Journal of Information Technology and Computer Science (INTECOMS)
Volume 9 Nomor 3, Tahun 2026
e-ISSN : 2614-1574
p-ISSN : 2621-3249

PERBANDINGAN KINERJA PRE-TRAINED LANGUAGE MODEL BAHASA INDONESIA BERBASIS TRANSFORMER UNTUK ANALISIS SENTIMEN ULASAN TOKOPEDIA GOOGLE PLAY

PERFORMANCE COMPARISON OF TRANSFORMER-BASED PRE-TRAINED INDONESIAN LANGUAGE MODELS FOR SENTIMENT ANALYSIS OF TOKOPEDIA GOOGLE PLAY REVIEWS

Muhammad Kevin[1], Hanafi[2]
MTI PJJ, Universitas Amikom Yogyakarta[1]
Fakultas Ilmu Komputer, Universitas Amikom Yogyakarta[2]
muhammadkevin404@amikom.ac.id[1], hanafi@amikom.ac.id[2]

ABSTRACT

Sentiment analysis of e-commerce application reviews on Google Play Store has become an important approach for understanding user perceptions. However, sentiment classification remains challenging due to the unstructured nature of review data, the use of informal language, abbreviations, and variations in regional dialects and local languages. This study compares the performance of three Indonesian Pre-trained Language Models (PLMs), namely BERT Indonesia, RoBERTa Indonesia, and IndoBERT, for classifying the sentiment of Tokopedia user reviews written in Indonesian. In addition, DistilBERT Indonesia is evaluated as a computational efficiency baseline to analyze the trade-off between classification performance and computational resource requirements. The dataset consists of 10,000 Tokopedia reviews collected through web scraping using the google_play_scraper library. The research methodology includes preprocessing (data cleaning, text normalization, and subword tokenization), sentiment labeling (positive, negative, and neutral), dataset splitting, fine-tuning using the Hugging Face Transformers library, and performance evaluation based on accuracy, precision, recall, and F1-score. The experimental results show that IndoBERT achieved the highest test accuracy of 92.65%, followed by BERT Indonesia (92.15%) and RoBERTa Indonesia (91.78%), indicating that the three models deliver relatively competitive classification performance with only minor differences in accuracy. Although it achieved slightly lower accuracy, DistilBERT Indonesia reached 91.78% while requiring the shortest training time, demonstrating superior computational efficiency compared with the other models. These findings suggest that selecting an appropriate Pre-trained Language Model for Indonesian sentiment analysis should consider not only classification accuracy but also architectural characteristics and computational efficiency according to the intended application.

Keywords: *Sentiment Analysis, Natural Language Processing, BERT, IndoBERT, Tokopedia*

ABSTRAK

Analisis sentimen terhadap ulasan aplikasi e-commerce pada Google Play Store merupakan salah satu pendekatan penting untuk memahami persepsi pengguna. Namun, karakteristik data yang tidak terstruktur, penggunaan bahasa informal, singkatan, serta variasi dialek dan bahasa daerah masih menjadi tantangan dalam proses klasifikasi sentimen. Penelitian ini membandingkan performa tiga Pre-trained Language Model (PLM) berbahasa Indonesia, yaitu BERT Indonesia, RoBERTa Indonesia, dan IndoBERT, dalam mengklasifikasikan sentimen ulasan Tokopedia berbahasa Indonesia. Sebagai pembanding terhadap aspek efisiensi komputasi, DistilBERT Indonesia turut dievaluasi untuk menganalisis trade-off antara performa klasifikasi dan kebutuhan sumber daya komputasi. Dataset penelitian terdiri atas 10.000 ulasan Tokopedia yang diperoleh melalui web scraping menggunakan google_play_scraper. Tahapan penelitian meliputi preprocessing (pembersihan data, normalisasi teks, dan tokenisasi subword), pelabelan sentimen (positif, negatif, dan netral), pembagian data, proses fine-tuning menggunakan pustaka Hugging Face Transformers, serta evaluasi menggunakan metrik accuracy, precision, recall, dan F1-score. Hasil eksperimen menunjukkan bahwa IndoBERT memperoleh performa terbaik dengan akurasi pengujian sebesar 92,65%, diikuti oleh BERT Indonesia (92,15%) dan RoBERTa Indonesia (91,78%), yang menunjukkan bahwa ketiga model menghasilkan performa klasifikasi yang relatif kompetitif dan selisih akurasi yang kecil. Meskipun memiliki akurasi yang sedikit lebih rendah, DistilBERT Indonesia mencapai akurasi 91,78% dengan waktu pelatihan tercepat, sehingga menawarkan efisiensi komputasi yang lebih baik dibandingkan model lainnya. Temuan ini menunjukkan bahwa pemilihan pretrained language model tidak hanya dipengaruhi oleh tingkat akurasi, tetapi juga oleh karakteristik arsitektur dan efisiensi komputasi yang diperlukan pada proses analisis sentimen berbahasa Indonesia.

Kata Kunci: Analisis Sentimen, Natural Language Processing, BERT Indonesia, IndoBERT, Tokopedia.

PENDAHULUAN

Perkembangan teknologi terus mendorong laju pertumbuhan platform e-commerce di Indonesia, salah satu marketplace terbesar di Indonesia yakni Tokopedia memainkan peran penting dalam ekosistem perdagangan berbasis digital di Indonesia. Di tengah persaingan yang ketat, perusahaan tidak hanya harus menawarkan tingkat layanan yang andal, tetapi juga perlu memiliki pemahaman mendalam tentang pendapat, persepsi, dan pengalaman pengguna. Dalam menggambarkan persepsi ini, salah satu sumber informasi yang cukup relevan adalah ulasan pengguna, yang ditemukan di platform distribusi aplikasi seperti Google Play Store. Ulasan ini mencerminkan pendapat pengguna tentang kualitas layanan, baik positif, negatif, maupun netral. Meskipun demikian, sebagian ulasan yang tersedia berpotensi mengandung *fake review* atau ulasan yang dibuat untuk memanipulasi persepsi pengguna terhadap suatu aplikasi, sehingga dapat memengaruhi kualitas informasi yang diperoleh dari analisis sentimen. Fenomena tersebut dapat ditemukan pada sejumlah ulasan positif yang menggunakan pola kalimat serupa atau berulang, sehingga mengindikasikan adanya kemungkinan ulasan yang tidak merepresentasikan pengalaman pengguna secara autentik.

Namun di sisi lain menjadi sebuah tantangan, dimana karakteristik data ulasan yang tidak terstruktur, jumlahnya yang besar, serta penggunaan bahasa informal dan tidak baku menjadikan analisis manual tidak efisien dan rentan terhadap bias. Oleh karena itu, pendekatan berbasis *Natural Language Processing* (NLP), khususnya analisis sentimen, menjadi solusi yang efektif untuk mengotomatisasi proses klasifikasi opini pengguna dalam skala besar (Mao dkk., 2024). Perkembangan metode *Natural Language Processing* (NLP) berbasis *Deep Learning*, khususnya model berbasis Transformer, telah menunjukkan peningkatan performa yang signifikan dibandingkan pendekatan *Machine Learning* tradisional dalam berbagai tugas klasifikasi teks. Hal ini disebabkan oleh kemampuan model Transformer dalam menangkap konteks dan representasi semantik secara lebih mendalam dibandingkan metode konvensional seperti *Support Vector Machine* (SVM), *Naive Bayes*, dan *Logistic Regression*. Sebagai contoh, pada penelitian terbaru mengenai analisis sentimen ulasan e-commerce berbahasa Indonesia, model IndoBERT mampu mencapai akurasi sebesar 94,1%, yang secara signifikan lebih tinggi dibandingkan SVM sebesar 89,5% dan *Naive Bayes* sebesar 84,2%. Temuan ini menunjukkan bahwa pendekatan berbasis Transformer tidak hanya unggul secara teoritis, tetapi juga terbukti secara empiris dalam meningkatkan kinerja klasifikasi teks pada data nyata yang kompleks dan tidak terstruktur (Widyananda dkk., 2025).

Model *Bidirectional Encoder Representations from Transformers* (BERT) merupakan salah satu pendekatan yang dominan dalam NLP modern karena kemampuannya dalam memahami konteks bahasa secara dua arah melalui mekanisme *bidirectional attention*, sehingga mampu menghasilkan representasi yang lebih kontekstual dalam berbagai tugas klasifikasi teks. Hal ini didukung oleh temuan bahwa model berbasis Transformer seperti BERT menunjukkan performa yang lebih baik dibandingkan pendekatan sebelumnya dalam tugas klasifikasi teks (Batra dkk., 2021).

Salah satu pengembangan dari BERT adalah RoBERTa (*Robustly Optimized BERT Pretraining Approach*) yang diperkenalkan oleh Liu dkk. (2019) sebagai hasil optimasi terhadap proses *pretraining* BERT. Berbeda dengan BERT, RoBERTa menghilangkan mekanisme *Next Sentence Prediction* (NSP), menerapkan *dynamic masking*, menggunakan ukuran *batch* yang lebih besar, serta memanfaatkan korpus pelatihan yang lebih banyak sehingga

mampu meningkatkan kualitas representasi bahasa pada berbagai tugas *Natural Language Understanding* tanpa mengubah arsitektur dasar Transformer.

Selain itu, pengembangan model berbasis BERT yang disesuaikan dengan bahasa tertentu, seperti IndoBERT, menunjukkan performa yang lebih baik pada teks bahasa Indonesia (Christian dkk., 2025), termasuk menangani kompleksitas linguistik lokal, dan variasi bahasa informal dan konteks budaya (Jazuli dkk., 2024).

Selain pengembangan melalui optimasi *pretraining*, dikembangkan pula model yang lebih ringan seperti DistilBERT melalui teknik *knowledge distillation*. Model ini mempertahankan sebagian besar kemampuan representasi BERT dengan jumlah parameter yang lebih sedikit sehingga lebih efisien dalam proses pelatihan maupun inferensi (Sanh dkk., 2020).

Keempat model tersebut merepresentasikan pendekatan yang berbeda dalam pengembangan *Pre-trained Language Model* berbasis Transformer. BERT menjadi model dasar yang banyak digunakan, RoBERTa menawarkan optimasi pada proses *pretraining*, DistilBERT berfokus pada efisiensi melalui *knowledge distillation*, sedangkan IndoBERT mengadaptasi arsitektur BERT menggunakan korpus Bahasa Indonesia. Perbedaan karakteristik tersebut menjadikan keempat model menarik untuk dibandingkan dalam tugas analisis sentimen berbahasa Indonesia.

Sejumlah penelitian sebelumnya telah mengeksplorasi penggunaan model Transformer dalam analisis sentimen dan klasifikasi teks. Studi menunjukkan bahwa IndoBERT dan DistilBERT mampu memberikan performa yang kompetitif dalam klasifikasi emosi pada ulasan produk e-commerce berbahasa Indonesia, dengan akurasi mencapai 78,76% untuk IndoBERT dan 76,38% untuk DistilBERT dan peningkatan mencapai 80% setelah proses *tuning* (Christian dkk., 2025). Penelitian lain menunjukkan bahwa DistilBERT memiliki efisiensi dilatih 1,8 lebih cepat dibandingkan dengan BERT dengan akurasi hingga 85% dengan F1-score sebesar 84% pada tugas analisis sentimen teks pendek (Asyaky dkk., 2025). Sementara itu IndoBERT juga mengungguli model klasik, mencapai 87,50% dan LSTM mencapai akurasi 77,93% (Purwanto, 2026). IndoBERT juga terbukti berhasil menangkap konteks lebih baik dengan *outperform* hingga +35,6 F1 Points (Saputra dkk., 2026). Pada perbandingan lain IndoBERT mencapai akurasi rata-rata 99%, sementara IndoRoBERTa di angka 94% (Apriansyah dkk., 2025). Penelitian lain pada bidang berita pariwisata dengan *Fine Tuning* IndoBERT mencapai 77% (Wijaya dkk., 2025). Selain itu, Apriansyah dkk. (2025) membandingkan performa IndoBERT dan IndoRoBERTa pada analisis sentimen komentar publik terhadap film dokumenter *Dirty Vote*. Hasil penelitian menunjukkan bahwa kedua model mampu memberikan performa yang kompetitif, dengan IndoBERT memperoleh akurasi rata-rata 99%, lebih tinggi dibandingkan IndoRoBERTa sebesar 94% pada dataset dan konteks penelitian tersebut. Temuan ini mengindikasikan bahwa karakteristik *pretraining* dan arsitektur model dapat memengaruhi performa analisis sentimen, sehingga masih diperlukan evaluasi pada dataset dan domain yang berbeda.

Meskipun demikian, sebagian besar penelitian tersebut berfokus pada satu atau dua model tertentu, atau tidak secara spesifik mengkaji perbandingan antara BERT Indonesia, RoBERTa Indonesia, dan IndoBERT dalam konteks analisis sentimen ulasan aplikasi e-commerce di Indonesia, khususnya pada platform Tokopedia.

Berdasarkan kajian literatur tersebut, sebagian besar penelitian terdahulu membandingkan model Transformer berbahasa Indonesia terutama dari sisi akurasi klasifikasi, sementara aspek efisiensi komputasi umumnya hanya dilaporkan secara terbatas, misalnya sebatas waktu pelatihan atau jumlah

parameter model, tanpa mengaitkannya dengan mekanisme penggunaan memori GPU selama proses *fine-tuning*. Penelitian ini memberikan kontribusi melalui dua aspek utama. Pertama, penelitian ini melakukan analisis komparatif terhadap empat model Transformer berbahasa Indonesia (BERT Indonesia, RoBERTa Indonesia, IndoBERT, dan DistilBERT Indonesia) dengan mengevaluasi tidak hanya performa klasifikasi (*accuracy, precision, recall*, dan *F1-score*), tetapi juga efisiensi komputasi melalui waktu pelatihan serta penggunaan *allocated* dan *reserved* GPU memory selama proses *fine-tuning*. Pendekatan ini memberikan gambaran *trade-off* antara performa klasifikasi dan biaya komputasi yang lebih komprehensif dibandingkan penelitian sebelumnya yang umumnya hanya berfokus pada salah satu aspek. Kedua, penelitian ini melengkapi evaluasi kuantitatif dengan analisis kesalahan klasifikasi (*error analysis*) secara kualitatif pada ulasan yang mengandung sarkasme dan makna implisit dalam domain e-commerce berbahasa Indonesia, sehingga memberikan pemahaman mengenai keterbatasan model Transformer dalam menangkap konteks pragmatik. Dengan demikian, penelitian ini mengintegrasikan evaluasi performa klasifikasi, efisiensi komputasi, dan analisis kesalahan kontekstual dalam satu kerangka evaluasi sehingga menghasilkan gambaran yang lebih menyeluruh terhadap karakteristik masing-masing model.

Permasalahan yang diangkat dalam penelitian ini adalah bagaimana tingkat performa masing-masing model dalam mengklasifikasikan sentimen ulasan pengguna, faktor-faktor apa saja yang memengaruhi hasil klasifikasi, serta model mana yang paling optimal untuk digunakan pada analisis sentimen berbahasa Indonesia. Berdasarkan permasalahan tersebut, hipotesis penelitian ini adalah bahwa terdapat perbedaan yang signifikan antara model BERT Indonesia, RoBERTa Indonesia, dan IndoBERT dalam analisis sentimen ulasan Tokopedia. Perbedaan tersebut diperkirakan dipengaruhi oleh karakteristik arsitektur model, strategi *pretraining*, dan representasi bahasa yang dipelajari oleh masing-masing model.

Penelitian ini bertujuan untuk menganalisis dan membandingkan performa model BERT Indonesia, RoBERTa Indonesia, dan IndoBERT dalam analisis sentimen ulasan pengguna Tokopedia, mengidentifikasi faktor-faktor yang mempengaruhi kinerja model, serta menentukan model yang paling optimal untuk digunakan dalam klasifikasi sentimen teks berbahasa Indonesia.

METODE

Penelitian ini menggunakan pendekatan kuantitatif dengan desain eksperimental yang bertujuan membandingkan performa tiga *Pre-trained Language Model* (PLM) berbasis Transformer, yaitu BERT Indonesia, RoBERTa Indonesia, dan IndoBERT, dalam melakukan analisis sentimen terhadap ulasan pengguna aplikasi Tokopedia pada Google Play Store. Selain itu, DistilBERT Indonesia digunakan sebagai model pembanding untuk mengevaluasi *trade-off* antara performa klasifikasi dan efisiensi komputasi. Penelitian ini mengikuti *pipeline* dari penelitian sebelumnya (Palomino & Aider, 2022), yang terdiri atas beberapa tahapan mulai dari pengumpulan data, *preprocessing*, *tokenization*, pembagian data, *fine-tuning* model, dan evaluasi. Namun, ada sedikit modifikasi pada tahap *preprocessing* untuk menyesuaikan karakteristik data yang digunakan, seperti yang disajikan pada gambar dibawah ini.

Gambar 1. Research Methodology

2.1 Data Collection

Dimulai dengan memanfaatkan data teks berbahasa Indonesia yang diperoleh

melalui proses *web scraping* menggunakan pustaka google_play_scraper. Data diambil berdasarkan ulasan terbaru dengan parameter bahasa Indonesia (lang='id') dan wilayah Indonesia (country='id'), sehingga merepresentasikan kondisi aktual persepsi pengguna. Dataset yang digunakan berjumlah 10.000 ulasan pengguna. Meskipun jumlah data pada review Google Play Store sangat melimpah dan setiap saat bertambah, penelitian ini hanya menggunakan 10.000 data untuk menjaga efisiensi proses pelatihan dan evaluasi model, sekaligus untuk mengurangi waktu dan beban proses pelabelan manual. Pendekatan ini umum digunakan dalam pembelajaran mesin untuk mengurangi biaya komputasi tanpa harus menggunakan seluruh dataset. Selain itu, pemilihan subset tetap mempertahankan distribusi data agar representatif terhadap keseluruhan dataset (Chandrasekaran dkk., 2020).

2.2 Data Labelling

Proses pelabelan sentimen dilakukan oleh peneliti secara manual terhadap seluruh dataset ulasan pengguna Tokopedia. Setiap data dikategorikan ke dalam tiga kelas sentimen, yaitu positif, negatif, dan netral, berdasarkan makna dominan dari isi ulasan. Pelabelan dilakukan dengan mempertimbangkan konteks kalimat secara keseluruhan untuk meminimalkan kesalahan interpretasi, khususnya pada ulasan yang mengandung bahasa informal, singkatan, maupun sarkasme.

2.3 Data Cleaning

Tahapan selanjutnya meliputi pembersihan data (*cleaning*), normalisasi kata tidak baku dan singkatan, penghapusan elemen *noise* seperti URL, simbol, dan tanda baca, serta proses *case folding*. Tahapan *preprocessing* ini bertujuan untuk meningkatkan kualitas data teks sebelum proses pemodelan. Tahapan *preprocessing* disesuaikan dengan karakteristik model Transformer agar informasi kontekstual yang dipelajari selama proses *pretraining* tetap dapat dimanfaatkan secara optimal (Siino dkk., 2024).

2.4 Tokenization

Tokenisasi dilakukan menggunakan *tokenizer* bawaan (*pretrained tokenizer*) yang sesuai dengan masing-masing model, yaitu BertTokenizer untuk BERT Indonesia dan IndoBERT, RobertaTokenizer untuk RoBERTa Indonesia, serta DistilBertTokenizer untuk DistilBERT Indonesia. Pendekatan ini memungkinkan teks dipecah menjadi unit yang lebih kecil (*subword*), sehingga mampu menangani kata yang jarang muncul serta meningkatkan representasi konteks dalam model berbasis Transformer (Feng dkk., 2022).

2.5 Data Splitting

Dataset yang telah dilabeli kemudian dibagi menjadi data latih, validasi, dan data uji dengan proporsi 80% untuk *training*, 10% untuk *validation*, dan 10% untuk *testing*. Pembagian data ini merupakan praktik umum dalam pengembangan model *Machine Learning* untuk memastikan model dapat dilatih secara optimal serta dievaluasi secara objektif pada data yang belum pernah dilihat sebelumnya. Selain itu, penggunaan data validasi berperan penting dalam proses pemilihan model dan membantu mencegah terjadinya *overfitting*, sehingga meningkatkan kemampuan generalisasi model (Joseph & Vakayil, 2022).

2.6 Model Training

Implementasi model dilakukan menggunakan *framework* Hugging Face Transformers dengan memanfaatkan *pretrained model* BERT Indonesia, RoBERTa Indonesia, IndoBERT, dan DistilBERT Indonesia sebagai *baseline* efisiensi komputasi. Proses pelatihan (*fine-tuning*) dilakukan dengan menyesuaikan parameter model terhadap dataset ulasan yang telah diproses, sehingga model mampu mengenali pola sentimen pada

konteks bahasa Indonesia (Devlin dkk., 2019; Sanh dkk., 2020; Wilie dkk., 2020).

Evaluasi performa model dilakukan menggunakan metrik klasifikasi standar, yaitu akurasi, *precision*, *recall*, dan *F1-score*, yang dihitung berdasarkan *confusion matrix* multiclass. Selain itu, waktu pelatihan dan efisiensi komputasi juga dianalisis sebagai parameter tambahan dalam membandingkan kinerja model, khususnya untuk menilai *trade-off* antara akurasi dan kompleksitas model.

Seluruh tahapan penelitian dilakukan pada lingkungan sebagai berikut:

Tabel 1. Research Environment

Component	Spesification
CPU	Amd Athlon 3000G
GPU	RTX 3050 6GB
RAM	16 GB
Framework	PyTorch
Library	Transformers

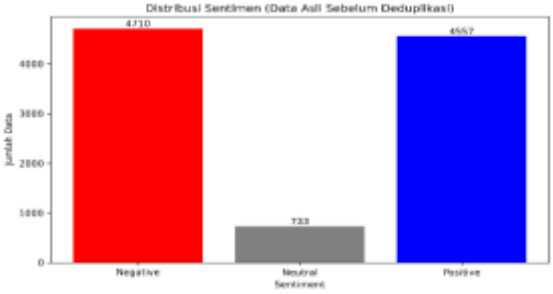
Seluruh tahapan penelitian dilaksanakan pada lingkungan eksperimen sebagaimana ditunjukkan pada Tabel 1, mulai dari pengumpulan data, *preprocessing*, proses *fine-tuning*, hingga evaluasi model untuk memperoleh perbandingan performa yang komprehensif antara BERT Indonesia, RoBERTa Indonesia, IndoBERT, serta DistilBERT Indonesia sebagai *baseline* efisiensi komputasi.

HASIL DAN PEMBAHASAN

Penelitian ini bertujuan menganalisis dan membandingkan performa tiga *Pre-trained Language Model* (PLM) berbasis Transformer, yaitu BERT Indonesia, RoBERTa Indonesia, dan IndoBERT, dalam analisis sentimen ulasan Tokopedia. Selain itu, DistilBERT Indonesia dievaluasi sebagai *baseline* untuk menganalisis *trade-off* antara performa klasifikasi dan efisiensi komputasi. Hasil pelatihan dan evaluasi model dianalisis secara menyeluruh untuk menjawab tujuan penelitian serta membandingkan performa klasifikasi dan efisiensi komputasi masing-masing model.

3.1 Data Preparation

Dataset sentimen berjumlah 10.000 data, yang dilabeli secara manual dengan sentimen positif, netral, dan negatif. Label sentimen positif dan negatif hampir setara, di 4000 lebih data, sementara label netral sedikit timpang, dimana dari 10.000 data hanya terdapat 733 data netral, seperti yang diilustrasikan pada Gambar 2.



Gambar 2. Sentiment Distribution

Pada dataset penelitian ini, terdapat banyak sekali kolom, mulai dari reviewId, content, userName, userImage, score, thumbsUpCount, dll. Dari data tersebut kami menambahkan satu kolom bernama Sentiment untuk *labeling*, dan membuang kolom yang tidak digunakan, sehingga menyisakan 2 kolom yakni content dan Sentiment. Berikut adalah contoh dari dataset kami:

Tabel 2. Example Entries From The Dataset

content	Sentiment
sejak pakai tokopedia, aplikasi sebelah saya uninstall	2
josss pokonya belanja di tokoped.. bnyk promo ..	2
mantap… barang datang sudah rusak semua	1
aplikasi tidak jelas buat pin susah sekali	1
pesanan perdana	0

Label 2 mewakili sentimen positif, label 1 mewakili sentimen negatif, dan label 0 mewakili sentimen netral. Penelitian ini menunjukkan ada beberapa dataset yang sarkas dan memiliki makna implisit, contohnya di baris ke 1 dan 3, kalimatnya seolah baik namun memiliki sentimen negatif. Selain itu bahasa yang digunakan juga kebanyakan informal, dengan bahasa gaul, singkatan, dan beberapa bahasa

daerah (misalnya, "bnyk", "mulu"). Struktur yang kasual, dengan tata bahasa yang setiap orang memiliki perbedaan, dan tidak sesuai KBBI, menghadirkan tantangan tersendiri.

Dataset tersebut kemudian dibagi menjadi beberapa set: mulai dari *training*, *validation*, dan *test* dengan rasio perbandingan 80:10:10. Rincian jumlah masing-masing set ada di Tabel 3.

Tabel 3. Data Splitting Overview

Total Sentiment	Training Set	Validation Set	Test Set
Negatif = 4710	3768	471	471
Positif = 4557	3645	456	456
Netral = 733	587	73	73

Sebelum melatih model, setiap set diproses terlebih dahulu, dimulai dengan menghapus data yang sentiment-nya tidak ada, NaN, dan hanya berisi spasi; mengubah label angka menjadi teks untuk mempermudah keterbacaannya; mengubah salah ketik dan singkatan menjadi kata-kata baku menggunakan satu file json yang telah dibuat dengan nama typoan.json; menghapus spasi berlebih; dan karena pelabelan dilakukan secara manual, kami menambahkan pengecekan label yang konflik, dimana content yang sama ada kemungkinan labelnya berbeda.

Untuk memvisualisasikan konten dataset yang sudah diperoleh, *word cloud* dibuat untuk menyoroti kata-kata yang paling dominan dalam setiap kategori sentimen, dan memberikan gambaran umum mengenai karakteristik masing-masing sentimen.

Gambar 3. Word cloud of Negatif Sentiment.

Gambar 4. Word cloud of Positif Sentiment.

Gambar 5. Word cloud of Neutral Sentiment.

Setelah data melewati beberapa proses diatas, dilanjutkan dengan proses tokenisasi, seperti yang sudah dibahas di awal, bahwa model IndoBERT menggunakan BertTokenizer, model RoBERTa Indonesia menggunakan RoBERTaTokenizer, model BERT Indonesia menggunakan BertTokenizer, DistilBERT Indonesia menggunakan DistilBertTokenizer.

3.2 Model Training and Evaluation

Pada penelitian ini, melibatkan 3 model antara lain: IndoBERT Indonesia, RoBERTa Indonesia, dan BERT Indonesia, sedangkan DistilBERT Indonesia digunakan sebagai *baseline* efisiensi komputasi. Selain menggunakan pengujian *performance metrics*, kami juga mengambil *training time*, *gpu usage log* untuk mengukur perbandingan efisiensi komputasi.

Pada model IndoBERT, penelitian ini menggunakan *tokenizer* BertTokenizer dengan *pretrained model* indobenchmark/indobert-base-p1 untuk melakukan tokenisasi teks. Proses tokenisasi dilakukan dengan menerapkan *truncation* dan *padding* secara otomatis dengan panjang maksimum token sebesar

128. Hasil tokenisasi kemudian digunakan pada proses *fine-tuning* menggunakan BertForSequenceClassification untuk tugas klasifikasi sentimen berbahasa Indonesia.

Sementara untuk RoBERTa Indonesia dan BERT Indonesia memiliki konfigurasi yang hampir sama, untuk lebih jelasnya ada pada Tabel 4.

Tabel 4. Hyperparameter

Parameter	IndoBERT	RoBERTa Indonesia	BERT Indonesia	DistilBERT Indonesia
Pretrained Model	indobenchmark/indobert-base-p1	cahya/roberta-base-indonesian-1.5G	cahya/bert-base-indonesian-522M	cahya/distilbert-base-indonesian
Tokenizer	BertTokenizer	RoBERTa Tokenizer	BertTokenizer	DistilbertTokenizer
Max Seq Length	128	128	128	128
Epoch	5	5	5	5
Batch Size	8	8	8	8
Learning Rate	5e-5	5e-5	5e-5	5e-5
Weight Decay	0.01	0.01	0.01	0.01
Warmup step	500	500	500	500
Evaluation Strategy	Epoch	Epoch	Epoch	Epoch
mixed precision	Yes	Yes	Yes	Yes

Seluruh model menggunakan konfigurasi *hyperparameter* yang relatif seragam untuk menjaga konsistensi dan keadilan dalam proses perbandingan performa klasifikasi sentimen. Perbedaan utama antar model terletak pada arsitektur *pretrained model* dan *tokenizer* yang digunakan.

Kemudian kami melakukan pengukuran performa pada model yang telah dilatih. Hasilnya, ketiga model berbasis Transformer mampu melakukan klasifikasi sentimen pada ulasan pengguna aplikasi Tokopedia dengan performa yang relatif tinggi. Hasil evaluasi menggunakan metrik akurasi, *precision*, *recall*, dan *F1-score*, IndoBERT memperoleh performa terbaik pada dataset penelitian ini. Meskipun demikian, selisih akurasi dengan BERT Indonesia (0,50%) dan RoBERTa Indonesia (0,87%) relatif kecil, menunjukkan bahwa ketiga PLM Bahasa Indonesia memberikan performa klasifikasi yang kompetitif. Menariknya, DistilBERT Indonesia memperoleh akurasi sebesar 91,78%, sama dengan RoBERTa Indonesia, namun dengan waktu pelatihan paling singkat. Hal ini menunjukkan bahwa DistilBERT layak digunakan ketika efisiensi komputasi menjadi prioritas.

Detail hasil perbandingan dari beberapa model tersebut kami sajikan dalam tabel dibawah ini.

Tabel 5. Classification Performance

Model	Accuracy	Precision	Recall	F1-Score
BERT Indonesia	92.15 %	92.00%	92.15%	92.00 %
RoBERTa Indonesia	91.78 %	92.00%	91.78%	91.45 %
IndoBERT	92.65 %	92.44%	92.65%	92.57 %
DistilBERT Indonesia	91.78 %	91.38%	91.78%	91.32 %

BERT Indonesia memperoleh akurasi tertinggi kedua dengan selisih hanya 0,50 poin persentase dibandingkan IndoBERT, menunjukkan bahwa kedua model memberikan performa klasifikasi yang relatif sebanding pada dataset penelitian ini. Pada posisi ketiga, DistilBERT menunjukkan performa yang kompetitif dibanding dengan dua model sebelumnya.

Selain itu kami juga membuat log pada hasil test yang salah prediksi, berikut pada Tabel 6.

Tabel 6. Wrong Prediction

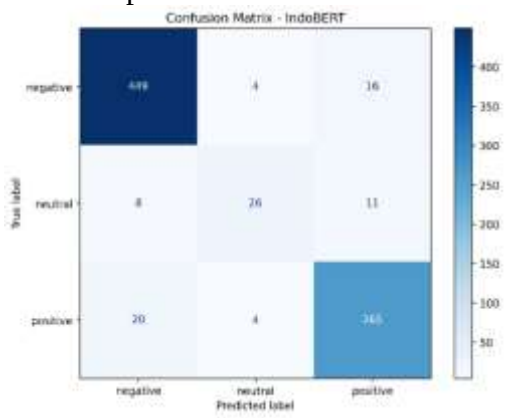
content	True Label	Prediction	Confidence
mantap… tinggal pindah ke kompetitor saja.	Negatif	Positif	0.90
mantap sekali barang datang sudah rusak semua	Negatif	ositif	.66

Dari Tabel 6, membuktikan bahwa model Transformer masih mengalami kesulitan dalam memahami kalimat yang mengandung sarkasme, ironi, maupun makna implisit. Sebagai contoh, ulasan "Terimakasih Tokopedia, order batal mulu jadi hemat" secara linguistik mengandung kata bernada positif, namun secara konteks merepresentasikan sentimen negatif. Kondisi ini menyebabkan model cenderung melakukan kesalahan prediksi karena keterbatasan dalam memahami konteks pragmatik dan emosi tersirat dalam teks (Bharti dkk., 2016).

Fenomena ini menunjukkan bahwa meskipun model Transformer seperti

IndoBERT memiliki kemampuan kontekstual yang kuat, model tersebut masih memiliki keterbatasan dalam memahami makna implisit yang bergantung pada konteks sosial dan interpretasi manusia. Sarkasme, sebagai bentuk ekspresi yang menyampaikan makna berlawanan dari kata-kata yang digunakan, menjadi salah satu faktor utama yang menyebabkan kesalahan klasifikasi. Hal ini juga diperkuat oleh karakteristik data ulasan aplikasi yang cenderung menggunakan bahasa informal, singkatan, serta campuran bahasa, sehingga menambah kompleksitas dalam proses pemahaman konteks.

Selain itu semua model mengalami penurunan performa pada kelas netral. Jumlah data netral yang lebih sedikit menjadi faktor utama, dimana totalnya tidak mencapai 10% dari total dataset yang tersedia. Terbukti F1 Neutral di bagian netral BERT hanya mencapai 59%, DistilBERT 54%, IndoBERT 64%, dan juga terbukti pada *confusion matrix* model IndoBERT pada Gambar 6.



Gambar 6. Confusion Matrix - IndoBERT

Berdasarkan *confusion matrix*, IndoBERT menunjukkan performa yang sangat baik pada kelas negatif dan positif, namun mengalami penurunan performa pada kelas netral. Hal ini terlihat dari cukup tingginya jumlah kesalahan klasifikasi sentimen netral menjadi positif maupun negatif. Kondisi tersebut diduga dipengaruhi oleh jumlah data netral yang relatif sedikit serta karakteristik sentimen netral yang cenderung ambigu dan tidak eksplisit.

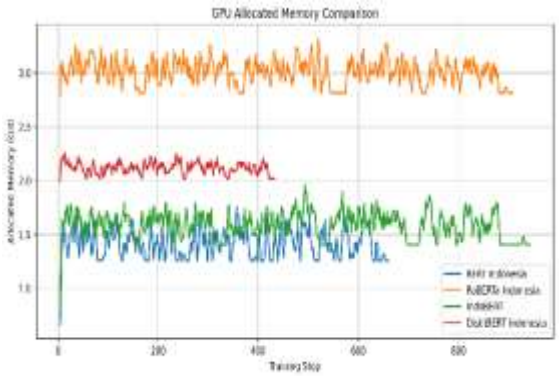
3.3 Trade-off Between Classification Performance and Computational Efficiency

Selain pengujian akurasi kami juga mencatat log dari pelatihan ke-3 model, seberapa besar *resources* yang digunakan dalam proses *training*.

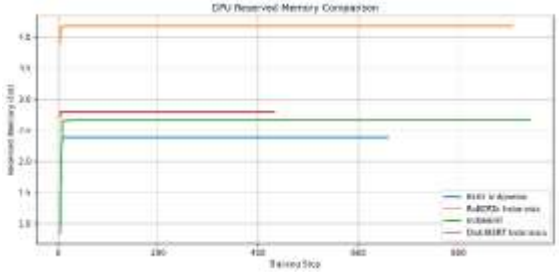
Tabel 7. Computational Efficiency

Model	Params (M)	Training Time	Avg GPU Memory	Max GPU Memory	Avg Reserved
BERT	110M	671.30 s (11.19 min)	1.42 GB	2.09 GB	2.38 GB
RoBERTa Indonesia	125M	938.30 s (15.64 min)	3.01 GB	3.76 GB	4.18 GB
IndoBERT	124.5 M	969.75 s (16.16 min)	1.58 GB	2.36 GB	2.66 GB
DistilBERT Indonesia	67M	443.69 s (7.39 min)	2.12 GB	2.37 GB	2.79 GB

Dari tabel diatas, sesuai dengan teori *knowledge distillation*, DistilBERT merupakan hasil *knowledge distillation* dari BERT dengan kecepatan inferensi sekitar 60% lebih cepat dibanding BERT (Sanh dkk., 2020). Meskipun DistilBERT Indonesia memiliki jumlah parameter paling sedikit (67 juta) dan waktu pelatihan tercepat, hasil eksperimen menunjukkan bahwa penggunaan memori GPU selama proses *fine-tuning* bukanlah yang paling rendah. Sebaliknya, BERT Indonesia menunjukkan penggunaan memori GPU paling efisien, sedangkan RoBERTa Indonesia memiliki konsumsi memori tertinggi. Temuan ini mengindikasikan bahwa efisiensi komputasi tidak dapat dinilai hanya berdasarkan jumlah parameter model, tetapi juga dipengaruhi oleh implementasi *framework*, mekanisme alokasi memori GPU, serta karakteristik proses *fine-tuning* masing-masing model. Pencatatan log lengkap penggunaan memori pada GPU ada pada Gambar 7 dan 8.



Gambar 7. GPU Memory Usage Comparison.



Gambar 8. GPU Memory Usage Comparison.

Gambar 7 diatas menunjukkan memori yang benar-benar digunakan oleh tensor / model. Sementara Gambar 8 adalah gambaran seberapa banyak memori yang dicadangkan PyTorch CUDA allocator agar training lebih cepat dan mengurangi alokasi ulang.

Meskipun DistilBERT secara arsitektur dirancang lebih ringan dibandingkan model Transformer lainnya, hasil monitoring GPU menunjukkan bahwa penggunaan memori selama proses *fine-tuning* tidak sepenuhnya sejalan dengan jumlah parameter model. Perbedaan penggunaan memori tersebut menunjukkan bahwa jumlah parameter bukan satu-satunya faktor yang menentukan konsumsi memori GPU selama *fine-tuning*. Aktivasi antar-layer, mekanisme penyimpanan gradien, implementasi *optimizer*, serta strategi alokasi memori pada PyTorch dan CUDA turut memengaruhi besarnya memori yang dialokasikan maupun dicadangkan selama proses pelatihan. Oleh karena itu, evaluasi efisiensi komputasi perlu mempertimbangkan waktu pelatihan dan penggunaan memori secara bersamaan, bukan hanya ukuran model. Visualisasi pada Gambar 7 dan Gambar 8 memperlihatkan bahwa penggunaan *allocated* memori dan *reserved* memori berbeda pada setiap model. Perbedaan tersebut mengindikasikan bahwa mekanisme *memory caching* pada CUDA dan PyTorch turut memengaruhi konsumsi memori GPU selama proses *fine-tuning*. Kondisi ini mengindikasikan bahwa mekanisme *caching* dan *memory reservation* pada CUDA melalui *framework* PyTorch mempertahankan blok memori GPU dalam jumlah lebih besar untuk mengurangi *overhead* alokasi ulang selama *training* berlangsung. Selain itu, perbedaan distribusi panjang token, *batch processing*, serta dinamika internal *optimizer* juga dapat memengaruhi pola penggunaan memori pada masing-masing model.

Jika dilihat dari *trade-off* antara performa klasifikasi dan efisiensi komputasi, IndoBERT memberikan akurasi tertinggi sebesar 92,65%, namun membutuhkan waktu pelatihan paling lama, yaitu 16,16 menit. Sebaliknya, DistilBERT Indonesia menyelesaikan proses pelatihan hanya dalam 7,39 menit, atau sekitar 54% lebih cepat dibandingkan IndoBERT, dengan penurunan akurasi yang relatif kecil, yaitu hanya 0,87 poin persentase. Hasil ini menunjukkan bahwa DistilBERT Indonesia merupakan alternatif yang menarik ketika efisiensi komputasi menjadi prioritas, sementara IndoBERT lebih sesuai digunakan ketika akurasi klasifikasi menjadi fokus utama.

SIMPULAN

Penelitian ini berhasil membandingkan performa tiga *Pre-trained Language Model* (PLM) berbasis Transformer, yaitu BERT Indonesia, RoBERTa Indonesia, dan IndoBERT, dalam analisis sentimen ulasan pengguna Tokopedia pada Google Play Store. Selain itu, DistilBERT Indonesia dievaluasi sebagai *baseline* untuk menganalisis *trade-off* antara performa klasifikasi dan efisiensi komputasi. Berdasarkan hasil evaluasi menggunakan metrik *accuracy, precision,*

recall, dan *F1-score*, IndoBERT memperoleh performa terbaik dengan akurasi sebesar 92,65%, diikuti oleh BERT Indonesia (92,15%), sedangkan RoBERTa Indonesia dan DistilBERT Indonesia sama-sama mencapai akurasi 91,78%. Temuan ini menunjukkan bahwa ketiga PLM berbahasa Indonesia memberikan performa klasifikasi yang relatif kompetitif dengan selisih akurasi kurang dari satu poin persentase.

Selain performa klasifikasi, penelitian ini juga menunjukkan adanya *trade-off* antara akurasi dan efisiensi komputasi. DistilBERT Indonesia memiliki jumlah parameter paling sedikit (67 juta) dan waktu pelatihan tercepat, sehingga menjadi alternatif yang menarik ketika efisiensi komputasi menjadi prioritas. Sebaliknya, IndoBERT membutuhkan waktu pelatihan paling lama, namun memberikan akurasi tertinggi pada dataset penelitian. Hasil penelitian juga menunjukkan bahwa jumlah parameter model yang lebih sedikit tidak selalu menghasilkan penggunaan memori GPU yang lebih rendah selama proses *fine-tuning*. Temuan ini mengindikasikan bahwa efisiensi komputasi dipengaruhi tidak hanya oleh ukuran model, tetapi juga oleh implementasi *framework* dan mekanisme alokasi memori GPU.

Faktor-faktor yang mempengaruhi performa model meliputi kualitas *preprocessing*, penggunaan bahasa informal, singkatan, ketidakseimbangan distribusi data, serta keberadaan sarkasme dan makna implisit pada ulasan pengguna. Seluruh model menunjukkan penurunan performa pada kelas sentimen netral karena jumlah data yang relatif sedikit dan karakteristik kalimat yang ambigu. Selain itu, model Transformer masih mengalami kesulitan dalam memahami konteks pragmatik dan ekspresi sarkastik yang bertentangan dengan makna literal kalimat.

Berdasarkan hasil penelitian, IndoBERT merupakan model yang paling sesuai ketika akurasi klasifikasi menjadi prioritas utama dalam analisis sentimen berbahasa Indonesia. Sebaliknya, DistilBERT Indonesia dapat dipertimbangkan sebagai alternatif ketika efisiensi waktu pelatihan lebih diutamakan. Penelitian selanjutnya dapat difokuskan pada peningkatan kemampuan model dalam memahami sarkasme, konteks implisit, serta penanganan ketidakseimbangan data agar performa klasifikasi menjadi lebih *robust* terhadap kompleksitas bahasa alami. Selain itu, evaluasi pada dataset dari domain yang berbeda, penerapan teknik penyeimbangan data, serta pengujian signifikansi statistik dapat dilakukan untuk memperkuat validitas hasil perbandingan. Penelitian lanjutan juga dapat mengeksplorasi hubungan antara karakteristik arsitektur model dan konsumsi memori GPU selama proses *fine-tuning* sehingga pemilihan model tidak hanya mempertimbangkan akurasi, tetapi juga efisiensi komputasi secara menyeluruh.

Penelitian ini memiliki beberapa keterbatasan, salah satunya adalah proses pelabelan sentimen yang dilakukan secara manual oleh satu orang peneliti sehingga masih berpotensi menimbulkan bias subjektivitas, terutama pada data yang mengandung sarkasme atau makna implisit. Penelitian selanjutnya disarankan melibatkan lebih dari satu anotator, menyusun pedoman anotasi yang lebih terstruktur, serta mengukur tingkat kesepakatan menggunakan *Cohen's Kappa* untuk meningkatkan reliabilitas proses pelabelan.

DAFTAR PUSTAKA

Apriansyah, F. M., Ramadhan, T. I., Hidayat, C. R., & Wijaya, A. K. (2025). Perbandingan IndoBERT dan IndoRoBERTa untuk analisis sentimen pada film dokumenter *Dirty Vote*. *Jurnal Informatika: Jurnal Pengembangan IT, 10*(3), 593–605. https://doi.org/10.30591/jpit.v10i3.8607

Asyaky, M. S., Al-Husaini, M., & Lukmana, H. H. (2025). Sentiment

analysis on short social media texts using DistilBERT. *Journal of Computer Networks, Architecture and High Performance Computing*, *7*(2), 524–533. https://doi.org/10.47709/cnahpc.v7i2.5836

Batra, H., Punn, N. S., Sonbhadra, S. K., & Agarwal, S. (2021). BERT-based sentiment analysis: A software engineering perspective. *Lecture Notes in Computer Science, 12923*, 138–148. https://doi.org/10.1007/978-3-030-86472-9_13

Bharti, S. K., Vachha, B., Pradhan, R. K., Babu, K. S., & Jena, S. K. (2016). Sarcastic sentiment detection in tweets streamed in real time: A big data approach. *Digital Communications and Networks, 2*(3), 108–121. https://doi.org/10.1016/j.dcan.2016.06.002

Chandrasekaran, J., Feng, H., Lei, Y., Kacker, R., & Kuhn, D. R. (2020). Effectiveness of dataset reduction in testing machine learning algorithms. Dalam *2020 IEEE International Conference on Artificial Intelligence Testing (AITest)* (hlm. 133–140). IEEE. https://doi.org/10.1109/AITEST49225.2020.00027

Christian, W., Adamlu, D., Yu, A., & Suhartono, D. (2025). *Leveraging IndoBERT and DistilBERT for Indonesian emotion classification in e-commerce reviews*. arXiv. https://doi.org/10.48550/arXiv.2509.14611

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding*. arXiv. https://doi.org/10.48550/arXiv.1810.04805

Feng, Z., dkk. (2022). *Pretraining without wordpieces: Learning over a vocabulary of millions of words*. arXiv. https://doi.org/10.48550/arXiv.2202.12142

Jazuli, A., Widowati, & Kusumaningrum, R. (2024). Optimizing aspect-based sentiment analysis using BERT for comprehensive analysis of Indonesian student feedback. *Applied Sciences*, *15*(1), 172. https://doi.org/10.3390/app15010172

Joseph, V. R., & Vakayil, A. (2022). SPlit: An optimal method for data splitting. *Technometrics, 64*(2), 166–176. https://doi.org/10.1080/00401706.2021.1921037

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A robustly optimized BERT pretraining approach*. arXiv. https://doi.org/10.48550/arXiv.1907.11692

Mao, Y., Liu, Q., & Zhang, Y. (2024). Sentiment analysis methods, applications, and challenges: A systematic literature review. *Journal of King Saud University - Computer and Information Sciences, 36*(4), 102048. https://doi.org/10.1016/j.jksuci.2024.102048

Palomino, M. A., & Aider, F. (2022). Evaluating the effectiveness of text pre-processing in sentiment analysis. *Applied Sciences, 12*(17), 8765. https://doi.org/10.3390/app12178765

Purwanto, D. D. (2026). Empirical evaluation of IndoBERT and LSTM for sentiment analysis of tourism reviews: A data-driven study on Kenjeran Park. *Jurnal Teknik Informatika (JUTIF)*, *7*(1), 463–474. https://doi.org/10.52436/1.jutif.2026.7.1.4901

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2020). *DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter*. arXiv.

https://doi.org/10.48550/arXiv.1910.01108

Saputra, M. A. A., Alamsyah, A., Ramadhani, D. P., Siadari, T. S., & Fakhrurroja, H. (2026). *IndoBERT-Sentiment: Context-conditioned sentiment classification for Indonesian text*. arXiv. https://doi.org/10.48550/arXiv.2604.07057

Siino, M., Tinnirello, I., & La Cascia, M. (2024). Is text preprocessing still worth the time? A comparative survey on the influence of popular preprocessing methods on Transformers and traditional classifiers. *Information Systems, 121*, 102342. https://doi.org/10.1016/j.is.2023.102342

Widyananda, W., Maskur, & Fauzi, A. (2025). Machine learning and transformer-based model for sentiment analysis of Indonesian e-commerce reviews. *Indonesian Journal of Computer Science, 14*(4). https://doi.org/10.33022/ijcs.v14i4.4980

Wijaya, I. N. S. W., Seputra, K. A., & Dewi, N. P. N. P. (2025). Fine tuning model IndoBERT untuk analisis sentimen berita pariwisata Indonesia. *Jurnal Pendidikan Teknologi dan Kejuruan, 22*(2), 195–204. https://doi.org/10.23887/jptk-undiksha.v22i2.104056

Wilie, B., Vincentio, K., Winata, G. I., Cahyawijaya, S., Li, X., Lin, Z., Lim, S., Kurniawan, S., Cipta, R., & Purwarianti, A. (2020). *IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding*. arXiv. https://doi.org/10.48550/arXiv.2009.05387