JITET (Jurnal Informatika dan Teknik Elektro Terapan)
Vol. 14 No. 3, pISSN: 2303-0577 eISSN: 2830-7062 http://dx.doi.org/10.23960/jitet.v14i3.10147

ANALISIS SENTIMEN TERHADAP APLIKASI DUOLINGO MENGGUNAKAN INDOBERTWEET

Muhammad Farhan Hadrawi[1*], Rizqa Raaiqa Bintana[2], Hasanatul Iftitah[3]

[1,2,3]Universitas Jambi, Jl. Lintas Jambi-Ma. Bulian, Muaro Jambi, Jambi, Indonesia, Telp. 0741-583377, Fax. 0741-583111

Keywords:
Analisis sentimen, Duolingo, IndoBERTweet, Ulasan Google Play Store.

Corespondent Email:
farhanhadrawi@gmail.com

Abstrak. Ulasan pengguna pada Google Play Store memuat informasi penting mengenai pengalaman, kepuasan, dan keluhan terhadap aplikasi pembelajaran bahasa seperti Duolingo. Namun, jumlah ulasan yang besar dan tidak terstruktur menyebabkan analisis secara manual menjadi kurang efisien. Penelitian ini bertujuan untuk mengklasifikasikan sentimen ulasan pengguna aplikasi Duolingo berbahasa Indonesia menggunakan model IndoBERTweet serta mengevaluasi kinerja model tersebut. Data awal terdiri atas 1.000 ulasan yang dikumpulkan melalui teknik *web scraping* dari Google Play Store. Setelah proses pelabelan dan validasi data, diperoleh 998 ulasan yang diklasifikasikan ke dalam sentimen positif, negatif, dan netral. Tahapan penelitian meliputi pengumpulan dan penyaringan data, pelabelan oleh tiga anotator, *preprocessing* teks, pembagian data dengan rasio 70:10:20, penyeimbangan data latih menggunakan *RandomOverSampler*, *fine-tuning* IndoBERTweet, dan evaluasi model. Hasil penelitian menunjukkan bahwa model memperoleh *accuracy* sebesar 0,9050, *F1-score macro* sebesar 0,7301, dan *F1-score weighted* sebesar 0,9155. Model menghasilkan performa terbaik pada sentimen positif, sedangkan sentimen netral menjadi kelas yang paling sulit dikenali. Hasil tersebut menunjukkan bahwa IndoBERTweet mampu mengklasifikasikan sentimen ulasan pengguna Duolingo dengan performa yang baik, meskipun masih dipengaruhi oleh ketidakseimbangan distribusi kelas.

Abstract. *User reviews on Google Play Store contain valuable information regarding users' experiences, satisfaction, and complaints about language-learning applications such as Duolingo. However, the large volume and unstructured nature of these reviews make manual analysis inefficient. This study aims to classify the sentiment of Indonesian-language Duolingo user reviews using the IndoBERTweet model and evaluate its classification performance. The initial dataset consisted of 1,000 reviews collected from Google Play Store through a web scraping technique. After the labeling and data validation processes, 998 reviews were classified into positive, negative, and neutral sentiments. The research stages included data collection and filtering, labeling by three annotators, text preprocessing, dataset splitting using a 70:10:20 ratio, training-data balancing using RandomOverSampler, IndoBERTweet fine-tuning, and model evaluation. The results show that the model achieved an accuracy of 0.9050, a macro F1-score of 0.7301, and a weighted F1-score of 0.9155. The model achieved its best performance on positive sentiment, while neutral sentiment was the most difficult class to identify. These results indicate that IndoBERTweet can effectively classify Indonesian Duolingo user-review sentiment, although its performance is still affected by class imbalance.*

1. PENDAHULUAN

Perkembangan teknologi digital telah memberikan pengaruh besar terhadap berbagai aspek kehidupan, termasuk bidang pendidikan. Salah satu bentuk pemanfaatan teknologi digital dalam pembelajaran bahasa adalah *Mobile-Assisted Language Learning* (MALL), yaitu pendekatan pembelajaran yang memanfaatkan perangkat *mobile* untuk mendukung proses belajar bahasa secara fleksibel tanpa batasan ruang dan waktu [1]. Penerapan MALL memungkinkan peserta didik mengakses materi pembelajaran, berlatih secara mandiri, serta menggunakan aplikasi pembelajaran bahasa melalui perangkat seluler secara lebih praktis [2]. Selain itu, penggunaan aplikasi berbasis MALL juga dapat mendukung pengembangan kosakata dan kemandirian belajar melalui interaksi digital yang mudah diakses oleh pengguna [3].

Duolingo merupakan salah satu aplikasi pembelajaran bahasa berbasis *mobile* yang banyak digunakan secara global, termasuk di Indonesia. Aplikasi ini menyediakan pembelajaran bahasa melalui latihan singkat, umpan balik langsung, dan sistem pembelajaran yang interaktif [4]. Salah satu karakteristik utama Duolingo adalah penerapan konsep *gamification*, seperti poin, level, *streak*, dan latihan berbasis permainan yang dapat meningkatkan motivasi dan keterlibatan pengguna [5]. Beberapa penelitian menunjukkan bahwa Duolingo dapat membantu meningkatkan penguasaan kosakata, kepercayaan diri, serta motivasi belajar bahasa asing [4], [5]. Meskipun demikian, penggunaan Duolingo masih memiliki beberapa kendala, seperti iklan yang mengganggu, sistem nyawa yang terbatas, serta kurangnya variasi konteks percakapan yang kompleks [6].

Opini, keluhan, dan saran pengguna terhadap aplikasi Duolingo banyak terekam dalam ulasan yang tersedia pada Google Play Store. Fitur rating dan ulasan pada Google Play Store memungkinkan pengguna menyampaikan pendapat mengenai aplikasi yang telah digunakan. Oleh karena itu, data ulasan dapat dimanfaatkan sebagai sumber informasi untuk memahami kepuasan maupun ketidakpuasan pengguna [7]. Penelitian sebelumnya juga menunjukkan bahwa ulasan aplikasi pada Google Play Store dapat digunakan sebagai sumber data dalam analisis sentimen untuk membantu pengembang memahami kualitas layanan dan aspek aplikasi yang perlu ditingkatkan [7], [8]. Namun, jumlah ulasan yang besar dan bentuk data yang tidak terstruktur menyebabkan proses analisis secara manual menjadi kurang efisien serta berpotensi menimbulkan subjektivitas. Oleh karena itu, diperlukan pendekatan komputasional berbasis *Natural Language Processing* (NLP) untuk mengolah dan memahami informasi yang terdapat dalam data teks ulasan pengguna [9].

Salah satu penerapan NLP yang banyak digunakan untuk menganalisis opini pengguna adalah analisis sentimen. Analisis sentimen bertujuan untuk mengidentifikasi dan mengklasifikasikan opini atau emosi yang terkandung dalam teks ke dalam kategori tertentu, seperti positif, negatif, dan netral [10]. Melalui analisis sentimen, kecenderungan persepsi pengguna terhadap aplikasi dapat diketahui berdasarkan pengalaman dan penilaian yang disampaikan dalam ulasan. Informasi tersebut dapat membantu pengembang dalam mengevaluasi kualitas aplikasi dan menentukan aspek yang perlu dipertahankan maupun ditingkatkan.

Seiring berkembangnya metode *deep learning*, pendekatan analisis sentimen mengalami perkembangan dari metode pembelajaran mesin tradisional menuju model berbasis *transformer*. Salah satu model *transformer* yang banyak digunakan adalah BERT (*Bidirectional Encoder Representations from Transformers*). BERT mampu memahami konteks kata secara bidireksional dengan mempertimbangkan kata-kata yang berada sebelum dan sesudah suatu kata dalam kalimat [11]. Kemampuan tersebut membuat BERT lebih efektif dalam menangkap makna teks yang kompleks, termasuk negasi, ambiguitas, dan variasi struktur kalimat.

Dalam konteks bahasa Indonesia, model BERT telah dikembangkan dan diterapkan untuk memproses teks berbahasa Indonesia. Kusuma dan Mogi menerapkan IndoBERT pada analisis sentimen ulasan destinasi wisata Bali dan menunjukkan bahwa model berbasis BERT mampu menghasilkan kinerja yang baik dalam mengklasifikasikan ulasan berbahasa Indonesia [12]. Namun, ulasan pengguna aplikasi pada Google Play Store sering kali menggunakan bahasa informal, singkatan, kata tidak baku, *slang*, kesalahan pengetikan, serta

campuran bahasa. Karakteristik tersebut belum tentu dapat diproses secara optimal oleh model yang lebih banyak dilatih menggunakan teks formal. Untuk menangani karakteristik tersebut, IndoBERTweet dapat digunakan karena model ini dikembangkan melalui proses *continual pretraining* pada korpus Twitter berbahasa Indonesia sehingga lebih adaptif terhadap ragam bahasa informal [13].

Penelitian sebelumnya menunjukkan bahwa IndoBERTweet mampu memberikan kinerja yang baik dalam analisis teks informal berbahasa Indonesia. Setiawan et al. menggunakan IndoBERTweet untuk menganalisis ulasan pengguna aplikasi TikTok di Google Play Store dan memperoleh hasil yang lebih baik dibandingkan model LSTM [14]. Penelitian lainnya yang dilakukan oleh Khairani et al. menunjukkan bahwa IndoBERTweet memperoleh akurasi sebesar 92,54% dalam mendeteksi emosi pada komentar Instagram dan mengungguli IndoBERT yang memperoleh akurasi sebesar 88,81% [15]. Penelitian tersebut juga menunjukkan bahwa IndoBERTweet dapat menghasilkan kinerja yang lebih baik tanpa penerapan *stemming* dan *stopword removal*, karena kedua tahapan tersebut berpotensi menghilangkan informasi yang berperan dalam membentuk konteks emosional kalimat [15]. Temuan-temuan tersebut menunjukkan bahwa IndoBERTweet sesuai digunakan untuk menganalisis teks ulasan aplikasi yang cenderung singkat, tidak baku, dan bersifat informal.

Meskipun IndoBERTweet telah diterapkan pada berbagai data teks informal, seperti ulasan aplikasi TikTok dan komentar Instagram, penerapannya dalam klasifikasi sentimen ulasan pengguna aplikasi pembelajaran bahasa, khususnya Duolingo berbahasa Indonesia, masih terbatas. Selain itu, distribusi sentimen pada ulasan aplikasi umumnya tidak seimbang karena jumlah ulasan positif sering kali lebih dominan dibandingkan ulasan negatif dan netral. Ketidakseimbangan tersebut dapat menyebabkan model lebih mudah mengenali kelas mayoritas, tetapi memiliki kinerja yang lebih rendah pada kelas minoritas. Oleh karena itu, penelitian ini menerapkan metode *RandomOverSampler* hanya pada data pelatihan untuk mengurangi pengaruh ketidakseimbangan kelas tanpa mengubah distribusi asli data validasi dan data pengujian.

Berbeda dengan penelitian sebelumnya, penelitian ini berfokus pada klasifikasi sentimen ulasan pengguna aplikasi Duolingo berbahasa Indonesia dengan menggunakan label yang ditentukan secara manual oleh tiga anotator. Penelitian ini juga mengevaluasi beberapa teknik penanganan ketidakseimbangan kelas dan menggunakan *macro F1-score* sebagai dasar pemilihan metode penyeimbangan serta konfigurasi model terbaik. Penggunaan *macro F1-score* dilakukan agar kemampuan model dalam mengenali setiap kelas sentimen dapat dinilai secara lebih seimbang dan tidak hanya dipengaruhi oleh kelas mayoritas.

Berdasarkan permasalahan tersebut, penelitian ini bertujuan untuk mengklasifikasikan sentimen ulasan pengguna aplikasi Duolingo berbahasa Indonesia menggunakan model IndoBERTweet serta mengevaluasi kinerja model tersebut. Ulasan diklasifikasikan ke dalam tiga kategori sentimen, yaitu positif, negatif, dan netral. Kinerja model dievaluasi menggunakan metrik *accuracy*, *precision*, *recall*, *F1-score*, serta *confusion matrix*. Penelitian ini diharapkan dapat memberikan gambaran mengenai kecenderungan sentimen pada sampel ulasan pengguna Duolingo sekaligus menunjukkan kemampuan IndoBERTweet dalam mengklasifikasikan ulasan berbahasa Indonesia yang bersifat informal dan memiliki distribusi kelas yang tidak seimbang.

2. TINJAUAN PUSTAKA

2.1. *Natural Language Processing*

Natural Language Processing (NLP) merupakan bidang dalam kecerdasan buatan yang berfokus pada pemrosesan, pemahaman, dan analisis bahasa alami manusia oleh komputer [9]. NLP digunakan untuk mengolah data teks tidak terstruktur agar dapat dipahami dan dianalisis secara komputasional. Dalam penelitian ini, NLP diterapkan untuk memproses ulasan pengguna aplikasi Duolingo yang berbentuk teks bebas sehingga dapat diklasifikasikan ke dalam kategori sentimen positif, negatif, dan netral.

Text mining merupakan proses penggalian informasi dari kumpulan data teks yang tidak terstruktur untuk memperoleh pola atau

pengetahuan yang bermanfaat. Dalam konteks ulasan aplikasi, *text mining* dapat digunakan untuk mengubah komentar pengguna menjadi informasi yang lebih terstruktur mengenai kecenderungan sentimen pengguna. Tahapan umum dalam *text mining* meliputi pengumpulan data, *preprocessing*, representasi teks, klasifikasi, dan interpretasi hasil.

2.2. Analisis Sentimen

Analisis sentimen merupakan salah satu penerapan NLP dan *text mining* yang bertujuan untuk mengidentifikasi opini, sikap, atau emosi yang terkandung dalam teks. Kategori sentimen yang umum digunakan meliputi sentimen positif, negatif, dan netral [10]. Sentimen positif menunjukkan adanya pujian, kepuasan, atau pengalaman yang menyenangkan. Sentimen negatif menunjukkan keluhan, kekecewaan, atau pengalaman yang tidak menyenangkan, sedangkan sentimen netral menunjukkan pernyataan yang tidak memiliki kecenderungan emosi yang kuat atau mengandung penilaian campuran.

Dalam konteks ulasan aplikasi, analisis sentimen dapat membantu mengetahui kecenderungan persepsi pengguna terhadap kualitas fitur dan layanan aplikasi. Penelitian sebelumnya yang dilakukan oleh Nurian dan Sari [7] menunjukkan bahwa ulasan pengguna pada Google Play Store dapat dimanfaatkan sebagai sumber data untuk menganalisis kepuasan dan ketidakpuasan pengguna. Hasil analisis tersebut dapat digunakan sebagai bahan masukan bagi pengembang dalam mengevaluasi dan meningkatkan kualitas aplikasi.

2.3. *Preprocessing Teks*

Preprocessing teks merupakan tahapan awal dalam pengolahan data teks yang bertujuan untuk membersihkan dan menyeragamkan data sebelum digunakan sebagai masukan model. Tahapan *preprocessing* dapat meliputi *cleaning*, *case folding*, tokenisasi, normalisasi, *stopword removal*, dan *stemming* [10].

Tahap *cleaning* dilakukan untuk menghapus elemen yang tidak relevan, seperti URL, tag HTML, emoji, simbol, tanda baca, karakter khusus, dan spasi berlebih. *Case folding* digunakan untuk mengubah seluruh huruf menjadi huruf kecil agar bentuk penulisan kata menjadi seragam. Sementara itu, normalisasi dilakukan untuk mengubah kata tidak baku, singkatan, dan bahasa gaul menjadi bentuk yang lebih baku serta memperbaiki kesalahan pengetikan.

Pada penelitian ini, tahapan *preprocessing* yang diterapkan meliputi *cleaning*, *case folding*, dan normalisasi. Tokenisasi tidak dilakukan secara manual karena proses tersebut dilakukan secara otomatis oleh *tokenizer* IndoBERTweet ketika teks diubah menjadi masukan numerik. Tahapan *stopword removal* dan *stemming* juga tidak diterapkan karena dapat menghilangkan kata-kata yang memiliki peran penting dalam membentuk konteks emosional suatu ulasan. Hal ini sejalan dengan penelitian Khairani et al. [14] yang menunjukkan bahwa IndoBERTweet dapat menghasilkan kinerja lebih baik pada teks informal tanpa penerapan *stemming* dan *stopword removal*.

2.4. *IndoBERTweet*

BERT (*Bidirectional Encoder Representations from Transformers*) merupakan model bahasa berbasis *transformer* yang dirancang untuk memahami konteks kata secara dua arah. Model ini mempertimbangkan kata-kata yang terdapat sebelum dan sesudah suatu kata sehingga dapat menghasilkan representasi teks yang lebih kontekstual [11]. Kemampuan tersebut membuat BERT dapat menangkap makna yang kompleks, termasuk negasi, ambiguitas, dan hubungan antarkata dalam sebuah kalimat.

IndoBERTweet merupakan model bahasa berbasis BERT yang dikembangkan untuk menangani teks informal berbahasa Indonesia. Model ini dikembangkan melalui proses *continual pretraining* menggunakan korpus Twitter berbahasa Indonesia sehingga lebih adaptif terhadap kata tidak baku, singkatan, bahasa gaul, dan ekspresi yang umum digunakan dalam percakapan daring [12].

Karakteristik bahasa pada ulasan Google Play Store memiliki kemiripan dengan teks media sosial karena cenderung singkat, informal, dan mengandung berbagai variasi penulisan. Oleh karena itu, IndoBERTweet dinilai sesuai untuk digunakan dalam mengklasifikasikan sentimen ulasan pengguna aplikasi Duolingo.

Penelitian Setiawan et al. [13] menunjukkan bahwa IndoBERTweet mampu memberikan

kinerja yang baik dalam analisis sentimen ulasan aplikasi TikTok berbahasa Indonesia. Selain itu, penelitian Khairani et al. [14] menunjukkan bahwa IndoBERTweet menghasilkan performa yang lebih baik dibandingkan IndoBERT dalam mendeteksi emosi pada komentar Instagram. Temuan tersebut mendukung penggunaan IndoBERTweet untuk analisis teks informal berbahasa Indonesia.

2.5. Evaluasi Model

Evaluasi model dilakukan untuk mengetahui kemampuan IndoBERTweet dalam mengklasifikasikan ulasan pengguna ke dalam kelas sentimen yang sesuai. Evaluasi dilakukan dengan membandingkan label hasil prediksi model dengan label sebenarnya pada data pengujian. Metrik evaluasi yang digunakan meliputi *accuracy*, *precision*, *recall*, dan *F1-score* [10].

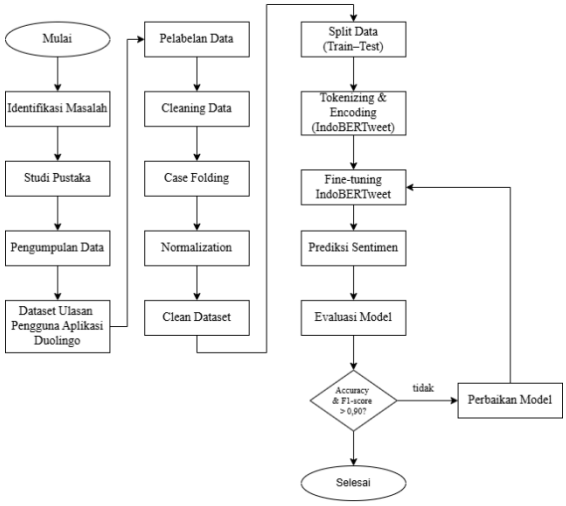
Accuracy menunjukkan proporsi seluruh data yang berhasil diprediksi dengan benar oleh model. *Precision* menunjukkan tingkat ketepatan model ketika memprediksi suatu kelas sentimen. *Recall* menunjukkan kemampuan model dalam menemukan seluruh data yang sebenarnya termasuk ke dalam suatu kelas. Sementara itu, *F1-score* merupakan rata-rata harmonik antara *precision* dan *recall* sehingga dapat digunakan untuk melihat keseimbangan kedua metrik tersebut.

Penelitian ini menggunakan nilai *macro average* dan *weighted average*. *Macro average* menghitung nilai rata-rata setiap kelas dengan bobot yang sama sehingga dapat menggambarkan kemampuan model pada seluruh kelas secara seimbang. Sementara itu, *weighted average* menghitung nilai rata-rata berdasarkan jumlah data pada setiap kelas sehingga hasilnya lebih dipengaruhi oleh kelas yang memiliki jumlah data lebih besar.

Selain metrik tersebut, evaluasi juga dilakukan menggunakan *confusion matrix*. *Confusion matrix* digunakan untuk menunjukkan jumlah prediksi benar dan salah pada setiap kelas sentimen. Melalui matriks tersebut, dapat diketahui kelas yang mampu dikenali dengan baik oleh model serta kelas yang paling sering mengalami kesalahan prediksi.

3. METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif berbasis *text mining* untuk menganalisis sentimen ulasan pengguna aplikasi Duolingo pada Google Play Store. Tahapan penelitian meliputi pengumpulan dan penyaringan data, pelabelan sentimen, *preprocessing* teks, pembagian dan penyeimbangan data, klasifikasi sentimen menggunakan IndoBERTweet, serta evaluasi model klasifikasi.



Gambar 1. Tahapan Penelitian

3.1. Pengumpulan Data

Data yang digunakan dalam penelitian ini berupa ulasan pengguna aplikasi Duolingo berbahasa Indonesia yang diperoleh dari Google Play Store. Pengumpulan data dilakukan melalui teknik *web scraping* menggunakan bahasa pemrograman Python pada lingkungan Google Colab dengan bantuan pustaka *google-play-scraper*. Parameter pengambilan data meliputi aplikasi Duolingo, bahasa Indonesia, wilayah Indonesia, dan pengurutan berdasarkan ulasan terbaru.

Data hasil *scraping* memuat atribut teks ulasan, rating pengguna, dan tanggal ulasan. Data tersebut kemudian disaring berdasarkan rentang waktu tahun 2025, bahasa Indonesia, duplikasi, ulasan kosong, dan jumlah kata. Ulasan yang memiliki kurang dari tiga kata dihapus karena dianggap belum cukup merepresentasikan opini pengguna secara bermakna.

Setelah proses penyaringan, diperoleh 56.543 ulasan yang memenuhi kriteria. Selanjutnya, dilakukan *random sampling* untuk memilih 1.000 ulasan sebagai dataset

penelitian. Dataset tersebut kemudian disimpan dalam format CSV untuk digunakan pada tahap pelabelan dan pengolahan data berikutnya.

3.2. Pelabelan Data

Tahap pelabelan dilakukan untuk menentukan kategori sentimen pada setiap ulasan. Label sentimen yang digunakan terdiri dari tiga kelas, yaitu positif, negatif, dan netral. Sentimen positif diberikan pada ulasan yang berisi pujian, kepuasan, atau pengalaman menyenangkan terhadap aplikasi. Sentimen negatif diberikan pada ulasan yang memuat keluhan, kekecewaan, masalah teknis, atau pengalaman tidak menyenangkan. Sementara itu, sentimen netral diberikan pada ulasan yang bersifat deskriptif, campuran, atau tidak menunjukkan kecenderungan emosi yang kuat.

Pelabelan dilakukan secara manual oleh tiga annotator untuk mengurangi subjektivitas. Penentuan label akhir dilakukan menggunakan metode majority voting, yaitu label yang dipilih oleh minimal dua annotator ditetapkan sebagai label akhir. Apabila terdapat perbedaan label pada seluruh annotator, dilakukan diskusi untuk menentukan label yang paling sesuai dengan konteks ulasan.

3.3. *Preprocessing Teks*

Tahap *preprocessing* dilakukan untuk membersihkan dan menyeragamkan teks ulasan sebelum digunakan sebagai masukan model. Tahapan yang diterapkan meliputi *cleaning*, *case folding*, dan normalisasi.

Tahap *cleaning* dilakukan dengan menghapus URL, tag HTML, emoji, simbol, tanda baca, karakter khusus, dan spasi berlebih. Selanjutnya, *case folding* dilakukan dengan mengubah seluruh karakter menjadi huruf kecil agar bentuk penulisan kata menjadi seragam.

Tahap normalisasi dilakukan untuk mengubah kata tidak baku, singkatan, dan bahasa gaul menjadi bentuk yang lebih baku. Normalisasi dilakukan menggunakan kamus kata tidak baku serta pustaka SymSpell untuk memperbaiki kesalahan pengetikan berdasarkan tingkat kemiripan kata.

Pada penelitian ini, *stopword removal* dan *stemming* tidak diterapkan karena berpotensi menghilangkan kata yang berperan dalam membentuk konteks emosional ulasan. Tokenisasi juga tidak dilakukan secara terpisah karena proses tersebut dilakukan secara otomatis oleh *tokenizer* IndoBERTweet pada tahap pembentukan masukan model.

3.4. Klasifikasi Sentimen

Klasifikasi sentimen dilakukan menggunakan model IndoBERTweet. Dataset yang telah melalui tahap pelabelan dan *preprocessing* dibagi menjadi data latih, data validasi, dan data uji dengan perbandingan 70:10:20. Pembagian data dilakukan menggunakan teknik *stratified split* agar proporsi setiap kelas sentimen tetap seimbang pada masing-masing subset data.

Sebelum masuk ke model, teks ulasan diproses menggunakan *tokenizer* IndoBERTweet. Proses ini mengubah teks menjadi token, kemudian mengonversinya ke dalam bentuk numerik berupa *input ids* dan *attention mask*. Model IndoBERTweet selanjutnya dilatih melalui proses *fine-tuning* untuk mengklasifikasikan ulasan ke dalam tiga kelas sentimen, yaitu positif, negatif, dan netral. Untuk memperoleh konfigurasi terbaik, dilakukan *hyperparameter tuning* terhadap beberapa parameter, seperti *learning rate*, *batch size*, dan jumlah *epoch*.

3.5. Evaluasi Model

Evaluasi model dilakukan dengan membandingkan label hasil prediksi model terhadap label sebenarnya pada data uji. Metrik evaluasi yang digunakan meliputi *accuracy*, *precision*, *recall*, dan *F1-score*. Selain itu, digunakan *confusion matrix* untuk melihat sebaran prediksi benar dan salah pada setiap kelas sentimen. Hasil evaluasi ini digunakan untuk mengetahui kemampuan model IndoBERTweet dalam melakukan klasifikasi sentimen terhadap ulasan pengguna aplikasi Duolingo.

4. HASIL DAN PEMBAHASAN

4.1. Hasil Pengumpulan Data

Pengumpulan data dilakukan dengan teknik *web scraping* terhadap ulasan pengguna aplikasi Duolingo pada Google Play Store menggunakan bahasa pemrograman Python dan *library google-play-scraper*. Data yang diperoleh berupa ulasan pengguna yang memuat teks ulasan, rating, tanggal ulasan, dan atribut pendukung lainnya. Berdasarkan hasil *scraping*, diperoleh sebanyak 288.364 ulasan mentah.

Data hasil *scraping* kemudian disaring melalui beberapa tahap agar sesuai dengan kebutuhan penelitian. Penyaringan dilakukan berdasarkan rentang waktu tahun 2025, bahasa Indonesia, penghapusan data duplikat, penghapusan data kosong, serta penghapusan ulasan yang memiliki kurang dari tiga kata. Setelah seluruh proses penyaringan dilakukan, diperoleh 56.543 ulasan yang memenuhi kriteria. Selanjutnya, dilakukan *random sampling* sebanyak 1.000 ulasan untuk digunakan sebagai *dataset* penelitian.

Tabel 1. Hasil penyaringan data ulasan

Tahap Pembersihan	Jumlah Data
data mentah hasil scraping	288.364
filter tahun 2025	95.574
filter bahasa Indonesia	67.308
penghapusan data duplikat	58.399
penghapusan data kosong	58.399
filter minimal 3 kata	56.543
random sampling	1.000

Berdasarkan Tabel 1, jumlah data mengalami pengurangan pada setiap tahap penyaringan. Proses penyaringan ini dilakukan untuk memastikan bahwa data yang digunakan hanya berupa ulasan berbahasa Indonesia, relevan dengan periode penelitian, tidak berulang, dan memiliki isi ulasan yang cukup untuk dianalisis. Dengan demikian, sebanyak 1.000 ulasan hasil *random sampling* digunakan pada tahap pelabelan sentimen dan analisis selanjutnya.

4.2. Hasil Pelabelan

Setelah diperoleh 1.000 ulasan hasil *random sampling*, tahap selanjutnya adalah pelabelan data sentimen. Pelabelan dilakukan secara manual oleh tiga *annotator* untuk mengurangi subjektivitas dalam penentuan label. Setiap ulasan diklasifikasikan ke dalam tiga kategori sentimen, yaitu positif, negatif, dan netral.

Sentimen positif diberikan pada ulasan yang berisi pujian, kepuasan, atau pengalaman menyenangkan terhadap aplikasi Duolingo. Sentimen negatif diberikan pada ulasan yang memuat keluhan, kekecewaan, permasalahan teknis, atau pengalaman tidak menyenangkan. Sementara itu, sentimen netral diberikan pada ulasan yang bersifat deskriptif, campuran, atau tidak menunjukkan kecenderungan emosi yang kuat.

Pada tahap pelabelan, terdapat 2 ulasan yang dieliminasi karena tidak dapat diinterpretasikan secara semantik dan tidak memuat opini yang bermakna terhadap aplikasi Duolingo. Dengan demikian, jumlah data akhir yang digunakan setelah pelabelan adalah 998 ulasan.

Tabel 2. Distribusi hasil pelabelan sentimen

Label	Jumlah Data	Persentase
Positif	831	83,27%
Negatif	120	12,02%
Netral	47	4,71%
Total	998	100%

Berdasarkan Tabel 2, dapat diketahui bahwa data ulasan didominasi oleh sentimen positif sebanyak 831 ulasan atau 83,27%. Sentimen negatif berjumlah 120 ulasan atau 12,02%, sedangkan sentimen netral berjumlah 47 ulasan atau 4,71%. Hasil ini menunjukkan bahwa sebagian besar pengguna memberikan tanggapan positif terhadap aplikasi Duolingo. Namun, perbedaan jumlah data yang cukup besar antar kelas juga menunjukkan adanya ketidakseimbangan kelas atau *class imbalance*, sehingga perlu diperhatikan pada tahap pelatihan model.

4.3. Hasil *Preprocessing*

Tahap *preprocessing* dilakukan untuk membersihkan dan menyeragamkan teks ulasan sebelum digunakan pada proses klasifikasi sentimen. Data ulasan yang diperoleh dari Google Play Store masih berbentuk teks mentah, sehingga perlu diproses agar lebih terstruktur dan sesuai sebagai masukan model IndoBERTweet. Pada penelitian ini, tahapan *preprocessing* yang digunakan meliputi *cleaning*, *case folding*, dan normalisasi.

Tahap *cleaning* dilakukan untuk menghapus elemen yang tidak relevan, seperti URL, tag HTML, emoji, simbol, tanda baca, karakter khusus, dan spasi berlebih. Selanjutnya, *case folding* dilakukan dengan mengubah seluruh huruf menjadi huruf kecil agar bentuk kata menjadi seragam. Tahap normalisasi dilakukan untuk mengubah kata tidak baku, singkatan, dan *slang* menjadi bentuk yang lebih baku, serta memperbaiki kesalahan pengetikan

menggunakan kamus alay dan *library* SymSpell.

Pada penelitian ini, proses *stopword removal* dan *stemming* tidak diterapkan. Hal ini dilakukan karena kedua proses tersebut berpotensi menghilangkan konteks emosional yang penting dalam analisis sentimen teks informal. Selain itu, tokenisasi tidak dilakukan secara manual karena proses tersebut dilakukan secara otomatis oleh *tokenizer* IndoBERTweet pada tahap pembentukan masukan model.

Tabel 3. Hasil Preprocessing

Preprocessing	Content
data mentah	bagus bangett!!aku belajar bahasa Jepang di sini dan aku jadi hafal bahasa Jepang!! 💗 💗
cleaning	bagus bangett aku belajar bahasa Jepang di sini dan aku jadi hafal bahasa Jepang
case folding	bagus bangett aku belajar bahasa jepang di sini dan aku jadi hafal bahasa jepang
normalisasi	bagus banget aku belajar bahasa jepang di sini dan aku jadi hafal bahasa jepang

4.4. Hasil Klasifikasi Sentimen

Klasifikasi sentimen dilakukan menggunakan model IndoBERTweet terhadap data ulasan yang telah melalui tahap pelabelan dan *preprocessing*. Label sentimen terdiri dari tiga kelas, yaitu negatif, netral, dan positif. Sebelum digunakan pada proses pelatihan model, label sentimen dikonversi ke dalam bentuk numerik, yaitu negatif menjadi 0, netral menjadi 1, dan positif menjadi 2.

Dataset dibagi menjadi tiga bagian, yaitu data *training*, data *validation*, dan data *testing* dengan perbandingan 70:10:20. Pembagian data dilakukan menggunakan teknik *stratified split* agar proporsi setiap kelas sentimen tetap terjaga pada masing-masing subset data. Berdasarkan hasil pembagian, diperoleh 698 data *training*, 100 data *validation*, dan 200 data *testing*.

Berdasarkan hasil pelabelan, distribusi kelas sentimen menunjukkan adanya ketidakseimbangan data atau *class imbalance*, karena kelas positif memiliki jumlah data yang jauh lebih banyak dibandingkan kelas negatif dan netral. Untuk mengatasi hal tersebut, penelitian ini menggunakan metode *RandomOverSampler* (ROS) pada data *training*. Metode ini bekerja dengan menambahkan sampel pada kelas minoritas secara acak hingga distribusi data menjadi lebih seimbang.

Penerapan ROS hanya dilakukan pada data *training*, sedangkan data *validation* dan data *testing* tetap menggunakan distribusi asli. Hal ini dilakukan agar proses evaluasi model tetap objektif dan tidak mengalami kebocoran data atau *data leakage*. Setelah proses *oversampling*, jumlah data *training* meningkat dari 698 menjadi 1.743 data dengan distribusi kelas yang lebih seimbang.

Tahap berikutnya adalah tokenisasi menggunakan *tokenizer* IndoBERTweet. Teks ulasan diubah menjadi representasi numerik berupa *input ids* dan *attention mask* agar dapat diproses oleh model. Panjang maksimum token ditetapkan sebesar 128 token. Setelah itu, dilakukan proses *fine-tuning* model IndoBERTweet untuk menyesuaikan model dengan tugas klasifikasi sentimen ulasan pengguna Duolingo.

Untuk memperoleh konfigurasi pelatihan terbaik, dilakukan *hyperparameter tuning* menggunakan pendekatan *grid search* secara manual. Parameter yang diuji meliputi *learning rate* 2e-5 dan 3e-5, *batch size* 16 dan 32, serta jumlah *epoch* 2, 3, dan 5. Berdasarkan hasil pengujian, konfigurasi terbaik diperoleh pada *learning rate* 2e-5, *batch size* 32, dan *epoch* 3, dengan nilai *F1-score macro* terbaik pada data *validation* sebesar 0,7214. Konfigurasi ini kemudian digunakan pada proses pelatihan model final.

4.5. Evaluasi Model Klasifikasi

Evaluasi model dilakukan untuk mengetahui kemampuan IndoBERTweet dalam mengklasifikasikan sentimen ulasan pengguna aplikasi Duolingo. Evaluasi dilakukan menggunakan data *testing* sebanyak 200 ulasan yang tidak digunakan pada proses pelatihan maupun validasi. Metrik evaluasi yang digunakan meliputi *accuracy*, *precision*, *recall*, dan *F1-score*. Hasil evaluasi model ditampilkan pada Gambar 2.

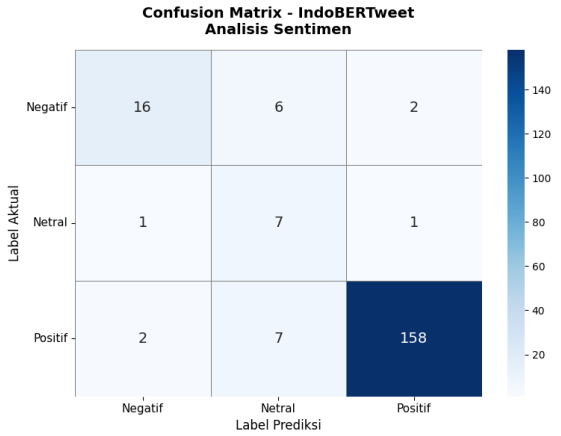
```
===== Hasil Evaluasi pada Data Test =====
Accuracy           : 0.9050
Precision (macro)  : 0.7245
Precision (weighted): 0.9362
Recall    (macro)  : 0.7969
Recall    (weighted): 0.9050
F1-Score  (macro)  : 0.7301
F1-Score  (weighted): 0.9155

Classification Report:
              precision    recall  f1-score   support

     Negatif       0.84      0.67      0.74        24
      Netral       0.35      0.78      0.48         9
     Positif       0.98      0.95      0.96       167

    accuracy                           0.91       200
   macro avg       0.72      0.80      0.73       200
weighted avg       0.94      0.91      0.92       200
```

Gambar 2. Hasil evaluasi model klasifikasi

Berdasarkan Gambar 2, model IndoBERTweet memperoleh nilai *accuracy* sebesar 0,9050, *precision macro* sebesar 0,7245, *recall macro* sebesar 0,7969, dan *F1-score macro* sebesar 0,7301. Selain itu, model memperoleh *precision weighted* sebesar 0,9362, *recall weighted* sebesar 0,9050, dan *F1-score weighted* sebesar 0,9155. Hasil tersebut menunjukkan bahwa model memiliki performa yang baik dalam mengklasifikasikan sentimen ulasan pengguna Duolingo secara keseluruhan.

Nilai *F1-score weighted* yang lebih tinggi dibandingkan *F1-score macro* menunjukkan bahwa performa model masih dipengaruhi oleh distribusi data yang tidak seimbang. Kelas positif memiliki jumlah data paling banyak, sehingga model lebih mudah mengenali pola pada kelas tersebut dibandingkan kelas negatif dan netral yang jumlah datanya lebih sedikit. Berdasarkan hasil evaluasi per kelas, sentimen positif memperoleh hasil terbaik dengan *F1-score* sebesar 0,96, sedangkan sentimen negatif memperoleh *F1-score* sebesar 0,74. Sementara itu, sentimen netral memperoleh *F1-score* sebesar 0,48, yang menunjukkan bahwa kelas netral masih menjadi kelas yang paling sulit dikenali oleh model.

Selain metrik evaluasi, performa model juga dianalisis menggunakan *confusion matrix* sebagaimana ditampilkan pada Gambar 3.



Gambar 3. Confusion matrix

Berdasarkan Gambar 3, model berhasil memprediksi dengan benar sebanyak 158 data positif, 16 data negatif, dan 7 data netral. Hasil ini menunjukkan bahwa model paling baik dalam mengenali kelas positif karena jumlah data positif pada data uji lebih dominan dibandingkan kelas lainnya. Kesalahan prediksi paling banyak terjadi pada ulasan positif dan negatif yang diprediksi sebagai netral. Hal ini menunjukkan bahwa kelas netral cukup sulit dibedakan karena sering memuat pernyataan yang bersifat ambigu, campuran, atau tidak menunjukkan kecenderungan emosi yang kuat.

Secara keseluruhan, hasil evaluasi menunjukkan bahwa model IndoBERTweet mampu mengklasifikasikan sentimen ulasan pengguna Duolingo dengan baik. Nilai *accuracy* sebesar 0,9050 dan *F1-score weighted* sebesar 0,9155 menunjukkan bahwa model memiliki performa yang tinggi secara umum. Meskipun demikian, model masih memiliki keterbatasan dalam mengenali kelas minoritas, khususnya kelas netral.

5. KESIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan, dapat disimpulkan bahwa:

1. Penelitian ini berhasil mengklasifikasikan sentimen ulasan pengguna aplikasi Duolingo berbahasa Indonesia menggunakan model IndoBERTweet. Data akhir yang digunakan berjumlah 998 ulasan, terdiri atas 831 ulasan positif, 120 ulasan negatif, dan 47 ulasan netral. Distribusi tersebut menunjukkan bahwa ulasan pengguna aplikasi Duolingo didominasi oleh sentimen positif dengan persentase sebesar 83,27%.

2. Model IndoBERTweet memperoleh nilai *accuracy* sebesar 0,9050, *F1-score macro* sebesar 0,7301, dan *F1-score weighted* sebesar 0,9155. Hasil tersebut menunjukkan bahwa IndoBERTweet mampu mengklasifikasikan sentimen ulasan pengguna Duolingo dengan baik secara keseluruhan. Sentimen positif memperoleh performa terbaik dengan *F1-score* sebesar 0,96, sedangkan sentimen netral memperoleh *F1-score* sebesar 0,48 dan menjadi kelas yang paling sulit dikenali oleh model.
3. Perbedaan antara nilai *F1-score macro* dan *F1-score weighted* menunjukkan bahwa kinerja model masih dipengaruhi oleh ketidakseimbangan distribusi kelas. Meskipun metode *RandomOverSampler* telah diterapkan pada data pelatihan, jumlah data asli pada kelas negatif dan netral masih jauh lebih sedikit dibandingkan kelas positif. Penelitian selanjutnya disarankan menggunakan dataset yang lebih besar dan lebih seimbang serta membandingkan IndoBERTweet dengan model *transformer* lain untuk memperoleh performa klasifikasi yang lebih merata pada setiap kelas sentimen.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada dosen pembimbing yang telah memberikan arahan, masukan, dan bimbingan selama proses penyusunan penelitian ini. Penulis juga menyampaikan terima kasih kepada Program Studi Sistem Informasi, Fakultas Sains dan Teknologi, Universitas Jambi, serta seluruh pihak yang telah memberikan dukungan dan bantuan sehingga penelitian ini dapat diselesaikan dengan baik.

DAFTAR PUSTAKA

[1] W. T. Rahmawati, Y. M. Harahap, R. Ginting, P. R. Anggraini, and A. A. Maulana, "Pengenalan Mobile Assisted Language Learning (MALL) untuk Berlatih Pronunciation," Journal of Training and Community Service Adpertisi (JTCSA), vol. 3, no. 2, pp. 50–56, 2023, doi: 10.62728/jtcsa.v3i2.454.

[2] N. W. N. Suryati, K. M. C. Dewi, P. Rusanti, and I. G. A. G. Wismadewi, "Penggunaan Mobile Assisted Language Learning (MALL) pada Pembelajaran Bahasa Inggris Keperawatan," EDUKATIF: Jurnal Ilmu Pendidikan, vol. 6, no. 4, pp. 4008–4017, 2024, doi: 10.31004/edukatif.v6i4.7255.

[3] V. Handayani and I. Damayanti, "Mobile-Assisted Language Learning for Vocabulary Development: Insights from Indonesian Pre-Service English Teachers Using HelloTalk," Jurnal Tahuri, vol. 19, no. 2, pp. 99–114, 2022, doi: 10.30598/tahurivol19issue2page99-114.

[4] H. Hakimantieq, D. Suherdi, and W. Gunawan, "Duolingo as A Mobile-Assisted Language Learning: A New Supplementary of Learning Basic English Reading for EFL Students," EDUKATIF: Jurnal Ilmu Pendidikan, vol. 4, no. 6, pp. 7548–7558, 2022, doi: 10.31004/edukatif.v4i6.4104.

[5] S. D. Puspita, S. L. Amanda, V. N. Swara, and Z. A. Shaffa, "Pemanfaatan Aplikasi Duolingo dalam Meningkatkan Penguasaan Kosakata Bahasa Inggris," Lencana: Jurnal Inovasi Ilmu Pendidikan, vol. 3, no. 2, pp. 168–178, 2025, doi: 10.55606/lencana.v3i2.5030.

[6] T. Salsabila, N. Nafilah, F. Patangga, S. Zulfa, and N. Listyaningsih, "Literature Review: Efektivitas Penggunaan Aplikasi Duolingo Terhadap Motivasi Belajar Bahasa Inggris," Jurnal Empati, vol. 13, no. 4, pp. 302–312, 2024, doi: 10.14710/empati.2024.46691.

[7] A. Nurian and B. N. Sari, "Analisis Sentimen Ulasan Pengguna Aplikasi Google Play Menggunakan Naïve Bayes," JITET (Jurnal Informatika dan Teknik Elektro Terapan), vol. 11, no. 3 S1, pp. 829–835, 2023, doi: 10.23960/jitet.v11i3%20s1.3348.

[8] D. F. Nawulansih, N. C. Santi, and I. A. Sa'ida, "Analisis Sentimen Ulasan Aplikasi DANA di Google Play Store: Penerapan Support Vector Machine dan Synthetic Minority Over-sampling Technique," Jurnal Pendidikan dan Teknologi Indonesia, vol. 5, no. 9, pp. 2660–2671, 2025, doi: 10.52436/1.jpti.1053.

[9] M. Amien, "Sejarah dan Perkembangan Teknik Natural Language Processing (NLP) Bahasa Indonesia: Tinjauan tentang sejarah, perkembangan teknologi, dan aplikasi NLP dalam bahasa Indonesia," 2023. [Online]. Available: http://arxiv.org/abs/2304.02746

[10] R. Abdillah, E. Haerani, and R. M. Candra, "Analisis Sentimen Ulasan Aplikasi Wetv Untuk Peningkatan Layanan Menggunakan Metode Support Vector Machine," Journal of Information System Research (JOSH), vol. 4, no. 3, pp. 865–873, 2023, doi: 10.47065/josh.v4i3.3353.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language

Understanding," 2019, doi: 10.48550/arXiv.1810.04805.
[12] T. B. Kusuma and I. K. A. Mogi, "Implementasi BERT pada Analisis Sentimen Ulasan Destinasi Wisata Bali," JELIKU (Jurnal Elektronik Ilmu Komputer Udayana), vol. 12, no. 2, pp. 409–420, 2023, doi: 10.24843/JLK.2023.v12.i02.p19.
[13] F. Koto, J. H. Lau, and T. Baldwin, "INDOBERTWEET: A Pretrained Language Model for Indonesian Twitter with Effective Domain-Specific Vocabulary Initialization," 2021, doi: 10.18653/v1/2021.emnlp-main.833.
[14] J. C. Setiawan, K. M. Lhaksmana, and B. Bunyamin, "Sentiment Analysis of Indonesian TikTok Review Using LSTM and IndoBERTweet Algorithm," JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika), vol. 8, no. 3, pp. 774–780, 2023, doi: 10.29100/jipi.v8i3.3911.
[15] U. Khairani, V. Mutiawani, and H. Ahmadian, "Pengaruh Tahapan Preprocessing Terhadap Model IndoBERT dan IndoBERTweet Untuk Mendeteksi Emosi Pada Komentar Akun Berita Instagram," Jurnal Teknologi Informasi dan Ilmu Komputer, vol. 11, no. 4, pp. 887–894, 2024, doi: 10.25126/jtiik.1148315.