# Deteksi Ujaran Kebencian Bahasa Indonesia Menggunakan Fine-Tuning IndoBERT Berbasis Transformer

**Meiskey Rambu Dini Naomi[1], Minarwati*[2]**
*[1,2,] Sistem Informasi, STMIK EL RAHMA Yogyakarta*
*e-mail: [1]meiskeynaomi20@gmail.com, [2] minarwati@stmikelrahma.ac.id*
**Correspondence author*

***Abstrak***

*Ujaran kebencian pada media sosial berbahasa Indonesia terus meningkat sehingga diperlukan sistem deteksi otomatis yang mampu melakukan klasifikasi teks secara akurat. Penelitian ini mengusulkan deteksi ujaran kebencian menggunakan model transformer pre-trained IndoBERT melalui proses fine-tuning. Berbeda dengan penelitian sebelumnya yang mengombinasikan transformer dengan arsitektur tambahan seperti CNN atau BiLSTM, penelitian ini mengevaluasi kemampuan IndoBERT secara langsung tanpa lapisan deep learning tambahan. Dataset yang digunakan terdiri dari 13.169 twit berbahasa Indonesia yang telah dianotasi ke dalam kategori Ujaran Kebencian (UK) dan Non-Ujaran Kebencian (Non-UK). Dataset dibagi menjadi 80% data pelatihan, 10% data validasi, dan 10% data pengujian. Model dilatih selama 3 epoch menggunakan optimizer AdamW dengan learning rate 2e-5 dan batch size 8. Hasil pengujian menunjukkan bahwa model IndoBERT mencapai accuracy sebesar 90,43%, precision sebesar 90,11%, recall sebesar 90,36%, dan F1-Score sebesar 90,23%. Hasil tersebut menunjukkan bahwa fine-tuning langsung pada IndoBERT mampu memberikan kinerja yang baik dengan arsitektur yang lebih sederhana dan kinerja yang baik. Penelitian ini menunjukkan bahwa model transformer berbasis bahasa Indonesia efektif untuk mendukung moderasi konten digital secara otomatis.*

***Kata Kunci****—Ujaran Kebencian; Bahasa Indonesia; IndoBERT; Transformer; Fine-tuning.*

## 1. PENDAHULUAN

Ujaran kebencian (hate speech) merupakan bentuk komunikasi yang mengandung penghinaan, permusuhan, diskriminasi, atau hasutan terhadap individu maupun kelompok berdasarkan identitas tertentu, seperti suku, agama, ras, dan antargolongan (SARA). Fenomena ini menjadi salah satu permasalahan dalam ruang digital karena penyebarannya dapat berlangsung secara cepat melalui media social[1], [2]

Perkembangan teknologi informasi dan meningkatnya penggunaan media sosial memungkinkan informasi dan konten negatif disebarkan secara real-time serta menjangkau banyak pengguna dalam waktu singkat[3], [4] .

Karakteristik teks media sosial berbahasa Indonesia menimbulkan tantangan tersendiri dalam proses deteksi ujaran kebencian. Penggunaan bahasa tidak baku, singkatan, variasi ejaan, campuran bahasa, tagar, mention, serta ekspresi yang bersifat implisit menyebabkan suatu teks tidak selalu dapat diklasifikasikan hanya berdasarkan keberadaan kata-kata ofensif[5]. Ujaran kebencian juga dapat disampaikan secara tidak langsung melalui sindiran, stereotip, atau konteks tertentu sehingga pendekatan berbasis aturan memiliki keterbatasan dalam memahami hubungan semantik antarkata [6].

Penelitian terdahulu telah menggunakan berbagai pendekatan untuk melakukan deteksi ujaran kebencian. Pendekatan machine learning seperti Multinomial Naïve Bayes dan Multinomial Logistic Regression memanfaatkan karakteristik tekstual untuk melakukan klasifikasi [7], [8]. Pendekatan deep learning seperti LSTM dan BiGRU kemudian digunakan untuk meningkatkan kemampuan pemodelan konteks teks [9], [10], [11]. Perkembangan Transformer selanjutnya menghasilkan model BERT yang mampu membangun representasi kontekstual dengan memanfaatkan hubungan antarkata secara bidireksional [12], [13]. Untuk

bahasa Indonesia, IndoBERT dikembangkan menggunakan korpus bahasa Indonesia sehingga dapat digunakan untuk berbagai tugas pemrosesan bahasa alami[14].

Meskipun model berbasis Transformer telah digunakan dalam penelitian deteksi ujaran kebencian, sejumlah penelitian masih menggunakan arsitektur tambahan di atas model bahasa. Marpaung et al. menggunakan kombinasi IndoBERT dan BiGRU, sedangkan Kusuma dan Chowanda menggunakan IndoBERTweet dan BiLSTM [15], [16] Penambahan lapisan tersebut dapat menjadi strategi untuk membangun model klasifikasi, tetapi menghasilkan arsitektur yang lebih kompleks dibandingkan penggunaan model bahasa secara langsung. Kondisi tersebut menimbulkan pertanyaan penelitian apakah representasi kontekstual yang telah dipelajari oleh IndoBERT dapat dimanfaatkan secara langsung untuk klasifikasi ujaran kebencian tanpa penambahan arsitektur deep learning lainnya.

**Tabel 1**. Perbandingan penelitian terdahulu dengan penelitian ini

| Penelitian | Model | Arsitektur tambahan | Dataset |
|---|---|---|---|
| Marpaung et al. (2021) | IndoBERT + BiGRU | Ya | 13.169 twit |
| Kusuma & Chowanda (2023) | IndoBERTweet + BiLSTM | Ya | 13.169 twit |
| Penelitian ini | IndoBERT fine-tuned murni | Tidak | 13.169 twit |

Berdasarkan penelitian terdahulu tersebut, research gap penelitian ini terletak pada kebutuhan untuk mengevaluasi kemampuan IndoBERT secara langsung tanpa penambahan CNN, LSTM, BiGRU, atau arsitektur deep learning lainnya pada tugas klasifikasi ujaran kebencian berbahasa Indonesia. Penelitian ini tidak mengusulkan arsitektur Transformer baru, tetapi memberikan evaluasi empiris terhadap penggunaan IndoBERT sebagai model utama dengan classification head untuk klasifikasi biner Ujaran Kebencian dan Non-Ujaran Kebencian.

Dengan demikian, kontribusi penelitian ini terletak pada evaluasi penerapan metode *fine-tuning* IndoBERT secara langsung pada dataset ujaran kebencian berbahasa Indonesia yang terdiri atas 13.169 twit. Selain itu, penelitian ini memberikan analisis komprehensif terhadap performa model menggunakan metrik *Accuracy*, *Precision*, *Recall*, dan *F1-Score* untuk mengukur efektivitas klasifikasi. Penelitian ini juga mengkaji kesalahan klasifikasi melalui *confusion matrix* serta analisis karakteristik *False Positive* dan *False Negative* guna mengidentifikasi kelemahan model. Selanjutnya, hasil yang diperoleh dibandingkan dengan beberapa pendekatan pada penelitian terdahulu sebagai dasar untuk menilai posisi kinerja IndoBERT serta menunjukkan efektivitas pendekatan yang diusulkan.

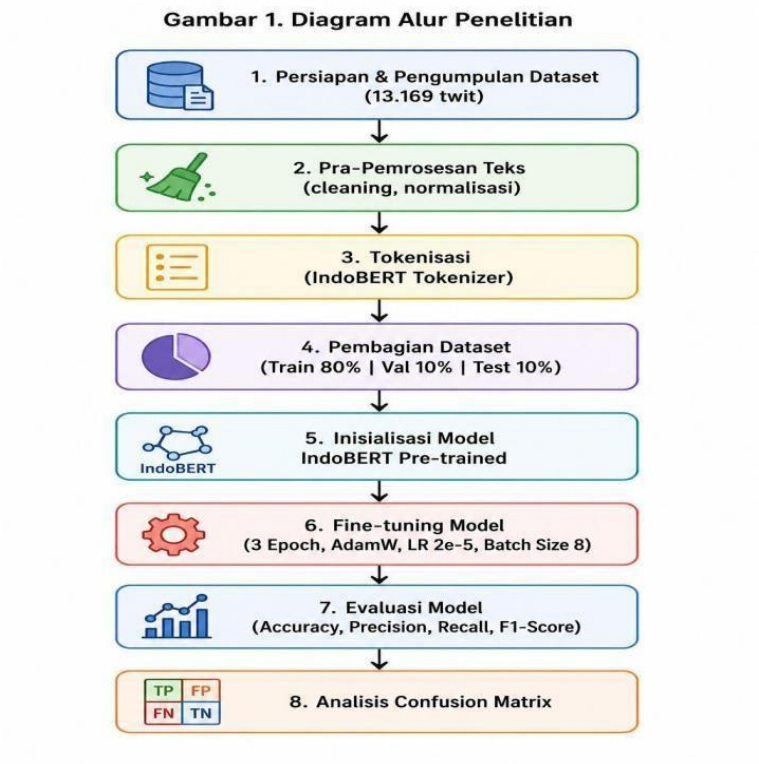
## 2. METODE PENELITIAN

Penelitian ini menggunakan metode eksperimen (*experimental research*) dengan pendekatan *supervised learning* untuk mengembangkan dan mengevaluasi model klasifikasi ujaran kebencian berbahasa Indonesia menggunakan arsitektur transformer IndoBERT. Pendekatan ini dipilih karena proses pembelajaran model dilakukan menggunakan data yang telah memiliki label kelas sehingga hubungan antara teks masukan dan kategori keluaran dapat dipelajari secara optimal.

Tujuan utama penelitian adalah mengevaluasi kemampuan model IndoBERT dalam mengidentifikasi ujaran kebencian pada teks media sosial berbahasa Indonesia melalui proses *fine-tuning* tanpa penambahan lapisan deep learning tambahan seperti CNN, LSTM, maupun BiGRU. Dengan pendekatan tersebut, penelitian berupaya mengetahui sejauh mana representasi kontekstual yang dimiliki IndoBERT mampu melakukan klasifikasi secara efektif menggunakan arsitektur yang lebih sederhana.

Tahapan penelitian terdiri dari proses pengumpulan dataset, pra-pemrosesan data,

tokenisasi teks, pembagian dataset, proses *fine-tuning* model IndoBERT, serta evaluasi performa model menggunakan berbagai metrik klasifikasi. Diagram alur penelitian ditunjukkan pada Gambar 1.




**Gambar 1.** Diagram Alur Penelitian

*Dataset*

Dataset yang digunakan dalam penelitian ini merupakan *Indonesian Hate Speech Dataset* yang diperkenalkan oleh Ibrohim dan Budi (2019). Dataset terdiri dari 13.169 twit berbahasa Indonesia yang telah diberi label ke dalam dua kelas, yaitu Ujaran Kebencian (UK) dan Non-Ujaran Kebencian (Non-UK).

Sebelum digunakan dalam proses fine-tuning, dataset dibagi menjadi tiga subset, yaitu data training, validation, dan testing dengan rasio 80%:10%:10%. Pembagian dilakukan menggunakan fungsi train_test_split dari pustaka scikit-learn. Tahap pertama dilakukan dengan menetapkan 80% data sebagai data training dan 20% sebagai data sementara. Selanjutnya, data sementara tersebut dibagi kembali menjadi dua bagian yang sama besar sehingga diperoleh 10% data validation dan 10% data testing.

Pembagian dataset menggunakan random_state=42 agar hasil pembagian dapat direproduksi. Parameter stratify juga digunakan berdasarkan label kelas untuk mempertahankan proporsi kelas Ujaran Kebencian dan Non-Ujaran Kebencian pada masing-masing subset. Dengan demikian, pembagian data tidak hanya mempertahankan jumlah data sesuai rasio yang ditentukan, tetapi juga menjaga distribusi kelas agar tetap proporsional.

Hasil pembagian menghasilkan 10.535 data training, 1.317 data validation, dan 1.317 data testing. Data training digunakan untuk proses fine-tuning dan pembaruan parameter model. Data validation digunakan untuk mengevaluasi perkembangan performa model pada setiap epoch serta memantau kemungkinan terjadinya overfitting. Sementara itu, data testing tidak digunakan selama proses pelatihan maupun pemilihan checkpoint dan hanya digunakan pada tahap akhir untuk mengukur kemampuan generalisasi model terhadap data yang belum pernah dilihat model.

Distribusi dataset setelah pembagian ditampilkan pada Tabel 2.

**Tabel 2.** Distribusi Dataset Ujaran Kebencian Berbahasa Indonesia

| Kategori | Training (80%) | Validasi (10%) | Pengujian (10%) |
|---|---|---|---|
| Ujaran Kebencian (UK) | 4.449 | 556 | 556 |
| Non-Ujaran Kebencian (Non-UK) | 6.086 | 761 | 761 |
| Total | 10.535 | 1.317 | 1.317 |

*Sumber: Ibrohim & Budi (2019), diolah*

*Persiapan dan Pra-Pemrosesan Data*

Pra-pemrosesan data dilakukan untuk mempersiapkan teks agar dapat digunakan sebagai masukan bagi model IndoBERT. Dataset yang telah dimuat kemudian diseleksi dengan mengambil dua kolom utama, yaitu kolom Tweet sebagai teks masukan dan kolom HS sebagai label. Kedua kolom tersebut kemudian diseragamkan menjadi text dan label untuk menyesuaikan format data dengan proses klasifikasi menggunakan Hugging Face Transformers.

Pada penelitian ini, tidak dilakukan proses pembersihan teks secara manual seperti penghapusan URL, mention, hashtag, maupun normalisasi kata secara terpisah. Teks diproses menggunakan tokenizer bawaan IndoBERT sehingga bentuk teks dapat dikonversi menjadi representasi token yang dapat diterima oleh model.

Proses tokenisasi dilakukan menggunakan tokenizer indobenchmark/indobert-base-p1. Pada tahap ini digunakan tiga konfigurasi utama, yaitu truncation=True, padding="max_length", dan max_length=128. Parameter truncation digunakan untuk memotong teks yang memiliki panjang lebih dari 128 token. Parameter padding="max_length" digunakan untuk menambahkan padding pada teks yang memiliki panjang kurang dari batas maksimum sehingga seluruh input memiliki ukuran yang seragam. Sementara itu, max_length=128 menetapkan panjang maksimum setiap sekuens masukan sebesar 128 token.

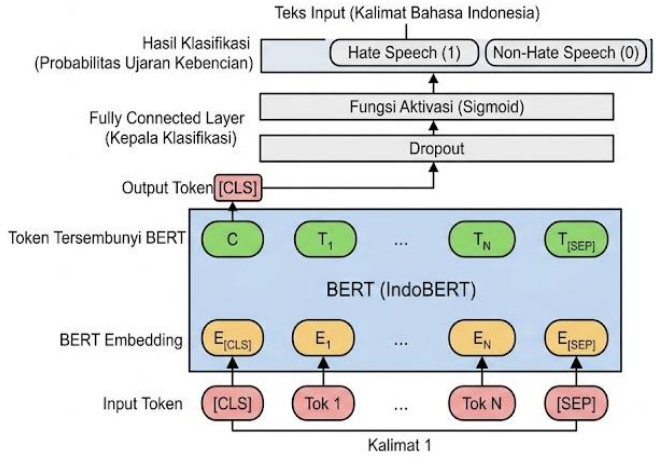
Dengan demikian, tahapan pra-pemrosesan dalam penelitian ini terdiri atas penyesuaian struktur dataset, tokenisasi menggunakan tokenizer IndoBERT, pemotongan teks yang melebihi batas maksimum, dan padding terhadap teks yang lebih pendek. Hasil dari proses tersebut kemudian digunakan sebagai input pada tahap fine-tuning model IndoBERT.

*Arsitektur Model IndoBERT*

Model yang digunakan dalam penelitian ini adalah indobenchmark/indobert-base-p1 yang diimplementasikan menggunakan AutoModelForSequenceClassification dari pustaka Hugging Face Transformers. Model dikonfigurasi untuk melakukan klasifikasi terhadap dua kelas dengan parameter num_labels=2, yaitu Ujaran Kebencian (UK) dan Non-Ujaran Kebencian (Non-UK).

Arsitektur model IndoBERT dalam penelitian ini terdiri atas beberapa tahapan, yaitu teks masukan, tokenisasi, pemrosesan melalui lapisan *Transformer Encoder*, pembentukan representasi kontekstual, dan *classification head*. Teks yang telah melalui proses tokenisasi diubah menjadi representasi token dan selanjutnya diproses oleh arsitektur IndoBERT untuk memperoleh representasi kontekstual dari teks. Representasi tersebut kemudian diteruskan ke *classification head* untuk menghasilkan prediksi terhadap dua kelas, yaitu Ujaran Kebencian dan Non-Ujaran Kebencian.

Model IndoBERT yang telah melalui tahap *pre-training* kemudian dilakukan *fine-tuning* menggunakan dataset penelitian. Pada tahap ini, parameter model disesuaikan berdasarkan data *training* agar model dapat mempelajari karakteristik ujaran kebencian dalam bahasa Indonesia. Proses fine-tuning dan alur klasifikasi model IndoBERT yang digunakan dalam penelitian ini ditunjukkan pada Gambar 2.




**Gambar 2.** Diagram Arsitektur IndoBERT

*Lingkungan Eksperimen*

Seluruh eksperimen dilakukan menggunakan Google Colaboratory (Google Colab). Implementasi model menggunakan bahasa pemrograman Python dengan framework PyTorch dan pustaka Hugging Face Transformers. Pengecekan lingkungan dilakukan menggunakan pustaka PyTorch untuk memastikan ketersediaan GPU sebelum proses fine-tuning.

Google Colab digunakan sebagai lingkungan komputasi karena menyediakan GPU yang membantu mempercepat proses fine-tuning model transformer. Model dan tokenizer juga disimpan setelah proses pelatihan selesai sehingga hasil eksperimen dapat digunakan kembali.

*Konfigurasi Fine-Tunning dan Hyperparameter*

Proses *fine-tuning* model IndoBERT dilakukan menggunakan pustaka Hugging Face Transformers melalui Trainer. Model dilatih menggunakan data *training* dan dievaluasi menggunakan data validasi pada setiap *epoch*. Konfigurasi pelatihan ditentukan melalui TrainingArguments dengan beberapa *hyperparameter*, yaitu *learning rate* sebesar 2e-5, *training batch size* sebesar 8, *validation batch size* sebesar 8, *weight decay* sebesar 0,01, dan jumlah *epoch* sebanyak 3. Model menggunakan optimizer AdamW untuk melakukan pembaruan parameter selama proses pelatihan. Panjang maksimum input ditetapkan sebesar 128 token sesuai dengan konfigurasi tokenizer IndoBERT.

Evaluasi dilakukan pada setiap *epoch* menggunakan data validasi untuk memantau perkembangan kinerja model selama proses *fine-tuning*. Selain itu, model disimpan pada setiap *epoch* sehingga *checkpoint* hasil pelatihan dapat digunakan untuk proses evaluasi. Konfigurasi *fine-tuning* dan *hyperparameter* yang digunakan dalam penelitian ini ditampilkan pada Tabel 3.

Pemilihan learning rate sebesar 2e-5 didasarkan pada rekomendasi standar fine-tuning model BERT yang telah terbukti efektif pada berbagai penelitian klasifikasi teks[12]. Optimizer AdamW digunakan karena mampu melakukan pembaruan parameter dengan regularisasi weight decay yang lebih stabil dibandingkan optimizer Adam standar. Meskipun eksplorasi hyperparameter melalui grid search tidak dilakukan karena keterbatasan komputasi, konfigurasi yang dipilih merupakan titik awal yang umum digunakan (learning rate antara 1e-5 hingga 5e-5, epoch 2-4) dan hasilnya menunjukkan konvergensi yang baik pada data validasi. Penelitian selanjutnya dapat mengeksplorasi pencarian hyperparameter yang lebih sistematis untuk potensi peningkatan performa lebih lanjut.

Weight decay sebesar 0,01 digunakan sebagai regularisasi untuk membantu mengurangi kecenderungan model mengalami overfitting. Panjang maksimum sekuens ditetapkan sebesar 128 token agar seluruh input memiliki ukuran yang seragam selama proses pelatihan.

Penelitian ini tidak melakukan grid search atau pencarian hyperparameter secara sistematis. Oleh karena itu, konfigurasi yang digunakan merupakan satu skenario eksperimen yang terdiri atas learning rate 2e-5, batch size 8, weight decay 0,01, dan 3 epoch.

**Tabel 3.** Konfigurasi Fine-Tuning dan Hyperparameter IndoBERT

| Parameter | Nilai |
|---|---|
| Model | *indobenchmark/indobert-base-p1* |
| Optimizer | AdamW |
| Learning Rate | 2e-5 |
| Training Batch Size | 8 |
| Validation Batch Size | 8 |
| Weight Decay | 0,01 |
| Maximum Sequence Length | 128 token |
| Epoch | 3 |
| Jumlah Kelas | 2 |
| Evaluasi | Setiap Epoch |
| Penyimpanan Model | Setiap Epoch |
| Platform | Google Colaboratory |

*Metrik Evaluasi*

Kinerja model dievaluasi menggunakan empat metrik utama dengan pendekatan macro-average: akurasi, presisi, recall, dan F1-Score. Selain itu, confusion matrix digunakan untuk analisis kesalahan yang lebih mendalam. Rumus masing-masing metrik adalah sebagai berikut:

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

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$F1\text{-}Score = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (4)$$

Pada keempat metrik tersebut, True Positive (TP) adalah jumlah data positif yang diprediksi positif dengan benar, True Negative (TN) adalah jumlah data negatif yang diprediksi negatif dengan benar, False Positive (FP) adalah jumlah data negatif yang salah diprediksi sebagai positif, dan False Negative (FN) adalah jumlah data positif yang salah diprediksi sebagai negatif.

## 3. HASIL DAN PEMBAHASAN

*Hasil Pra-Pemrosesan Data*

Dataset yang digunakan terdiri dari 13.169 data teks berbahasa Indonesia. Pada tahap persiapan data, kolom teks dan label disesuaikan dengan format yang digunakan dalam proses klasifikasi. Teks kemudian diproses menggunakan tokenizer indobenchmark/indobert-base-p1 dengan konfigurasi truncation=True**,** padding="max_length"**,** dan max_length=128**.**

Konfigurasi tersebut memastikan setiap masukan memiliki panjang maksimum 128 token. Teks yang melebihi batas tersebut akan dipotong, sedangkan teks yang lebih pendek akan diberi padding hingga mencapai panjang yang ditentukan. Hasil tokenisasi selanjutnya digunakan

sebagai masukan pada proses fine-tuning IndoBERT.

*Hasil Pelatihan Model*

Model IndoBERT dilakukan proses fine-tuning selama 3 epoch menggunakan optimizer AdamW dengan learning rate 2e-5 dan batch size sebesar 8. Evaluasi dilakukan pada setiap epoch menggunakan data validasi untuk melihat perkembangan kemampuan model dalam melakukan klasifikasi ujaran kebencian.

Tabel 4 menyajikan hasil evaluasi model pada data validasi selama proses pelatihan di setiap epoch. Perlu dicatat bahwa angka-angka pada Tabel 4 merupakan performa validasi, sedangkan hasil akhir pengujian pada data uji (test set) disajikan secara terpisah pada Tabel 6.

**Tabel 4.** Hasil Evaluasi Model IndoBERT pada Data Validasi (per Epoch) selama pelatihan

| Model / Konfigurasi | Akurasi (%) | Presisi (%) | Recall (%) | F1-Score (%) |
|---|---|---|---|---|
| IndoBERT (epoch 1) | 85,87% | 85,44% | 86,31% | 85,46% |
| IndoBERT (epoch 2) | 88,46% | 88,06% | 88,54% | 88,25% |
| IndoBERT (epoch 3) | 88,00% | 87,65% | 87,83% | 87,73% |

Berdasarkan Tabel 4, performa model pada data validasi meningkat dari epoch pertama ke epoch kedua. Nilai accuracy dan F1-Score tertinggi diperoleh pada epoch kedua, masing-masing sebesar 88,46% dan 88,25%. Pada epoch ketiga, nilai tersebut menurun menjadi 88,00% untuk accuracy dan 87,73% untuk F1-Score. Hasil ini menunjukkan bahwa peningkatan performa validasi tidak berlanjut pada epoch ketiga. Meskipun demikian, eksperimen dilakukan sesuai konfigurasi pelatihan yang telah ditetapkan, yaitu sebanyak 3 epoch, sehingga hasil pengujian pada data uji dilaporkan sebagai hasil dari skenario pelatihan tersebut. Pada data uji, model memperoleh accuracy sebesar 90,43%, precision sebesar 90,11%, recall sebesar 90,36%, dan F1-Score sebesar 90,23%. Selanjutnya, training loss dan validation loss dianalisis untuk melihat perubahan nilai loss selama proses fine-tuning, sebagaimana disajikan pada Tabel 5.

**Tabel 5.** Nilai Training Loss dan Validation Loss Selama Proses Pelatihan

| Epoch | Training Loss | Validation Loss |
|---|---|---|
| 1 | 0,380030 | 0,436364 |
| 2 | 0,233251 | 0,480488 |
| 3 | 0,109644 | 0,608882 |

Berdasarkan Tabel 5, nilai training loss mengalami penurunan secara konsisten dari 0,380030 pada epoch pertama menjadi 0,109644 pada epoch ketiga. Penurunan tersebut menunjukkan bahwa model semakin mampu mempelajari pola yang terdapat pada data pelatihan selama proses fine-tuning.

Di sisi lain, nilai validation loss mengalami peningkatan dari 0,436364 pada epoch pertama menjadi 0,608882 pada epoch ketiga. Kondisi ini menunjukkan bahwa meskipun performa model pada data pelatihan terus membaik, peningkatan performa pada data validasi tidak berlangsung secara konsisten. Hal tersebut mengindikasikan adanya gejala overfitting pada epoch-epoch akhir pelatihan.

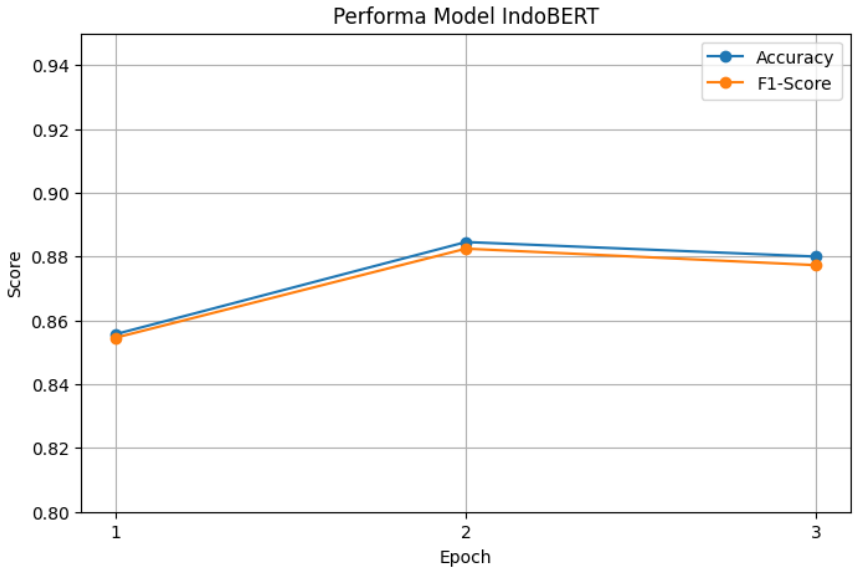
Berdasarkan hasil evaluasi pada data validasi yang ditunjukkan pada Tabel 4, model memperoleh performa terbaik pada epoch kedua dengan nilai accuracy sebesar 88,46% dan F1-score sebesar 88,25%. Pada epoch ketiga, nilai accuracy dan F1-score sedikit menurun menjadi 88,00% dan 87,73%. Oleh karena itu, epoch kedua dipilih sebagai checkpoint terbaik berdasarkan nilai F1-score. Hasil tersebut menunjukkan bahwa model IndoBERT mampu mempelajari pola ujaran kebencian pada teks berbahasa Indonesia dengan performa yang baik.

*Grafik Performa Model*

Untuk memberikan gambaran yang lebih jelas mengenai perkembangan kemampuan model selama proses pelatihan, dilakukan visualisasi perubahan nilai accuracy dan F1-Score pada setiap epoch sebagaimana ditunjukkan pada Gambar 3. Visualisasi ini bertujuan untuk memantau

performa model pada data validasi selama proses fine-tuning serta mengevaluasi kestabilan model dalam melakukan klasifikasi ujaran kebencian.

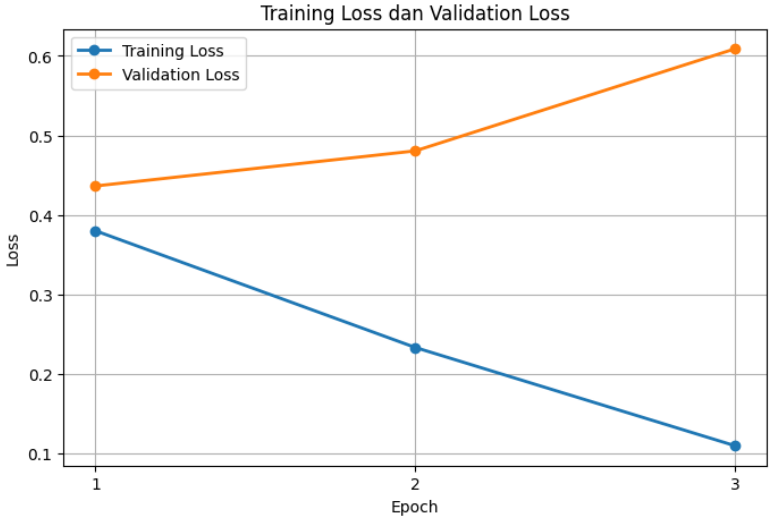
Accuracy digunakan untuk mengukur proporsi prediksi yang berhasil diklasifikasikan dengan benar terhadap seluruh data validasi, sedangkan F1-Score digunakan untuk mengevaluasi keseimbangan antara precision dan recall. Penggunaan kedua metrik tersebut penting karena dataset ujaran kebencian umumnya memiliki distribusi kelas yang tidak sepenuhnya seimbang sehingga evaluasi model tidak cukup hanya berdasarkan nilai akurasi.




**Gambar 3**. Perkembangan Accuracy dan F1-Score selama proses pelatihan

Berdasarkan Gambar 3, nilai accuracy dan F1-Score meningkat dari epoch pertama ke epoch kedua, kemudian mengalami sedikit penurunan pada epoch ketiga. Hasil tersebut menunjukkan bahwa performa validasi tertinggi dalam skenario pelatihan yang dilakukan diperoleh pada epoch kedua.

Setelah analisis accuracy dan F1-score, evaluasi dilanjutkan pada nilai loss. Gambar 4 menunjukkan grafik training loss dan validation loss selama proses fine-tuning, yang mengindikasikan bahwa model berhasil belajar dengan baik dan memiliki generalisasi yang cukup baik.




**Gambar 4.** Perkembangan Training Loss dan Validation Loss Selama Proses Pelatihan

Setelah analisis accuracy dan F1-Score, evaluasi dilanjutkan dengan mengamati perubahan training loss dan validation loss. Gambar 4 menunjukkan bahwa training loss mengalami penurunan secara konsisten, sedangkan validation loss justru mengalami peningkatan selama proses fine-tuning. Perbedaan pola tersebut menjadi indikator bahwa model mulai mengalami overfitting pada epoch akhir.

*Hasil Pengujian Model*

Setelah proses *fine-tuning(pelatihan)* selesai dilakukan, model IndoBERT kemudian diuji menggunakan data testing untuk mengetahui kemampuan model dalam melakukan klasifikasi ujaran kebencian pada data yang belum pernah digunakan selama proses pelatihan. Evaluasi dilakukan menggunakan metrik Accuracy, Precision, Recall, dan F1-Score yang umum digunakan pada tugas klasifikasi teks. Hasil pengujian model ditunjukkan pada Tabel 6.

**Tabel 6.** Hasil Evaluasi Akhir Model IndoBERT pada Data Uji (Test Set)

| Metrik | Nilai |
|---|---|
| Accuracy | 90,43% |
| Precision | 90,11% |
| Recall | 90,36% |
| F1-Score | 90,23% |

Berdasarkan hasil tersebut, model berhasil memperoleh nilai accuracy sebesar 90,43%, precision sebesar 90,11%, recall sebesar 90,36%, dan F1-Score sebesar 90,23%. Nilai accuracy yang tinggi menunjukkan bahwa model mampu melakukan klasifikasi dengan tingkat ketepatan yang baik terhadap keseluruhan data pengujian.

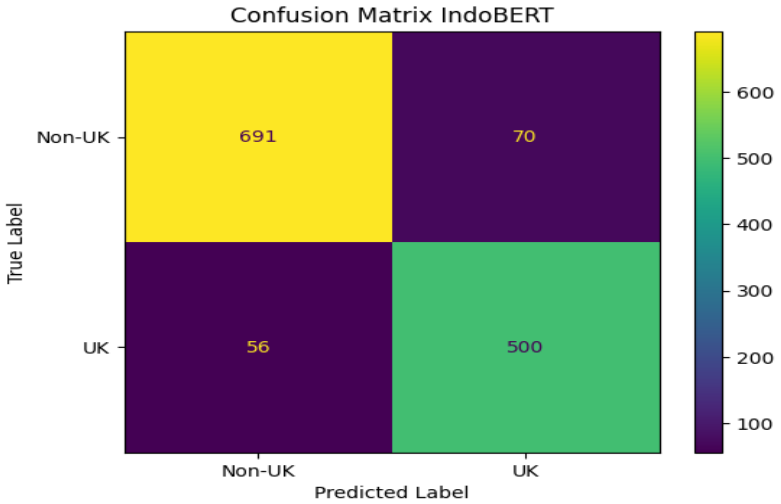
Selain itu, nilai precision dan recall yang seimbang menunjukkan bahwa model memiliki kemampuan yang baik dalam mengidentifikasi ujaran kebencian maupun non-ujaran kebencian. Nilai F1-Score sebesar 90,23% juga mengindikasikan bahwa model mampu menjaga keseimbangan antara precision dan recall sehingga menghasilkan performa klasifikasi yang konsisten. Secara keseluruhan, hasil pengujian menunjukkan bahwa model IndoBERT efektif digunakan untuk mendeteksi ujaran kebencian dalam teks berbahasa Indonesia.

Hasil tersebut menunjukkan bahwa model tidak hanya mampu memperoleh ketepatan klasifikasi secara keseluruhan, tetapi juga memiliki keseimbangan antara kemampuan menemukan data Ujaran Kebencian dan ketepatan prediksi yang dihasilkan. Hal ini berkaitan dengan kemampuan IndoBERT dalam membentuk representasi kontekstual teks berbahasa Indonesia. Dalam tugas deteksi ujaran kebencian, konteks menjadi penting karena kata yang bersifat negatif atau ofensif tidak selalu menunjukkan bahwa suatu teks termasuk Ujaran Kebencian. Hubungan antarkata dan konteks keseluruhan kalimat turut menentukan makna yang disampaikan.

*Analisis Confusion Matrix*

*Confusion matrix* model IndoBERT pada data uji sebanyak 1.317 sampel disajikan pada Gambar 5. Hasil pengujian menunjukkan bahwa model berhasil mengklasifikasikan dengan benar 691 data non-ujaran kebencian dan 500 data ujaran kebencian. Sementara itu, kesalahan klasifikasi terdiri atas 70 data non-ujaran kebencian yang diprediksi sebagai ujaran kebencian (*False Positive*) dan 56 data ujaran kebencian yang diprediksi sebagai non-ujaran kebencian (*False Negative*). Hasil tersebut menunjukkan bahwa sebagian besar data uji telah diklasifikasikan dengan benar, meskipun masih terdapat kesalahan pada kedua kelas. Kesalahan tersebut mengindikasikan bahwa model masih mengalami kesulitan dalam membedakan beberapa sampel yang memiliki karakteristik bahasa serupa atau ambigu antar kelas. Untuk mengevaluasi

kinerja model secara lebih rinci pada masing-masing kelas, Tabel 7 menyajikan *classification report* yang mencakup nilai *precision*, *recall*, dan *F1-score*.




**Gambar 5.** Confusion Matrix IndoBERT Fine-tuned pada Data Uji

**Tabel 7.** Classification Report Per Kelas Model IndoBERT pada Data Uji

| Kelas | Presisi (%) | Recall (%) | F1-Score (%) | Support (n) |
|---|---|---|---|---|
| Non-Ujaran Kebencian (Non-UK) | 92,50 | 90,80 | 91,64 | 761 |
| Ujaran Kebencian (UK) | 87,72 | 89,93 | 88,81 | 556 |
| **Macro Average** | **90,11** | **90,36** | **90,23** | **1.317** |

Berdasarkan Tabel 6, terdapat perbedaan performa antara kedua kelas. Kelas Non-UK menunjukkan performa yang lebih tinggi dengan F1-Score sebesar 91,64%, sedangkan kelas UK memperoleh F1-Score sebesar 88,81%. Perbedaan tersebut dapat dipengaruhi oleh jumlah data pada masing-masing kelas, yaitu 6.086 data Non-UK dan 4.449 data UK pada data pelatihan. Selain jumlah data, karakteristik ujaran kebencian juga dapat memengaruhi tingkat kesulitan klasifikasi karena ujaran kebencian dapat disampaikan secara eksplisit maupun implisit. Meskipun demikian, F1-Score kelas UK sebesar 88,81% menunjukkan bahwa model tetap mampu mengenali kelas Ujaran Kebencian dengan performa yang baik.

*Analisis Kesalahan Model*

Meskipun model IndoBERT memperoleh performa yang baik dengan accuracy sebesar 90,43% dan F1-Score sebesar 90,23%, masih terdapat sejumlah kesalahan klasifikasi yang ditunjukkan oleh confusion matrix, yaitu 70 data False Positive dan 56 data False Negative. Kesalahan tersebut menunjukkan bahwa model belum sepenuhnya mampu memahami konteks dan makna ujaran dalam data berbahasa Indonesia. Untuk mengetahui karakteristik kesalahan yang terjadi, dilakukan analisis terhadap beberapa contoh data False Positive dan False Negative yang dihasilkan pada data pengujian. Contoh kesalahan klasifikasi tersebut disajikan pada Tabel 8.

*Faktor-Faktor yang Memengaruhi Performa Model*

Performa model IndoBERT dalam penelitian ini dipengaruhi oleh beberapa faktor, yaitu distribusi dataset, jumlah *epoch*, konfigurasi *hyperparameter*, dan karakteristik bahasa pada data.

Distribusi jumlah data antar kelas dapat memengaruhi proses pembelajaran model, sedangkan jumlah *epoch* berkaitan dengan kemampuan model dalam mempelajari pola data. Hal ini terlihat dari meningkatnya *training loss* dan menurunnya performa validasi pada epoch akhir yang menunjukkan adanya indikasi *overfitting*.

Selain itu, karakteristik bahasa ujaran kebencian yang dapat disampaikan secara langsung maupun tidak langsung turut memengaruhi hasil klasifikasi. Kesalahan *false positive* dan *false negative* pada *confusion matrix* menunjukkan bahwa model masih mengalami kesulitan dalam membedakan teks yang memiliki konteks atau makna yang ambigu. Konfigurasi *hyperparameter* seperti *learning rate*, *batch size*, *weight decay*, dan jumlah *epoch* juga berpengaruh terhadap proses pembelajaran dan kemampuan generalisasi model.

**Tabel 8.** Contoh Kesalahan Klasifikasi Model pada Data Uji

| Jenis Kesalahan | Contoh Teks | Label Aktual | Prediksi | Analisis |
|---|---|---|---|---|
| False Positive | "Bentar lagi pilkada ya gaiis? Buat yg mau nyolok... Kalo milih pemimpin itu yg sesuai mata hati, bukan sesuai mata uang!" | Non-UK | UK | Konteks politik dan kosakata bernada negatif menyebabkan teks diprediksi sebagai UK. |
| False Positive | "ahok tuding lurah nakal manfaatkan pasukan oranye" | Non-UK | UK | Model kemungkinan terpengaruh oleh penggunaan kata bernada negatif sehingga teks kritik tersebut diprediksi sebagai UK. |
| False Negative | "Para pembajak agama ada di semua agama... Betapa mudahnya membajak agama yang bersifat inklusif..." | UK | Non-UK | Model gagal mengenali konteks yang berkaitan dengan agama dan kelompok tertentu. |
| False Negative | "Aset indonesia katanya di ambil orang asing freeport di ambil amrik" | UK | Non-UK | Model gagal mengenali konteks yang mengarah pada penyebutan dan generalisasi terhadap kelompok asing. |

**Tabel 9.** Keunggulan dan Keterbatasan Penelitian

| Aspek | Keunggulan | Keterbatasan |
|---|---|---|
| Performa | Accuracy 90,43% dan F1-Score 90,23% | Masih terdapat *false positive* dan *false negative* |
| Arsitektur | Menggunakan IndoBERT dengan *classification head* sederhana | Belum menggunakan metode PEFT |
| Dataset | Menggunakan dataset publik | Dataset relatif lama |
| Implementasi | Mudah direplikasi | Belum diuji pada dataset eksternal |

*Kelebihan dan Keterbatasan Penelitian*

Penelitian ini memiliki beberapa keunggulan, yaitu menggunakan arsitektur sederhana berupa IndoBERT dengan *classification head* tanpa lapisan tambahan, namun mampu menghasilkan accuracy sebesar 90,43% dan F1-Score sebesar 90,23%. Selain itu, penggunaan IndoBERT memungkinkan model memanfaatkan representasi kontekstual bahasa Indonesia

dalam proses klasifikasi. Meskipun demikian, penelitian ini masih memiliki beberapa keterbatasan, seperti penggunaan dataset Twitter tahun 2018 yang belum sepenuhnya merepresentasikan kondisi bahasa pada media sosial saat ini, pengujian hanya pada satu konfigurasi *hyperparameter* dan satu skenario pembagian data (80:10:10), serta fokus pada klasifikasi biner. Ringkasan keunggulan dan keterbatasan penelitian disajikan pada Tabel 9.

## 4. KESIMPULAN

Penelitian ini berhasil menunjukkan bahwa model transformer pre-trained IndoBERT yang di-fine-tune secara langsung, tanpa penambahan lapisan arsitektur deep learning, mampu menghasilkan performa yang baik untuk tugas deteksi ujaran kebencian berbahasa Indonesia. Berdasarkan hasil pengujian pada 1.317 data uji, model memperoleh accuracy sebesar 90,43%, precision sebesar 90,11%, recall sebesar 90,36%, dan F1-Score sebesar 90,23%. Hasil tersebut menunjukkan bahwa fine-tuning langsung pada IndoBERT mampu melakukan klasifikasi teks ke dalam kelas Ujaran Kebencian (UK) dan Non-Ujaran Kebencian (Non-UK) dengan performa yang baik.

Hasil evaluasi juga menunjukkan bahwa model masih mengalami kesalahan klasifikasi, yaitu sebanyak 70 False Positive dan 56 False Negative. Analisis kesalahan menunjukkan bahwa False Positive dapat terjadi pada teks yang mengandung kosakata bernada negatif atau konteks politik tertentu, sedangkan False Negative ditemukan pada teks yang memiliki konteks agama, kelompok tertentu, atau makna yang lebih kompleks. Hal tersebut menunjukkan bahwa pemahaman konteks semantik dan karakteristik bahasa yang beragam masih menjadi tantangan dalam deteksi ujaran kebencian berbahasa Indonesia.

Kontribusi penelitian ini adalah memberikan bukti empiris mengenai efektivitas fine-tuning IndoBERT secara langsung untuk mendeteksi ujaran kebencian pada dataset berbahasa Indonesia yang terdiri dari 13.169 twit. Penggunaan arsitektur yang relatif sederhana tanpa penambahan CNN atau BiLSTM juga memberikan pendekatan yang lebih praktis untuk pengembangan sistem deteksi ujaran kebencian. Untuk penelitian selanjutnya, dapat dilakukan eksplorasi terhadap teknik *parameter-efficient fine-tuning* seperti LoRA atau DoRA untuk meningkatkan efisiensi proses pelatihan, augmentasi data untuk memperkaya variasi data, serta pengembangan klasifikasi *multilabel* untuk mengidentifikasi target dan jenis ujaran kebencian secara lebih spesifik

## DAFTAR PUSTAKA

[1] K. Saha, E. Chandrasekharan, and M. de Choudhury, “Prevalence and psychological effects of hateful speech in online college communities,” 2019. doi: 10.1145/3292522.3326032.

[2] G. Kovács, P. Alonso, and R. Saini, “Challenges of hate speech detection in social media,” 2021, doi: 10.1007/s42979-021-00457-3.

[3] H. Margono, M. Saud, and A. Ashfaq, “Dynamics of hate speech in social media: Insights from Indonesia,” 2024, doi: 10.1108/GKMC-11-2023-0464.

[4] M. O. Ibrohim and I. Budi, “Multi-label hate speech and abusive language detection in Indonesian Twitter,” 2019. doi: 10.18653/v1/W19-3506.

[5] S. D. A. Putri, M. O. Ibrohim, and I. Budi, “Abusive language and hate speech detection for Javanese and Sundanese languages in tweets,” 2021. doi: 10.18178/wcse.2021.02.011.

[6] J. Patihullah and E. Winarko, “Hate speech detection for Indonesia tweets using word embedding and gated recurrent unit,” 2019, doi: 10.22146/ijccs.40125.

[7] T. T. A. Putri and S. Sriadhi, "A comparison of classification algorithms for hate speech detection," 2020. doi: 10.1088/1757-899X/830/3/032006.
[8] P. S. B. Ginting, B. Irawan, and C. Setianingsih, "Hate speech detection on Twitter using multinomial logistic regression classification method," 2019. doi: 10.1109/IoTaIS47347.2019.8980379.
[9] T. L. Sutejo and D. P. Lestari, "Indonesia hate speech detection using deep learning," 2018. doi: 10.1109/IALP.2018.8629154.
[10] D. A. N. Erlani and E. B. Setiawan, "Hate comment detection on Twitter using Long Short Term Memory (LSTM) with Genetic Algorithm (GA)," 2024, doi: 10.59188/eduvest.v4i11.1758.
[11] Q. Sifak and E. B. Setiawan, "Hate Speech Detection using CNN and BiGRU with Attention Mechanism on Twitter," in 2023 IEEE International Conference on Communication, Networks and Satellite (COMNETSAT), IEEE, Nov. 2023, pp. 170–175. doi: 10.1109/COMNETSAT59769.2023.10420628.
[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2019. doi: 10.18653/v1/N19-1423.
[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] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," in Proceedings of the 28th International Conference on Computational Linguistics, Stroudsburg, PA, USA: International Committee on Computational Linguistics, 2020, pp. 757–770. doi: 10.18653/v1/2020.coling-main.66.
[15] A. Marpaung, R. Rismala, and H. Nurrahmi, "Hate speech detection in Indonesian Twitter texts using Bidirectional Gated Recurrent Unit," 2021. doi: 10.1109/KST51265.2021.9415760.
[16] J. F. Kusuma and A. Chowanda, "Indonesian hate speech detection using IndoBERTweet and BiLSTM on Twitter," 2023, doi: 10.30630/joiv.7.3.1035.