

**IMPLEMENTASI MODEL DEBERTA-V3 UNTUK
MENGOREKSI SUSUNAN KATA TERTUKAR
DALAM KALIMAT BAHASA INDONESIA**



SKRIPSI

KEVIN CHAWANDI

00000045270

**PROGRAM STUDI INFORMATIKA
FAKULTAS TEKNIK DAN INFORMATIKA
UNIVERSITAS MULTIMEDIA NUSANTARA
TANGERANG
2026**

**IMPLEMENTASI MODEL DEBERTA-V3 UNTUK
MENGOREKSI SUSUNAN KATA TERTUKAR
DALAM KALIMAT BAHASA INDONESIA**



SKRIPSI

Diajukan sebagai salah satu syarat untuk memperoleh
Gelar Sarjana Komputer (S.Kom.)

KEVIN CHAWANDI

00000045270

UMN

UNIVERSITAS

MULTIMEDIA

NUSANTARA

**PROGRAM STUDI INFORMATIKA
FAKULTAS TEKNIK DAN INFORMATIKA
UNIVERSITAS MULTIMEDIA NUSANTARA**

TANGERANG

2026

HALAMAN PERNYATAAN TIDAK PLAGIAT

Dengan ini saya,

Nama : Kevin Chawandi

Nomor Induk Mahasiswa : 00000045270

Program Studi : Informatika

Skripsi dengan judul:

Implementasi Model DeBERTa-V3 untuk Mengoreksi Susunan Kata Tertukar dalam Kalimat Bahasa Indonesia

merupakan hasil karya saya sendiri bukan plagiat dari laporan karya tulis ilmiah yang ditulis oleh orang lain, dan semua sumber, baik yang dikutip maupun dirujuk, telah saya nyatakan dengan benar serta dicantumkan di Daftar Pustaka.

Jika di kemudian hari terbukti ditemukan kecurangan/penyimpangan, baik dalam pelaksanaan maupun dalam penulisan laporan karya tulis ilmiah, saya bersedia menerima konsekuensi dinyatakan TIDAK LULUS untuk mata kuliah yang telah saya tempuh.

Tangerang, 8 Januari 2026



(Kevin Chawandi)

UMMN
UNIVERSITAS
MULTIMEDIA
NUSANTARA

HALAMAN PERNYATAAN PENGGUNAAN BANTUAN KECERDASAN ARTIFISIAL (AI)

Saya yang bertanda tangan di bawah ini:

Nama Lengkap : Kevin Chawandi
NIM : 00000045270
Program Studi : Informatika
Judul Laporan : Implementasi Model DeBERTa-V3 untuk
Mengoreksi Susunan Kata Tertukar dalam Kalimat
Bahasa Indonesia

Dengan ini saya menyatakan secara jujur menggunakan bantuan Kecerdasan Artifisial (AI) dalam pengerjaan Laporan sebagai berikut:

- ☐ Menggunakan AI sebagaimana diizinkan untuk membantu dalam menghasilkan ide-ide utama saja.
- ☐ Menggunakan AI sebagaimana diizinkan untuk membantu menghasilkan teks pertama saja.
- ☒ Menggunakan AI untuk menyempurnakan sintaksis dan tata bahasa untuk pengumpulan tugas.
- ☐ Karena tidak diizinkan: Tidak menggunakan bantuan AI dengan cara apa pun dalam pembuatan tugas

Saya juga menyatakan bahwa:

- (1) Menyerahkan secara lengkap dan jujur penggunaan perangkat AI yang diperlukan dalam tugas melalui Formulir Penggunaan Perangkat Kecerdasan Artifisial (AI).
- (2) Mengakui telah menggunakan bantuan AI dalam tugas saya baik dalam bentuk kata, parafrase, penyertaan ide, atau fakta penting yang disarankan oleh AI dan telah dicantumkan dalam sitasi serta referensi.
- (3) Terlepas dari pernyataan di atas, tugas ini sepenuhnya merupakan karya saya sendiri.

Tangerang, 8 Januari 2026



(Kevin Chawandi)

HALAMAN PENGESAHAN

Skripsi dengan judul

IMPLEMENTASI MODEL DEBERTA-V3 UNTUK MENGOREKSI SUSUNAN KATA TERTUKAR DALAM KALIMAT BAHASA INDONESIA

oleh

Nama : Kevin Chawandi
NIM : 00000045270
Program Studi : Informatika
Fakultas : Fakultas Teknik dan Informatika

Telah diujikan pada hari Rabu, 14 Januari 2026

Pukul 13.00 s/d 15.00 dan dinyatakan

LULUS

Dengan susunan penguji sebagai berikut

Ketua Sidang

Penguji

(Marlinda Vasty Overbeek, S.Kom., M.Kom.)

(Alethea Saryadibrata, S.Kom., M.Eng.)

NIDN: 0818038501

NIDN: 0322099201

Pembimbing

(Moeljono Widjaja, B.Sc., M.Sc., Ph.D.)

NIDN: 0311106903

Ketua Program Studi Informatika,

(Arya Wicaksana, S.Kom., M.Eng.Sc., OCA)

NIDN: 0315109103

HALAMAN PERSETUJUAN PUBLIKASI KARYA ILMIAH UNTUK KEPENTINGAN AKADEMIS

Yang bertanda tangan di bawah ini:

Nama : Kevin Chawandi
NIM : 00000045270
Program Studi : Informatika
Jenjang : S1
Judul Karya Ilmiah : Implementasi Model DeBERTa-
V3 untuk Mengoreksi Susunan
Kata Tertukar dalam Kalimat
Bahasa Indonesia

Menyatakan dengan sesungguhnya bahwa saya bersedia (**pilih salah satu**):

- ☒ Saya bersedia memberikan izin sepenuhnya kepada Universitas Multimedia Nusantara untuk mempublikasikan hasil karya ilmiah saya ke dalam repositori Knowledge Center sehingga dapat diakses oleh Sivitas Akademika UMN/Publik. Saya menyatakan bahwa karya ilmiah yang saya buat tidak mengandung data yang bersifat konfidensial.
- ☐ Saya tidak bersedia mempublikasikan hasil karya ilmiah ini ke dalam repositori Knowledge Center, dikarenakan: dalam proses pengajuan publikasi ke jurnal/konferensi nasional/internasional (dibuktikan dengan *letter of acceptance*) **.
- ☐ Lainnya, pilih salah satu:
 - ☐ Hanya dapat diakses secara internal Universitas Multimedia Nusantara
 - ☐ Embargo publikasi karya ilmiah dalam kurun waktu tiga tahun.

Tangerang, 8 Januari 2026

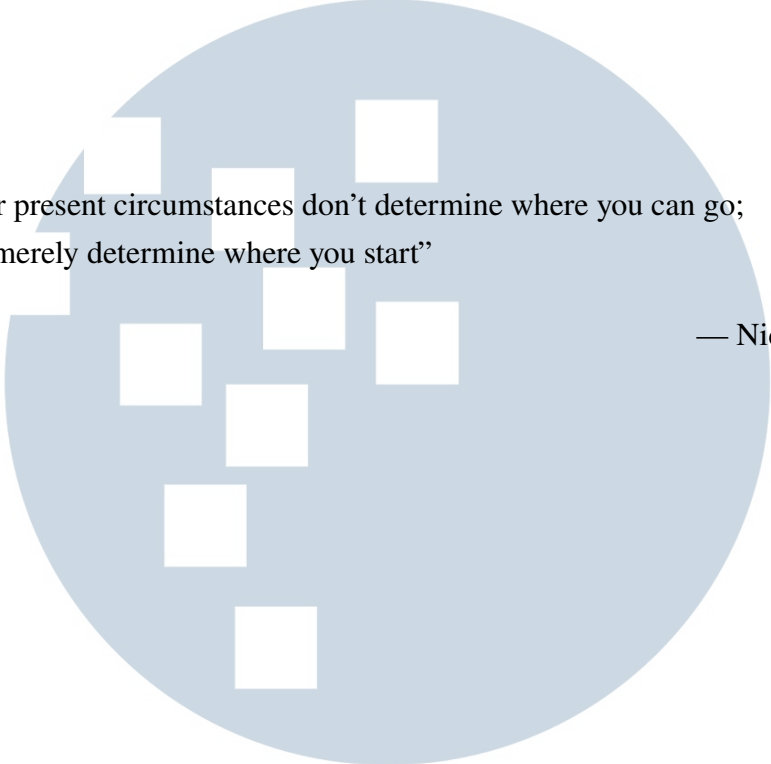
Yang menyatakan



Kevin Chawandi

**Jika tidak bisa membuktikan LoA jurnal/HKI, saya bersedia mengizinkan penuh karya ilmiah saya untuk dipublikasikan ke KC UMN dan menjadi hak institusi UMN.

HALAMAN PERSEMBAHAN / MOTTO



“Your present circumstances don’t determine where you can go;
they merely determine where you start”

— Nido Qubein

UMN
UNIVERSITAS
MULTIMEDIA
NUSANTARA

KATA PENGANTAR

Puji Syukur atas berkat dan rahmat kepada Tuhan Yang Maha Esa, atas selesainya penulisan laporan skripsi ini dengan judul: Implementasi Model DeBERTa-V3 untuk Mengoreksi Susunan Kata Tertukar dalam Kalimat Bahasa Indonesia. Saya menyadari bahwa, tanpa bantuan dan bimbingan dari berbagai pihak sangat sulit bagi saya untuk menyelesaikan laporan ini.

Oleh karena itu, saya mengucapkan terima kasih kepada:

1. Bapak Dr. Ir. Andrey Andoko, M.Sc., selaku Rektor Universitas Multimedia Nusantara.
2. Bapak Dr. Eng. Niki Prastomo, S.T., M.Sc., selaku Dekan Fakultas Teknik dan Informatika Universitas Multimedia Nusantara.
3. Bapak Arya Wicaksana, S.Kom., M.Eng.Sc., OCA, selaku Ketua Program Studi Informatika Universitas Multimedia Nusantara.
4. Bapak Moeljono Widjaja, B.Sc., M.Sc., Ph.D., sebagai Pembimbing yang telah memberikan bimbingan, arahan, dan motivasi atas terselesainya tugas akhir ini.
5. Keluarga saya yang telah memberikan bantuan dukungan material dan moral, sehingga penulis dapat menyelesaikan tugas akhir ini.

Semoga karya ilmiah ini dapat memberikan manfaat bagi para pembaca dan dapat menjadi referensi untuk penelitian selanjutnya.

Tangerang, 8 Januari 2026



Kevin Chawandi

IMPLEMENTASI MODEL DEBERTA-V3 UNTUK MENGOREKSI SUSUNAN KATA TERTUKAR DALAM KALIMAT BAHASA INDONESIA

Kevin Chawandi

ABSTRAK

Kesalahan tata bahasa dalam teks Bahasa Indonesia dapat menurunkan kualitas komunikasi dan efisiensi penulisan, terutama pada media berita dengan jumlah artikel yang besar. Penelitian ini bertujuan membangun model *Large Language Model* berbasis DeBERTa-V3 untuk mengoreksi posisi kata yang tertukar dalam kalimat Bahasa Indonesia. Model di-*pre-training* menggunakan korpus berita KoPI-CC News dan di-*fine-tuning* dengan dataset sintetik berupa penukaran posisi kata. Evaluasi dilakukan menggunakan akurasi, *F1-score*, dan BLEU, menghasilkan akurasi 60,07%, *F1-score* 60,09%, dan BLEU 92,57%. Hasil ini menunjukkan bahwa model DeBERTa-V3 memiliki potensi untuk memperbaiki kesalahan pertukaran posisi kata secara efektif dalam teks Bahasa Indonesia.

Kata kunci: DeBERTa-V3, koreksi pertukaran kata, *Large Language Model*, *pre-training*



Implementation of DeBERTa-V3 Model for Word-Swapping Correction in Indonesian Sentences

Kevin Chawandi

ABSTRACT

Grammatical errors in Indonesian text can reduce communication quality and writing efficiency, especially in news media with high article volume. This study aims to develop a DeBERTa-V3-based Large Language Model to correct swapped word positions in Indonesian sentences. The model was pre-trained on the KoPI-CC News corpus and fine-tuned with a synthetic dataset of word swaps. Sentence-level evaluation shows an accuracy of 60.07%, an F1-score of 60.09%, and a BLEU score of 92.57%. These results demonstrate that DeBERTa-V3 has the potential to effectively correct word order errors in Indonesian text.

Keywords:

DeBERTa-V3, Large Language Model, pre-training, word-swapping correction



DAFTAR ISI

HALAMAN JUDUL	i
PERNYATAAN TIDAK MELAKUKAN PLAGIAT	ii
HALAMAN PERNYATAAN PENGGUNAAN BANTUAN AI	iii
HALAMAN PENGESAHAN	iv
HALAMAN PERSETUJUAN PUBLIKASI KARYA ILMIAH	v
HALAMAN PERSEMBAHAN/MOTO	vi
KATA PENGANTAR	vii
ABSTRAK	viii
ABSTRACT	ix
DAFTAR ISI	x
DAFTAR TABEL	xii
DAFTAR GAMBAR	xiii
DAFTAR KODE	xiv
DAFTAR RUMUS	xv
DAFTAR LAMPIRAN	xvi
BAB 1 PENDAHULUAN	1
1.1 Latar Belakang Masalah	1
1.2 Rumusan Masalah	2
1.3 Batasan Permasalahan	3
1.4 Tujuan Penelitian	3
1.5 Manfaat Penelitian	3
1.6 Sistematika Penulisan	4
BAB 2 LANDASAN TEORI	5
2.1 Mekanisme Attention	5
2.2 Transformer	6
2.2.1 Encoder-Decoder	6
2.2.2 Attention	7
2.3 BERT (Bidirectional Encoder Representations from Transformers)	8
2.4 DeBERTa	9
2.4.1 DeBERTa-V2	9
2.4.2 DeBERTa-V3	10
2.5 Metrik Evaluasi	10
2.5.1 Accuracy	10
2.5.2 F1-score	11
2.5.3 BLEU	12
BAB 3 METODOLOGI PENELITIAN	13
3.1 Metodologi Penelitian	13
3.2 Pengumpulan Data	13
3.2.1 KoPI-CC_News	14
3.2.2 idn-news-az	14
3.3 Pre-Training Model	14
3.3.1 Preprocessing Data	15
3.3.2 Training Pre-trained Model	15
3.3.3 Pengujian/Evaluasi	16
3.4 Fine-tuning Model	16
3.4.1 Preprocessing Data	16
3.4.2 Training Fine-tuned Model	17
3.4.3 Pengujian/Evaluasi	18

BAB 4	HASIL DAN DISKUSI	19
4.1	Proses Pre-Training Model	19
4.1.1	Pengumpulan dan Preprocessing Data	19
4.1.2	Training Model	24
4.2	Evaluasi Pre-Trained Model	29
4.2.1	Evaluasi IndoNLU	29
4.2.2	Analisis Faktor Penyebab Performa	30
4.3	Proses Fine-tuning Model	31
4.3.1	Pembuatan Dataset Kesalahan Sintetis	31
4.3.2	Preprocessing Dataset	34
4.3.3	Training Model GEC	44
4.4	Evaluasi Model GEC	54
4.4.1	Metric Evaluasi	54
4.4.2	Hasil Evaluasi	54
BAB 5	SIMPULAN DAN SARAN	69
5.1	Simpulan	69
5.2	Saran	69
DAFTAR PUSTAKA		70



DAFTAR TABEL

Tabel 4.1	Hasil benchmark IndoNLU (classification)	30
Tabel 4.2	Hasil benchmark IndoNLU (sequence labeling)	30
Tabel 4.3	Hasil evaluasi model berdasarkan jumlah kesalahan	55
Tabel 4.4	Contoh kalimat tanpa kesalahan dengan prediksi benar . . .	56
Tabel 4.5	Contoh kalimat tanpa kesalahan dengan perubahan salah kata	58
Tabel 4.6	Contoh kalimat tanpa kesalahan dengan pertukaran salah .	59
Tabel 4.7	Contoh kalimat satu kesalahan dengan prediksi benar . . .	59
Tabel 4.8	Contoh kalimat satu kesalahan dengan perubahan salah kata	60
Tabel 4.9	Contoh kalimat satu kesalahan dengan pertukaran salah . .	61
Tabel 4.10	Contoh kalimat dua kesalahan dengan prediksi benar . . .	62
Tabel 4.11	Contoh kalimat dua kesalahan dengan perubahan salah kata	63
Tabel 4.12	Contoh kalimat dua kesalahan dengan pertukaran salah . .	63
Tabel 4.13	Contoh kalimat dua kesalahan yang sulit	64
Tabel 4.14	Contoh kalimat tiga kesalahan dengan prediksi benar . . .	64
Tabel 4.15	Contoh kalimat tiga kesalahan dengan perubahan salah kata	65
Tabel 4.16	Contoh kalimat tiga kesalahan dengan pertukaran salah . .	66
Tabel 4.17	Contoh kalimat tiga kesalahan yang sulit	67



DAFTAR GAMBAR

Gambar 2.1	Arsitektur <i>Encoder-Decoder</i> (a) tradisional (b) dengan <i>attention</i>	6
Gambar 2.2	Arsitektur model <i>Transformer</i>	7
Gambar 2.3	<i>Attention</i> dalam <i>Transformer</i>	8
Gambar 3.1	Flowchart metodologi penelitian	13
Gambar 3.2	Flowchart proses pre-training model	15
Gambar 3.3	Flowchart proses fine-tuning model	16
Gambar 4.1	Load dataset <i>acul3/KoPI-CC_News</i>	19
Gambar 4.2	Contoh isi dataset <i>acul3/KoPI-CC_News</i>	20
Gambar 4.3	Hasil <i>deduplicate</i> dataset	20
Gambar 4.4	Split dataset (train-test)	21
Gambar 4.5	Mengubah <i>test</i> set menjadi <i>validation</i> set	21
Gambar 4.6	Load <i>tokenizer</i> <i>DebertaV2TokenizerFast</i>	22
Gambar 4.7	Output <i>tokenize_data</i> dan <i>save file dataset</i>	24
Gambar 4.8	Isi folder dataset yang telah di download	25
Gambar 4.9	Isi konfigurasi <i>hyperparameter</i> model	26
Gambar 4.10	Grafik train dan validation loss	29
Gambar 4.11	Probabilitas distribusi menentukan jumlah kesalahan	32
Gambar 4.12	Hasil pembuatan dataset sintetis	34
Gambar 4.13	Hasil setiap kalimat menjadi <i>row</i> pada dataset sintetis	35
Gambar 4.14	Hasil <i>deduplicate</i> dataset sintetis	36
Gambar 4.15	Statistik deskriptif panjang teks untuk dataset	36
Gambar 4.16	Jumlah dataset terfilter berdasarkan panjang teks	37
Gambar 4.17	Statistik deskriptif panjang teks untuk dataset setelah difilter	37
Gambar 4.18	Dataset setelah tokenisasi	38
Gambar 4.19	Statistik deskriptif panjang token untuk dataset	39
Gambar 4.20	Jumlah dataset terfilter berdasarkan panjang token	40
Gambar 4.21	Statistik deskriptif panjang token untuk dataset setelah filter	40
Gambar 4.22	Range <i>bins</i> panjang token	40
Gambar 4.23	Mengubah <i>bins</i> menjadi <i>IntervalIndex</i> range panjang token	40
Gambar 4.24	Output saat proses <i>stratified splitting</i>	43
Gambar 4.25	Dataset hasil <i>stratified splitting</i>	44
Gambar 4.26	Statistik deskriptif untuk dataset hasil <i>stratified splitting</i>	44
Gambar 4.27	Load dataset sintetis	45
Gambar 4.28	Arsitektur model bagian encoder	47
Gambar 4.29	Arsitektur model bagian decoder	48
Gambar 4.30	Distribusi dataset pada berbagai panjang token	49
Gambar 4.31	Dataset setelah difilter berdasarkan panjang token	49
Gambar 4.32	Grafik progres <i>training</i> model	53

DAFTAR KODE

Kode 4.1	<i>Deduplicate</i> dengan <code>drop_duplicates</code>	20
Kode 4.2	Fungsi <i>tokenize_data</i> untuk <i>tokenize</i> setiap set	22
Kode 4.3	Menghasilkan <i>output</i> file untuk dataset <i>pre-training</i>	23
Kode 4.4	Download dataset untuk <i>pre-training</i>	25
Kode 4.5	Konfigurasi Training Argument	27
Kode 4.6	Isi dari <code>cliargs</code>	27
Kode 4.7	Download checkpoint dari iterasi sebelumnya	28
Kode 4.8	Konfigurasi tambahan untuk melanjutkan training	28
Kode 4.9	Class untuk membuat error sintetis	32
Kode 4.10	Pembuatan dataset sintetis dari dataset <i>esteler-ai/idn-news-az</i>	33
Kode 4.11	Membuat setiap kalimat menjadi <i>row</i> data	34
Kode 4.12	<i>Deduplicate</i> dengan <code>drop_duplicates</code>	35
Kode 4.13	Menghitung panjang teks pada dataset	36
Kode 4.14	Filter panjang teks pada dataset	37
Kode 4.15	Tokenisasi dataset dan metadata kesalahan sintetis	37
Kode 4.16	Menghitung panjang token pada dataset	39
Kode 4.17	Filter dataset berdasarkan panjang token	39
Kode 4.18	Fungsi untuk menentukan rasio dinamis pembagian dataset	41
Kode 4.19	Proses <i>splitting</i> dataset secara <i>stratified</i> manual	41
Kode 4.20	Menggabungkan kembali hasil <i>stratified splitting</i>	44
Kode 4.21	Penyesuaian konfigurasi encoder dan decoder agar memiliki ukuran dan tokenizer yang sama	45
Kode 4.22	Pengaturan batasan panjang token output	49
Kode 4.23	Konfigurasi <i>training arguments</i> untuk Seq2SeqTrainer	49
Kode 4.24	<i>Data collator</i> untuk <i>training</i>	50
Kode 4.25	Fungsi metrik untuk <i>training</i>	51
Kode 4.26	Menyiapkan fungsi trainer Seq2SeqTrainer	52



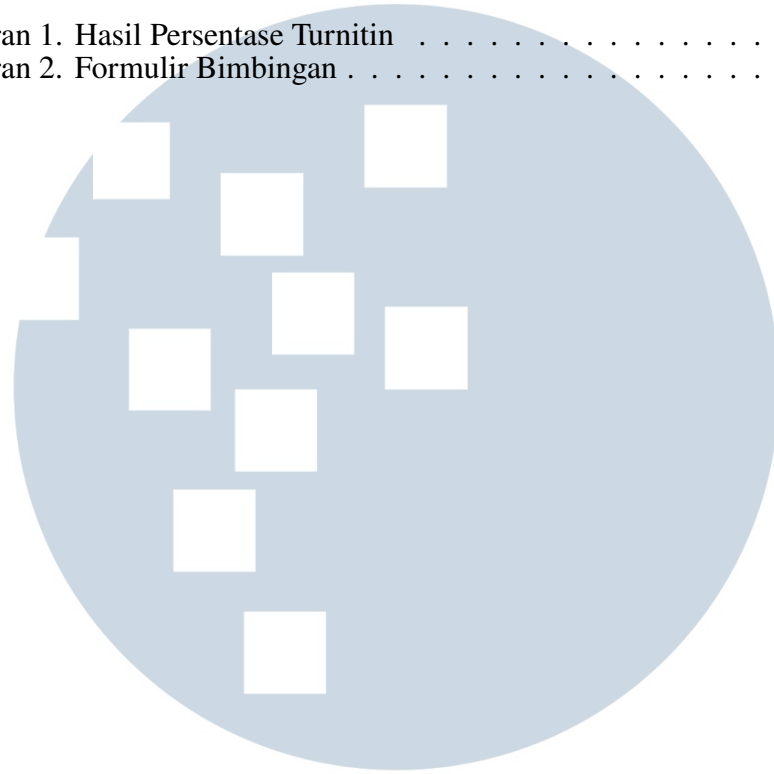
DAFTAR RUMUS

Rumus 2.0	<i>Scaled dot-product attention</i>	7
Rumus 2.3	<i>Accuracy</i>	11
Rumus 2.4	<i>Precision</i>	11
Rumus 2.5	<i>Recall</i>	11
Rumus 2.6	<i>F1-score</i>	11
Rumus 2.7	<i>Brevity Penalty (BLEU)</i>	12
Rumus 2.8	<i>BLEU score</i>	12



DAFTAR LAMPIRAN

Lampiran 1. Hasil Persentase Turnitin	76
Lampiran 2. Formulir Bimbingan	78



UMN
UNIVERSITAS
MULTIMEDIA
NUSANTARA