Terbit *online* pada laman : http://teknosi.fti.unand.ac.id/

Jurnal Nasional Teknologi dan Sistem Informasi

| ISSN (Print) 2460-3465 | ISSN (Online) 2476-8812 |

Artikel Penelitian

Optimalisasi DoRA untuk Deteksi Ujaran Kebencian Berbahasa Indonesia Berbasis Transformer

David Suharjanto [a,*], *Sumarsono* [b]

[a,b] *Informatika, Universitas Islam Negeri Sunan Kalijaga,Yogyakarta, Indonesia*

INFORMASI ARTIKEL

Sejarah Artikel:
Diterima Redaksi: 10 Juli 2025
Revisi Akhir: 19 Desember 2025
Diterbitkan *Online*: 14 Januari 2026

KATA KUNCI

Ujaran Kebencian,
Bahasa Indonesia,
Transformer ,
DoRA,
Fine-tuning

KORESPONDENSI

E-mail: davidsuharjanto7@gmail.com*

A B S T R A C T

Ujaran kebencian (UK) merupakan fenomena meresahkan yang cepat menyebar melalui media sosial, menimbulkan dampak negatif pada kohesi sosial dan kesehatan mental individu. Di Indonesia, peningkatan kasus UK menuntut pengembangan sistem deteksi otomatis yang cepat dan akurat. Penelitian sebelumnya telah memanfaatkan model transformer, namun sering kali disertai dengan penambahan arsitektur *deep learning* seperti CNN atau BiLSTM, yang justru meningkatkan kompleksitas model. Penelitian ini bertujuan untuk mengoptimalkan kinerja deteksi UK berbahasa Indonesia dengan menerapkan teknik *Weight-Decomposed Low-Rank Adaptation* (DoRA) pada model transformer *pre-trained* IndoBERT-Base, IndoBERT-Large, dan IndoBERTweet-Base. Efektivitas DoRA dibandingkan dengan teknik *full fine-tuning* dievaluasi menggunakan dataset berisi 13.169 twit berbahasa Indonesia yang telah dianotasi. Hasil eksperimen menunjukkan bahwa DoRA secara konsisten meningkatkan kinerja pada semua model yang diuji. Model IndoBERTweet-Base dengan DoRA mencapai F1-score tertinggi sebesar 89,64%, melampaui *full fine-tuning* IndoBERTweet (88,18%) serta hasil terbaik dari studi sebelumnya yang menggunakan arsitektur lebih kompleks, seperti IndoBERTweet + CNN (87,60%) dan IndoBERTweet + BiLSTM (88,30%). Temuan ini menunjukkan bahwa *fine-tuning* model transformer menggunakan DoRA merupakan strategi yang efektif untuk deteksi UK dalam Bahasa Indonesia, tanpa memerlukan penambahan arsitektur *deep learning* yang kompleks.

1. PENDAHULUAN

Ujaran kebencian (UK) merupakan pernyataan secara eksplisit maupun implisit yang menyerang individu atau kelompok atas dasar identitas tertentu, seperti etnis atau agama [1]. Ini adalah proyeksi kebencian dan ekspresi permusuhan yang mendorong diskriminasi, bias emosional, serta prasangka yang pada akhirnya dapat menyebabkan konsekuensi buruk seperti desensitisasi dan viktimisasi [2]. UK telah menjadi fenomena global yang meresahkan, dengan cepat menyebar melalui platform media sosial dan menciptakan dampak negatif signifikan pada kohesi sosial serta kesehatan mental individu [3], [4]. Di Indonesia, kasus ujaran kebencian terus menunjukkan peningkatan dari waktu ke waktu [5]. Fenomena ini kerap hadir dalam bentuk hoaks yang memanfaatkan kemudahan akses internet dan media sosial, serta berpotensi memicu tindakan-tindakan yang merugikan bagi individu maupun masyarakat luas [6], [7], [8].

Penyebaran ujaran kebencian yang masif dan berlangsung cepat di media sosial menjadikan deteksi dan penanganan secara manual tidak lagi efisien [9]. Kompleksitas konten media sosial, penggunaan bahasa yang tidak baku, serta keberadaan elemen paralinguistik seperti emotikon dan tagar turut memperumit proses identifikasi. Di samping itu, konteks yang sangat bervariasi dan ketiadaan definisi yang disepakati secara umum semakin menyulitkan deteksi ujaran kebencian, bahkan oleh manusia sekalipun [10]. Oleh karena itu, diperlukan sistem deteksi otomatis yang mampu mengidentifikasi konten bermuatan ujaran kebencian secara cepat dan akurat.

Upaya untuk mendeteksi ujaran kebencian telah dilakukan melalui berbagai pendekatan. Sebagai contoh, Shofianina *et al*.

https://doi.org/10.25077/TEKNOSI.v11i3.2025.341-349

[11] mengimplementasikan model *machine learning* tradisional seperti *Naïve Bayes* (NB), *Support Vector Machine* (SVM), dan *Random Forest Decision Tree* (RFDT) untuk mendeteksi ujaran kebencian pada bahasa lokal Indonesia (Jawa dan Sunda) dari 5.656 cuitan Twitter. Penelitian tersebut menghasilkan nilai F-measure tertinggi di atas 60% untuk kedua bahasa lokal tersebut. Selain itu, Putri *et al*. [12] menerapkan algoritma *machine learning* seperti *Naïve Bayes*, *Multi Level Perceptron*, *AdaBoost Classifier*, *Decision Tree*, dan *Support Vector Machine* pada 4.002 twit berbahasa Indonesia yang mencakup isu politik hingga agama. Studi ini menorehkan hasil akurasi 71,2% dengan recall tertinggi 93,2% menggunakan algoritma *Multinomial Naïve Bayes* tanpa SMOTE. Sementara itu, Ginting *et al*. [13] menggunakan metode *Multinomial Logistic Regression* pada 1.067 data ujaran kebencian dari Twitter berbahasa Indonesia. Hasil yang dicapai cukup mumpuni dengan presisi 80,02%, recall 82%, dan akurasi 87,68%.

Seiring perkembangan teknologi, penelitian deteksi ujaran kebencian semakin banyak beralih ke pendekatan berbasis *neural network* yang lebih canggih. Misalnya, Sutejo dan Lestari [14], mengembangkan model deteksi ujaran kebencian menggunakan *Long Short-Term Memory* (LSTM) dengan fitur tekstual (2.273 data) dan akustik (2.469 data). Model mereka yang hanya menggunakan fitur tekstual mencapai F1-score 87,98%, menunjukkan keunggulan fitur tekstual dibandingkan fitur akustik atau kombinasi keduanya. Selanjutnya, Patihullah dan Winarko [15], mengusulkan penggunaan *Gated Recurrent Unit* (GRU) yang dikombinasikan dengan Word2vec untuk deteksi ujaran kebencian yang berisi 260 twit UK dan 453 twit Non-UK berbahasa indonesia. Hasil eksperimen mereka menunjukkan akurasi terbaik dari GRU dengan fitur Word2vec adalah 92,96%. Tidak hanya itu, Erlani dan Setiawan [16] juga mengimplementasikan LSTM yang dioptimalkan dengan *Genetic Algorithm* (GA) untuk deteksi ujaran kebencian di Twitter data berbahasa Indonesia. Mereka menggunakan teknik ekstraksi fitur TF-IDF dan GloVe, serta mencapai akurasi hingga 92,91% dan F1-Score 91,10% pada dataset yang berisi 48.920 cuitan dari Twitter berbahasa Indonesia.

Dalam beberapa tahun terakhir, model transformer *pre-trained* telah mendominasi bidang *Natural Language Processing* (NLP) berkat kemampuannya menangkap konteks dan semantik bahasa yang kompleks, termasuk untuk tugas deteksi ujaran kebencian. Koto *et al*. [17], memperkenalkan IndoBERT, sebuah *Bidirectional Encoder Representations from Transformers* (BERT) versi Indonesia yang dilatih khusus pada dataset berbahasa Indonesia dari berbagai sumber seperti Wikipedia, Kompas, Tempo, hingga korpus web indonesia. Selain itu, Koto *et al*. [18], juga mengembangkan IndoBERTweet, sebuah model transformer yang dilatih khusus pada data Twitter berbahasa Indonesia. Keduanya menunjukkan potensinya untuk digunakan sebagai model transformer untuk deteksi ujaran kebencian. Sementara itu, Marpaung *et al*. [19], mengimplementasikan IndoBERT yang dikembangkan oleh Lintang [20], yang dikolaborasikan dengan *Bidirectional Gated Recurrent Unit* (BiGRU) untuk deteksi ujaran kebencian pada dataset berisi 13.169 twit berbahasa Indonesia, mencapai akurasi tertinggi sebesar 84,77%. Lebih lanjut, Kusuma dan Chowanda [21], mengusulkan model gabungan IndoBERTweet dengan lapisan tambahan BiLSTM atau CNN untuk meningkatkan kinerja deteksi ujaran kebencian di Twitter berbahasa Indonesia. Mereka melaporkan akurasi hingga 93,7% pada dataset berisi 713 data menggunakan model IndoBERTweet+BiLSTM. Sementara itu, pada dataset yang lebih besar yaitu 13.169 data, model yang sama menorehkan akurasi 88,6%.

Penggunaan model-model pada dataset berisi 13.169 data dalam penelitian Kusuma dan Chowanda [21], menunjukkan bahwa penambahan lapisan *neural network* setelah model transformer dapat meningkatkan performa klasifikasi. Namun, integrasi arsitektur transformer dengan komponen *deep learning* tambahan seperti BiGRU, CNN, atau BiLSTM, justru meningkatkan kompleksitas model. Kompleksitas ini dapat menyulitkan proses *fine-tuning* dan interpretasi model. Oleh karena itu, terdapat celah penelitian yang penting untuk mengeksplorasi metode adaptasi yang lebih efisien dari segi arsitektur, namun tetap optimal untuk model transformer.

Optimalisasi melalui teknik *fine-tuning* seperti W*eight-Decomposed Low-Rank Adaptation* (DoRA) menjadi sangat relevan [22]. Pendekatan ini memungkinkan *fine-tuning* pada model transformer dengan hanya memodifikasi sejumlah kecil parameter tambahan sehingga tetap mempertahankan kekuatan representasi dari model *pre-trained*. Selain itu, DoRA memungkinkan adaptasi yang efektif terhadap tugas spesifik tanpa perlu menambahkan lapisan *deep learning* atau membangun arsitektur kompleks dari awal [23].

Berdasarkan latar belakang dan celah penelitian yang telah diidentifikasi, penelitian ini memiliki beberapa tujuan. Pertama, penelitian ini bertujuan untuk mengevaluasi kinerja model transformer *pre-trained* berbahasa Indonesia, khususnya IndoBERT dan IndoBERTweet dalam tugas deteksi UK. Kedua, menganalisis efektivitas teknik *Weight-Decomposed Low-Rank Adaptation* (DoRA) sebagai metode adaptasi alternatif dalam mengoptimalkan kinerja UK. Ketiga, penelitian ini juga akan membandingkan performa model yang telah dioptimalkan menggunakan DoRA dengan hasil yang diperoleh dari penelitian-penelitian sebelumnya yang cenderung menggunakan arsitektur model yang lebih kompleks.

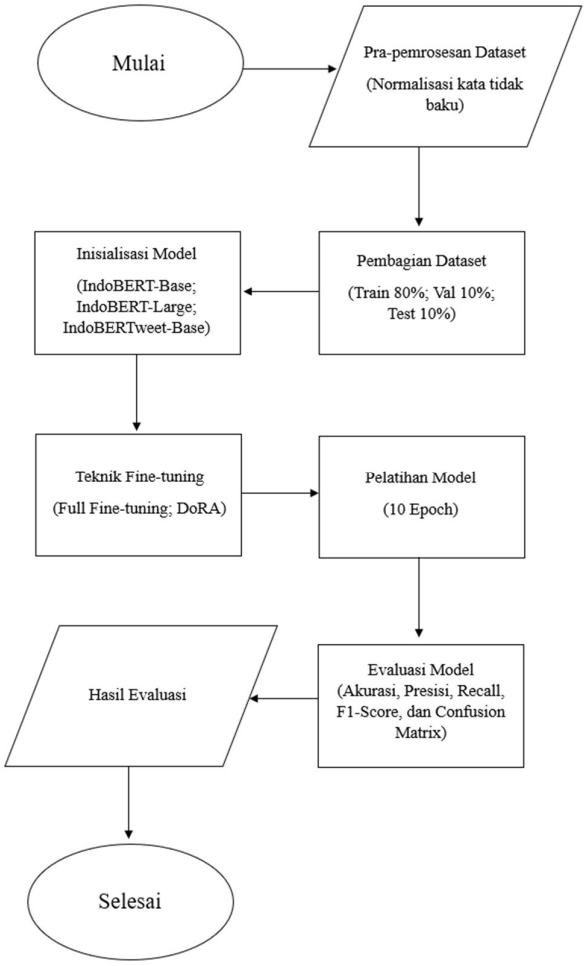
Selanjutnya artikel ini diorganisir sebagai berikut. Bagian 2 menguraikan metode penelitian, termasuk detail arsitektur model, dataset yang digunakan, lingkungan pengembangan, serta konfigurasi *fine-tuning* dan metrik evaluasinya. Bagian 3 menyajikan hasil eksperimen yang diperoleh. Bagian 4 membahas dan menganalisis temuan-temuan penelitian secara komprehensif, termasuk batasan dan saran untuk penelitian mendatang. Terakhir, Bagian 5 merangkum kesimpulan utama dari penelitian ini.

2. METODE

Bagian ini menguraikan secara detail desain eksperimen, arsitektur model, dataset yang digunakan, lingkungan pengembangan, serta konfigurasi pelatihan dan metrik evaluasi yang diterapkan dalam penelitian ini. Selain itu, Penelitian ini mengikuti diagram alir penelitian yang digambarkan secara komprehensif pada Gambar 1. Proses dimulai dengan persiapan dataset dengan pra-pemrosesan mengenai kata tidak baku, kemudian pembagian dataset, diikuti oleh tahapan inisialisasi dan

https://doi.org/10.25077/TEKNOSI.v11i3.2025.341-349

pelatihan model transformer menggunakan teknik *fine-tuning* yang berbeda, dan diakhiri dengan evaluasi serta analisis hasil untuk mendapatkan temuan penelitian.



Gambar 1. Diagram Alir Penelitian

2.1. Arsitektur Model

Penelitian ini memanfaatkan tiga model transformer *pre-trained* berbahasa Indonesia yang berbeda untuk tugas deteksi ujaran kebencian. Model-model tersebut meliputi IndoBERT-base-uncased dengan 124,5 juta parameter, IndoBERT-Large-p2 sebagai versi yang lebih besar dengan 335 juta parameter, dan IndoBERTweet-base-p2 dengan 110 juta parameter [17], [18]. Ketiga model ini dipilih karena kemampuannya dalam memahami nuansa Bahasa Indonesia yang kompleks dan performa yang telah terbukti dalam berbagai tugas *Natural Language Processing* (NLP).

2.2. Dataset

Dataset yang digunakan dalam penelitian ini adalah dataset ujaran kebencian berbahasa Indonesia yang dikumpulkan oleh Ibrohim dan Budi [24]. Dataset ini terdiri dari total 13.169 twit yang diambil dari Twitter dalam rentang waktu Maret hingga September 2018. Dataset ini menggunakan Bahasa Indonesia dan telah dianotasi secara manual. Data kemudian diklasifikasikan ke dalam dua kategori, yaitu Ujaran Kebencian (UK) dan Non-Ujaran Kebencian (Non-UK). Pembagian dataset dilakukan secara acak dengan proporsi yaitu 80% untuk *training set*, 10% untuk *validation set*, dan 10% untuk *test set*. Detail pembagian dataset dapat dilihat pada Tabel 1.

Tabel 1. Dataset Ujaran Kebencian Berbahasa Indonesia

Kategori	*Train*	*Validation*	*Test*
UK	4.449	556	556
Non-UK	6.086	761	761
Total	10.535	1.317	1.317

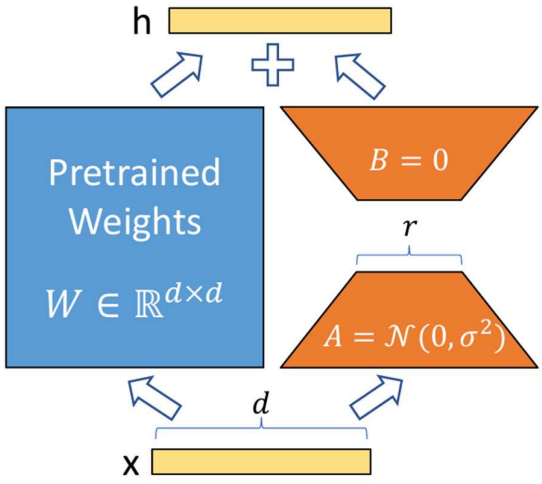
2.3. Lingkungan Pengembangan

Seluruh eksperimen dan proses pelatihan model dilakukan menggunakan lingkungan Google Colaboratory yang menyediakan akses ke GPU T4 dengan VRAM sebesar 15 GB. Framework deep learning yang digunakan adalah PyTorch, dengan bantuan pustaka Hugging Face Transformers untuk implementasi model transformer dan tokenisasi.

2.4. Konfigurasi *Fine-tuning*

Proses *fine-tuning* model dilakukan dengan dua teknik utama: *Full fine-tuning* (Full FT) dan *Weight-Decomposed Low-Rank* Adaptation (DoRA). Konfigurasi umum hyperparameter yang diterapkan pada kedua teknik ini disajikan pada Tabel 2. Selain konfigurasi umum, terdapat pengaturan spesifik untuk masing-masing teknik *fine-tuning*. Untuk DoRA, parameter r (*rank*) diatur ke 8, lora alpha ke 16, dan lora dropout ke 0.1. *Learning rate* yang digunakan untuk DoRA adalah 2e-4, sedangkan untuk *full fine-tuning* adalah 2e-5, disesuaikan dengan praktik umum dalam *fine-tuning* model transformer besar. Detail konfigurasi spesifik ini dapat dilihat pada Tabel 3.

Untuk memberikan gambaran visual mengenai mekanisme DoRA, Gambar 2 berikut menyajikan arsitektur penyisipan adaptasi DoRA pada lapisan model *pre-trained*. Dalam pendekatan ini, bobot awal dari model ($W \in \mathbb{R}^{d\times d}$) tetap dipertahankan, sementara adaptasi parameter dilakukan melalui penyisipan matriks *low-rank*. DoRA menginisialisasi matriks A secara acak dari distribusi normal $N(0, \sigma^2)$ dan menetapkan $B = 0$. Output akhir kemudian diperoleh dengan menjumlahkan hasil dari bobot *pre-trained* dan kontribusi arah dari matriks A [22].



Gambar 2. Mekanisme DoRA dalam adaptasi parameter pada model *pre-trained* [22]

Tabel 2. Konfigurasi Umum Hyperparameter

Hyperparameter	Nilai
Optimizer	AdamW
LR scheduler	Cosine
Training batch size	16
Validation batch size	16
Weight decay	0,01
Max length	256
Epoch	10

Tabel 3. Konfigurasi Spesifik Hyperparameter

Teknik	*Learning rate*	r	Lora alpha	Lora dropout
DoRA	2e-4	8	16	0.1
Full FT	2e-5	N/A	N/A	N/A

2.5. Metrik Evaluasi

Kinerja model dievaluasi menggunakan metrik klasifikasi standar (akurasi, presisi, recall, dan F1-score) dengan pendekatan *macro-average*, serta *confusion matrix* untuk analisis kesalahan. Selain metrik kinerja klasifikasi, efektivitas DoRA juga akan dievaluasi berdasarkan kebutuhan VRAM dan durasi pelatihan untuk menilai efisiensi komputasi dibandingkan dengan *full fine-tuning*. Analisis komparatif antara *full fine-tuning* dan DoRA juga akan dianalisis untuk setiap model transformer yang diuji guna mengidentifikasi pola peningkatan yang konsisten pada metrik-metrik tersebut.

Secara umum, *confusion matrix* terdiri dari empat komponen utama:

- *True Positive* (TP):
  Data yang seharusnya termasuk kelas positif dan berhasil diprediksi sebagai positif.
- *True Negative* (TN):
  Data yang seharusnya termasuk kelas negatif dan berhasil diprediksi sebagai negatif.
- *False Positive* (FP):
  Data yang sebenarnya negatif, tetapi salah diprediksi sebagai positif.
- *False Negative* (FN):
  Data yang sebenarnya positif, tetapi salah diprediksi sebagai negatif.

Rumus masing-masing metrik dituliskan sebagai berikut:

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

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

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

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

3. HASIL

Bagian ini menyajikan hasil eksperimen deteksi ujaran kebencian menggunakan model transformer *pre-trained* berbahasa Indonesia yang diadaptasi dengan teknik *full fine-tuning* dan *Weight-Decomposed Low-Rank Adaptation* (DoRA). Tabel 4 merangkum kinerja deteksi ujaran kebencian dari model-model transformer yang diuji, meliputi metrik akurasi, presisi, recall, dan F1-Score berdasarkan rumus pada bagian metode penelitian.

Tabel 4. Perbandingan Kinerja Deteksi Ujaran Kebencian

Model	Teknik	Akurasi	Presisi	Recall	F1-score
IndoBERT-Base	Full FT	85,95	85,79	86,66	85,84
	DoRA	88,38	88,61	87,52	87,94
IndoBERT-Large	Full FT	88,46	88,11	88,32	88,21
	DoRA	89,60	89,41	89,23	89,32
IndoBERTweet-Base	Full FT	88,46	88,15	88,22	88,18
	DoRA	**89,90**	**89,67**	**89,61**	**89,64**

4. PEMBAHASAN

Bagian ini menyajikan pembahasan yang mengacu langsung pada tiga tujuan utama penelitian. Pertama, dianalisis performa dari berbagai model transformer pre-trained berbahasa Indonesia, yaitu IndoBERT dan IndoBERTweet, dalam mendeteksi UK. Kedua, menganalisis efektivitas teknik *Weight-Decomposed Low-Rank Adaptation* (DoRA) sebagai metode *fine-tuning* alternatif dalam mengoptimalkan kinerja model. Ketiga, membandingkan hasil eksperimen dengan temuan dari studi sebelumnya. Selain itu, pembahasan ini juga mencakup interpretasi *confusion matrix* untuk masing-masing model, serta analisis efisiensi sumber daya komputasi berdasarkan penggunaan VRAM dan durasi pelatihan antara DoRA dan *full fine-tuning*.

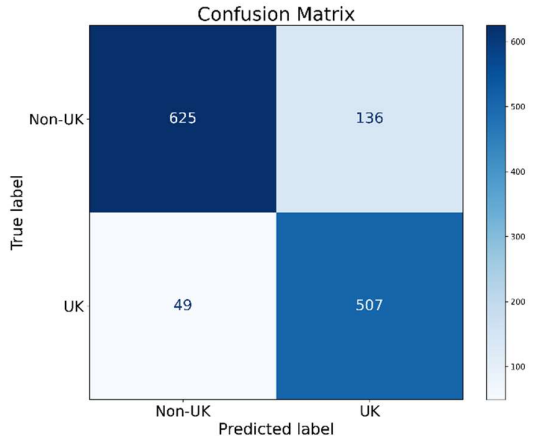
4.1. Analisis Kinerja Model *Full Fine-tuning*

Bagian ini membahas kinerja deteksi ujaran kebencian dari masing-masing model transformer *pre-trained* berbahasa Indonesia (IndoBERT-Base, IndoBERT-Large, dan IndoBERTweet-Base) yang diuji dalam penelitian ini, menggunakan teknik *full fine-tuning* sebagai baseline awal. Model IndoBERT-Base yang dilatih dengan *full fine-tuning* menunjukkan F1-score sebesar 88,21%, dengan akurasi 88,46%, presisi 88,11%, dan recall 88,32%. Kinerja ini cukup optimal untuk model berukuran *base*. Adapun *confusion matrix* (Gambar 3) menunjukkan bahwa dari 1.317 data uji, model mampu mengklasifikasikan 625 Non-UK dengan benar (*True Negative*) dan 507 UK dengan benar (*True Positive*). Namun, terdapat 136 *false positive* (Non-UK diklasifikasikan sebagai UK) dan 49 *false negative* (UK diklasifikan sebagai Non-UK).

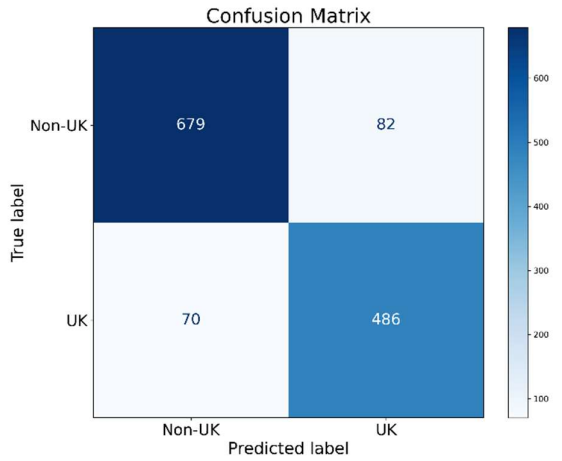
Sementara itu, IndoBERT-Large dengan *full fine-tuning* mencatatkan hasil lebih tinggi dibandingkan versi *base*-nya yaitu F1-score 88,21%, akurasi 88,46%, presisi 88,11%, dan recall 88,32%. *Confusion matrix* untuk IndoBERT-Large *full fine-*

https://doi.org/10.25077/TEKNOSI.v11i3.2025.341-349

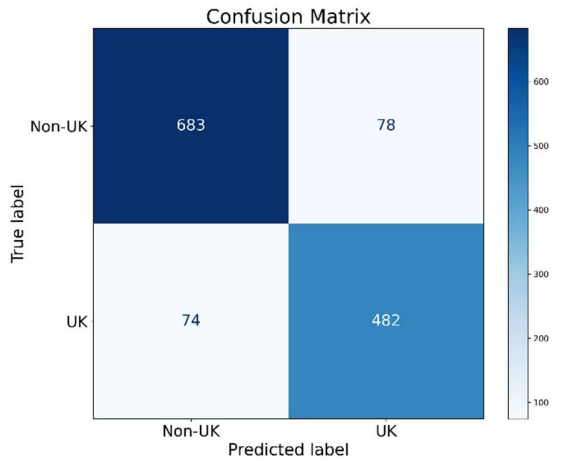
tuning (Gambar 4) menunjukkan 679 *true negative* dan 486 *true positive*, dengan 82 *false positive* dan 70 *false negative*. Kinerja yang lebih baik pada model *large* ini, dibandingkan versi *base*, mengindikasikan kemampuannya untuk memanfaatkan kapasitas parameternya yang lebih besar dalam proses *full fine-tuning*.



Gambar 3. *Confusion Matrix* IndoBERT-Base dengan *Full fine-tuning*



Gambar 4. *Confusion Matrix* IndoBERT-Large dengan *Full fine-tuning*

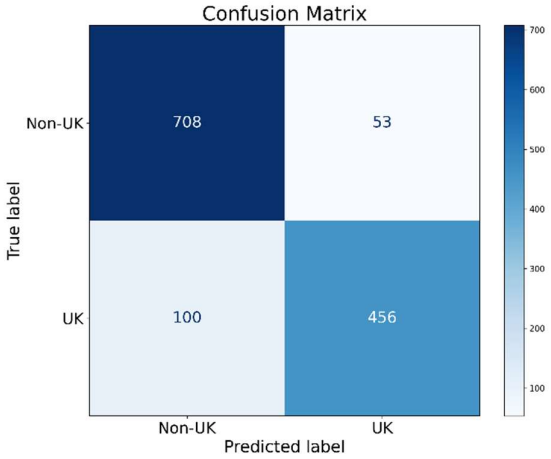


Gambar 5. *Confusion matrix* IndoBERTweet-Base dengan *Full fine-tuning*

Terakhir, IndoBERTweet-Base dengan *full fine-tuning* mencapai F1-score 88,18%, akurasi 88,46%, presisi 88,15%, dan recall 88,22%. Kinerja ini sangat kompetitif, bahkan sebanding dengan IndoBERT-Large yang memiliki parameter jauh lebih besar. *Confusion matrix* (Gambar 5) menunjukkan 683 *true negative* dan 482 *true positive*, dengan 78 *false positive* dan 74 *false negative*. Performa yang mengesankan ini menggarisbawahi bahwa model *pre-training* pada data spesifik Twitter sangat relevan dan menguntungkan. Meskipun memiliki jumlah parameter yang lebih sedikit, IndoBERTweet-Base mampu menyamai bahkan sedikit mengungguli kinerja IndoBERT-Large dalam *full fine-tuning*, menunjukkan keefektifan dan relevansi data *pre-training*. Ini sangat bermanfaat ketika diaplikasikan pada dataset ujaran kebencian yang juga berasal dari aplikasi sosial media pada umumnya.

4.2. Analisis Efektivitas DoRA vs. *Full fine-tuning* dalam Optimalisasi Kinerja

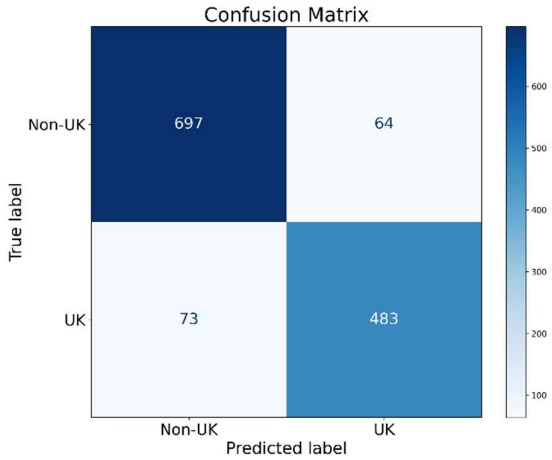
Setelah penerapan *Weight-Decomposed Low-Rank Adaptation* (DoRA), diperoleh F1-score pada model IndoBERT-Base meningkat secara signifikan menjadi 87,94%, dengan akurasi 88,38%, presisi 88,61%, dan recall 87,52%. *Confusion matrix* pada Gambar 6 memperlihatkan adanya perubahan dalam pola klasifikasi. Jumlah *true negative* meningkat menjadi 708, sementara *true positive* adalah 456. *False positive* menurun drastis menjadi 53, namun *false negative* meningkat menjadi 100. Peningkatan F1-score ini menunjukkan efektivitas DoRA dalam mengoptimalkan performa model berukuran *base*. Meskipun terjadi peningkatan *false negative*, pengurangan *false positive* yang substansial dan peningkatan akurasi serta presisi keseluruhan berkontribusi pada peningkatan F1-score secara keseluruhan, mengindikasikan model menjadi lebih presisi dalam identifikasi UK.



Gambar 6. *Confusion matrix* IndoBERT-Base dengan DoRA

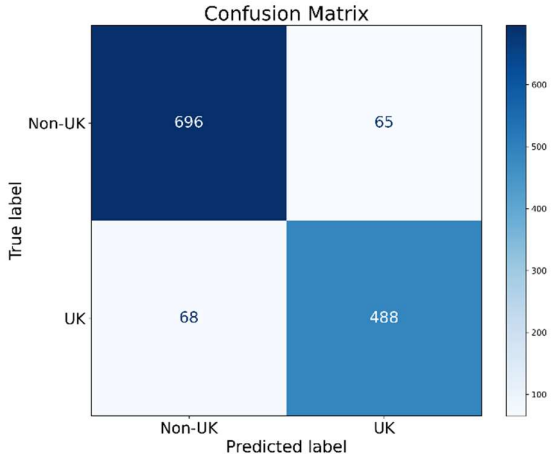
Penerapan DoRA pada IndoBERT-Large berhasil meningkatkan F1-score menjadi 89,32%, dengan akurasi 89,60%, presisi 89,41%, dan recall 89,23%. *Confusion matrix* pada Gambar 7 menunjukkan peningkatan *true negative* menjadi 697 dan *true positive* menjadi 483. *False positive* menurun menjadi 64, sementara *false negative* meningkat sedikit menjadi 73. Peningkatan kinerja ini mengindikasikan bahwa DoRA mampu mengoptimalkan model dengan jumlah parameter yang lebih

besar, terutama dalam mengurangi kesalahan klasifikasi *false positive*. DoRA tampaknya membantu model *large* mencapai potensi penuhnya dalam proses *fine-tuning*.



Gambar 7. *Confusion matrix* IndoBERT-Large dengan DoRA

Model IndoBERTweet-Base menunjukkan kinerja terbaik secara keseluruhan dalam penelitian ini, dan DoRA berhasil memaksimalkan potensinya. Dengan DoRA, F1-score-nya melonjak signifikan menjadi 89,64%, diikuti oleh akurasi 89,90%, dan presisi 89,67%. *Confusion matrix* pada Gambar 8 menunjukkan peningkatan *true negative* menjadi 696 dan *true positive* menjadi 488, serta penurunan *false positive* dari 78 menjadi 65 dan *false negative* dari 74 menjadi 68. Hasil yang superior ini menegaskan bahwa DoRA sangat efektif dalam memaksimalkan potensi adaptasi IndoBERTweet untuk tugas deteksi ujaran kebencian, menghasilkan peningkatan kinerja yang lebih baik pada kedua kelas klasifikasi (UK dan Non-UK) secara lebih seimbang, dengan penurunan *false positive* dan *false negative*.



Gambar 8. *Confusion matrix* IndoBERTweet-Base dengan DoRA

4.3. Perbandingan Performa Model dengan Studi Sebelumnya

Berdasarkan Tabel 4 dan 5, Model IndoBERTweet-Base dengan DoRA yang mencapai F1-score 89,64% menunjukkan keunggulan yang signifikan dibandingkan hasil dari studi sebelumnya. Sebagai contoh, kinerja ini melampaui F1-score 88,10% yang dicapai oleh Kusuma dan Chowanda [21] saat menggunakan *full fine-tuning* pada IndoBERTweet. Bahkan, model-model penelitian ini dapat melampaui hasil kombinasi arsitektur yang lebih kompleks seperti IndoBERTweet + BiLSTM (F1-Score 88,30%) dan IndoBERTweet + CNN (F1-Score 87,60%) dari Kusuma dan Chowanda [21]. Selain itu, model IndoBERT-Large dengan DoRA yang menghasilkan F1-score 89,32% juga menunjukkan performa yang lebih tinggi dibandingkan dengan beberapa studi terdahulu. Di mana model tersebut dapat melampaui hasil akurasi 84,77% yang dicapai oleh Marpaung *et al*. [19] dengan arsitektur model IndoBERT + BiGRU, dan juga lebih tinggi dari akurasi 87,50% dari *full fine-tuned* IndoBERTweet oleh Koto *et al*. [18].

Tabel 5. Perbandingan Kinerja DoRA dengan Studi Sebelumnya

Model	Teknik	Akurasi	Presisi	Recall	F1-score
IndoBERT-Base	DoRA	88,38	88,61	87,52	87,94
IndoBERT-Large	DoRA	89,60	89,41	89,23	89,32
IndoBERTweet-Base	DoRA	**89,90**	**89,67**	**89,61**	**89,64**
IndoBERTweet [21]	Full FT	88,50	88,00	88,30	88,10
IndoBERTweet + BiLSTM [21]	Full FT	88,60	88,10	88,50	88,30
IndoBERTweet + CNN [21]	Full FT	88,00	87,50	87,70	87,60
IndoBERT + BiGRU [19]	Full FT	84,77	-	-	-
IndoBERTweet [18]	Full FT	87,50	-	-	-

Temuan ini secara kolektif mengindikasikan bahwa fokus pada optimalisasi adaptasi model transformer *pre-trained* dengan teknik seperti DoRA adalah strategi yang sangat efektif untuk deteksi ujaran kebencian berbahasa Indonesia. Keberhasilan DoRA menunjukkan bahwa peningkatan kinerja dapat dicapai tidak selalu dengan menambah kompleksitas arsitektur model, melainkan dapat dengan cara lain yaitu dengan melakukan *fine-tuning* pada model dasar yang sudah kuat. Hal ini dapat membuka jalan bagi pengembangan sistem deteksi ujaran kebencian yang lebih akurat dan berpotensi lebih mudah diimplementasikan.

4.4. Batasan dan Penelitian Mendatang

Meskipun penelitian ini berhasil menunjukkan keunggulan DoRA dalam meningkatkan kinerja deteksi ujaran kebencian, terdapat beberapa batasan yang perlu diakui. Salah satunya terkait dengan efisiensi komputasi. Meskipun DoRA dirancang untuk mengurangi jumlah parameter yang dilatih sehingga secara teoretis lebih efisien dalam hal jumlah parameter yang diperbarui,

https://doi.org/10.25077/TEKNOSI.v11i3.2025.341-349

dan memungkinkan *fine-tuning* tanpa penambahan arsitektur *deep learning* yang kompleks. Hasil eksperimen justru menunjukkan DoRA memerlukan penggunaan VRAM yang lebih tinggi dan durasi pelatihan yang relatif lebih lama dibandingkan *full fine-tuning*, seperti yang tertera pada Tabel 6. Sebagai contoh, IndoBERT-Large dengan DoRA membutuhkan 11,3 GB VRAM dan durasi pelatihan 2 jam 43 menit 55 detik, sementara *full fine-tuning* hanya memerlukan 8,8 GB VRAM dan durasi 1 jam 19 menit 31 detik. Pola serupa juga terlihat pada IndoBERT-Base (DoRA: 4,4 GB VRAM, 1 jam 4 menit 12 detik; Full FT: 3,4 GB VRAM, 27 menit 4 detik) dan IndoTweetBERT-Base (DoRA: 4,3 GB VRAM, 59 menit 58 detik; Full FT: 3,3 GB VRAM, 25 menit 6 detik).

Tabel 6. Perbandingan VRAM dan Waktu Pelatihan

Model	Teknik	VRAM	Waktu Pelatihan
IndoBERT-Base	Full FT	3,4	00:27:04
	DoRA	4,4	01:04:12
IndoBERT-Large	Full FT	8,8	01:19:31
	DoRA	11,3	02:43:55
IndoBERTweet-Base	Full FT	3,3	00:25:06
	DoRA	4,3	00:59:58

Fenomena ini dapat terjadi karena meskipun DoRA melatih lebih sedikit parameter, implementasi adapter pada arsitektur transformer mungkin dapat menimbulkan *overhead* komputasi atau memori tambahan selama pelaksanaan pelatihan. *Overhead* ini mungkin disebabkan oleh akses memori yang kurang optimal, *re-computation* sebagian *gradient*, atau kompleksitas *forward/backward pass* yang spesifik dari desain DoRA itu sendiri, sehingga meniadakan keuntungan efisiensi komputasi yang diharapkan dari pengurangan jumlah *trainable parameters* pada model-model yang digunakan dalam penelitian ini. Oleh karena itu, terdapat *trade-off* yang signifikan antara peningkatan kinerja deteksi ujaran kebencian yang dapat dicapai dengan DoRA dengan kebutuhan sumber daya komputasi yang relatif lebih tinggi.

Berdasarkan batasan yang ditemukan, penelitian mendatang dapat berfokus pada eksplorasi hyperparameter DoRA, seperti r, lora alpha, dan lora dropout, secara lebih mendalam. Tujuannya adalah untuk menemukan konfigurasi optimal yang tidak hanya memaksimalkan kinerja, tetapi juga mengidentifikasi *trade-off* terbaik antara akurasi dan efisiensi sumber daya (VRAM dan waktu pelatihan). Selain itu, penting untuk melakukan analisis kualitatif terhadap *error model*, yang dapat melibatkan pemeriksaan *false positive* dan *false negative* secara manual guna mengidentifikasi pola kesalahan yang spesifik, seperti deteksi ujaran kebencian pada teks yang mengandung sarkasme, metafora, atau *hate speech* implisit yang masih menjadi tantangan bagi model untuk mempelajarinya.

Penelitian selanjutnya juga dapat menguji hasil model *full fine-tuning* maupun DoRA yang sudah di-*fine-tune* dalam aplikasi nyata untuk mengukur parameter efisiensi seperti waktu inferensi dan kebutuhan sumber daya komputasi saat model digunakan. Selain itu, selanjutnya juga dapat memperluas eksplorasi pada teknik *fine-tuning* lainnya, seperti LoRA [25], QloRA [26], ReFT [27], atau kombinasi dari teknik-teknik tersebut. Pendekatan ini dapat membantu mengidentifikasi apakah terdapat metode lain yang mampu memberikan kinerja setara atau lebih baik dengan efisiensi sumber daya yang lebih optimal.

5. KESIMPULAN

Penelitian ini berhasil mengeksplorasi dan mengoptimalkan deteksi ujaran kebencian berbahasa Indonesia menggunakan model transformer *pre-trained* dengan teknik *Weight-Decomposed Low-Rank Adaptation* (DoRA). Hasil eksperimen menunjukkan bahwa DoRA secara konsisten meningkatkan kinerja deteksi ujaran kebencian pada semua model transformer yang diuji, melampaui pendekatan *full fine-tuning* standar. Secara khusus, model IndoBERTweet-Base dengan DoRA mencapai F1-score tertinggi 89,64%, yang secara signifikan melampaui hasil terbaik dari studi-studi sebelumnya yang menggunakan arsitektur model lebih kompleks, seperti IndoBERTweet + CNN (87,60%) dan IndoBERTweet + BiLSTM (88,30%) [21].

Temuan ini mengkonfirmasi bahwa DoRA dapat menjadi strategi yang efektif untuk *fine-tuning* model transformer pada tugas deteksi ujaran kebencian berbahasa Indonesia. Keberhasilan ini menunjukkan bahwa peningkatan kinerja model dapat dicapai melalui metode adaptasi efisien pada model dasar yang sudah kuat, yaitu dengan hanya memodifikasi sebagian kecil parameter tanpa perlu menambah lapisan arsitektur *deep learning* yang kompleks. Namun, perlu dicatat bahwa implementasi DoRA ditemukan memiliki *trade-off* dalam efisiensi komputasi, seperti kebutuhan VRAM yang lebih tinggi dan durasi pelatihan yang lebih lama dibandingkan *full fine-tuning*. Kontribusi utama penelitian ini adalah memberikan bukti empiris tentang superioritas kinerja DoRA dalam konteks deteksi ujaran kebencian berbahasa Indonesia, menawarkan wawasan mengenai penerapan teknik *fine-tuning* alternatif yang optimal dari segi arsitektur, dan menyediakan rekomendasi praktis untuk pengembangan sistem deteksi ujaran kebencian yang lebih akurat dan efisien dalam implementasi, dengan mempertimbangkan *trade-off* sumber daya komputasi.

DAFTAR PUSTAKA

[1] F. Ihsan, I. Iskandar, N. S. Harahap, and S. Agustian, "Algoritme decision tree untuk mendeteksi ujaran kebencian dan bahasa kasar multilabel pada Twitter berbahasa Indonesia," Jurnal Teknologi dan Sistem Komputer, vol. 9, no. 4, pp. 199–204, Oct. 2021, doi: 10.14710/jtsiskom.2021.13907.

[2] H. Margono, M. Saud, and A. Ashfaq, "Dynamics of hate speech in social media: insights from Indonesia," Global Knowledge, Memory and Communication, vol. ahead-of-print, no. ahead-of-print, Jan. 2024, doi: 10.1108/GKMC-11-2023-0464.

[3] K. Saha, E. Chandrasekharan, and M. de Choudhury, "Prevalence and Psychological Effects of Hateful Speech in Online College Communities," in Proceedings of the 2019 ACM Web Science Conference, Association for Computing Machinery, 2019, pp. 255–264. doi: 10.1145/3292522.3326032.

[4] Y. P. Setianto, H. Nurjuman, and U. R. Handaningtias, "Remaja, Media Sosial Dan Ujaran Kebencian: Studi

Konsumsi Online Religious Content Di Banten," Interaksi: Jurnal Ilmu Komunikasi, vol. 12, no. 1, pp. 125–145, Jun. 2023, doi: 10.14710/interaksi.12.1.125-144.

[5] A. B. F. Cahyono, A. Khalisah, L. Safitri, T. Lestari, and Y. N. Hudaya, "Ujaran Kebencian di Media Sosial: Ditinjau dari Kematangan Emosi Dengan Kecerdasan Moral sebagai Mediator," Jurnal Psikologi Integratif, vol. 11, no. 2, pp. 205–218, Oct. 2023, doi: 10.14421/jpsi.v11i2.2750.

[6] L. al Hakim and S. H. Anshori, "Konektivitasi Hate Speech, Hoaks, Media Mainstream dan Pengaruhnya Bagi Sosial Islam Indonesia," Jurnal Dakwah dan Komunikasi, vol. 6, no. 2, pp. 149–168, Dec. 2021, doi: 10.29240/jdk.v6i2.3675.

[7] I. Abdullah, H. Jubba, S. Z. Qudsy, M. Pabbajah, and Z. H. Prasojo, "The Use and Abuse of Internet Spaces: Fitna, Desacralization, and Conflict in Indonesia's Virtual Reality," Cosmopolitan Civil Societies: An Interdisciplinary Journal, vol. 16, no. 3, pp. 1–12, Dec. 2024, doi: 10.5130/ccs.v16.i3.8962.

[8] Y. Lestari, N. Elian, Diego, A. Anindya, and R. F. Helmi, "The Relationship Between Social Media Usage and Responses to Hoax and Hate Speech in Padang," Studies in Media and Communication, vol. 12, no. 3, pp. 393–404, Sep. 2024, doi: 10.11114/smc.v12i3.6682.

[9] L. Espinosa Anke et al., "Hate speech detection: A solved problem? The challenging case of long tail on Twitter," Semant. Web, vol. 10, no. 5, pp. 925–945, Jan. 2019, doi: 10.3233/SW-180338.

[10] G. Kovács, P. Alonso, and R. Saini, "Challenges of Hate Speech Detection in Social Media," SN Computer Science, vol. 2, no. 2, p. 95, 2021, doi: 10.1007/s42979-021-00457-3.

[11] S. D. A. Putri, M. O. Ibrohim, and I. Budi, "Abusive language and hate speech detection for Javanese and Sundanese languages in tweets: Dataset and preliminary study," in 2021 11th International Workshop on Computer Science and Engineering, WCSE 2021, in 2021 11th International Workshop on Computer Science and Engineering, WCSE 2021. International Workshop on Computer Science and Engineering (WCSE), 2021, pp. 461–465. doi: 10.18178/wcse.2021.02.011.

[12] T. T. A. Putri, S. Sriadhi, R. D. Sari, R. Rahmadani, and H. D. Hutahaean, "A comparison of classification algorithms for hate speech detection," IOP Conference Series: Materials Science and Engineering, vol. 830, no. 3, p. 32006, Apr. 2020, doi: 10.1088/1757-899X/830/3/032006.

[13] P. S. Br Ginting, B. Irawan, and C. Setianingsih, "Hate Speech Detection on Twitter Using Multinomial Logistic Regression Classification Method," in 2019 IEEE International Conference on Internet of Things and Intelligence System (IoTaIS), 2019, pp. 105–111. doi: 10.1109/IoTaIS47347.2019.8980379.

[14] T. L. Sutejo and D. P. Lestari, "Indonesia Hate Speech Detection Using Deep Learning," in 2018 International Conference on Asian Language Processing (IALP), 2018, pp. 39–43. doi: 10.1109/IALP.2018.8629154.

[15] J. Patihullah and E. Winarko, "Hate Speech Detection for Indonesia Tweets Using Word Embedding and Gated Recurrent Unit," IJCCS (Indonesian Journal of Computing and Cybernetics Systems), vol. 13, no. 1, pp. 43–52, Jan. 2019, doi: 10.22146/ijccs.40125.

[16] D. A. N. Erlani and E. B. Setiawan, "Hate Comment Detection on Twitter Using Long Short Term Memory (LSTM) With Genetic Algorithm (GA)," Eduvest – Journal of Universal Studies, vol. 4, no. 11, pp. 10191–10201, Nov. 2024, doi: 10.59188/eduvest.v4i11.1758.

[17] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," CoRR, vol. abs/2011.00677, 2020, [Online]. Available: https://arxiv.org/abs/2011.00677

[18] F. Koto, J. H. Lau, and T. Baldwin, "IndoBERTweet: A Pretrained Language Model for Indonesian Twitter with Effective Domain-Specific Vocabulary Initialization," in Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, M.-F. Moens, X. Huang, L. Specia, and S. W. Yih, Eds., Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 10660–10668. doi: 10.18653/v1/2021.emnlp-main.833.

[19] A. Marpaung, R. Rismala, and H. Nurrahmi, "Hate Speech Detection in Indonesian Twitter Texts using Bidirectional Gated Recurrent Unit," in 2021 13th International Conference on Knowledge and Smart Technology (KST), 2021, pp. 186–190. doi: 10.1109/KST51265.2021.9415760.

[20] S. Lintang, "IndoBERT: Transformer-based Model for Indonesian Language," Yogyakarta, 2020. [Online]. Available: https://etd.repository.ugm.ac.id/penelitian/detail/190630.

[21] J. F. Kusuma and A. Chowanda, "Indonesian Hate Speech Detection Using IndoBERTweet and BiLSTM on Twitter," International Journal on Informatics Visualization, vol. 7, no. 3, pp. 773–780, Sep. 2023, doi: 10.30630/joiv.7.3.1035.

[22] S.-Y. Liu et al., "DoRA: weight-decomposed low-rank adaptation," in Proceedings of the 41st International Conference on Machine Learning, in ICML'24. JMLR.org, 2024.

[23] V. B. Parthasarathy, A. Zafar, A. I. khan, and A. Shahid, "The Ultimate Guide to *Fine-tuning* LLMs from Basics to BreakthrougUK: An Exhaustive Review of Technologies, Research, Best Practices, Applied Research Challenges and Opportunities," ArXiv, vol. abs/2408.13296, 2024, [Online]. Available:https://api.semanticscholar.org/CorpusID:271956978.

[24] M. O. Ibrohim and I. Budi, "Multi-label Hate Speech and Abusive Language Detection in Indonesian Twitter," in Proceedings of the Third Workshop on Abusive Language Online, S. T. Roberts, J. Tetreault, V. Prabhakaran, and Z. Waseem, Eds., Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 46–57. doi: 10.18653/v1/W19-3506.

[25] E. J. Hu et al., "LoRA: Low-Rank Adaptation of Large Language Models," CoRR, vol. abs/2106.09685, 2021, [Online]. Available: https://arxiv.org/abs/2106.09685.

[26] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLORA: efficient finetuning of quantized LLMs," in Proceedings of the 37th International Conference on Neural Information Processing Systems, in NIPS '23. Red Hook, NY, USA: Curran Associates Inc., 2023.

[27] Z. Wu et al., "ReFT: Representation Finetuning for Language Models." 2024. [Online]. Available: https://arxiv.org/abs/2404.03592.

BIODATA PENULIS

Penulis Pertama
David Suharjanto memperoleh gelar Sarjana di bidang Fisika dari Universitas Negeri Yogyakarta pada tahun 2024. Saat ini, ia sedang menempuh pendidikan Magister Informatika di UIN Sunan Kalijaga Yogyakarta. Minat penelitiannya adalah *deep learning*.

Penulis Kedua
Sumarsono memperoleh gelar Sarjana dari Universitas Ahmad Dahlan pada tahun 2000, Magister Komputer dari Universitas Gadjah Mada pada tahun 2004, dan gelar Doktor dari Universitas Islam Negeri Sunan Kalijaga pada tahun 2023. Saat ini, beliau merupakan dosen pada Program Studi Informatika, Universitas Islam Negeri Sunan Kalijaga. Minat penelitian beliau adalah sistem informasi.