

Perbandingan Kinerja Model IndoBERT Base-P1 dan IndoBERT Base-P2 dalam Klasifikasi Sentimen Ulasan Aplikasi Flip

Siti Nur Salamah

Program Studi Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Indonesia
siti.nur26@ui.ac.id

ARTICLE INFO	ABSTRACT
 Siti Nur Salamah – Matematika 2022	<p>NIM 2206048833</p>
<hr/>	
<p>Keywords: IndoBERT Analisis Sentimen Klasifikasi Teks Aplikasi Flip</p>	
<p>Dalam era ekonomi digital, aplikasi <i>fintech</i> seperti Flip berperan penting dalam mempermudah transaksi keuangan masyarakat Indonesia. Melalui fitur transfer antarbank tanpa biaya administrasi dan layanan keuangan digital lainnya, Flip telah menarik jutaan pengguna dan menghasilkan banyak ulasan di platform seperti <i>Google Play Store</i>. Ulasan-ulasan ini menjadi sumber data penting untuk memahami persepsi pengguna yang perlu dianalisis secara sistematis melalui pendekatan analisis sentimen berbasis kecerdasan buatan. Penelitian ini bertujuan untuk membandingkan kinerja dua model <i>pretrained</i> IndoBERT, yaitu IndoBERT-base-p1 dan IndoBERT-base-p2, dalam tugas klasifikasi sentimen ulasan aplikasi Flip. Model dilatih menggunakan <i>optimizer Adam</i> dengan <i>learning rate</i> 3e-6 dan <i>batch size</i> 32 selama 5 <i>epoch</i>, dengan pembagian data sebesar 70% <i>training</i>, 20% validasi, dan 10% <i>testing</i>. Evaluasi menggunakan metrik akurasi, presisi, <i>recall</i>, dan F1-score menunjukkan bahwa IndoBERT-base-p2 memiliki performa lebih unggul dengan akurasi 87% dan <i>macro average</i> F1-score 0.87, dibandingkan IndoBERT-base-p1 yang mencapai akurasi 85% dan <i>macro average</i> F1-score 0.84. Dengan demikian, IndoBERT-base-p2 dinilai lebih efektif dalam mengklasifikasikan sentimen ulasan berbahasa Indonesia.</p>	
<p>Copyright © 2024. All rights reserved.</p>	

I. Pendahuluan

Dalam era ekonomi digital, teknologi finansial atau *financial technology (fintech)* memainkan peran yang semakin penting dalam kehidupan masyarakat Indonesia. *Fintech* adalah inovasi pada layanan keuangan yang menggabungkan teknologi informasi untuk menciptakan transaksi yang lebih efisien, aman, dan modern [1]. Salah satu bentuk nyata dari layanan *fintech* adalah hadirnya aplikasi dompet digital seperti Flip. Flip adalah aplikasi keuangan digital yang memungkinkan pengguna melakukan transfer antarbank secara gratis tanpa biaya administrasi yang biasanya berkisar antara Rp2.500 hingga Rp6.500 per transaksi [2]. Selain fitur transfer beda bank tanpa biaya, Flip juga menyediakan berbagai layanan tambahan seperti pembelian pulsa dan paket data, pembayaran tagihan listrik dan air, *top-up* dompet digital seperti Gopay dan OVO, serta pengiriman uang ke luar negeri [3]. Dengan beragam fitur unggulan tersebut, Flip berhasil menarik perhatian masyarakat luas dan mengalami pertumbuhan signifikan. Berdasarkan data dari *Google Play Store* hingga April 2025, aplikasi ini telah diunduh lebih dari 10 juta kali dan memperoleh rating sebesar 4,8 dari 5,0 berdasarkan 675 ribu ulasan pengguna.

Ulasan pengguna di platform seperti *Google Play Store* merupakan sumber data berharga yang dapat mencerminkan persepsi, kepuasan, dan keluhan mereka terhadap sebuah layanan. Pendapat-pendapat ini berfungsi sebagai umpan balik yang penting bagi pengembang aplikasi untuk meningkatkan kualitas layanan dan menciptakan inovasi yang relevan [4]. Namun, jumlah ulasan yang sangat besar dapat menyulitkan pihak pengembang untuk menganalisis secara manual, sehingga dibutuhkan pendekatan sistematis dan otomatis melalui analisis sentimen. Analisis sentimen adalah proses memahami, mengekstrak, dan mengelompokkan opini pengguna dari data teks menjadi kategori sentimen tertentu, seperti positif, negatif, atau netral [5]. Metode ini dapat memberikan

wawasan mengenai emosi dan persepsi pengguna terhadap layanan digital, serta membantu dalam pengambilan keputusan strategis. Lebih lanjut, pendekatan seperti *Aspect-Based Sentiment Analysis* (ABSA) atau biasa disebut sebagai sentimen analisis telah dikembangkan untuk mengklasifikasikan opini berdasarkan aspek layanan tertentu agar hasilnya lebih akurat dan bermanfaat secara praktis [3].

Penelitian terdahulu mengenai analisis sentimen telah banyak dilakukan pada berbagai *platform* media sosial dan *e-commerce* yang menunjukkan keunggulan berbagai metode *machine learning*. Pada penelitian oleh M. Isnain et al. [6], penggunaan model *Support Vector Machine* (SVM) berhasil mencapai F1-score terbaik dengan akurasi sebesar 80%. Sementara itu, penelitian oleh M. E. Purbaya et al. [7] menunjukkan bahwa model SVM mencapai akurasi terbaik sebesar 89,60%. Penelitian lain oleh Z. A. Diekson et al. [8], T. Wilianto et al. [9], dan M. J. Hossain et al. [10] menggunakan dataset ulasan *e-commerce* dan menemukan bahwa model SVM memberikan akurasi terbaik berturut-turut sebesar 84,5%, 78%, dan 94%.

Penelitian khusus mengenai aplikasi Flip juga telah dilakukan dengan hasil yang beragam. Penelitian oleh Sri Rahayu et al. menunjukkan bahwa penggunaan algoritma *K-Nearest Neighbor* (KNN) dengan pembobotan TF-IDF pada ulasan Flip menghasilkan akurasi klasifikasi sebesar 76,68%, dengan *precision* 82,67% dan *recall* 86,92% [11]. Penelitian lainnya oleh Kusmayanti Solecha dan Oky Irnawati menunjukkan bahwa penggunaan algoritma *Naive Bayes* berbasis *Particle Swarm Optimization* menghasilkan akurasi sebesar 88,24%, sedangkan SVM berbasis *Particle Swarm Optimization* mencapai akurasi tertinggi sebesar 88,61% [12].

Namun, seiring berkembangnya teknologi pemrosesan bahasa alami *Natural Language Processing* (NLP), pendekatan berbasis *deep learning* semakin banyak digunakan untuk meningkatkan akurasi dan efisiensi dalam tugas klasifikasi sentimen. Salah satu model *deep learning* berbasis *transformer* yang dirancang khusus untuk bahasa Indonesia adalah IndoBERT [13]. Model ini merupakan varian dari *Bidirectional Encoder Representations from Transformers* (BERT) yang dilatih menggunakan korpus berbahasa Indonesia. Penelitian sebelumnya telah menunjukkan bahwa IndoBERT mampu memberikan kinerja yang sangat baik dalam berbagai tugas NLP berbahasa Indonesia. Penelitian oleh M. A. Hadiwijaya et al. [14] menemukan bahwa IndoBERT mengungguli model *Logistic Regression* dan *Support Vector Classification* (SVC) dengan akurasi sebesar 91%. Selain itu, penelitian lain yang mengombinasikan IndoBERT dengan RCNN juga menghasilkan akurasi tinggi sebesar 95,16% [15], dan penelitian oleh W. M. Baihaqi dan A. Munandar [16] menemukan bahwa IndoBERT mengungguli model *Naïve Bayes* dengan akurasi sebesar 85%.

Berdasarkan latar belakang tersebut, dalam penelitian ini akan digunakan dua varian dari IndoBERT, yaitu IndoBERT-base-p1 dan IndoBERT-base-p2, untuk melakukan analisis sentimen terhadap ulasan pengguna aplikasi Flip. Penggunaan kedua model ini bertujuan untuk membandingkan performa masing-masing dalam klasifikasi sentimen positif, negatif, dan netral, serta menilai keandalan pendekatan berbasis *pretrained transformer* dalam konteks *fintech* Indonesia.

II. Literature Review

A. Preprocessing Data Teks

Data *preprocessing* bertujuan untuk mengubah data mentah tidak terstruktur menjadi format yang terstruktur dan siap untuk diproses lebih lanjut. Tahapan ini meliputi beberapa proses, yaitu:

- *Case Folding*: Mengubah semua huruf dalam teks menjadi huruf kecil [17].
- *Data cleaning*: Proses untuk mengurangi *noise* pada data dengan cara membersihkan elemen-elemen teks yang tidak diperlukan [18].
- *Tokenization*: Memecah teks menjadi token (kata, tanda baca, dan ekspresi bermakna) sesuai kaidah bahasa [19].
- *Normalization*: Mengubah teks informal atau slang menjadi bentuk baku [17].

B. Analisis Sentimen

Analisis sentimen merupakan subbidang NLP untuk memahami opini, perasaan, atau emosi dalam teks. Tujuannya adalah mengekstrak, mengelompokkan, dan menganalisis informasi untuk mengetahui sudut pandang atau emosi penulis [20].

C. Bidirectional Encoder Representations from Transformers (BERT)

BERT adalah model representasi bahasa berbasis *transformer* dua arah yang memahami hubungan kata dari konteks kiri dan kanan secara simultan [21]. BERT menggunakan mekanisme *self-attention* untuk membentuk representasi kontekstual dan unggul dalam tugas transfer *learning*. Model ini terdiri dari token [CLS], [SEP], *mask ID*, *segment ID*, dan *positional embedding* dalam proses tokenisasinya.

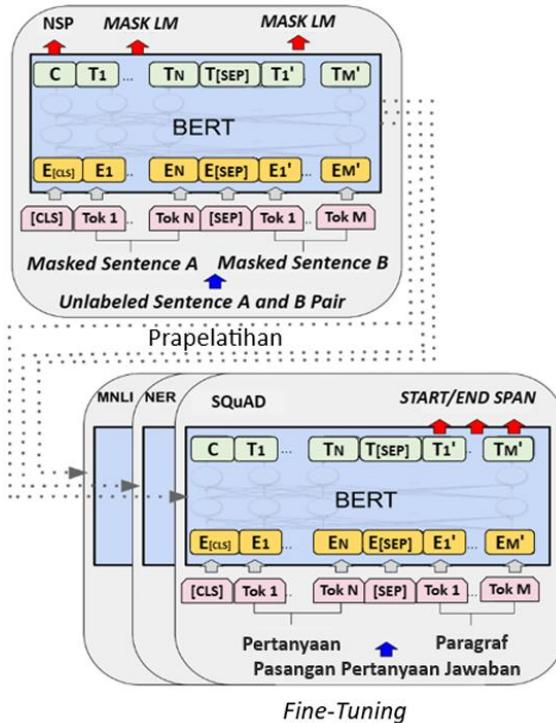


Fig. 1. Prapelatihan dan *Fine-Tuning* BERT

D. IndoBERT

IndoBERT adalah adaptasi BERT yang dilatih pada korpus besar bahasa Indonesia menggunakan 2,4 juta langkah atau 180 *epoch*. Model ini dibangun dengan arsitektur 12 lapisan *transformer* dan kosakata khusus bahasa Indonesia sebanyak 31.923 token [22].

E. IndoBERT-Base

IndoBERT-base adalah model dasar dari IndoBERT yang telah dilatih dengan korpus 5,5 miliar kata yang mencakup beberapa bentuk teks bahasa Indonesia. Model ini dapat digunakan untuk berbagai tugas NLP. Model ini terdiri atas 12 lapisan transformator dengan 12 *heads* per lapisan dan 125 juta parameter. Berikut ini jenis-jenis IndoBERT-base [22], [23].

1) IndoBERT-base-p1

Dengan menggunakan teknik transfer *learning*, model ini dilatih pada kumpulan data teks dalam bahasa Indonesia yang sangat besar dari berbagai sumber, termasuk artikel berita, Wikipedia, dan media sosial. Model ini juga dapat digunakan untuk melakukan berbagai tugas NLP.

2) IndoBERT-base-p2

Dibandingkan dengan IndoBERT-base-p1, model ini telah dilatih pada *dataset* yang lebih kompleks dan dapat menghasilkan keluaran yang lebih akurat. Hasilnya adalah model ini sesuai untuk tugas-tugas yang membutuhkan pemahaman bahasa yang lebih dalam, seperti klasifikasi dokumen dan analisis sentimen.

III. Metodologi Penelitian

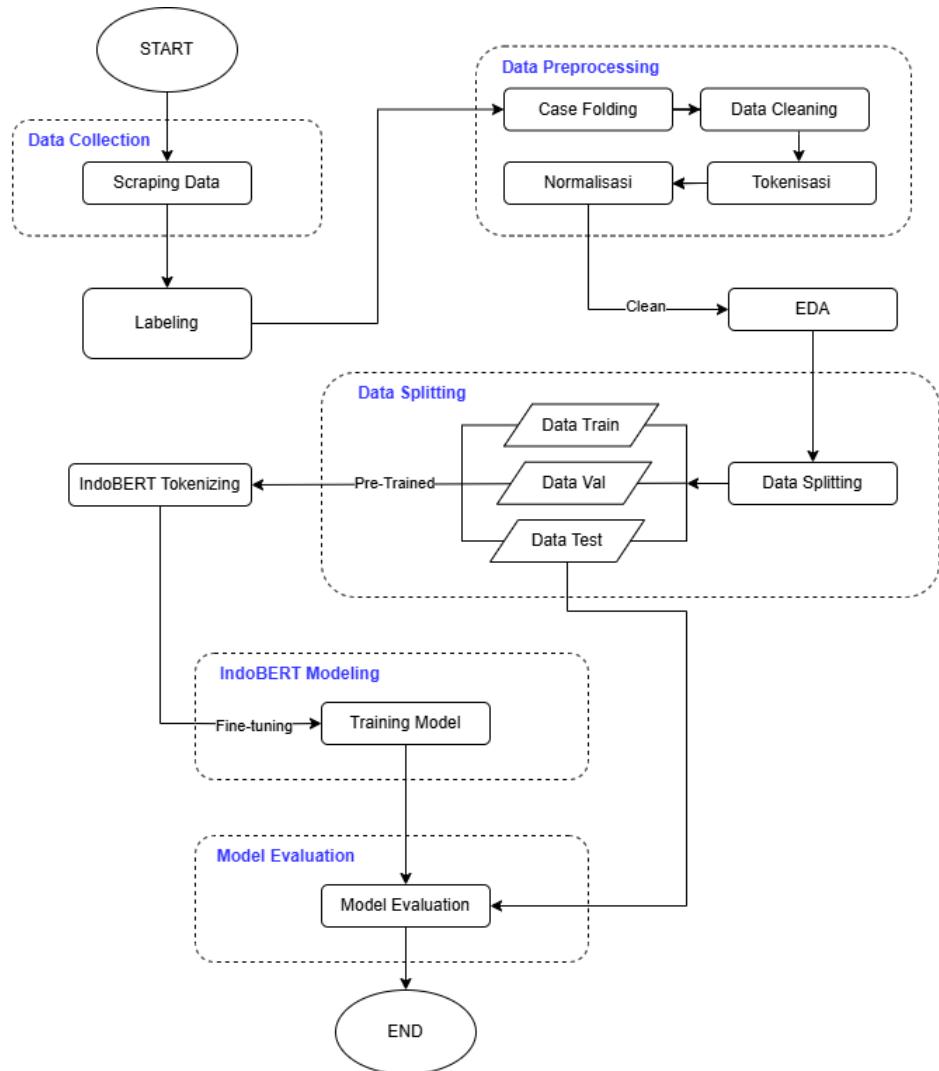


Fig. 2. Alur proses analisis sentimen menggunakan IndoBERT

Penelitian ini menggunakan *dataset* ulasan pengguna aplikasi Flip yang tersedia secara publik dan telah melalui proses *scraping* serta *labeling* oleh pihak lain [24]. *Dataset* yang digunakan telah dilabeli ke dalam tiga kategori sentimen, yaitu positif, netral, dan negatif. Tahap selanjutnya adalah data *preprocessing*, yaitu proses pembersihan dan persiapan data teks agar layak digunakan dalam pelatihan model. Proses ini mencakup empat langkah utama, yaitu *case folding*, *data cleaning*, *tokenisasi*, dan *normalisasi* kata tidak baku. Setelah melalui tahap *preprocessing*, dilakukan EDA dan data dibagi menjadi tiga bagian, yakni data latih (*training data*), data validasi (*validation data*), dan data uji (*test data*) melalui proses *data splitting*. Setelah pembagian data, masing-masing bagian diproses menggunakan *tokenizer* khusus IndoBERT dan masuk ke tahap *modeling*. Pada tahap ini, dilakukan proses *fine-tuning* terhadap model *pre-trained IndoBERT* yang sebelumnya telah dilatih menggunakan korpus 5,5 miliar kata yang mencakup beberapa bentuk teks bahasa Indonesia [25]. Proses pelatihan ini bertujuan untuk mengadaptasi model terhadap konteks ulasan aplikasi Flip. Langkah terakhir dalam alur ini adalah evaluasi model yang dilakukan menggunakan metrik-metrik seperti akurasi, *precision*, *recall*, dan F1-score untuk mengukur performa model dalam mengklasifikasikan sentimen. Secara keseluruhan, tahapan penelitian ini dijelaskan secara visual pada Fig 2.

A. Data Collection

Dataset yang digunakan dalam penelitian ini merupakan data ulasan pengguna aplikasi Flip yang telah tersedia secara publik dan telah melalui proses *scraping* serta *labeling* oleh pihak lain [24].

Proses *scraping* yang dilakukan dalam penyusunan *dataset* tersebut memanfaatkan Pustaka *google-play-scrapers*, yaitu sebuah antarmuka API berbasis *Python* yang memungkinkan ekstraksi data dari *Google Play* secara sistematis [26]. Untuk menghasilkan *dataset* yang seimbang, proses pengambilan data dilakukan berdasarkan kategori rating bintang (1 hingga 5), dengan proporsi yang disesuaikan. Selanjutnya, dilakukan penyeleksian fitur dengan membuang beberapa kolom seperti *id*, *username*, dan tanggal karena dianggap tidak relevan dalam proses analisis sentimen. *Dataset* ini terdiri dari 2820 ulasan.

B. Labeling

Dalam analisis sentimen berbasis metode *supervised learning*, dibutuhkan *dataset* yang telah dianotasi atau dilabeli terlebih dahulu. Labelisasi ini penting karena metode *supervised* memerlukan data yang telah diberi contoh untuk proses pelatihan model. Pada *dataset* yang digunakan dalam penelitian ini, proses pelabelan telah dilakukan oleh pihak penyusun *dataset* asli [24]. Label diberikan berdasarkan skor rating bintang yang diberikan pengguna di *Google Play Store*, dengan ketentuan sebagai berikut: bintang 1 dan 2 dikategorikan sebagai sentimen *negatif*, bintang 3 sebagai *netral*, dan bintang 4 serta 5 sebagai *positif*.

Table 1. Contoh Kalimat Hasil Labeling

Ulasan	Category
Mudah kirim ke berbagai bank dan proses aman dan cepat ,yg pastinya gratis biaya admin,flip muach... BdW mau tanya kira kira brp batas maksimal untuk transaksi transfer,tq	Positif
Cara top up nya gimana	Netral
Saya mengalami kendala dalam konfirmasi email...sampai sekarang blm bisa saya pake aplikasi nya...konfirmasi emailnya lamaaaaaaaa banget ga ada keterangan apapun pas saya tekan konfirmasi email,malah biru layar saya...	Negatif

Proses pelabelan diatas dilakukan secara semi-otomatis dengan memanfaatkan fungsi *VLOOKUP* di *Microsoft Excel* untuk mempercepat pengklasifikasian data berdasarkan nilai bintang. Setelah tahap tersebut, penyusun *dataset* juga melakukan pengecekan manual untuk memastikan kesesuaian antara label yang diberikan dengan isi teks ulasan. Langkah ini penting dilakukan karena terdapat sejumlah ulasan yang tidak relevan dengan skor bintang yang tercantum, serta adanya data yang bersifat ambigu. Pemahaman terhadap proses pelabelan ini menjadi krusial dalam mengevaluasi kualitas dan validitas *dataset* yang digunakan dalam proses pelatihan model.

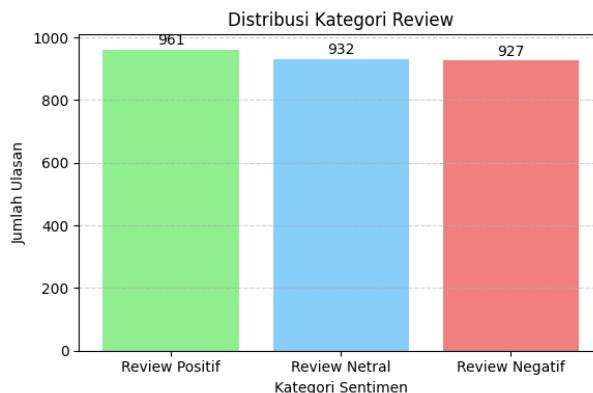


Fig. 3.Distribusi Hasil Labeling Semi-Otomatis dari Penyusun Dataset

Fig 3 menunjukkan hasil *labeling* yang terdiri dari 961 ulasan positif, 932 ulasan netral, dan 927 ulasan negatif.

C. Data Preprocessing

Pada tahap ini diterapkan proses *preprocessing* untuk memudahkan proses klasifikasi. *Preprocessing* merupakan langkah penting dalam membersihkan data mentah dengan tujuan meningkatkan akurasi dan efisiensi model [27]. Pada penelitian ini, *preprocessing* dilakukan untuk mengubah *dataset* yang tidak terstruktur menjadi lebih terstruktur, sehingga mempermudah data untuk diproses lebih lanjut. Tahapan yang dilakukan meliputi *case folding*, *data cleaning*, *tokenisasi*, serta normalisasi kata tidak baku. Dengan penerapan tahapan ini, kualitas data yang akan digunakan dalam proses klasifikasi dapat ditingkatkan secara signifikan.

1) Case folding

Case folding merupakan proses yang dilakukan untuk mengubah setiap kata yang ada di dalam *dataset* menjadi huruf kecil [28]. Tahapan ini dilakukan karena data yang diperoleh tidak selalu terstruktur dan konsisten dalam penggunaan huruf kapital sehingga *case folding* dilakukan untuk menyamaratakan penggunaan huruf kapital.

Table 2. *Case Folding*

<i>Original Teks</i>	<i>Case Folding</i>
Mantap tanpa biaya admin  terbaik	mantap tanpa biaya admin  terbaik
Error terus aplikasi nya ga bs di buka	error terus aplikasi nya ga bs di buka

2) Data Cleaning

Data cleaning merupakan proses untuk mengurangi *noise* pada data dengan cara membersihkan elemen-elemen teks yang tidak diperlukan [18]. Pada tahap ini, dilakukan penghapusan angka, simbol tertentu, URL, *username* (contoh: @username), *hashtag* (contoh: #tagar), spasi berlebih, tanda baca, *emoji*, serta karakter yang muncul berulang dalam sebuah kata. Proses ini menggunakan pola *regular expression* (regex) untuk mendeteksi dan menghapus karakter-karakter tersebut secara otomatis sehingga menghasilkan data teks yang lebih bersih dan siap untuk diproses lebih lanjut.

Table 3. *Data Cleaning*

<i>Case Folding</i>	<i>Data Cleaning</i>
Mantap tanpa biaya admin  terbaik	mantap tanpa biaya admin terbaik
error terus aplikasi nya ga bs di buka	error terus aplikasi nya ga bs di buka

3) Tokenisasi

Tokenisasi adalah proses memecah kalimat menjadi kata, tanda baca, dan ekspresi bermakna lainnya sesuai dengan aturan dalam suatu bahasa [29]. Hasil dari proses ini disebut sebagai token [30]. Dalam penelitian ini, tokenisasi dilakukan untuk memecah setiap kalimat menjadi daftar kata (*list of words*). Proses ini dilakukan menggunakan fungsi *word_tokenize* yang disediakan oleh *library Natural Language Toolkit* (NLTK) dalam Python. Selain itu, tahap ini juga didukung oleh penggunaan *regular expression* untuk menemukan karakter yang akan dihapus.

Table 4. Tokenisasi

<i>Data Cleaning</i>	<i>Tokenisasi</i>
mantap tanpa biaya admin terbaik	[mantap, tanpa, biaya, admin, terbaik]
error terus aplikasi nya ga bs di buka	[error, terus, aplikasi, nya, ga, bs, di, buka]

4) Normalisasi

Normalisasi adalah proses mengubah kata singkatan atau slang menjadi bentuk baku dalam bahasa Indonesia [31]. Tahapan ini penting untuk menyamakan kata-kata seperti ga, enggak, ngga, dan gak yang memiliki makna sama, agar tidak dianggap berbeda oleh sistem. Dengan normalisasi, konsistensi data dapat terjaga sehingga mendukung analisis yang lebih akurat.

Table 5. Normalisasi

Tokenisasi	Normalisasi
[mantap, tanpa, biaya, admin, terbaik]	[mantap, tanpa, biaya, admin, terbaik]
[error, terus, aplikasi, nya, ga, bs, di, buka]	[error, terus, aplikasi, nya, enggak, bisa, di, buka]

Proses normalisasi pada *Table 5* mengubah kata singkatan seperti “ga” dan “bs” menjadi “enggak” dan “bisa” sesuai bentuk normal bahasa Indonesia.

D. Exploratory Data Analysis (EDA)

Sebagai bagian dari tahapan eksplorasi data, dilakukan analisis visual menggunakan *WordCloud* untuk masing-masing kategori sentimen, yaitu negatif, netral, dan positif. *WordCloud* merupakan metode analisis deskriptif yang berfungsi menampilkan kata-kata yang paling sering muncul dalam kumpulan teks secara visual. Teknik ini memanfaatkan *Term Document Matrix* untuk mengubah frekuensi kata menjadi ukuran visual yang menarik dan informatif [32]. *WordCloud* dari ulasan positif ditampilkan pada Fig. 4, negatif pada Fig. 5, dan netral pada Fig. 6.



Fig. 4. WordCloud Sentimen Positif

Sentimen positif banyak mengandung kata-kata seperti “sangat”, “membantu”, “mantap”, “bagus”, “terima kasih”, dan “hemat”. Hal ini menunjukkan bahwa pengguna merasa puas terhadap manfaat aplikasi, khususnya terkait efisiensi biaya dan kemudahan transaksi antarbank.



Fig. 5. WordCloud Sentimen Negatif

Sentimen negatif didominasi oleh kata-kata seperti “saya”, “sudah”, “enggak”, “bisa”, “aplikasi”, dan “transaksi”. Kata-kata ini mengindikasikan keluhan atau ketidakpuasan pengguna, terutama terkait kegagalan dalam transaksi dan kendala teknis pada aplikasi.

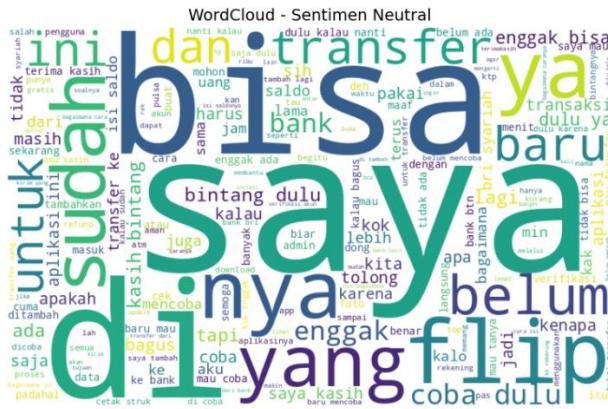


Fig. 6. WordCloud Sentimen Netral

Sentimen netral memuat kata-kata seperti “bisa”, “transfer”, “flip”, “saldo”, dan “bank”. Kata-kata ini cenderung bersifat informatif atau deskriptif tanpa ekspresi emosi yang kuat, sehingga tidak mencerminkan keberpihakan pengguna terhadap layanan.

E. Data Splitting

Data splitting atau pembagian *dataset* merupakan proses penting untuk memastikan kinerja model yang optimal. Tahapan ini bertujuan untuk mencegah *overfitting* dan memvalidasi kemampuan generalisasi model terhadap data yang belum pernah dilihat sebelumnya [33].

Table 6. Hasil Pembagian Data

<i>Split Data</i>	Total
Data <i>Training</i>	1963
Data <i>Validation</i>	564
Data <i>Test</i>	278

Pada *Table 6*, penelitian ini membagi data menjadi tiga proporsi, yaitu data *training* untuk pelatihan dan *fine-tuning* model, data *validation* digunakan untuk mengevaluasi kinerja model

selama proses pelatihan dan mencegah *overfitting*, dan data *testing* untuk mengevaluasi kinerja akhir model setelah proses pelatihan selesai. Dataset dibagi dengan proporsi 70% sebagai *train set*, 20% sebagai *validation set*, dan 10% sebagai *test set*.

F. IndoBERT Tokenization

Tokenisasi merupakan tahap penting dalam pemrosesan data teks sebelum dilakukan pelatihan model berbasis *transformer* seperti IndoBERT. Pada penelitian ini, digunakan IndoBERT *tokenizer* yang didasarkan pada arsitektur BERT dan telah disesuaikan untuk memahami struktur dan kosakata bahasa Indonesia [34]. Tokenisasi bertujuan untuk memecah kalimat menjadi potongan kata (*tokens*) menggunakan metode *WordPiece*. Setiap token kemudian dikonversi menjadi representasi numerik (token ID) yang sesuai dengan kosakata model. Token khusus seperti [CLS] dan [SEP] ditambahkan untuk menandai awal dan akhir kalimat, sementara [PAD] digunakan untuk menyamakan panjang input [35]. Tokenisasi ini memastikan setiap teks terstruktur sesuai format yang dapat dikenali oleh model BERT sehingga dapat dilanjutkan ke tahap *encoding* dan pelatihan model klasifikasi sentimen. Berikut adalah ilustrasi dari IndoBERT *tokenization*.

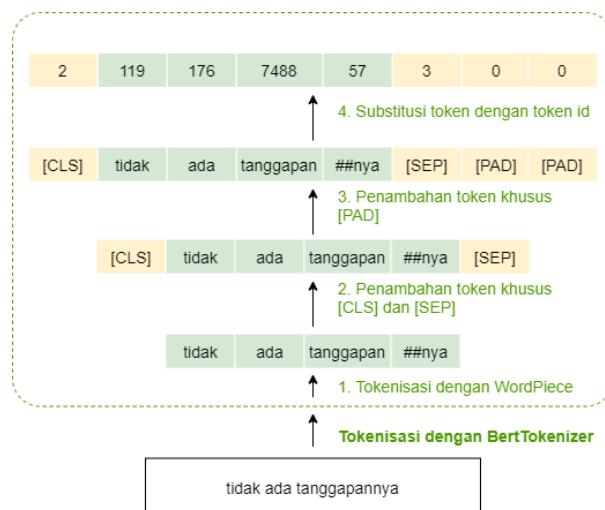


Fig. 7. IndoBERT Tokenization

G. IndoBERT Modeling

Dalam tahap pemodelan menggunakan IndoBERT, penelitian ini menggunakan Indobert-base-p1 dan Indobert-base-p2, dua varian *pre-trained* model IndoBERT khusus bahasa Indonesia yang dikembangkan dengan menggunakan arsitektur Bidirectional Encoder Representations from Transformers (BERT). Pada proses pelatihan model menggunakan pendekatan *trial and error* atau mencoba berbagai konfigurasi secara berulang untuk menemukan kombinasi parameter yang paling optimal [36]. Setelah menemukan kombinasi parameter, dilakukan *fine-tuning* pada model *pre-trained* IndoBERT. *Fine tuning* merupakan proses penyesuaian lebih lanjut dari model *pre-trained* dengan tujuan meningkatkan performa model sehingga model mampu memberikan hasil prediksi yang lebih akurat dalam analisis sentimen [37]. Berikut adalah parameter yang digunakan *fine-tuning* model IndoBERT.

Table 7. Training Model Parameter

Parameter	IndoBERT-base-p1	IndoBERT-base-p2
Batch size	32	32
Optimizer	Adam	Adam
Learning rate	3e-6(0.000003)	3e-6(0.000003)

Pada *Table 7*, terdapat hasil kombinasi parameter yang diterapkan pada dua varian *pre-trained* yaitu IndoBERT-base-p1 dan IndoBERT-base-p2 dengan kombinasi parameter yang setara di antara keduanya.

H. Model Evaluation

Penelitian ini menggunakan beberapa metrik evaluasi, yaitu *precision*, *recall*, F1-score, *accuracy*, *macro average*, dan *weighted average* [38][39]. *Precision* mengukur sejauh mana prediksi positif yang dihasilkan oleh model benar-benar relevan, sedangkan *recall* menilai kemampuan model dalam menemukan seluruh kasus positif yang ada. Selanjutnya, F1-score menggabungkan kedua aspek performa tersebut *precision* dan *recall* ke dalam satu nilai agregat untuk memberikan gambaran yang lebih komprehensif terhadap kinerja model. *Accuracy* mengukur jumlah total prediksi yang benar dibandingkan dengan seluruh prediksi yang dilakukan terhadap satu himpunan data.

Macro Average, sesuai dengan namanya, menghitung metrik seperti *precision*, *recall*, F1-score, dan *accuracy* secara independen untuk setiap kelas. Pendekatan ini tidak memperhitungkan ketidakseimbangan distribusi kelas dan memberikan bobot yang sama untuk setiap kelas [40]. Rata-rata *macro* dihitung dengan mengambil nilai rata-rata tidak berbobot dari metrik tiap kelas tersebut. Pendekatan ini memberikan evaluasi yang tidak bias terhadap kinerja model, dengan memperlakukan setiap kelas secara setara sehingga dapat mengurangi bias akibat ketimpangan jumlah data antar kelas.

Sebaliknya, *Weighted Average* dirancang khusus untuk menangani masalah ketidakseimbangan kelas dengan memberikan bobot yang berbeda pada setiap kelas secara proporsional terhadap jumlah sampel di kelas tersebut. Dengan demikian, metrik ini memberikan representasi kinerja model yang lebih akurat karena memberi penekanan lebih besar pada kelas-kelas yang memiliki jumlah data lebih banyak. Metrik seperti *Weighted Precision*, *Weighted Recall*, *Weighted F1-Score*, dan *Weighted Accuracy* dihitung berdasarkan kontribusi bobot tersebut. *Weighted Average* menjadi pilihan yang lebih praktis dalam skenario di mana ketidakseimbangan kelas merupakan masalah signifikan, karena memberikan cerminan nyata atas kegunaan model dalam aplikasi dunia nyata [41][42].

Table 8. *Confusion Matrix*

Confusion Matrix		Actual Classes	
		Yes	No
Predicted Classes	Yes	True Positive (TP)	False Positive (FP)
	No	False Negative (FN)	True Negative (TN)

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \quad (1)$$

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

$$F1 - Score = \frac{2 \times precision \times recall}{precision + recall} \quad (4)$$

IV. Hasil dan Pembahasan

A. Performa Model

Evaluasi performa model dilakukan dengan menggunakan kurva pembelajaran (*learning curve*), yang membandingkan akurasi antara data pelatihan (*training accuracy*) dan data validasi (*validation accuracy*) selama proses pelatihan. Kurva ini berfungsi untuk memantau dan memahami sejauh mana model mampu belajar dari data serta mendeteksi potensi masalah seperti *underfitting* atau *overfitting* selama *training* [43].

Fig. 8 dan Fig. 9 berikut menampilkan kurva pembelajaran berdasarkan metrik akurasi untuk dua model yang digunakan dalam penelitian ini, yaitu IndoBERT-base-p1 dan IndoBERT-base-p2 yang keduanya merupakan model *pretrained* berbasis arsitektur BERT untuk bahasa Indonesia. Konfigurasi pelatihan untuk kedua model ini telah disesuaikan sebagaimana dijelaskan pada *Table 7*.

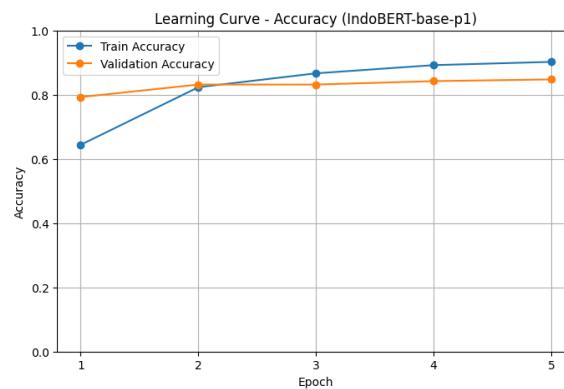


Fig. 8. Learning Curve IndoBERT-base-p1

Pada Fig. 8, model IndoBERT-base-p1 menunjukkan peningkatan akurasi pelatihan yang konsisten dari *epoch* ke-1 hingga *epoch* ke-5, dengan nilai yang hampir mencapai 90%. Sementara itu, akurasi validasi juga meningkat secara stabil hingga *epoch* ke-3 dan kemudian cenderung stagnan, menandakan bahwa model telah mulai mencapai kestabilan performa tanpa indikasi *overfitting* yang signifikan.

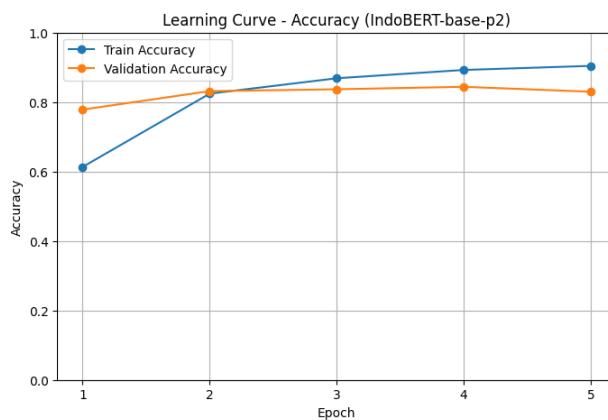


Fig. 9. Learning Curve IndoBERT-base-p2

Sebaliknya, pada Fig. 9, model IndoBERT-base-p2 memperlihatkan pola peningkatan yang serupa pada akurasi pelatihan. Namun, akurasi validasi justru menurun sedikit pada *epoch* ke-5 setelah mencapai puncaknya di *epoch* ke-4. Hal ini dapat mengindikasikan potensi *overfitting* ringan, di mana model terlalu menyesuaikan terhadap data latih dan sedikit menurun performanya saat diuji terhadap data yang belum pernah dilihat.

B. Evaluasi Model

Hasil evaluasi model menggunakan sejumlah metrik evaluasi yang relevan. Metrik tersebut mencakup *confusion matrix*, *precision*, *recall*, F1-score, *accuracy*, *macro average* dan *weighted average* yang bertujuan untuk memberikan gambaran komprehensif terhadap performa klasifikasi sentimen model.

1) Hasil Evaluasi IndoBERT-base-p1

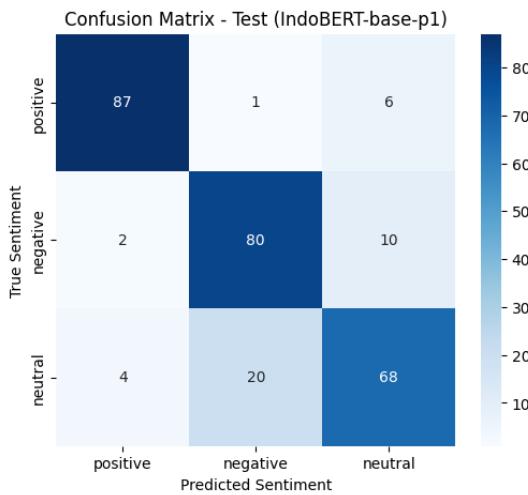


Fig. 10. *Confusion Matrix* IndoBERT-base-p1

Pada Fig.10 merupakan *confusion matrix* dari hasil pengujian model dengan *pre-trained* IndoBERT-base-p1 pada data *testing*. Model berhasil mengklasifikasikan dengan benar 87 dari total 94 data berlabel positif, serta 80 dari 92 data berlabel negatif. Untuk kelas netral, model memprediksi secara akurat sebanyak 68 dari 92 data. Hasil ini menunjukkan bahwa model mampu mempertahankan performa klasifikasi yang stabil pada data yang belum pernah dilihat sebelumnya.

Namun demikian, masih ditemukan beberapa kesalahan klasifikasi, terutama antara kelas netral dan negatif, di mana terdapat 20 data netral yang salah diklasifikasikan sebagai negatif, serta 10 data negatif yang diprediksi sebagai netral. Selain itu, terdapat sedikit kesalahan prediksi pada kelas positif ke netral (6 data). Temuan ini kembali mengindikasikan bahwa model masih menghadapi tantangan dalam membedakan konteks yang bersifat netral dan negatif, khususnya dalam kalimat-kalimat yang bermuansa samar atau tidak eksplisit menyampaikan sentimen.

Table 9. Laporan Klasifikasi IndoBERT-base-p1

	Precision	Recall	F1-Score	Support
Positif	0.94	0.93	0.93	94
Negatif	0.79	0.87	0.83	92
Netral	0.81	0.74	0.77	92
<i>Accuracy</i>			0.85	278
<i>Macro avg</i>	0.85	0.84	0.84	278
<i>Weighted avg</i>	0.85	0.85	0.84	278

Table 9 menunjukkan hasil laporan klasifikasi model IndoBERT dengan *pretrained* IndoBERT-base-p1 menggunakan metrik evaluasi berupa *precision*, *recall*, F1-score, dan *support* untuk setiap kelas. Model menunjukkan performa yang cukup stabil dengan nilai akurasi keseluruhan sebesar 85%. Kelas positif memperoleh F1-score tertinggi yaitu 0.93 yang menunjukkan kemampuan model dalam mengklasifikasikan sentimen positif dengan sangat baik.

Sementara itu, kelas negatif memperoleh F1-score sebesar 0.83, dan kelas netral memiliki nilai F1-score terendah yaitu 0.77 yang mengindikasikan bahwa model masih memiliki kesulitan dalam membedakan ekspresi netral secara konsisten. Nilai rata-rata makro (*macro avg*) dan rata-rata tertimbang (*weighted avg*) masing-masing berada di angka 0.84 yang mendukung kesimpulan bahwa model memiliki performa klasifikasi yang seimbang namun tetap dapat ditingkatkan, khususnya pada kelas yang lebih ambigu seperti netral.

2) Hasil Evaluasi IndoBERT-base-p2

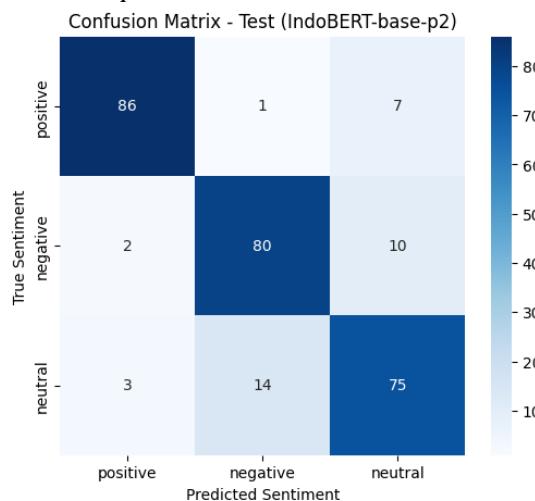


Fig. 11. *Confusion Matrix* IndoBERT-base-p2

Pada Fig. 11 merupakan *confusion matrix* dari hasil pengujian model dengan *pre-trained* IndoBERT-base-p2 pada data *testing*. Model berhasil mengklasifikasikan dengan benar 86 dari total 94 data berlabel positif, serta 80 dari 92 data berlabel negatif. Untuk kelas netral, model mampu memberikan prediksi yang akurat sebanyak 75 dari 92 data. Hasil ini mengindikasikan bahwa model IndoBERT-base-p2 tidak hanya stabil tetapi juga sedikit lebih unggul dibandingkan P1 dalam hal generalisasi terhadap data baru.

Meskipun demikian, masih terdapat beberapa kesalahan klasifikasi. Sebanyak 14 data netral salah diprediksi sebagai negatif, dan 10 data negatif diklasifikasikan sebagai netral. Selain itu, terdapat 7 data positif yang diprediksi sebagai netral, meskipun proporsinya relatif kecil. Kesalahan ini menunjukkan bahwa model masih menghadapi kesulitan dalam membedakan antara sentimen netral dan negatif yang sering kali memiliki konteks kalimat yang serupa atau implisit, namun performanya tetap menunjukkan peningkatan dibandingkan model sebelumnya.

Table 10. Laporan Klasifikasi IndoBERT-base-p2

	Precision	Recall	F1-Score	Support
Positif	0.95	0.91	0.93	94
Negatif	0.84	0.87	0.86	92

Netral	0.82	0.82	0.82	92
Accuracy			0.87	278
Macro avg	0.87	0.87	0.87	278
Weighted avg	0.87	0.87	0.87	278

Table 10 menunjukkan hasil laporan klasifikasi model IndoBERT dengan *pretrained* IndoBERT-base-p2 menggunakan metrik evaluasi berupa *precision*, *recall*, F1-score, dan *support* untuk masing-masing kelas. Model menunjukkan performa yang sangat baik dengan akurasi keseluruhan sebesar 87%. Kelas positif mencatat F1-score tertinggi yaitu 0.93 yang menunjukkan bahwa model sangat efektif dalam mengenali ulasan dengan sentimen positif.

Kelas negatif memiliki F1-score sebesar 0.86, sedangkan kelas netral memperoleh F1-score 0.82. Meskipun nilai F1 pada kelas netral lebih rendah dibanding dua kelas lainnya, namun masih dalam kategori baik dan konsisten. Nilai *macro average* dan *weighted average* dari seluruh metrik berada di angka 0.87 yang mengindikasikan bahwa model ini tidak hanya akurat secara keseluruhan, tetapi juga mampu menjaga keseimbangan performa antarkelas, termasuk dalam kondisi distribusi kelas yang tidak seimbang.

3) Perbandingan Kinerja Model

Table 11 dan Fig. 12 berikut menyajikan perbandingan metrik performa antara kedua model:

Table 11. Perbandingan Kinerja Model IndoBERT-base-p1 dan IndoBERT-base-p2

Metrik	P1 Validasi	P1 Test	P2 Validasi	P2 Test
Accuracy	0.85	0.85	0.83	0.87
F1-Score (Positif)	0.93	0.93	0.92	0.93
F1-Score (Negatif)	0.82	0.83	0.80	0.86
F1-Score (Netral)	0.78	0.77	0.76	0.82
Macro avg F1	0.85	0.84	0.83	0.87

Perbandingan Metrik Evaluasi IndoBERT-base-p1 vs IndoBERT-base-p2

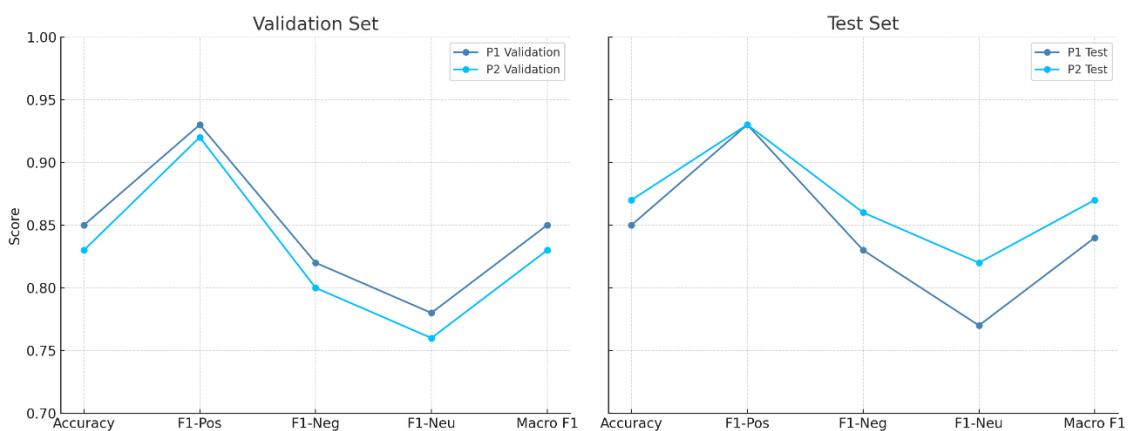


Fig. 12. Perbandingan Metrik Evaluasi IndoBERT-base-p1 dan IndoBERT-base-p2

Secara umum, model **IndoBERT-base-p2** menunjukkan performa yang **lebih unggul** pada data uji dibandingkan IndoBERT-base-p1, meskipun pencapaiannya pada data validasi sedikit lebih rendah. Temuan ini mengindikasikan bahwa model IndoBERT-base-p2 memiliki **kemampuan generalisasi yang lebih baik**, sehingga lebih dapat diandalkan ketika diterapkan pada data baru yang belum pernah dilihat sebelumnya.

V. Kesimpulan

Penelitian ini membandingkan kinerja dua varian model *pretrained* IndoBERT, yaitu IndoBERT-base-p1 dan IndoBERT-base-p2, dalam klasifikasi sentimen ulasan berbahasa Indonesia pada aplikasi Flip. Model dilatih menggunakan *optimizer* Adam dengan *learning rate* sebesar 3e-6 (0.000003), *batch size* 32, dan dilatih selama 5 *epoch*. Dataset dibagi dengan rasio 70% untuk data *training*, 20% untuk data *validation*, dan 10% untuk data *testing*.

Hasil evaluasi menunjukkan bahwa model IndoBERT-base-p2 secara umum memiliki performa lebih baik, khususnya pada data uji. Model IndoBERT-base-p2 mencapai akurasi 87% pada data uji, sedangkan IndoBERT-base-p1 mencapai 85%. Selain itu, pada metrik lain seperti *precision*, *recall*, dan F1-score, model IndoBERT-base-p2 juga menunjukkan performa yang lebih konsisten dan unggul, terutama dalam mengklasifikasikan sentimen negatif dan netral yang umumnya lebih sulit dibedakan. Meskipun performa validasi IndoBERT-base-p2 sedikit lebih rendah dibandingkan IndoBERT-base-p1, hasil pengujian memperlihatkan bahwa IndoBERT-base-p2 memiliki kemampuan generalisasi yang lebih kuat terhadap data baru. Berdasarkan analisis tersebut, dapat disimpulkan bahwa penggunaan model IndoBERT dengan varian IndoBERT-base-p2 lebih efektif dalam mencapai akurasi dan kinerja yang optimal untuk tugas analisis sentimen, khususnya pada ulasan aplikasi berbahasa Indonesia.

VI. Saran

Berdasarkan hasil penelitian, terdapat beberapa saran untuk pengembangan studi di masa mendatang. Penelitian selanjutnya disarankan untuk mengeksplorasi konfigurasi *fine-tuning* yang lebih dalam, seperti penyesuaian jumlah *epoch*, penggunaan *learning rate scheduler*, serta penerapan teknik regularisasi tambahan untuk meningkatkan performa model. Selain itu, perlu perhatian khusus terhadap peningkatan akurasi pada kelas netral yang dalam penelitian ini cenderung lebih rendah, misalnya melalui penggunaan teknik *class weighting*, *data augmentation*, atau penerapan *focal loss*. Penelitian berikutnya juga dapat membandingkan performa model dengan arsitektur transformer lain yang lebih ringan atau lebih baru, seperti IndoBERT-lite, IndoBART, atau XLM-R. Penggunaan *dataset* yang lebih besar, lebih beragam, atau berasal dari berbagai platform ulasan lain juga diharapkan dapat meningkatkan generalisasi model. Selain itu, evaluasi kinerja model sebaiknya tidak hanya mengandalkan metrik tradisional, tetapi juga mempertimbangkan metrik tambahan seperti ROC-AUC, PR-AUC, dan analisis *confidence interval* untuk memberikan gambaran yang lebih komprehensif mengenai performa dan ketahanan model.

References

- [1] U. Bina, S. Informatika, F. Ekonomi, D. Bisnis, and C. Sitasi, “Analisis SWOT Technology Financial (FinTech) Terhadap Industri Perbankan. Cakrawala,” vol. 19, no. 1, pp. 55–60, 2019, doi: 10.31294/jc.v19i1.
- [2] K. Solecha and O. Irnawati, “Komparasi Algoritma Support Vector Machine Dan Naïve Bayes Berbasis Particle Swarm Optimization Pada Analisis Sentimen Ulasan Aplikasi Flip,” 2023.I.S. Jacobs and C.P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G.T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.
- [3] S. A. Helmayanti, F. Hamami, and R. Y. Fa’rifah, “PENERAPAN ALGORITMA TF-IDF DAN NAÏVE BAYES UNTUK ANALISIS SENTIMEN BERBASIS ASPEK ULASAN APLIKASI FLIP PADA GOOGLE PLAY STORE,” *Jurnal Indonesia : Manajemen Informatika dan Komunikasi*, vol. 4, no. 3, pp. 1822–1834, Sep. 2023, doi: 10.35870/jimik.v4i3.415.R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.
- [4] Suwanda Aditya Saputra, Didi Rosiyadi, Windu Gata, and Syepry Maulana Husain, “Analisis Sentimen E-Wallet Pada Google Play Menggunakan Algoritma Naïve Bayes Berbasis Particle Swarm

- Optimization," *masa berlaku mulai*, vol. 1, no. 3, pp. 377–382, 2017.M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.
- [5] N. Fitriyah, B. Warsito, D. Asih, and I. Maruddani, "ANALISIS SENTIMEN GOJEK PADA MEDIA SOSIAL TWITTER DENGAN KLASIFIKASI SUPPORT VECTOR MACHINE (SVM)," *JURNAL GAUSSIAN*, vol. 9, no. 3, pp. 376–390, 2020, [Online]. Available: <https://ejournal3.undip.ac.id/index.php/gaussian>
 - [6] M. Isnan, G. N. Elwirehardja, and B. Pardamean, "Sentiment Analysis for TikTok Review Using VADER Sentiment and SVM Model," *Procedia Comput. Sci.*, vol. 227, pp. 168–175, 2023, doi: 10.1016/j.procs.2023.10.514.
 - [7] M. E. Purbaya, D. Putra Rakhmadani, M. Puspa Arum, and L. Zian Nasifah, "Comparison of Kernel Support Vector Machines in Conducting Sentiment Analysis Review of Buying Chips on the Shopee E-Marketplace in Indonesian," in 2022 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS), IEEE, Nov. 2022, pp. 435–440. doi: 10.1109/ICIMCIS56303.2022.10017546.
 - [8] Z. A. Diekson, M. R. B. Prakoso, M. S. Q. Putra, M. S. A. F. Syaputra, S. Achmad, and R. Sutoyo, "Sentiment analysis for customer review: Case study of Traveloka," *Procedia Comput. Sci.*, vol. 216, pp. 682–690, 2023, doi: 10.1016/j.procs.2022.12.184.
 - [9] T. Willianto, Supryadi, and A. Wibowo, "Sentiment Analysis on E-commerce Product using Machine Learning and Combination of TF-IDF and Backward Elimination," *Int. J. Recent Technol. Eng.*, vol. 8, no. 6, pp. 2862–2867, 10.35940/ijrte.F7889.038620. Mar. 2020, doi: 10.35940/ijrte.F7889.038620.
 - [10] M. J. Hossain, D. Das Joy, S. Das, and R. Mustafa, "Sentiment Analysis on Reviews of E-commerce Sites Using Machine Learning Algorithms," in 2022 International Conference on Innovations in Science, Engineering and Technology (ICISET), IEEE, Feb. 2022, pp. 522–527. 10.1109/ICISET54810.2022.9775846.
 - [11] S. Rahayu, Y. Mz, J. E. Bororing, and R. Hadiyat, "Implementasi Metode K-Nearest Neighbor (K-NN) untuk Analisis Sentimen Kepuasan Pengguna Aplikasi Teknologi Finansial FLIP," *Edumatic: Jurnal Pendidikan Informatika*, vol. 6, no. 1, pp. 98–106, 2022.
 - [12] O. Irnawati and K. Solecha, "Komparasi Algoritma Support Vector Machine dan Naïve Bayes Berbasis Particle Swarm Optimization pada Analisis Sentimen Ulasan Aplikasi Flip," *JIEET (Journal of Information Engineering and Educational Technology)*, vol. 7, no. 1, pp. 10–15, 2023.
 - [13] B. Wilie *et al.*, "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," 2020. [Online]. Available: <https://github.com/annisanurulazhar/absa-playground>
 - [14] M. A. Hadiwijaya, F. P. Pirdaus, D. Andrews, S. Achmad, and R. Sutoyo, "Sentiment Analysis on Tokopedia Product Reviews using Natural Language Processing," in 2023 International Conference on Informatics, Multimedia, Cyber and Informations System (ICIMCIS), IEEE, Nov. 2023, pp. 380–386. doi: 10.1109/ICIMCIS60089.2023.10348996.
 - [15] H. Jayadianti, W. Kaswidjanti, A. T. Utomo, S. Saifullah, F. A. Dwiyanto, and R. Drezewski, "Sentiment analysis of Indonesian reviews using fine tuning IndoBERT and R-CNN," *Ilk. J. Ilm.*, vol. 14, no. 3, pp. 348–354, Dec. 10.33096/ilkom.v14i3.1505.348–354.
 - [16] W. M. Baihaqi and A. Munandar, "Sentiment Analysis of Student Comment on the College Performance Evaluation Questionnaire Using Naïve Bayes and IndoBERT," *JUITA J. Inform.*, vol. 11, no. 2, p. 213, Nov. 2023, doi: 10.30595/juita.v11i2.17336.
 - [17] B. Kurniawan, A. A. Aldino, and A. R. Isnain, "Sentimen Analisis Terhadap Kebijakan Penyelenggara Sistem Elektronik (PSE) Menggunakan Algoritma Bidirectional Encoder Representations From Transformers (BERT)," *Jurnal Teknologi dan Sistem Informasi*, vol. 3, pp. 98–106, 2022.
 - [18] M. Khader, A. Awajan, and G. Al-Naymat, "The Effects of Natural Language Processing on Big Data Analysis: Sentiment Analysis Case Study," in 2018 International Arab Conference on Information Technology (ACIT), IEEE, Nov. 2018, pp. 1–7. doi: 10.1109/ACIT.2018.8672697.
 - [19] N. M. S. Hadna, P. I. Santosa, and W. W. Winarno, "Studi Literatur tentang Perbandingan Metode untuk Proses Analisis Sentimen di Twitter," *Seminar Nasional Teknologi Informasi dan Komunikasi*, pp. 57–64, 2016.
 - [20] I. P. Cvijikj and F. Michahelles, "Understanding Social Media Marketing: A Case Study on Topics, Categories and Sentiment on a Facebook Brand Page," in *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments*, pp. 175–182, 2011.
 - [21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *Proc. 2019 Conf. N. Am. Chapter Assoc. Comput. Linguist., Hum. Lang. Technol.*, 2019, hal. 4171–4186, doi: 10.18653/v1/N19-1423.

- [22] F. Koto, A. Rahimi, J.H. Lau, dan T. Baldwin, "IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP," Proc. 28th Int. Conf. Comput. Linguist., 2020, hal. 757--770, doi: 10.18653/v1/2020.coling-main.66.
- [23] S.M. Isa, G. Nico, dan M. Permana, "IndoBERT for Indonesian fake news detection," ICIC Express Lett., vol. 16, no. 3, hal. 289-297, Mar. 2022, doi: 10.24507/icicel.16.03.289.
- [24] E. Zaaputra, "Sentiment Analysis Using BERT," GitHub, [Online]. Available: https://github.com/ezaaputra/Sentiment-Analysis-Using-BERT/blob/main/flip_cat_balance.tsv. [Accessed: 21-Apr-2025].
- [25] Anugerah Simanjuntak *et al.*, "Research and Analysis of IndoBERT Hyperparameter Tuning in Fake News Detection," *Jurnal Nasional Teknik Elektro dan Teknologi Informasi*, vol. 13, no. 1, pp. 60–67, Feb. 2024, doi: 10.22146/jnteti.v13i1.8532.
- [26] P. Kaur, "Sentiment analysis using web scraping for live news data with machine learning algorithms," Mater. Today Proc., vol. 65, pp. 3333–3341, 2022, doi: 10.1016/j.matpr.2022.05.409.
- [27] S. G. C. G and B. S. -, "Grid Search Tuning of Hyperparameters in Random Forest Classifier for Customer Feedback Sentiment Prediction," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 9, 2020, doi: 10.14569/IJACSA.2020.0110920.
- [28] C. Slamet, A. R. Atmadja, D. S. Maylawati, R. S. Lestari, W. Darmalaksana, and M. A. Ramdhani, "Automated Text Summarization for Indonesian Article Using Vector Space Model," IOP Conf. Ser. Mater. Sci. Eng., vol. 288, p. 012037, Jan. 2018, doi: 10.1088/1757-899X/288/1/012037.
- [29] B. Kurniawan, A. A. Aldino, and A. R. Isnain, "Sentimen Analisis Terhadap Kebijakan Penyelenggara Sistem Elektronik (PSE) Menggunakan Algoritma Bidirectional Encoder Representations From Transformers (BERT)," *Jurnal Teknologi dan Sistem Informasi*, vol. 3, pp. 98–106, 2022.
- [30] N. M. S. Hadna, P. I. Santosa, and W. W. Winarno, "Studi Literatur tentang Perbandingan Metode untuk Proses Analisis Sentimen di Twitter," *Seminar Nasional Teknologi Informasi dan Komunikasi*, pp. 57–64, 2016.
- [31] F. Hemmatian and M. K. Sohrabi, "A survey on classification techniques for opinion mining and sentiment analysis," Artif. Intell. Rev., vol. 52, no. 3, pp. 1495–1545, Oct. 2019, doi: 10.1007/s10462-017-9599-6.
- [32] G. Pradana, "Penggunaan Fitur Wordcloud dan Document Term Matrix dalam Text Mining," J. Ilm. Inform., vol. 8, no. 1, pp. 38–43, 2020.
- [33] Y. Xu and R. Goodacre, "On Splitting Training and Validation Set: A Comparative Study of Cross Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning," J. Anal. Test., vol. 2, no. 3, pp. 249–262, Jul. 2018, doi: 10.1007/s41664-018-0068-2.
- [34] J. H. Computer, S. M. Honova, V. P. Computer, C. A. Setiawan, I. H. Parmonangan, and Diana, "Sentiment Analysis of Skincare Product Reviews in Indonesian Language using IndoBERT and LSTM," in 2023 IEEE 9th Information Technology International Seminar (ITIS), IEEE, Oct. 2023, pp. 1–6. doi: 10.1109/ITIS59651.2023.10420222.
- [35] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf., vol. 1, no. Mlm, pp. 4171–4186, 2019.
- [36] H. D. Sharma and P. Goyal, "An Analysis of Sentiment: Methods, Applications, and Challenges," in RAiSE-2023, Basel Switzerland: MDPI, Dec. 2023, p. 68. doi: 10.3390/engproc2023059068.
- [37] K. S. Nugroho, A. Y. Sukmadewa, H. W. DW, F. A. Bachtiar, and N. Yudistira, "BERT Fine-Tuning for Sentiment Analysis on Indonesian Mobile Apps Reviews," Jul. 2021, doi: 10.1145/3479645.3479679.
- [38] M. Hossin and M. N. Sulaiman, "A Review on Evaluation Metrics for Data Classification Evaluations," *International Journal of Data Mining & Knowledge Management Process (IJDKP)*, vol. 5, no. 2, pp. 1–11, 2015.
- [39] Ž. Vujović, "Classification Model Evaluation Metrics," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 6, pp. 599–606, 2021. doi: 10.14569/IJACSA.2021.0120670.
- [40] K. Takahashi, K. Yamamoto, A. Kuchiba, and T. Koyama, "Confidence interval for micro-averaged F1 and macro-averaged F1 scores," *Applied Intelligence*, vol. 52, no. 5, pp. 4961–4972, Mar. 2022.
- [41] D. J. Hand, P. Christen, and N. Kirielle, "F*: an interpretable transformation of the F-measure," *Machine Learning*, vol. 110, no. 3, pp. 451–456, Mar. 2021.
- [42] M. Grandini, E. Bagli, and G. Visani, "Metrics for multi-class classification: an overview," *arXiv preprint*, arXiv:2008.05756, Aug. 2020.
- [43] M. Totox and H. F. Pardede, "Exploring the Effectiveness of Deep Learning in Analyzing Review Sentiment," *JIKO (Jurnal Inform. dan Komputer)*, vol. 6, no. 2, Aug. 2023, doi: 10.33387/jiko.v6i2.6372.

Lampiran

Dalam penelitian ini, proses training dan evaluasi model dilakukan menggunakan Google Colaboratory. Proses ini mencakup tahap preprocessing data, tokenisasi menggunakan IndoBERT tokenizer, pelatihan model IndoBERT-base-p1 dan IndoBERT-base-p2, serta evaluasi model menggunakan metrik akurasi, presisi, recall, dan F1-score.

Berikut ini adalah link menuju file Google Colab yang digunakan dalam penelitian:

Analisis Sentimen Flip menggunakan IndoBERT