scheduler StepLR*. Pendekatan BERT Multilanguage dirancang untuk memahami konteks bahasa secara global. Langkah-langkah teknis yang diterapkan ditunjukkan pada Gambar 3.



Gambar 3. Alur pengembangan model BERT untuk pemrosesan teks

Proses pengembangan model klasifikasi teks berbasis BERT multilingual dalam penelitian ini mengikuti alur kerja yang terdiri dari enam tahap utama. Dimulai dari *Data Preprocessing*, di mana data teks mentah dibersihkan dan ditransformasi menjadi format standar melalui normalisasi, penghapusan *noise*, dan strukturisasi agar siap diproses. Selanjutnya, teks di-tokenisasi menggunakan BERT Multilingual Tokenizer yang mampu menangani berbagai bahasa dengan membagi teks menjadi subword units dan menambahkan token khusus seperti [CLS] dan [SEP] sesuai kebutuhan arsitektur BERT [18]. Setelah tokenisasi, dataset dibagi menjadi subset *Train/Validation Split* secara acak untuk memungkinkan pelatihan model dan evaluasi performa selama proses belajar, sehingga dapat mendeteksi potensi *overfitting* dan memastikan generalisasi yang baik. Tahap berikutnya adalah *BERT Model Initialization*, di mana model diinisialisasi dengan *pre-trained weights* dari BERT multilingual, memberikan fondasi linguistik yang kuat sebelum *fine-tuning* pada tugas spesifik [19]. Proses pelatihan dilakukan menggunakan *Hugging Face Trainer*, sebuah *library* yang menyediakan antarmuka tingkat tinggi untuk melatih model transformer secara efisien, dengan optimasi parameter melalui *backpropagation* dan validasi periodik untuk memantau konvergensi [20]. Terakhir, model dievaluasi secara komprehensif menggunakan metrik seperti akurasi, presisi, recall, F1-score, serta hasilnya divisualisasikan untuk memudahkan interpretasi performa, identifikasi pola kesalahan, dan analisis perbandingan antar kelas. Alur kerja ini mencerminkan pendekatan *end-to-end* yang terstruktur, memanfaatkan kekuatan model pra-pelatihan dan alat modern untuk membangun sistem NLP.

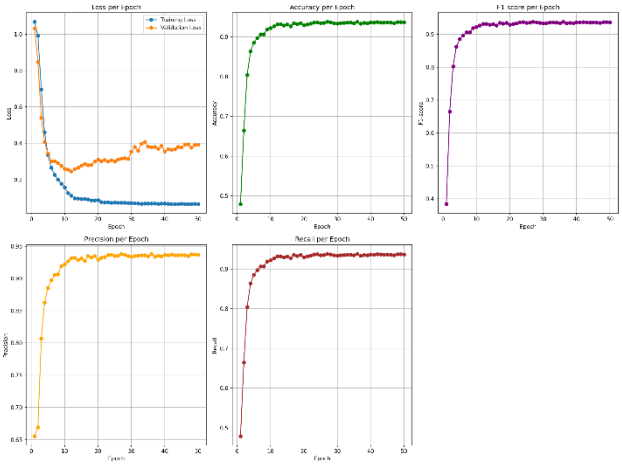
III. HASIL DAN PEMBAHASAN

Penelitian ini bertujuan untuk membandingkan kinerja dua arsitektur deep learning, yaitu *Long Short-Term Memory* (LSTM) dan *Bidirectional Encoder Representations from Transformers Multilanguage* (BERT *Multilanguage*) dalam tugas klasifikasi sentimen ulasan pengguna game mobile berbasis *direct ads*. Dataset yang digunakan telah melalui proses pembersihan dan *balancing* untuk memastikan distribusi kelas yang seimbang, sehingga hasil evaluasi dapat diandalkan [21]. Evaluasi dilakukan selama 50 epoch dengan metrik utama yaitu loss, akurasi, F1-score, presisi, dan recall. Berikut adalah analisis terhadap hasil pelatihan masing-masing model.

A. Analisis Model LSTM

Gambar 4. menampilkan perubahan nilai loss (fungsi kerugian) baik pada data pelatihan (*training loss*) maupun data validasi (*validation loss*) sepanjang 50 epoch. Pada awal pelatihan (epoch 0-5), terjadi penurunan drastis pada kedua kurva yang menunjukkan bahwa model cepat belajar dari data. Setelah epoch ke-20, *training loss* stabil di sekitar nilai 0.08, sementara *validation loss* berfluktuasi di sekitar 0.39. Perbedaan yang cukup signifikan antara *training loss* dan

validation loss mengindikasikan adanya potensi *overfitting*, di mana model menjadi terlalu spesifik pada data pelatihan dan tidak sepenuhnya generalisasi ke data baru [22]. Namun, karena *validation loss* tidak meningkat secara drastis, *overfitting* ini masih dalam batas wajar.



Gambar 4. Hasil analisis Model LSTM.

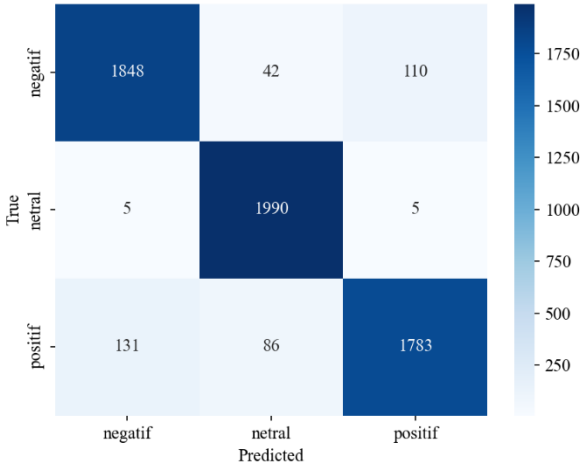
Akurasi model LSTM yang ditunjukkan pada Gambar 4. Memberikan gambaran peningkatan yang pesat sejak awal pelatihan dari sekitar 0.47 pada epoch pertama menjadi lebih dari 0.94 pada epoch ke-10. Setelah itu, akurasi terus naik secara perlahan dan stabil hingga mencapai puncaknya di sekitar 0.95 pada epoch ke-50. Ini menunjukkan bahwa model berhasil belajar untuk mengklasifikasikan sentimen dengan sangat baik, meskipun ada sedikit fluktuasi di akhir pelatihan yang bisa disebabkan oleh *noise* atau kompleksitas data.

F1-score, yang merupakan rata-rata harmonik dari presisi dan *recall*, juga menunjukkan tren peningkatan yang kuat. Nilainya melonjak dari sekitar 0.38 pada epoch pertama menjadi lebih dari 0.92 pada epoch ke-10, dan kemudian stabil di atas 0.93 hingga akhir pelatihan. Nilai *F1-score* yang tinggi ini menegaskan bahwa model tidak hanya mampu mengidentifikasi kelas mayoritas dengan baik, tetapi juga menjaga keseimbangan antara presisi dan *recall* untuk semua kelas.

Presisi model LSTM meningkat secara signifikan dari 0.66 pada epoch awal menjadi sekitar 0.93 pada epoch ke-15, lalu stabil di kisaran 0.94-0.95. Hal ini menunjukkan bahwa ketika model memprediksi suatu kelas, probabilitas prediksinya benar sangat tinggi. Dengan kata lain, model memiliki tingkat kesalahan positif palsu yang rendah.

Recall juga mengalami peningkatan yang dramatis, dari sekitar 0.47 pada epoch pertama menjadi lebih dari 0.93 pada epoch ke-15, dan kemudian stabil di atas 0.94. Nilai recall yang tinggi menunjukkan bahwa model mampu mendeteksi sebagian besar sampel yang benar-benar termasuk dalam suatu kelas. Hal ini penting dalam konteks analisis sentimen, di mana kita ingin memastikan bahwa keluhan atau pujian penting dari pengguna tidak terlewatkan.

Secara keseluruhan, model LSTM menunjukkan performa yang sangat baik, dengan akurasi akhir di atas 95% dan *F1-score* di atas 93%. Model ini cepat konvergen dan mampu mempelajari pola dalam teks ulasan dengan efektif. Namun, perbedaan antara *training loss* dan *validation loss* memberi sinyal bahwa model mungkin sedikit *overfit*, meskipun hal ini tidak terlalu merugikan karena performa validasi tetap sangat tinggi.



Gambar 5. *Confusion Matrix* model LSTM pada dataset validasi

Berdasarkan Gambar 5, model LSTM menunjukkan performa yang solid pada seluruh kelas sentimen dengan performa paling unggul terlihat pada kelas netral yang ditunjukkan dengan klasifikasi yang benar sebanyak 1990 dari 2000 ulasan netral (akurasi kelas ≈ 99,5%). Hanya 5 ulasan salah yang diklasifikasikan sebagai negatif dan 5 lainnya sebagai positif. Hal ini menunjukkan bahwa LSTM sangat andal dalam mengenali ulasan yang tidak mengandung ekspresi emosional kuat.

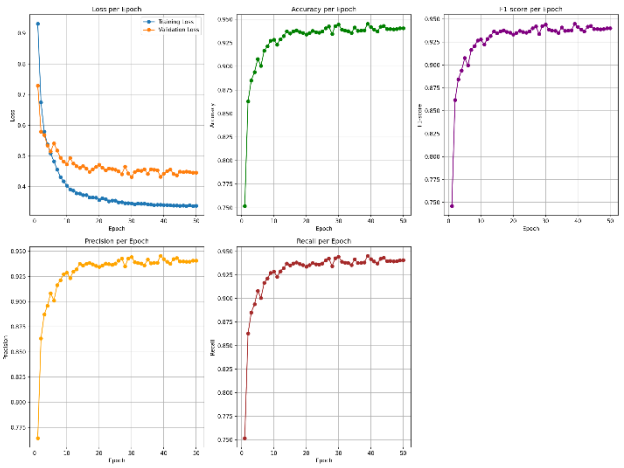
Pada kelas negatif, model berhasil mengklasifikasikan 1848 dari 2000 ulasan negatif dengan benar (akurasi kelas ≈ 92,4%). Kesalahan utama terjadi ketika 110 ulasan negatif diklasifikasikan sebagai positif dan 42 sebagai netral. Pola ini mengindikasikan bahwa LSTM cenderung salah menafsirkan keluhan yang disampaikan secara halus atau ambigu sebagai sentimen positif.

Akurasi kelas untuk sentimen positif mencapai 89,2% (1783 dari 2000). Sebanyak 131 ulasan positif salah diklasifikasikan sebagai negatif, dan 86 sebagai netral. Ini menunjukkan bahwa LSTM lebih sensitif terhadap kata-kata negatif dalam konteks campuran (misalnya, “game-nya seru tapi iklannya ganggu”), sehingga cenderung mengklasifikasikan ulasan tersebut sebagai negatif.

B. Analisis Model BERT Multilanguage

Mirip dengan LSTM, model BERT juga menunjukkan penurunan *loss* yang cepat pada awal pelatihan. *Training loss* turun dari sekitar 1.07 menjadi stabil di bawah 0.10, sementara validation *loss* turun dari sekitar 0.85 menjadi stabil di sekitar 0.39 (Gambar 6.). Perbedaan antara *training loss* dan *validation loss* juga terlihat, namun lebih kecil dibandingkan dengan LSTM, yang menunjukkan bahwa

BERT memiliki kemampuan generalisasi yang lebih baik. Kurva *validation loss* yang relatif stabil tanpa peningkatan menunjukkan bahwa model tidak mengalami *overfitting* yang parah.



Gambar 6. Grafik hasil analisis model *BERT Multilanguage.*

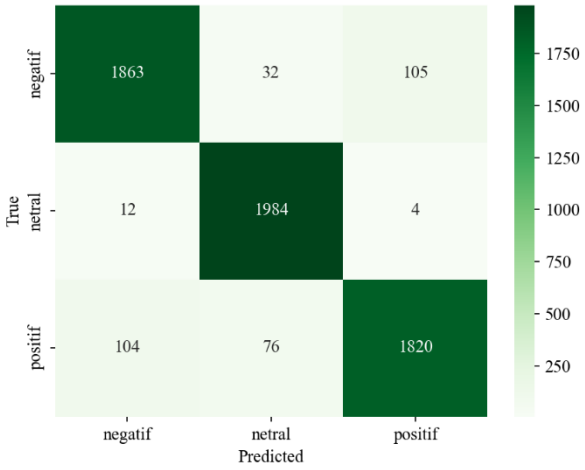
Akurasi model BERT meningkat sangat cepat, dari sekitar 0.47 pada *epoch* pertama menjadi lebih dari 0.92 pada *epoch* ke-10, dan kemudian stabil di atas 0.94 hingga akhir pelatihan. Performa akurasi ini sangat kompetitif dengan LSTM, menunjukkan bahwa representasi kontekstual yang dihasilkan oleh BERT efektif untuk tugas klasifikasi sentimen.

F1-score model BERT juga menunjukkan tren peningkatan yang kuat, dari sekitar 0.38 pada *epoch* pertama menjadi lebih dari 0.92 pada *epoch* ke-10, dan kemudian stabil di atas 0.94. Nilai *F1-score* yang tinggi ini menegaskan bahwa model BERT mampu mencapai keseimbangan yang baik antara presisi dan *recall* untuk semua kelas sentimen.

Precision model BERT meningkat dari 0.65 pada *epoch* awal menjadi sekitar 0.93 pada *epoch* ke-15, lalu stabil di kisaran 0.94-0.95. Nilai presisi yang tinggi ini menunjukkan bahwa prediksi model BERT sangat andal, dengan tingkat kesalahan positif palsu *(false positive)* yang rendah.

Recall model BERT juga meningkat secara signifikan, dari sekitar 0.47 pada *epoch* pertama menjadi lebih dari 0.93 pada *epoch* ke-15, dan kemudian stabil di atas 0.94. Nilai recall yang tinggi ini menunjukkan bahwa model BERT mampu mendeteksi sebagian besar sampel yang benar-benar termasuk dalam suatu kelas, yang sangat penting untuk memastikan tidak ada sentimen penting yang terlewatkan.

Model BERT *Multilanguage* menunjukkan performa yang sangat impresif, dengan akurasi akhir di atas 94% dan F1-score di atas 94%. Kemampuan model ini untuk memahami konteks linguistik yang kompleks, bahkan dalam teks multibahasa, membuatnya sangat cocok untuk tugas ini. Meskipun waktu pelatihan dan kebutuhan sumber daya komputasi jauh lebih besar dibandingkan LSTM, performa akhirnya sangat kompetitif dan bahkan sedikit lebih unggul dalam hal stabilitas dan kemampuan generalisasi.



Gambar 7. *Confusion Matrix* model BERT Multilingual pada dataset validasi

Berdasarkan Gambar 7, model BERT Multilingual menunjukkan kemampuan superior dalam memahami konteks linguistik kompleks. BERT unggul pada kelas netral, dengan 1984 dari 2000 ulasan diklasifikasikan dengan benar (akurasi kelas ≈ 99,2%). Terdapat 12 ulasan salah sebagai negatif dan 4 sebagai positif. Meski sedikit lebih banyak kesalahan dibanding LSTM, performa tetap sangat tinggi.

BERT mengklasifikasikan 1863 dari 2000 ulasan negatif dengan benar (akurasi kelas ≈ 93,2%), sedikit lebih baik daripada LSTM. Kesalahan utama terdapat pada 105 ulasan negatif yang diklasifikasikan sebagai positif dan 32 sebagai netral, jumlah tersebut lebih rendah dibanding LSTM yang menunjukkan pemahaman konteks yang lebih baik. Pada kelas positif BERT mencapai akurasi kelas 91,0% (1820 dari 2000). Hanya 104 ulasan positif salah sebagai negatif dan 76 sebagai netral, lebih sedikit dibanding LSTM, terutama pada kesalahan positif → negatif.

C. *Perbandingan*

TABEL I
Perbandingan Hasil Analisis *LSTM & BERT Multilangguage* pada *epoch* 50

Metrik Evaluasi	***LSTM***	***BERT Multilingual***
Validation Loss	0.44	0.39
Accuracy	0.94	0.94
F1-score	0.94	0.93
Precision	0.94	0.94
Recall	0.94	0.93

Secara keseluruhan, kedua model menunjukkan performa yang sangat baik dalam mengklasifikasikan sentimen ulasan game mobile. Baik LSTM maupun BERT Multilanguage mencapai akurasi dan *F1-score* yang sangat tinggi (di atas 93%). Perbedaan performa antara keduanya sangat kecil,

dengan BERT memiliki keunggulan tipis dalam hal stabilitas dan kemampuan generalisasi. Dari segi kecepatan konvergensi, kedua model konvergen dengan cepat, peningkatan signifikan dalam metrik performa terjadi dalam 10-15 epoch pertama.

Model LSTM menunjukkan indikasi overfitting yang lebih jelas dibandingkan BERT, meskipun tidak terlalu parah. Model BERT, dengan arsitektur transformer-nya yang didesain untuk menangkap dependensi jangka panjang, menunjukkan kemampuan generalisasi yang lebih baik. Meskipun tidak ditampilkan dalam grafik, perlu dicatat bahwa model BERT jauh lebih berat secara komputasi dibandingkan LSTM. Pelatihan BERT memerlukan lebih banyak waktu dan sumber daya GPU. Untuk implementasi di lingkungan dengan keterbatasan sumber daya, LSTM bisa menjadi pilihan yang lebih praktis tanpa mengorbankan akurasi yang signifikan.

Model BERT secara konsisten menunjukkan akurasi per kelas yang lebih tinggi untuk kelas *positif* dan *negatif*, terutama dalam menghindari kesalahan antara keduanya. Ini mencerminkan kemampuannya dalam menangkap nuansa semantik dan konteks global melalui mekanisme *attention*. Sedangkan LSTM sedikit lebih unggul dalam klasifikasi kelas netral, dengan 6 kesalahan lebih sedikit dibanding BERT. Ini menunjukkan bahwa arsitektur berbasis RNN masih efektif untuk teks netral yang strukturnya sederhana.

Kedua model mengalami kesulitan serupa dalam mengklasifikasikan ulasan campuran (misalnya, pujian terhadap gameplay tetapi keluhan terhadap iklan). Namun, BERT lebih mampu memisahkan komponen positif dan negatif dalam satu kalimat, sehingga menghasilkan prediksi yang lebih akurat.

Performa BERT Multilingual dibandingkan LSTM dalam klasifikasi sentimen ulasan multilingual (terutama campuran Bahasa Indonesia dan Inggris) dapat dijelaskan secara teknis melalui dua mekanisme inti arsitekturnya. Pertama, *self-attention global* memungkinkan BERT menangkap hubungan semantik antar kata secara simultan tanpa terikat urutan linear, sehingga mampu mengenali bahwa frasa seperti "iklan mulu" atau "annoying banget" mendominasi sentimen meskipun muncul bersama pujian [23]. Kedua, tokenizer *WordPiece* dan representasi embedding multilingual yang dipelajari selama pra-pelatihan pada 104 bahasa memungkinkan BERT menyelaraskan makna lintas bahasa. Misalnya, "good" dan "bagus" memiliki vektor yang berdekatan sehingga ulasan campuran seperti "game nya seru tapi iklannya annoying" diproses sebagai satu unit semantik koheren, bukan sebagai entitas terpisah [24]. Sebaliknya, LSTM hanya mengandalkan sequential processing dan embedding lokal yang tidak memiliki pengetahuan lintas bahasa, sehingga kurang efektif dalam menangani kode-*switching* dan ambiguitas kontekstual yang umum dalam data nyata. Kombinasi kemampuan ini menjelaskan mengapa BERT tidak hanya unggul dalam akurasi, tetapi juga lebih stabil dalam generalisasi pada tugas sentimen multilingual.

Temuan tersebut memperkuat kesimpulan bahwa BERT Multilingual lebih cocok untuk aplikasi yang memprioritaskan akurasi tinggi dan pemahaman konteks, sementara LSTM tetap menjadi alternatif yang efisien jika sumber daya komputasi terbatas dan fokus utama adalah pada deteksi sentimen netral.

IV. KESIMPULAN

Penelitian ini berhasil membandingkan kinerja model LSTM dan BERT Multilangual dalam klasifikasi sentimen ulasan game mobile berbahasa Indonesia dan campuran dengan skema *direct ads*. Hasil menunjukkan bahwa BERT *Multilanguage* secara konsisten unggul dalam akurasi, stabilitas pelatihan, dan kecepatan konvergensi, dengan metrik evaluasi (akurasi, presisi, recall, dan F1-score) yang stabil di atas 0,93. Sementara itu, LSTM juga menunjukkan performa yang baik (mendekati 0,93–0,94), namun membutuhkan lebih banyak epoch untuk konvergen dan menunjukkan fluktuasi yang lebih besar selama pelatihan. Berdasarkan temuan ini, disarankan bagi pengembang game engan skema *direct ads* yang memprioritaskan akurasi tinggi dan kemampuan pemahaman konteks linguistik kompleks untuk menggunakan BERT Multilanguage, sedangkan LSTM dapat menjadi alternatif yang layak jika terdapat keterbatasan sumber daya komputasi atau kebutuhan inferensi cepat dengan konsekuensi akurasi yang minimal. Lebih lanjut, rekomendasi pemilihan model dapat dikontekstualisasikan ke dalam skenario nyata seperti startup game kecil yang terbatas sumber daya dapat mengadopsi LSTM untuk efisiensi komputasi tanpa mengorbankan akurasi secara signifikan, sedangkan perusahaan game besar dengan infrastruktur canggih sebaiknya memilih BERT Multilingual untuk memaksimalkan pemahaman konteks multibahasa dan stabilitas prediksi. Untuk penelitian lanjutan, disarankan untuk menguji model pada data yang lebih beragam, menerapkan teknik augmentasi data, serta mengevaluasi efisiensi inferensi secara eksplisit di lingkungan produksi.

DAFTAR PUSTAKA

[1] "Royal Match," Dream Games, Ltd. Accessed: Nov. 09, 2025. [Online]. Available: https://play.google.com/store/apps/details?id=com.dreamgames.royalmatch&hl=en

[2] S. Adigüzel, "The Effect Of In-Game Advertising As A Marketing Technique On The Purchasing Behavior Of Generation Z In Turkey.," *Airlangga International Journal of Islamic Economics & Finance*, vol. 7, no. 2, 2024.

[3] K. Kopanov, "Comparative Performance of Advanced NLP Models and LLMs in Multilingual Geo-Entity Detection," in *ACM International Conference Proceeding Series*, Association for Computing Machinery, May 2024, pp. 106–110. doi: 10.1145/3660853.3660878.

[4] U. B. Mahadevaswamy and P. Swathi, "Sentiment Analysis using Bidirectional LSTM Network," *Procedia Comput. Sci.*, vol. 218, pp. 45–56, 2023, doi: https://doi.org/10.1016/j.procs.2022.12.400.

[5] S. M. Al-Selwi, M. F. Hassan, S. J. Abdulkadir, and A. Muneer, "LSTM inefficiency in long-term dependencies regression problems," *Journal of Advanced Research in Applied Sciences and Engineering Technology*, vol. 30, no. 3, pp. 16–31, 2023.

[6] S. S. Khan, P. K. Mondal, S. Shaqib, N. Ahmed, N. N. I. Prova, and A. Sattar, "Performance Analysis of LSTM and Bi-LSTM Model with Different Optimizers in Bangla Sentiment Analysis," in *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 2024, pp. 1–7. doi: 10.1109/ICCCNT61001.2024.10726142.

[7] N. M. Gardazi, A. Daud, M. K. Malik, A. Bukhari, T. Alsahfi, and B. Alshemaimri, "BERT applications in natural language processing: a review," *Artif. Intell. Rev.*, vol. 58, no. 6, Jun. 2025, doi: 10.1007/s10462-025-11162-5.

[8] J. Shobana and M. Murali, "An improved self attention mechanism based on optimized BERT-BiLSTM model for accurate polarity prediction," *Comput. J.*, vol. 66, no. 5, pp. 1279–1294, 2023.

[9] A. De Varda and R. Zamparelli, "Multilingualism Encourages Recursion: a Transfer Study with mBERT," in *Proceedings of the 4th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, E. Vylomova, E. Ponti, and R. Cotterell, Eds., Seattle, Washington: Association for Computational Linguistics, Jul. 2022, pp. 1–10. doi: 10.18653/v1/2022.sigtyp-1.1.

[10] M. S. Jailani, "Teknik pengumpulan data dan instrumen penelitian ilmiah pendidikan pada pendekatan kualitatif dan kuantitatif," *IHSAN: Jurnal Pendidikan Islam*, vol. 1, no. 2, pp. 1–9, 2023.

[11] T. Maeno *et al.*, "PanDA: Production and Distributed Analysis System," Dec. 01, 2024, *Springer Nature*. doi: 10.1007/s41781-024-00114-3.

[12] I. Harouni, "The Modern Methods of Data Analysis in Social Research: Python Programming Language and its Pandas Library as an Example-a Theoretic Study," *Social Empowerment Journal*, vol. 6, no. 1, pp. 56–70, 2024, Accessed: Dec. 18, 2025. [Online]. Available: https://www.ajol.info/index.php/sej/article/view/274626/259255

[13] F. Ridzuan and W. M. N. Wan Zainon, "A review on data cleansing methods for big data," in *Procedia Computer Science*, Elsevier B.V., 2019, pp. 731–738. doi: 10.1016/j.procs.2019.11.177.

[14] N. Kumar and B. R. Hanji, "Combined sentiment score and star rating analysis of travel destination prediction based on user preference using morphological linear neural network model with correlated topic modelling approach," *Multimed. Tools Appl.*, vol. 83, no. 22, pp. 61347–61378, 2024, doi: 10.1007/s11042-023-17995-y.

[15] B. Bala and S. Behal, "A Brief Survey of Data Preprocessing in Machine Learning and Deep Learning Techniques," in *2024 8th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, 2024, pp. 1755–1762. doi: 10.1109/I-SMAC61858.2024.10714767.

[16] H. Li *et al.*, "Enhancing Large Language Model Performance with Gradient-Based Parameter Selection," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 23, pp. 24431–24439, Apr. 2025, doi: 10.1609/aaai.v39i23.34621.

[17] M. Zhang, X. Ye, Q. Liu, P. Ren, S. Wu, and Z. Chen, "Uncovering overfitting in large language model editing," *arXiv preprint arXiv:2410.07819*, 2024, doi: https://doi.org/10.48550/arXiv.2410.07819.

[18] A. N. Azhar and M. L. Khodra, "Fine-tuning Pretrained Multilingual BERT Model for Indonesian Aspect-based Sentiment Analysis," in *2020 7th International Conference on Advance Informatics: Concepts, Theory and Applications (ICAICTA)*, 2020, pp. 1–6. doi: 10.1109/ICAICTA49861.2020.9428882.

[19] H. Y. Kim, N. Balasubramanian, and B. Kang, "On initializing transformers with pre-trained embeddings," *arXiv preprint arXiv:2407.12514*, 2024.

[20] S. M. Jain, "Hugging Face," in *Introduction to Transformers for NLP: With the Hugging Face Library and Models to Solve Problems*, S. M. Jain, Ed., Berkeley, CA: Apress, 2022, pp. 51–67. doi: 10.1007/978-1-4842-8844-3_4.

[21] I. F. Ilyas and T. Rekatsinas, "Machine learning and data cleaning: Which serves the other?," *ACM Journal of Data and Information Quality (JDIQ)*, vol. 14, no. 3, pp. 1–11, 2022.

[22] T. Liu *et al.*, "Mitigating heterogeneous token overfitting in llm knowledge editing," *arXiv preprint arXiv:2502.00602*, 2025.

[23] N. Kitaev, S. Cao, and D. Klein, "Multilingual Constituency Parsing with Self-Attention and Pre-Training," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, A. Korhonen, D. Traum, and L. Màrquez, Eds., Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 3499–3505. doi: 10.18653/v1/P19-1340.

[24] X. Song, A. Salcianu, Y. Song, D. Dopson, and D. Zhou, "Fast WordPiece Tokenization," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, M.-F. Moens, X. Huang, L. Specia, and S. W. Yih, Eds., Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 2089–2103. doi: 10.18653/v1/2021.emnlp-main.160.