

Analisis Sentimen Ulasan Pengguna Aplikasi Duolingo Menggunakan IndoBERT

Sentiment Analysis of User Reviews of The Duolingo Application Using IndoBERT

Abstrak

Banyaknya ulasan pengguna Duolingo di Indonesia membuat proses evaluasi secara manual menjadi sulit dilakukan. Oleh karena itu, penelitian ini bertujuan untuk menganalisis sentimen pengguna guna mengetahui tanggapan dan pengalaman mereka terhadap aplikasi Duolingo. Penelitian ini menggunakan model IndoBERT untuk mengelompokkan ulasan ke dalam tiga kategori sentimen, yaitu positif, netral, dan negatif. Hasil pengujian menunjukkan bahwa model IndoBERT memiliki kinerja yang sangat baik dengan tingkat Accuracy sebesar 92% dan Macro F1-Score sebesar 71%. Sebagian besar ulasan yang dianalisis termasuk sentimen positif (83,47%), yang menunjukkan tingginya tingkat kepuasan pengguna terhadap aplikasi. Kelas sentimen positif memperoleh F1-Score sebesar 97%, sedangkan kelas netral memiliki Macro F1-Score terendah sebesar 44% karena banyak ulasan netral yang singkat dan ambigu. Hasil penelitian ini menunjukkan bahwa IndoBERT efektif dan andal dalam memetakan opini publik secara otomatis serta memberikan gambaran mengenai persepsi pengguna terhadap aplikasi Duolingo.

Kata kunci: Analisis Sentimen, IndoBERT, Duolingo, Ulasan Pengguna, Klasifikasi Sentimen.

Abstract

The large volume of user reviews for Duolingo in Indonesia makes the manual evaluation process difficult to conduct. Therefore, this study aims to analyze user sentiment to understand their responses and experiences with the Duolingo application. This study utilizes the IndoBERT model to classify reviews into three sentiment categories: positive, neutral, and negative. The evaluation results demonstrate that the IndoBERT model performs exceptionally well, achieving an Accuracy rate of 92% and a Macro F1-Score of 71%. The majority of the analyzed reviews fall into the positive sentiment category (83.47%), indicating a high level of user satisfaction with the application. The positive sentiment class achieves an F1-Score of 97%, whereas the neutral class yields the lowest Macro F1-Score at 44% due to many neutral reviews being brief and ambiguous. The findings of this study show that IndoBERT is effective and reliable in automatically mapping public opinion and providing an overview of user perception toward the Duolingo application.

Keywords: Sentiment Analysis, IndoBERT, Duolingo, User Reviews, Sentiment Classification.

1. PENDAHULUAN

Kemampuan berbahasa asing menjadi kebutuhan penting di era globalisasi untuk mendukung pendidikan, pekerjaan, dan komunikasi internasional [1]. Namun, keterbatasan waktu, biaya kursus, dan akses terhadap lembaga pembelajaran formal masih menjadi kendala bagi sebagian masyarakat. Perkembangan teknologi digital menghadirkan solusi melalui aplikasi pembelajaran bahasa berbasis mobile atau *Mobile-Assisted Language Learning* (MALL), yang memungkinkan proses belajar dilakukan secara fleksibel melalui *smartphone*

1180

[2].

Salah satu aplikasi pembelajaran bahasa yang paling populer adalah Duolingo. Berdasarkan data Google Play Store tahun 2025, Duolingo telah diunduh lebih dari 500 juta kali secara global, jauh melampaui aplikasi sejenis seperti Babbel, Busuu, dan Memrise [3]. Popularitas tersebut didukung oleh penerapan gamifikasi, seperti sistem level, poin pengalaman (*XP*), papan peringkat, dan *streak* harian yang mampu meningkatkan motivasi belajar pengguna [4]. Tingginya jumlah pengguna menghasilkan volume ulasan yang sangat besar, yang berisi berbagai pengalaman, saran, maupun keluhan terkait penggunaan aplikasi [5].

Besarnya jumlah ulasan membuat analisis manual menjadi kurang efektif sehingga diperlukan pendekatan otomatis berbasis *Natural Language Processing* (NLP). Dalam analisis sentimen berbahasa Indonesia, model IndoBERT terbukti mampu memberikan performa yang lebih baik dibandingkan metode *machine learning* konvensional karena dapat memahami konteks semantik secara lebih mendalam. Penelitian sebelumnya menunjukkan bahwa IndoBERT mampu menghasilkan akurasi yang lebih tinggi dibandingkan beberapa model multibahasa dan metode klasifikasi tradisional pada data teks berbahasa Indonesia [6]. Selain itu, IndoBERT memiliki nilai loss paling rendah serta waktu komputasi yang lebih cepat, sehingga lebih efektif dalam memproses teks berbahasa Indonesia karena dilatih menggunakan korpus berita berbahasa Indonesia. Penelitian lain oleh Nur et al. (2025) juga membuktikan efektivitas fine-tuning IndoBERT dalam menganalisis persepsi konsumen terhadap merek teknologi di media sosial X merepresentasikan permasalahan produk dan layanan secara jelas [7]. Meskipun berbagai penelitian telah membuktikan keunggulan IndoBERT, namun secara khusus menerapkan metode tersebut pada ulasan pengguna aplikasi pembelajaran bahasa berbahasa Indonesia masih terbatas.

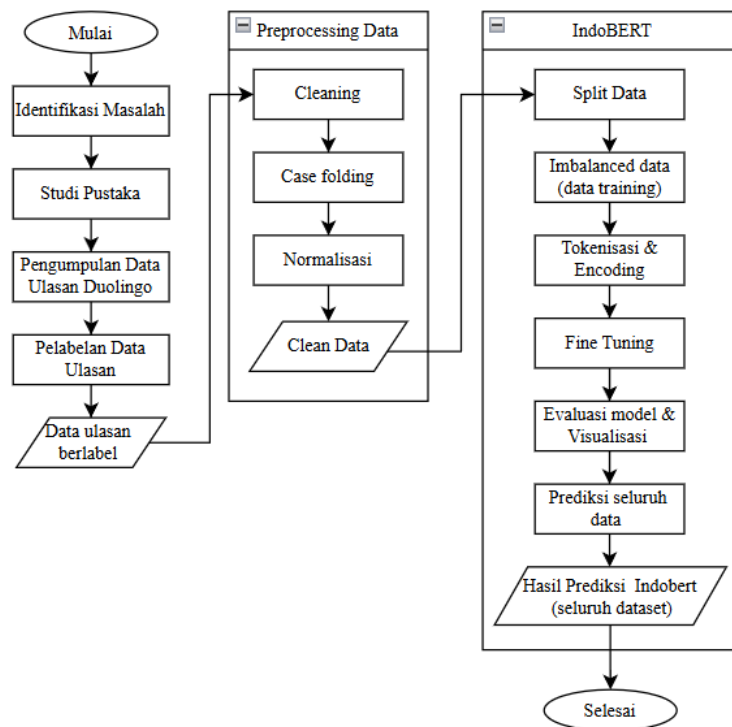
Berdasarkan uraian tersebut, penelitian ini bertujuan untuk menganalisis sentimen ulasan pengguna aplikasi Duolingo menggunakan model IndoBERT. Pendekatan ini diharapkan mampu mengubah data ulasan publik menjadi informasi yang terstruktur dan bermakna, sehingga dapat memberikan rekomendasi berbasis data dalam peningkatan kualitas pengalaman belajar. Secara akademis, penelitian ini berkontribusi dalam memperluas penerapan model NLP berbasis Transformer pada analisis teks pendek berbahasa Indonesia serta memperkuat literatur mengenai integrasi analisis sentimen dalam konteks aplikasi pembelajaran digital.

2. METODOLOGI PENELITIAN

Penelitian ini menggunakan metode IndoBERT untuk analisis sentimen pada ulasan pengguna aplikasi Duolingo. Tahapan penelitian meliputi pengumpulan data, *data understanding*, pelabelan data, *preprocessing*, analisis sentimen, dan evaluasi hasil. Data diperoleh melalui teknik *scraping* dari Google Play Store menggunakan package *google_play_scraper* dengan bahasa pemrograman Python pada platform Google Colab. Data dikumpulkan pada periode Januari–Desember 2025, data diseleksi dengan menyaring ulasan berbahasa Indonesia, menghapus data duplikat dan nilai kosong, serta mempertahankan ulasan yang terdiri dari minimal tiga kata. Lalu dilakukan pelabelan data secara manual oleh 3 anotator.

Tahap *preprocessing* dilakukan melalui proses *cleaning*, *case folding*, dan

normalization. Selanjutnya, analisis sentimen dilakukan menggunakan model IndoBERT yang telah di-*fine-tuning* untuk mengklasifikasikan ulasan ke dalam kategori positif, negatif, dan netral. Evaluasi model sentimen dilakukan menggunakan *Confusion Matrix* dengan metrik *Accuracy*, *Precision*, *Recall*, dan *Macro F1-score*. Tahapan penelitian ditunjukkan pada Gambar 1.



Gambar 1. Diagram Alur Tahapan Penelitian Analisis Sentimen Duolingo

2.1. Pengumpulan Data

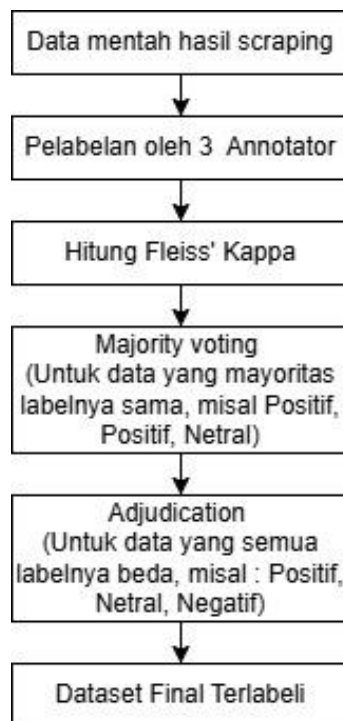
Data penelitian dikumpulkan dengan metode *scraping* ulasan pengguna aplikasi *Duolingo* di *Google Play Store* menggunakan paket *google_play_scraper* dalam Python yang dijalankan di *Google Colab*. Dari total 288.364 ulasan yang berbahasa Indonesia, penelitian ini mengambil 95.574 ulasan yang diterbitkan antara 1 Januari hingga 31 Desember 2025. Tahap selanjutnya adalah *understanding data* dengan menyaring ulasan berbahasa Indonesia (probabilitas $\geq 0,80$), menghapus data yang duplikat dan nilai yang hilang, serta mempertahankan ulasan yang memiliki minimal tiga kata. Sebanyak 1.000 ulasan diambil secara acak untuk dijadikan sampel dalam menguji semua tahap penelitian sebelum diterapkan pada seluruh dataset. Dataset yang diperoleh dari *scraping* mencakup berbagai atribut, tetapi studi ini hanya memanfaatkan kolom *content*, *score*, dan *at* sebagai variabel analisis dan disimpan dalam format CSV dengan nama *dataset_duolingo_clean_labeled.csv*.

... Data ditemukan untuk periode 2025-01-01 s/d 2025-12-31: 95574
=== RINGKASAN DATA UNDERSTANDING ===

	Tahap Pembersihan	Jumlah Data
0	Data Mentah Tahun 2025	95574
1	Filter Bahasa Indonesia (≥ 0.80)	67308
2	Penghapusan Data Duplikat	58399
3	Penghapusan Nilai Kosong	58399
4	Filter Minimal 3 Kata	56543

Gambar 2. Hasil Pengumpulan Data dan Ringkasan Tahap *Data Understanding*

2.2. Pelabelan Data



Gambar 3. Diagram Alur Proses Pelabelan Manual, *Majority Voting* dan *Adjudication*

Pelabelan sentimen dilakukan secara manual oleh tiga anotator terhadap 1.000 ulasan pengguna Duolingo ke dalam kategori positif (2), negatif(0), dan netral(1). Setelah proses pelabelan, terdapat 2 ulasan ambigu lalu dihapus sehingga diperoleh 998 data untuk penelitian.

Tabel 1. Contoh Pelabelan Ulasan oleh Anotator

Ulasan	Anotator 1	Anotator 2	Anotator 3
ihhh ada kursus caturnya 😍. kalo bisa tambahkan juga yang coding 🙌🙌🤖	2	2	1

Reliabilitas pelabelan dievaluasi menggunakan *Fleiss' Kappa* yang digunakan untuk mengukur tingkat kesepakatan antar anotator dalam proses pelabelan data [8]. Tahapan pengujian *Fleiss' Kappa* pada penelitian ini meliputi:

1. Mengumpulkan hasil pelabelan dari seluruh anotator.
2. Mengonversi hasil label ke dalam bentuk numerik.
3. Menghitung proporsi kesepakatan antar anotator.
4. Menghitung nilai *Fleiss' Kappa*.
5. Menginterpretasikan nilai *Kappa* berdasarkan kategori interpretasi kesepakatan.

Nilai *Fleiss' Kappa* berada pada rentang -1 hingga 1, di mana semakin tinggi nilainya menunjukkan tingkat kesepakatan yang semakin baik. Berikut merupakan tabel interpretasi *kappa*:

Tabel 2. Interpretasi Nilai Indeks *Fleiss' Kappa* (Sumber: [9])

Indeks Kappa	Agreement
<0	<i>no agreement</i>
0.0 – 0.20	<i>Slight agreement</i>
0.21 – 0.40	<i>Fair agreement</i>
0.41 – 0.60	<i>Moderate agreement</i>
0.61 – 0.80	<i>Substantial agreement</i>
0.81 – 1.00	<i>Almost perfect agreement</i>

Selanjutnya jika terdapat perbedaan hasil pelabelan maka diselesaikan menggunakan *Majority Voting*, yaitu penentuan label berdasarkan suara terbanyak dari anotator [10]. Berikut contoh hasil *Majority Voting*:

Tabel 3. Contoh Hasil Penentuan Label Menggunakan *Majority Voting*

Data	Anotator1	Anotator2	Anotator3	Majority Voting	label
1	2	2	1	2	Positif
2	0	1	0	0	Negatif
3	1	2	1	1	Netral
4	1	2	0	Diskusi	Diskusi

kemudian apabila tidak diperoleh mayoritas seperti tampak pada Tabel 2 yaitu data ke 4, dilakukan *adjudication* melalui diskusi bersama untuk mencapai kesepakatan label akhir. Seperti pada Tabel 3 berikut:

Tabel 4. Contoh Hasil Penentuan Label Menggunakan *Adjudication*

Data	Anotator1	Anotator2	Anotator3	Majority Voting	Adjudication
4	1	2	0	Diskusi	2

2.3. *Preprocessing Data*

Setelah proses pelabelan data selesai, data ulasan memasuki tahap *preprocessing* untuk meningkatkan kualitas teks sebelum dipakai dalam analisis sentimen dan pemodelan tema. Tahapan *preprocessing* yang dilakukan sebagai berikut:

1. *Cleaning*

Tahap *cleaning* dilakukan dengan menghapus URL, tag HTML, *emoji*, angka, tanda baca, simbol khusus, karakter non-alfabet, serta spasi berlebih yang tidak memberikan informasi penting dalam analisis. Proses ini bertujuan untuk mengurangi *noise* dan menghasilkan data teks yang lebih bersih serta konsisten untuk tahap selanjutnya [11].

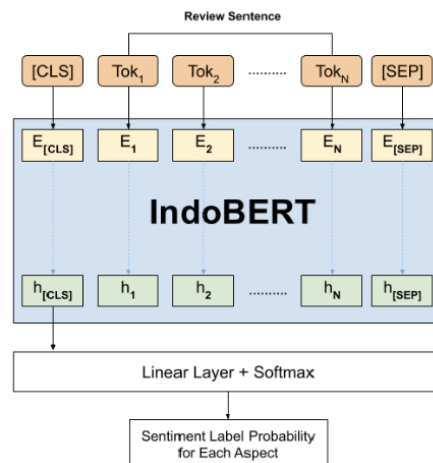
2. *Case Folding*

Tahap *case folding* dilakukan dengan mengubah seluruh huruf pada teks menjadi huruf kecil (*lowercase*) untuk menyeragamkan format penulisan. Proses ini bertujuan menghindari perbedaan representasi kata yang disebabkan oleh penggunaan huruf kapital, sehingga kata dengan makna yang sama dapat dikenali sebagai satu bentuk yang identik oleh model [12].

3. Normalisasi

Tahap normalisasi menggunakan gabungan kamus kata tidak baku (kamus alay) yang berasal dari dataset *Colloquial Indonesian Lexicon* yang dikembangkan oleh Nikmatun Aliyah Salsabila serta kamus normalisasi dari dataset *Multi-label Hate Speech and Abusive Language Detection in Indonesian Twitter* yang dikembangkan oleh Muhammad Okky Ibrohim dan Indra Budi. Penggabungan kedua kamus tersebut dilakukan untuk memperluas cakupan kata tidak baku, singkatan, bahasa gaul, serta variasi penulisan yang umum ditemukan pada ulasan pengguna aplikasi Duolingo [13] [14]. Dan kemudian untuk algoritma *SymSpell* dengan korpus frekuensi kata bahasa Indonesia (*id_50k.txt*) guna untuk memperbaiki kesalahan pengetikan (*typo correction*) [15].

a. Analisis Sentimen IndoBERT



Gambar 4. Arsitektur Klasifikasi Sentimen Menggunakan Model IndoBERT
 (Sumber : [16])

Analisis sentimen pada penelitian ini dilakukan menggunakan model IndoBERT yang telah dilatih pada korpus berbahasa Indonesia sehingga mampu memahami konteks bahasa secara lebih baik dibandingkan metode berbasis konvensional. Model ini digunakan untuk mengklasifikasikan ulasan pengguna Duolingo ke dalam tiga kategori sentimen, yaitu negatif (0), netral (1), dan positif (2).

Sebelum proses klasifikasi, data hasil *preprocessing* ditokenisasi menggunakan *tokenizer* bawaan IndoBERT sehingga menghasilkan *input IDs* dan *attention masks*. Dataset kemudian dibagi menjadi data pelatihan dan data pengujian dengan rasio 80:20 menggunakan metode *stratified split*. Selanjutnya, 10% dari data pelatihan dipisahkan sebagai data validasi yang digunakan untuk mengevaluasi performa model pada setiap epoch selama proses *fine-tuning*. Dengan demikian, dataset terdiri atas data pelatihan, data validasi, dan data pengujian yang masing-masing digunakan untuk pelatihan model, pemantauan proses pelatihan, serta evaluasi akhir model.

Untuk mengatasi ketidakseimbangan distribusi kelas, diterapkan metode *Class Weight* pada data pelatihan dengan memberikan bobot lebih besar pada kelas minoritas sehingga kesalahan klasifikasi pada kelas tersebut memperoleh penalti yang lebih tinggi selama proses pelatihan model. Berikut nilai bobot kelas yang dihasilkan oleh *class weight*:

```
=====
NILAI BOBOT KELAS YANG DIHASILKAN (AFTER)
=====
Negatif : 2.8000
Netral : 6.8205
Positif : 0.4006
```

Gambar 5. Nilai bobot Kelas (*class weight*)

Setelah itu, model IndoBERT di-*fine-tuning* menggunakan data pelatihan berlabel untuk mempelajari pola sentimen pada ulasan pengguna. Parameter pelatihan yang digunakan ditunjukkan pada Tabel 2.

Tabel 5. Parameter pelatihan (*Hyperparameters*) *Fine-Tuning* Model IndoBERT

Parameter	Nilai
Model	IndoBERT Base P1
Max Length	128
Batch Size	16
Epoch	10
Learning Rate	2e-5
Optimizer	AdamW
Scheduler	Linear Warmup
Warmup Ratio	10
Loss Function	CrossEntropyLoss/ Focal Loss
Data Split	70:10:20
Device	GPU T4

Kinerja model dievaluasi menggunakan *Confusion Matrix* dengan metrik *Accuracy*, *Precision*, *Recall*, dan *Macro F1-score*. Selain itu, *Macro F1-score* digunakan sebagai metrik utama karena mampu memberikan evaluasi yang lebih representatif pada dataset dengan distribusi kelas yang tidak seimbang. Model IndoBERT terbaik kemudian digunakan untuk memprediksi sentimen seluruh data ulasan hasil *preprocessing*.

b. Visualisasi dan Analisis Hasil

Tahap visualisasi dan analisis hasil dilakukan untuk menyajikan temuan analisis sentimen secara lebih mudah dipahami. Melalui visualisasi tersebut, pola opini pengguna terhadap aplikasi Duolingo dapat diinterpretasikan secara lebih jelas sehingga mendukung proses pembahasan dan penarikan kesimpulan penelitian.

4. HASIL DAN PEMBAHASAN

a. Hasil Pelabelan dan *Fleiss' Kappa*

Tabel 6. Distribusi Hasil Pelabelan Sentimen Manual

Label Sentimen	Jumlah Data	Persentase
Positif	830	83,2%
Negatif	119	11,9%
Netral	49	4,9%
Total	998	100%

Hasil pelabelan menghasilkan distribusi sentimen yang didominasi oleh kelas positif sebanyak 830 data, diikuti kelas negatif sebanyak 119 data dan kelas netral sebanyak 49 data sehingga total keseluruhan data ulasan sebanyak 998 ulasan. Pengujian reliabilitas anotator menggunakan *Fleiss' Kappa* memperoleh nilai 0,677 yang menunjukkan tingkat kesepakatan kategori (*Substantial agreement*), sehingga hasil pelabelan dinilai konsisten dan layak digunakan sebagai data pelatihan.

b. Majority Voting dan Adjudication

Metode ini menetapkan label berdasarkan kategori yang memperoleh suara terbanyak dari tiga anotator. Dari total 998 ulasan, sebanyak 985 ulasan memperoleh label akhir secara langsung melalui *Majority Voting*. Sementara itu, terdapat 13 ulasan yang tidak memperoleh mayoritas suara karena masing-masing anotator memberikan label yang berbeda. Untuk kasus tersebut dilakukan proses *adjudication* melalui diskusi bersama antar anotator hingga diperoleh kesepakatan terhadap label yang paling sesuai. Proses ini bertujuan meningkatkan konsistensi dan validitas label akhir yang digunakan dalam penelitian.

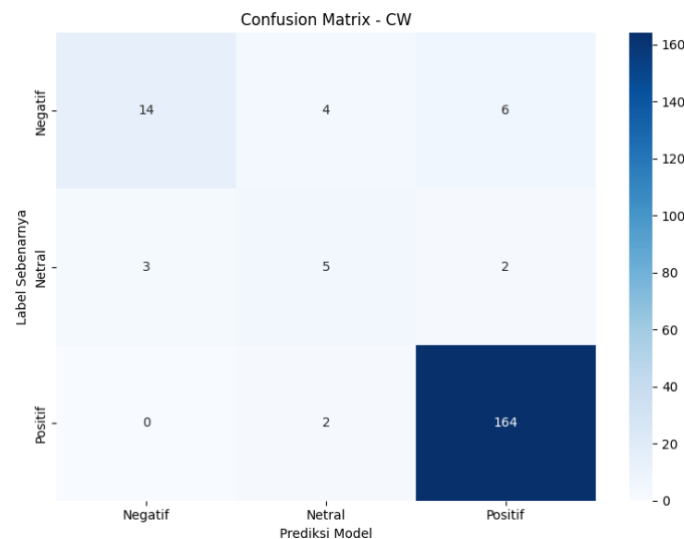
c. Hasil Analisis Sentimen IndoBERT

--- CLASSIFICATION REPORT ---				
	precision	recall	f1-score	support
Negatif	0.82	0.58	0.68	24
Netral	0.45	0.50	0.48	10
Positif	0.95	0.99	0.97	166
accuracy			0.92	200
macro avg	0.74	0.69	0.71	200
weighted avg	0.91	0.92	0.91	200

Gambar 6. *Classification Report* Hasil Pengujian Model IndoBERT
pada Dataset Uji

Model IndoBERT memberikan hasil yang sangat baik dengan *Macro F1-Score* yang mencapai 71% dan *Accuracy* yang berada di angka 92% pada data pengujian (dataset uji yang berjumlah 200 ulasan, mewakili 20% dari total 998 data). Penting untuk membedakan antara dataset uji (200 ulasan) yang digunakan untuk mengukur metrik performa model, dengan evaluasi prediksi pada seluruh dataset (998 ulasan) untuk memetakan polaritas secara menyeluruh. Kinerja terbaik ditunjukkan oleh kelas positif yang mencatat *Macro F1-score* ideal sebesar 97%. Sementara itu, kelas netral menunjukkan kemajuan dengan *Macro F1-score* yang mencapai 48%. Walau karakteristik dari ulasan netral masih memiliki tingkat ambiguitas yang lebih tinggi, penggunaan teknik *Class Weight* telah terbukti berhasil meningkatkan konsistensi prediksi model secara keseluruhan. Ulasan netral cenderung berupa pernyataan pendek yang minim emosi seperti 'oke', 'bagus', atau

sekadar kombinasi emoji tanpa teks deskriptif, sehingga menjadi tantangan tersendiri bagi model untuk mengenali bobot polaritasnya secara spesifik.



Gambar 7. Confusion Matrix

Berdasarkan *Confusion Matrix*, kesalahan klasifikasi terbesar model bergeser ke arah kelas sentimen lain, di mana model berhasil memprediksi secara tepat sebanyak 5 dari total 10 data ulasan netral. Karakteristik ulasan netral yang ambigu dan memiliki kemiripan semantik yang tinggi memang menjadi tantangan tersendiri bagi model, namun penerapan teknik *Class Weight* telah berhasil menjaga stabilitas performa makro model tetap optimal secara keseluruhan.

d. Prediksi Sentimen Seluruh Data

	normalisasi	label	prediksi_indobert	status_prediksi
0	ada kesalahan bahasa inggris	0	Negatif	SESUAI
1	permainan ini memberi saya ilmu dan membuat sa...	2	Positif	SESUAI
2	sudah benar apa yang saya bilang tapi dibuat s...	0	Negatif	SESUAI
3	bagus banget ini aplikasinya bisa buat kita be...	2	Positif	SESUAI
4	game buruk mendukung homo aku kasih bintang	0	Positif	SALAH
5	gila sih ini aplikasi keren banget gue bisa be...	2	Positif	SESUAI
6	aku suka aplikasi ini karena membantu saya unt...	2	Positif	SESUAI
7	pembelajaran yang efektif	2	Positif	SESUAI
8	game edukasi yang sangat menyenangkan	2	Positif	SESUAI
9	mantap bisa bertambah materi bahasa inggris	2	Positif	SESUAI

Gambar 8. Contoh Prediksi Sentimen Seluruh Data

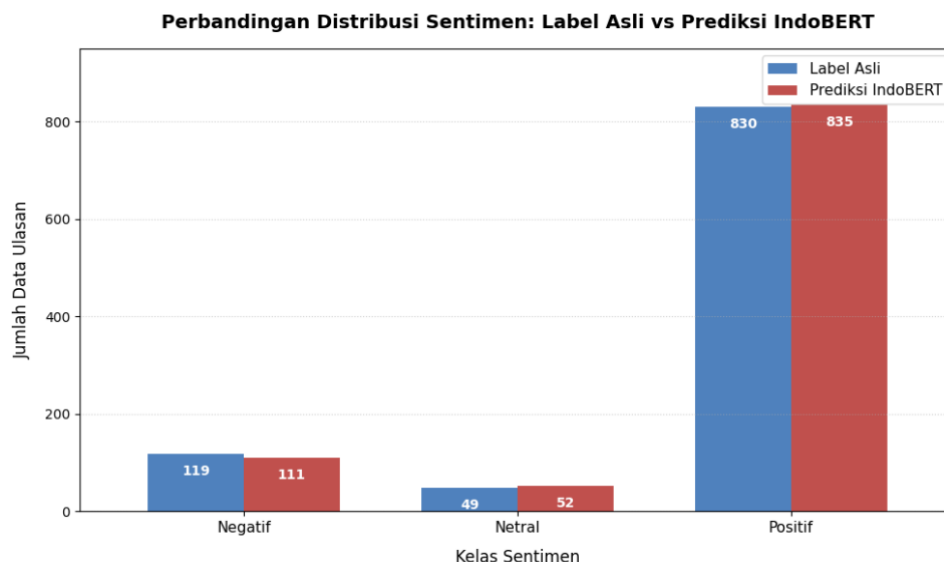
```
=== RINGKASAN HASIL PREDIKSI ===  
Jumlah prediksi per sentimen:  
prediksi_indobert  
Positif      835  
Negatif      111  
Netral       52  
Name: count, dtype: int64  
  
Jumlah akurasi prediksi (SESUAI/SALAH):  
status_prediksi  
SESUAI       977  
SALAH        21  
Name: count, dtype: int64
```

Gambar 9. Ringkasan Hasil Prediksi IndoBERT

Hasil prediksi sentimen dari seluruh total data ulasan yaitu 998 data menggunakan model IndoBERT menunjukkan dominasi yang kuat pada kelas sentimen positif dengan jumlah mencapai 835 ulasan (83,67%), diikuti oleh sentimen negatif sebanyak 111 ulasan (11,12%), dan sentimen netral dengan proporsi terendah yaitu sebanyak 52 ulasan (5,21%). Dari total data pengujian tersebut, model berhasil mengklasifikasikan 977 ulasan secara tepat (*SESUAI*), dan hanya ditemukan 21 ulasan yang mengalami kesalahan klasifikasi (*SALAH*). Hasil evaluasi ini membuktikan secara konkret bahwa model IndoBERT memiliki performa yang sangat baik dan tingkat keandalan yang tinggi dalam mengidentifikasi serta memetakan sentimen ulasan pengguna terhadap aplikasi Duolingo.

e. Visualisasi Performa model IndoBERT

1. Distribusi Hasil Prediksi Sentimen



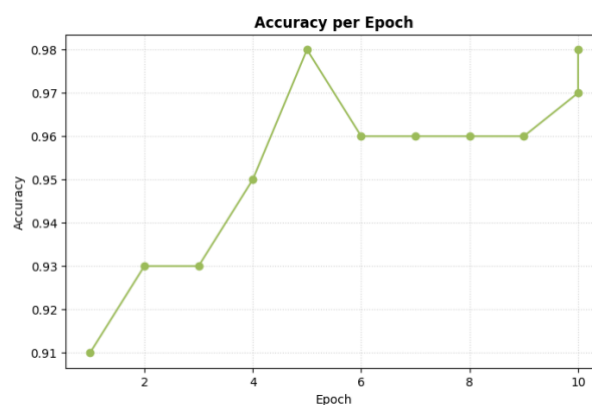
Gambar 10. Perbandingan Distribusi Sentimen Label Asli dan Hasil Prediksi IndoBERT

Gambar 10, menunjukkan bahwa distribusi sentimen hasil prediksi IndoBERT sangat mendekati distribusi label asli. Kelas negatif berjumlah 119 data pada label asli dan 118 pada hasil prediksi, kelas netral 49 dan 47 data, serta kelas positif 830 dan 833 data. Secara keseluruhan, kedekatan jumlah antara label asli dan hasil prediksi menunjukkan bahwa model IndoBERT mampu merepresentasikan distribusi sentimen dataset dengan baik.



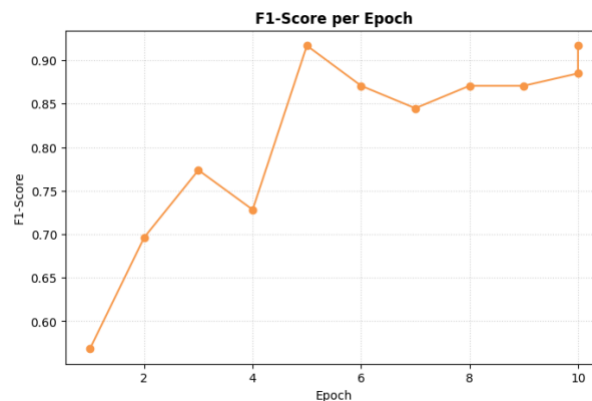
Gambar 11. Kurva *Loss Function*

Pada gambar 11, Kurva *Loss Function* menunjukkan bahwa *Training Loss* terus menurun hingga mendekati nol, menandakan model semakin baik mempelajari data pelatihan. Namun, *Validation Loss* hanya menurun hingga epoch ke-3 lalu meningkat kembali, yang mengindikasikan terjadinya *overfitting*. Oleh karena itu, performa optimal model dicapai pada epoch awal sebelum kemampuan generalisasi terhadap data validasi mulai menurun.



Gambar 12. Kurva Accuracy per Epoch

Kurva *Accuracy* diatas menunjukkan bahwa nilai *Accuracy* mengalami peningkatan dari 91% pada epoch pertama menjadi 98% pada epoch ke-5. Setelah itu, performa model relatif stabil pada kisaran 96%–98% hingga akhir pelatihan. Hal ini menunjukkan bahwa model mampu mempelajari pola data dengan baik dan mempertahankan tingkat akurasi yang tinggi selama proses fine-tuning.



Gambar 13. Kurva *Macro F1-Score* per Epoch

Gambar 13 menunjukkan bahwa nilai *Macro F1-Score* meningkat secara signifikan dari sekitar 0,57 pada epoch pertama menjadi 0,92 pada epoch ke-5. Setelah mengalami sedikit penurunan pada beberapa epoch berikutnya, nilai *Macro F1-Score* kembali stabil dan mencapai 0,92 pada epoch ke-10. Hasil ini menunjukkan bahwa model memiliki kemampuan yang semakin baik dan seimbang dalam mengklasifikasikan seluruh kelas sentimen, termasuk kelas minoritas.

KESIMPULAN

Penelitian ini berhasil menerapkan model IndoBERT untuk melakukan klasifikasi sentimen ulasan pengguna Duolingo berbahasa Indonesia ke dalam tiga kategori, yaitu positif, netral, dan negatif. Hasil pengujian menunjukkan bahwa model memiliki performa yang baik dengan tingkat akurasi sebesar 92% dan *Macro F1-Score* sebesar 71%. Distribusi sentimen didominasi oleh sentimen positif, yang menunjukkan bahwa sebagian besar pengguna memberikan tanggapan yang baik terhadap aplikasi Duolingo. Meskipun demikian, performa pada kelas netral masih relatif rendah dibandingkan kelas lainnya karena banyak ulasan netral yang bersifat singkat dan ambigu. Secara keseluruhan, penelitian ini membuktikan bahwa IndoBERT efektif digunakan untuk menganalisis sentimen ulasan aplikasi berbahasa Indonesia dan mampu memberikan gambaran mengenai persepsi pengguna terhadap Duolingo.

Penelitian selanjutnya disarankan untuk menerapkan teknik penanganan *imbalanced data* yang lebih baik guna meningkatkan performa pada kelas sentimen minoritas. Selain itu, kamus normalisasi dapat diperluas untuk mengatasi bahasa gaul dan kesalahan penulisan pada ulasan pengguna. Perbandingan dengan model *Transformer* lain atau *Large Language Models (LLMs)* juga dapat dilakukan untuk memperoleh hasil klasifikasi yang lebih optimal. Selain itu, analisis dapat dikembangkan dengan menggabungkan metode analisis sentimen dan pemodelan topik sehingga tidak hanya mampu mengidentifikasi polaritas sentimen pengguna, tetapi juga mengungkap isu dan topik utama yang mendasari opini pengguna terhadap aplikasi Duolingo.

DAFTAR PUSTAKA

- [1] L. Judijanto, "Tren Penelitian Mobile-Assisted Language Learning (MALL) di Era Digital," *Sanskara Pendidik. dan Pengajaran*, vol. 3, no. 02, pp. 100–113, 2025, doi: 10.58812/spp.v3i02.558.
- [2] F. Etfita, Elinur, Yulianto, A. Ahmad, Alber, and E. Sastriani, "Jurnal abdidas," *J. Abdidas*, vol. 6, no. 1, pp. 97–106, 2025, doi: <https://doi.org/10.31004/abdidas.v6i1.1113>.
- [3] M. Essafi, L. Belfakir, and M. Moubtassime, "Investigating Mobile-Assisted Language Learning Apps: Babbel, Memrise, and Duolingo as a Case Study," *J. Curric. Teach.*, vol. 13, no. 2, pp. 197–215, 2024, doi: 10.5430/jct.v13n2p197.
- [4] R. M. Putra, M. Rasuki, and I. Astutik, "AI- Driven Mobile Learning in Enhancing Learners ' Speaking Confidence in EFL Context," vol. 17, no. 1, pp. 131–138, 2026, doi: <https://doi.org/10.47200/aoej.v17i1.3338>.
- [5] D. Curry, "Duolingo Revenue and Usage Statistics (2026)," *Business of Apps*. Accessed: Mar. 02, 2026. [Online]. Available: <https://www.businessofapps.com/data/duolingo-statistics/>
- [6] B. Juarto and Yulianto, "Indonesian News Classification Using IndoBert," *Int. J. Intell. Syst. Appl. Eng.*, vol. 11, no. 2, pp. 454–460, 2023, [Online]. Available: <https://ijisae.org/index.php/IJISAE/article/view/2654>
- [7] M. R. Nur, Y. Wibisono, and R. Megasari, "Analisis Sentimen dan Pemodelan Topik pada Post tentang Merek Teknologi di X Menggunakan Fine-tuning IndoBERT dan BERTopic," *J. Komput. Teknol. Inf. Sist. Inf.*, vol. 4, no. 2, pp. 743–750, 2025, doi: 10.62712/juktisi.v4i2.508.
- [8] F. Moons and E. Vandervieren, "Measuring agreement among several raters classifying subjects into one or more (hierarchical) categories: A generalization of Fleiss' kappa," *Behav. Res. Methods*, vol. 57, no. 10, pp. 1--19, 2023, doi: 10.3758/s13428-025-02746-8.
- [9] Ssin122, "Understanding Diagnostic Metrics for Classifiers: Accuracy vs Cohen's Kappa vs ROC vs Precision and Recall vs F1," *Research Strategies and Skills*. Accessed: May 19, 2026. [Online]. Available: <https://my-hdr-blogs.shaveensingh.com/article/understanding-diagnostic-metrics-for-classifiers-accuracy-vs-cohens-kappa/>
- [10] D. Triyana, M. Muharrom Al Haromainy, and H. Maulana, "Implementasi Metode Ensemble Majority Vote Pada Algoritma Naive Bayes Dan Random Forest Untuk Analisis Sentimen Twitter Harga Tiket Pesawat Domestik," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 4, pp. 7885–7894, 2024, doi: 10.36040/jati.v8i4.10475.
- [11] A. F. Aufar, Mochamad Alfian Rosid, A. Eviyanti, and I. R. I. Astutik, "Optimizing

-
- Text Preprocessing for Accurate Sentiment Analysis on E-Wallet Reviews,” *JICTE (Journal Inf. Comput. Technol. Educ.*, vol. 7, no. 2, pp. 42–50, 2023, doi: 10.21070/jicte.v7i2.1650.
- [12] R. Budianoor, S. W. Saputro, F. Abadi, R. A. Nugroho, and A. Farmadi, “Quantifying the Impact of Text Preprocessing on IndoBERT Fine-Tuning for Indonesian Informal Culinary Sentiment Analysis,” *J. Comput. Theor. Appl.*, vol. 3, no. 4, pp. 564–581, 2026, doi: 10.62411/jcta.15980.
- [13] M. O. Ibrohim and I. Budi, “Multi-label Hate Speech and Abusive Language Detection in Indonesian Twitter,” pp. 46–57, 2019, doi: 10.18653/v1/w19-3506.
- [14] N. A. Salsabila, Y. A. Winatmoko, A. A. Septiandri, and A. Jamal, “Colloquial Indonesian Lexicon Nikmatun,” *Int. Conf. Asian Lang. Process.*, pp. 226–229, 2018, doi: 10.1109/IALP.2018.8629151.
- [15] Hermitdave, “FrequencyWords: Repository of word frequency lists (Version terbaru) [Computer software]. GitHub.” Accessed: May 05, 2026. [Online]. Available: <https://github.com/hermitdave/FrequencyWords>
- [16] R. Anadra, H. Wijayanto, and K. Sadik, “Sentiment Analysis of Tokopedia Customer Reviews Using BiLSTM and IndoBERT with Comparative Analysis of Preprocessing and Labeling Methods,” *Int. J. Adv. Data Inf. Syst.*, vol. 6, no. 3, pp. 773–788, 2025, doi: 10.59395/ijadis.v6i3.1458.