LAPORAN FINAL PROJECT
KECERDASAN BUATAN (LANJUT)

ANALISIS SENTIMEN ULASAN APLIKASI DUOLINGO DI GOOGLE PLAY STORE MENGGUNAKAN MODEL BERT

Kelompok 11

Anggota Kelompok:

23.11.5678	Sahila Amalia
23.11. 5674	Gema Satria Tama
23.11.5657	Rangga Firman Ade Syah Putra
23.11.5631	Yusuf Fahrudin

I. LATAR BELAKANG

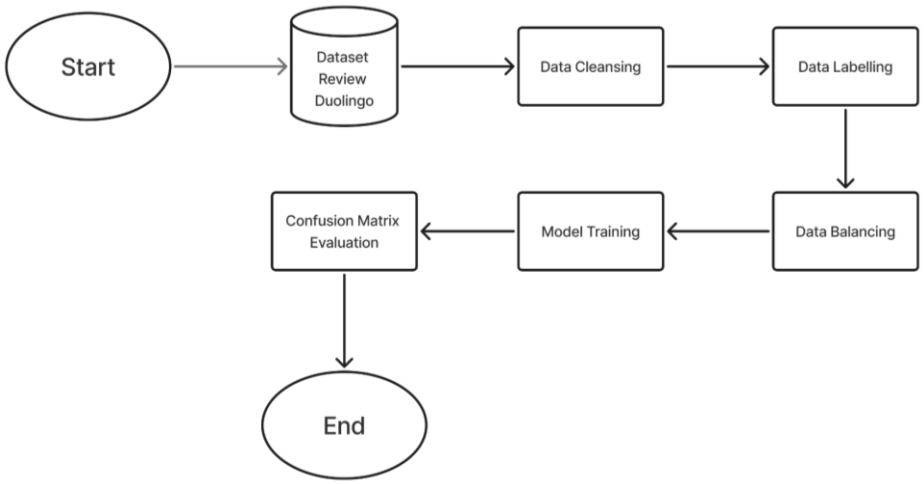
Kemajuan teknologi informasi telah mengubah metode pembelajaran bahasa, beralih dari cara konvensional ke platform digital yang lebih mudah untuk dijangkau. Salah satu inovasi dalam bidang ini adalah Duolingo, aplikasi edukasi yang mengintegrasikan metode pembelajaran interaktif berbasis tantangan untuk memudahkan pengguna belajar bahasa secara praktis melalui perangkat seluler. Aplikasi ini memanfaatkan berbagai elemen permainan seperti sistem level, perolehan poin pengalaman (experience points), serta peringkat pengguna berdasarkan prestasi mereka [1].

Duolingo kemudian semakin berkembang dengan adanya integrasi kecerdasan buatan (AI) untuk menghadirkan pengalaman belajar yang lebih personal, seperti fitur pengenalan suara. Hadirnya teknologi AI bertujuan untuk meningkatkan efektivitas pembelajaran melalui materi yang telah disesuaikan dengan kemampuan masing-masing pengguna [2].

Seiring dengan basis penggunanya yang terus tumbuh secara global, ulasan yang diberikan pengguna melalui layanan toko aplikasi juga semakin meningkat. Ulasan-ulasan tersebut mencakup berbagai sentimen yang mencerminkan pengalaman pengguna selama menggunakan aplikasi, termasuk tanggapan mereka terhadap inovasi AI tersebut. Analisis sentimen terhadap ulasan menjadi penting bagi pengembang aplikasi, karena hasilnya dapat digunakan untuk memahami fitur apa yang disukai pengguna dan apa yang perlu diperbaiki. Hal ini membantu pengembang dalam membuat keputusan yang tepat untuk meningkatkan kualitas aplikasi [3].

Namun, seiring meningkatnya jumlah ulasan, analisis sentimen secara manual menjadi kurang efisien dan memakan waktu yang cukup lama. Oleh karena itu, penelitian ini mengusulkan penggunaan model Bidirectional Encoder Representations from Transformers (BERT) untuk melakukan analisis sentimen. Metode berbasis BERT dipilih karena kemampuannya dalam memahami konteks kata dari dua arah dan menangkap makna yang kompleks dalam ulasan pengguna [4], sehingga tetap dapat merefleksikan persepsi pengguna terhadap inovasi dan fitur aplikasi Duolingo secara akurat.

II. METODE



A. Unduh Dataset Review Duolingo

Tahap pertama dalam final project ini adalah mengumpulkan dataset, yang dimulai dari pengunduhan dataset ulasan aplikasi Duolingo dari KaggleHub. Dataset yang digunakan dalam final project ini adalah duolingo app user review play store dataset 2025, yang merupakan kumpulan ulasan dari pengguna aplikasi Duolingo yang diambil dari platform Google Play Store. Setiap ulasan dalam dataset ini mencakup content (teks ulasan) beserta score (rating yang diberikan oleh pengguna), dan informasi tambahan lainnya seperti review_id, thumbs_up_count, dan created_at.

B. Data Cleansing

Setelah memuat dataset Duolingo dari Kagglehub, Tahap selanjutnya adalah data cleansing, dalam project ini terdapat kolom yang tidak diperlukan yang harus didrop yaitu review_id,thumbs_up_count, dan created_at. Selanjutnya tipe data kolom score dikonversi ke numerik. Integritas data kemudian dijaga dengan menghapus baris-baris yang mengandung nilai kosong (NaN) pada kolom content maupun score. Untuk memastikan konsistensi dan kualitas teks ulasan, review yang terdapat spasi berlebih pada kolom content dibersihkan. Terakhir, nama kolom

content diganti menjadi text untuk menyesuaikan dengan standar penamaan yang umum dalam “Natural Language Processing”.

C. Labelling Data

Langkah berikutnya adalah pemberian label pada data. Proses ini dilakukan untuk mengategorikan setiap ulasan ke dalam kelas sentimen yang relevan. Setiap ulasan akan dikategorikan menjadi tiga kelas sentimen utama yaitu: ‘Positive’ , ‘Neutral’ , atau ‘Negative’, yang ditentukan dari nilai score ulasan. Secara spesifik, sentimen ‘Positive’ diberi label dengan skor 4 dan 5, sentimen ‘Netral’ diberi label skor 3, dan sentimen ‘Negatif’ diberi label skor 1 dan 2. Pendekatan pemetaan skor ulasan 1–5 menjadi label sentimen positif, netral, dan negatif juga banyak digunakan pada penelitian analisis sentimen berbasis ulasan daring [5]. Kategori sentimen ini selanjutnya diubah menjadi label numerik yang sangat krusial untuk melatih model analisis sentimen yang akan diterapkan.

D. Data Balancing

Tahap ini bertujuan untuk mencapai distribusi kelas sentimen yang seimbang dalam dataset. Dataset ulasan awal menunjukkan ketidakseimbangan yang signifikan, dengan jumlah ulasan ‘Positive’ jauh lebih dominan dibandingkan ‘Neutral’ dan ‘Negative’. Untuk mengatasi ketidakseimbangan ini, digunakan teknik penyeimbangan data dari ‘sklearn.utils.resample’. Secara spesifik, penulis mengurangi jumlah ulasan 'Positive' dan 'Negative' yang terlalu (undersampling), sambil memperbanyak ulasan 'Neutral' yang jumlahnya sedikit (oversampling). Setelah proses ini, semua kategori sentimen digabungkan kembali dan diacak, memastikan bahwa dataset siap untuk melatih model dengan proporsi label yang seimbang.

E. Model Training

Proses pelatihan model dimulai dengan memisahkan dataset yang sudah seimbang menjadi dua bagian: data train dan data validation. Teks ulasan kemudian diubah menjadi bentuk angka melalui proses tokenisasi yang menggunakan

‘AutoTokenizer’ dari model bert-base-multilingual-cased. Setelah itu, model untuk klasifikasi model (trainer.train()) dilakukan menggunakan ‘Hugging Face Trainer’ dengan konfigurasi yang ditentukan, seperti learning rate dan ukuran batch. Selama pelatihan, kinerja model dievaluasi menggunakan metrik seperti akurasi dan F1-score pada data validasi. Setelah latihan selesai, model siap digunakan untuk melakukan prediksi sentimen pada ulasan baru.

F. Confusion Matrix Evaluation

Untuk mengukur kinerja model secara mendalam, penulis menggunakan Confusion Matrix. Matriks ini dapat memvisualisasikan seberapa baik model mengklasifikasikan setiap sentimen (‘Negative’, ‘Neutral’, ‘Positive’), menunjukkan jumlah ulasan yang diklasifikasikan dengan benar dan salah untuk setiap kategori, sehingga penulis dapat mengidentifikasi area di mana model mungkin mengalami kesulitan [6].

G. Arsitektur Model

Model yang digunakan dalam project ini adalah arsitektur BERT multilingual cased, sebuah model transformer yang telah dilatih sebelumnya dan kemudian diadaptasi secara khusus untuk tugas klasifikasi sentimen ulasan aplikasi menjadi kategori Positive, Neutral, atau Negative.

III. DATASET

Data yang digunakan dalam penelitian ini adalah Duolingo App User Review Play Store 2025 yang disusun oleh Bilal Akhtar yang bersumber dari platform Kaggle. Dataset ini berisi ulasan pengguna aplikasi Duolingo di Google Play Store yang dikumpulkan untuk kebutuhan analisis sentimen dan evaluasi ulasan pengguna. Secara keseluruhan, dataset ini mencakup 504.862 data ulasan.

Tabel berikut menyajikan contoh informasi yang terkandung dalam dataset tersebut:

review_id	content	score	thumbs_up_count	created_at
d88949e3-2726-4c7a-91f5-906d16472e0a	give us the heart system	1.0	9	2025-10-24T19:55:10.729Z
3947dc6b-6cf6-45f9-8cfc-64115ebc53cc	The energy system has really lowered my engagement.	2.0	9	2025-10-23T17:31:50.766Z
64ef0d63-1900-415a-8fa5-34ac57d86d19	Good app to learn a language but there are just so many ads	3.0	8	2025-07-10T21:18:34.820Z
f70e0bf9-da33-405e-b94d-512445d75f0b	Would love to be able to opt out of leagues!	4.0	893	2025-05-30T05:49:28.693Z
77d85078-c32f-40f1-b73e-8b3c3944c9e6	This is so good. you should try it.	5.0	9	2025-07-09T12:40:08.711Z

Struktur dataset ini terdiri dari beberapa kolom utama, yaitu:

1. **review_id** merupakan identitas unik untuk setiap ulasan yang masuk.
2. **content** berisi teks ulasan yang ditulis oleh pengguna.
3. **score** merupakan skor yang diberikan pengguna dalam skala 1 sampai 5.
4. **thumbs_up_count** merupakan jumlah pengguna lain yang menyukai ulasan tersebut.

5. **created_at** menunjukkan keterangan waktu kapan ulasan tersebut dipublikasikan oleh pengguna.

IV. HASIL PENGUJIAN

Pengujian dilakukan dengan model BERT yang dilatih selama 3 epoch. Hasil pengujian menunjukkan apakah model mampu mengklasifikasikan sentimen ulasan Duolingo ke dalam kategori Negatif, Netral, dan Positif.

A. Metrik Evaluasi

Aspek	Precision	Recall	F1-Score	Support
Negative	0.89	0.81	0.85	15310
Neutral	0.69	0.60	0.64	10000
Positive	0.81	0.92	0.86	20000
Accuracy			0.81	45310
Macro Avg	0.80	0.78	0.78	45310
Weighted Avg	0.81	0.81	0.81	45310

Secara garis besar, model BERT yang dikembangkan berhasil mencapai tingkat Akurasi (Accuracy) global sebesar 81%. Angka ini menunjukkan bahwa dari total 45.310 data uji, model mampu memprediksi label yang benar. Namun, nilai akurasi tidak dapat menggambarkan keseluruhan performa model dikarenakan adanya ketimpangan jumlah data (class imbalance).

Secara garis besar, model BERT yang dikembangkan berhasil mencapai tingkat Akurasi (Accuracy) global sebesar 81%. Angka ini menunjukkan bahwa dari total 45.310 data uji, model mampu memprediksi label yang benar. Namun, nilai akurasi tidak dapat menggambarkan keseluruhan performa model dikarenakan adanya ketimpangan jumlah data (class imbalance).

Pada kategori Sentimen Positif (Label 2), model menunjukkan performa paling dominan dengan nilai Recall tertinggi sebesar 0.92. Tingginya nilai Recall ini menunjukkan bahwa model mampu menemukan dan menangkap sebagian besar kasus, yaitu 92% dari seluruh ulasan positif yang ada di dalam dataset. Hal ini dipengaruhi oleh kolom Support, di mana kategori positif memiliki jumlah data terbanyak (20.000 sampel). Banyaknya contoh data latih ini membuat model dapat belajar dengan baik. Namun, nilai Precision sebesar 0.81 menunjukkan bahwa terkadang model tetap keliru memprediksi sebagai positif (false positive) untuk sentimen Netral dan Positif.

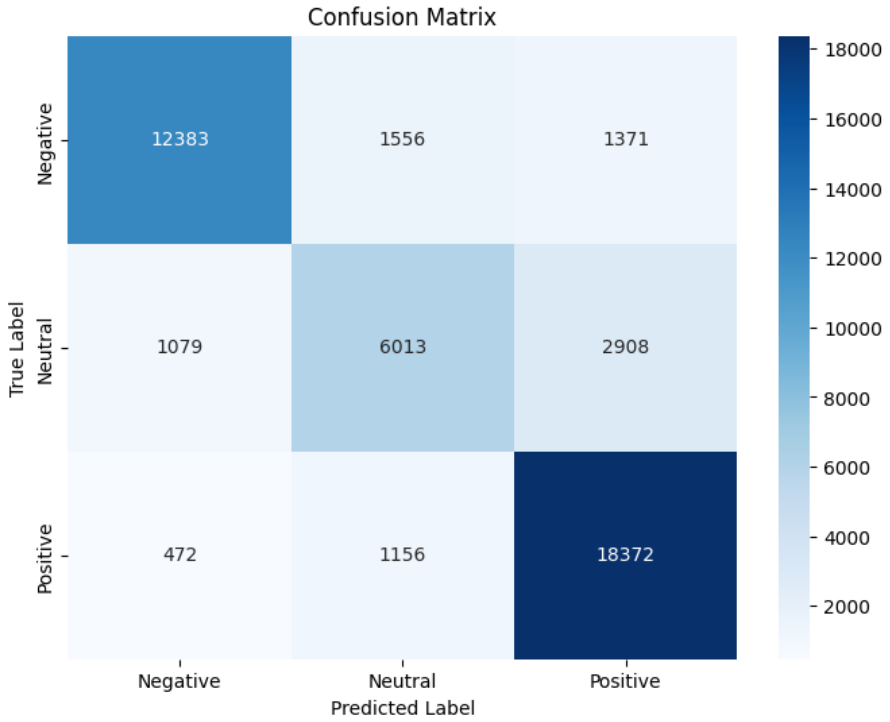
Sebaliknya, pada kategori Sentimen Negatif (Label 0), model menunjukkan hasil yang berbeda, yaitu unggul dalam aspek ketepatan prediksi dengan Precision sebesar 0.89. Angka ini merupakan nilai tertinggi dibandingkan kategori lainnya, artinya ketika model memprediksi sebuah ulasan sebagai "Negatif", tingkat kepercayaannya sangat tinggi (89% benar).. Namun, nilai Recall sebesar 0.81 menunjukkan masih ada sekitar 19% ulasan negatif asli yang lolos dari deteksi model (false negative) dan malah diprediksi menjadi sentimen lain.

Kelemahan paling signifikan terlihat pada kategori Sentimen Netral (Label 1), yang memiliki performa terendah di semua metrik (Precision 0.69, Recall 0.60, F1-Score 0.64). Hal ini kemungkinan terjadi karena nilai Support yang paling sedikit (10.000 sampel) dengan rendahnya Recall (0.60). Karena kurangnya referensi data, model gagal mengenali 40% dari ulasan netral yang sebenarnya. Rendahnya skor ini juga disebabkan oleh karakteristik ulasan netral yang seringkali ambigu atau memuat campuran opini (mixed feelings), sehingga model kesulitan membedakannya dari sentimen positif atau negatif. Nilai F1-Score yang rendah (0.64) mengonfirmasi bahwa model belum mencapai keseimbangan yang baik dalam menangani kategori ini.

Kesimpulannya, nilai Weighted Avg yang identik dengan akurasi (0.81) menunjukkan bahwa performa model secara umum didongkrak oleh kemampuannya mengenali kelas mayoritas (Positif dan Negatif). Namun, nilai Macro Avg (rata-rata tanpa pembobotan jumlah data) untuk Recall yang hanya 0.78

menjadi tanda bahwa model masih memiliki bias terhadap kelas mayoritas dan memerlukan strategi penanganan imbalanced data atau augmentation untuk meningkatkan kemampuannya dalam memahami ciri ulasan Netral..

B. Confusion Matrix



Visualisasi Confusion Matrix memberikan gambaran mengenai dimana tepatnya model melakukan kesalahan prediksi. Ditandai dengan kolom diagonal yang bernilai tinggi menunjukkan model semakin baik mengenali data.

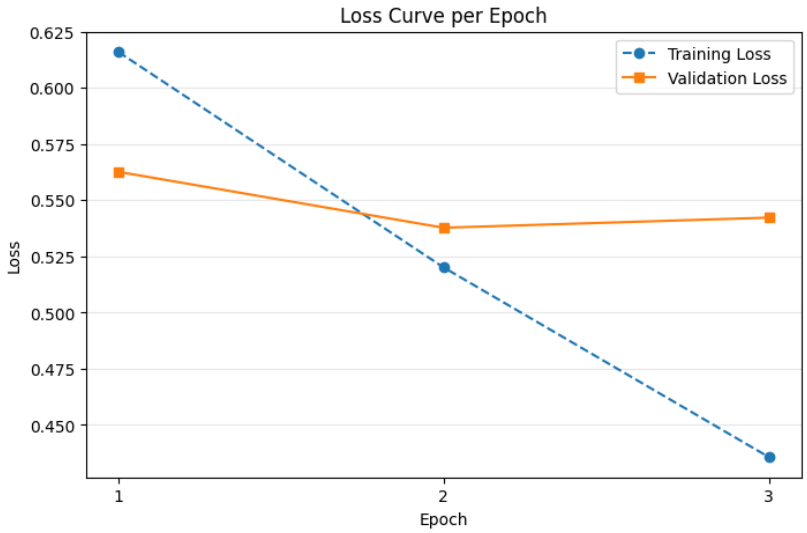
Jumlah data yang tinggi terdapat pada diagonal kelas Positif (18372) dan Negatif (12383) secara langsung menunjukkan tingginya nilai F1-Score pada kedua kelas tersebut.

Namun, Confusion Matrix juga menunjukkan kelemahan prediksi pada kelas Netral (6013). Sejalan dengan rendahnya nilai Recall (0.60) pada tabel metrik. Model gagal menangkap ulasan netral dan cenderung memasukkannya ke dalam kategori ekstrem (positif atau negatif). Hal ini menjelaskan mengapa Support (jumlah data) netral yang sedikit (10.000) yang berdampak besar.

V. ANALISA HASIL

A. Performa Per-Epoch

Selama proses pelatihan (3 epoch), nilai loss terus menurun sementara akurasi pada dataset validasi meningkat walaupun pada epoch kedua sedikit menurun.



B. Kekuatan Model

Model BERT menunjukkan keunggulan dalam memahami konteks ulasan sentimen negatif dan positif. Akurasi tertinggi ditemukan pada sentimen Positif ditunjukkan dengan nilai Recall tertinggi (0.92). Hal ini berkorelasi lurus dengan jumlah data latih (Support) terbanyak pada kategori ini (20.000 sampel). Mekanisme Self-Attention pada BERT berhasil menangkap pola kepuasan pengguna dengan sangat baik, serta mampu membedakan ulasan negatif dengan presisi tinggi (0.89).

C. Kelemahan Model

Performa terendah terdapat pada kategori Netral dengan nilai Recall dan F1-Score terendah yaitu 0.77 dan 0.79. Hal ini disebabkan oleh Faktor Ambiguitas, ulasan netral seringkali bersifat implisit atau campuran (misal: "Fiturnya lengkap, tapi mahal"). Model cenderung salah memprediksi ulasan ini sebagai positif atau

negatif (misclassification). Adapun, Faktor Ketidakseimbangan Data (Imbalance): Dengan jumlah data paling sedikit (10.000 sampel) dibandingkan kelas lain, model memiliki referensi belajar yang lebih terbatas untuk kategori Netral, menyebabkannya bias ke arah kelas dengan data lebih banyak (Positif/Negatif).

VI. KESIMPULAN

Berdasarkan penelitian ini, dapat disimpulkan bahwa analisis sentimen terhadap ulasan aplikasi Duolingo berhasil dilakukan menggunakan model BERT. Proses diawali dengan pengumpulan dataset, dilanjutkan dengan pembersihan data, pelabelan sentimen berdasarkan rating, serta penyeimbangan data untuk mengatasi ketimpangan kelas. Model mampu mengklasifikasikan ulasan ke dalam tiga kategori sentimen (Positif, Netral, dan Negatif) dengan performa yang diukur menggunakan *confusion matrix*, akurasi, dan *F1-score*.

Hasil pengujian menunjukkan akurasi model sebesar 81% pada data uji. Berdasarkan metrik evaluasi, model memiliki kemampuan yang sangat baik dalam mendeteksi sentimen negatif (*F1-score* 0,85) dan positif (*F1-score* 0,86). Namun, performa model pada sentimen netral masih kurang optimal (*F1-score* 0,64). Hal ini disebabkan oleh terbatasnya jumlah data netral pada dataset asli, sehingga pola sentimen netral lebih sulit dipelajari oleh model meskipun telah dilakukan upaya penyeimbangan data.

Secara keseluruhan, teknik *transfer learning* dengan arsitektur BERT terbukti efektif untuk tugas klasifikasi teks pada ulasan aplikasi. Model ini berpotensi menjadi alat analisis umpan balik otomatis yang dapat membantu pengembang dalam memahami pola kepuasan pengguna secara cepat dan akurat. Untuk pengembangan selanjutnya, kinerja pada sentimen netral dapat ditingkatkan dengan penambahan variasi data atau penggunaan teknik augmentasi data yang lebih kompleks.

VII. REFERENSI

[1] M. Shortt, S. Tilak, I. Kuznetcova, B. Martens, and B. Akinkuolie, "Gamification in mobile-assisted language learning: a systematic review of Duolingo literature from public release of 2012 to early 2020," *Comput Assist Lang Learn*, vol. 36, no. 3, pp. 517–554, 2023, doi: 10.1080/09588221.2021.1933540.

[2] T. Gajic, J. Nikolić, N. Maenza, and A. Gagić, "Artificial Intelligence in Mobile Language Learning: Duolingo and the Rise of a New Educational Era," *Sinteza 2025 - International Scientific Conference on Information Technology, Computer Science, and Data Science*, pp. 405–410, Jun. 2025, doi: 10.15308/SINTEZA-2025-405-410.

[3] F. Palomba *et al.*, "Crowdsourcing user reviews to support the evolution of mobile apps," *Journal of Systems and Software*, vol. 137, pp. 143–162, Mar. 2018, doi: 10.1016/J.JSS.2017.11.043.

[4] S. Alaparthi and M. Mishra, "Bidirectional Encoder Representations from Transformers (BERT): A sentiment analysis odyssey," Jul. 2020, Accessed: Jan. 04, 2026. [Online]. Available: https://arxiv.org/abs/2007.01127v1

[5] E. S. Alamoudi and N. S. Alghamdi, "Sentiment classification and aspect-based sentiment analysis on yelp reviews using deep learning and word embeddings," *J Decis Syst*, vol. 30, no. 2–3, pp. 259–281, Jul. 2021, doi: 10.1080/12460125.2020.1864106.

[6] A. Luque, A. Carrasco, A. Martín, and A. de las Heras, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix," *Pattern Recognit*, vol. 91, pp. 216–231, Jul. 2019, doi: 10.1016/J.PATCOG.2019.02.023.

VIII. KONTRIBUSI & DISTRIBUSI ANGGOTA KELOMPOK

Sahila Amalia	● Menyusun dokumentasi hasil pengujian ● Menganalisa hasil pengujian dan mengidentifikasi kelebihan serta kekurangan model
Gema Satria Tama	● Menyusun latar belakang penelitian ● Menangani pembersihan data, pelatihan model, serta eksperimen *hyperparameter* agar performa model lebih optimal.
Rangga Firman Ade Syah Putra	● Menyusun kesimpulan akhir berdasarkan hasil pengujian dan analisa model
Yusuf Fahrudin	● Menyusun metodologi penelitian ● Menyusun deskripsi dataset