

**KLASIFIKASI BERITA PALSU BERBAHASA INDONESIA  
MENGUNAKAN ALGORITMA BERT**

**SKRIPSI**

Diajukan untuk memenuhi salah satu syarat mencapai derajat Sarjana  
Program Studi Informatika



disusun oleh

**AFFAN ARDANA**

**20.11.3636**

Kepada

**FAKULTAS ILMU KOMPUTER  
UNIVERSITAS AMIKOM YOGYAKARTA  
YOGYAKARTA  
2023**

**KLASIFIKASI BERITA PALSU BERBAHASA INDONESIA  
MENGUNAKAN ALGORITMA BERT**

**SKRIPSI**

untuk memenuhi salah satu syarat mencapai derajat Sarjana  
Program Studi Informatika



disusun oleh

**AFFAN ARDANA**

**20.11.3636**

Kepada

**FAKULTAS ILMU KOMPUTER  
UNIVERSITAS AMIKOM YOGYAKARTA  
YOGYAKARTA  
2023**

**HALAMAN PERSETUJUAN**

**SKRIPSI**

**KLASIFIKASI BERITA PALSU BERBAHASA INDONESIA  
MENGUNAKAN ALGORITMA BERT**

yang disusun dan diajukan oleh

**Affan Ardana**

**20.11.3636**

telah disetujui oleh Dosen Pembimbing Skripsi  
pada tanggal 20 November 2023

**Dosen Pembimbing,**



**Theopilus Bayu Sasongko, S.Kom, M.Eng**  
**NIK. 190302375**

HALAMAN PENGESAHAN

SKRIPSI

**KLASIFIKASI BERITA PALSU BERBAHASA INDONESIA  
MENGUNAKAN ALGORITMA BERT**

yang disusun dan diajukan oleh

**Affan Ardana**

**20.11.3636**

Telah dipertahankan di depan Dewan Penguji  
pada tanggal 20 November 2023

**Susunan Dewan Penguji**

**Nama Penguji**

**Tanda Tangan**

**Dr. Ferry Wahyu Wibowo, S.Si., M.Cs.**  
NIK. 190302235

**Ferian Fauzi Abdulloh, M.Kom**  
NIK. 190302276

**Theopilus Bayu Sasongko, S.Kom, M.Eng**  
NIK. 190302375



Skripsi ini telah diterima sebagai salah satu persyaratan  
untuk memperoleh gelar Sarjana Komputer  
Tanggal 20 November 2023

**DEKAN FAKULTAS ILMU KOMPUTER**



**Hanif Al Fatta, S.Kom., M.Kom.**  
NIK. 190302096

## HALAMAN PERNYATAAN KEASLIAN SKRIPSI

Yang bertandatangan di bawah ini,

Nama mahasiswa : Affan Ardana  
NIM : 20.11.3636

Menyatakan bahwa Skripsi dengan judul berikut:

**Klasifikasi Berita Palsu Berbahasa Indonesia Menggunakan Algoritma BERT**

Dosen Pembimbing : Theopilus Bayu Sasongko, S.Kom, M.Eng

1. Karya tulis ini adalah benar-benar **ASLI** dan **BELUM PERNAH** diajukan untuk mendapatkan gelar akademik, baik di Universitas AMIKOM Yogyakarta maupun di Perguruan Tinggi lainnya.
2. Karya tulis ini merupakan gagasan, rumusan dan penelitian **SAYA** sendiri, tanpa bantuan pihak lain kecuali arahan dari Dosen Pembimbing.
3. Dalam karya tulis ini tidak terdapat karya atau pendapat orang lain, kecuali secara tertulis dengan jelas dicantumkan sebagai acuan dalam naskah dengan disebutkan nama pengarang dan disebutkan dalam Daftar Pustaka pada karya tulis ini.
4. Perangkat lunak yang digunakan dalam penelitian ini sepenuhnya menjadi tanggung jawab **SAYA**, bukan tanggung jawab Universitas AMIKOM Yogyakarta.
5. Pernyataan ini **SAYA** buat dengan sesungguhnya, apabila di kemudian hari terdapat penyimpangan dan ketidakbenaran dalam pernyataan ini, maka **SAYA** bersedia menerima **SANKSI AKADEMIK** dengan pencabutan gelar yang sudah diperoleh, serta sanksi lainnya sesuai dengan norma yang berlaku di Perguruan Tinggi.

Yogyakarta, 20 November 2023

Yang Menyatakan,



Affan Ardana

## KATA PENGANTAR

Puji syukur kehadirat Allah SWT atas limpahan rahmat dan hidayah-Nya, sehingga penulis dapat menyelesaikan penelitian berupa skripsi ini dengan judul **“Klasifikasi Berita Palsu Berbahasa Indonesia Menggunakan Algoritma BERT”** sebagai salah satu syarat menyelesaikan Program Sarjana Informatika Universitas Amikom Yogyakarta.

Skripsi ini dapat terselesaikan dengan baik tidak lepas dari bimbingan, bantuan, dan doa dari berbagai pihak. Maka dari itu, penulis mengucapkan terima kasih kepada:

1. Allah SWT yang telah memberikan rahmat kepada penulis sehingga dapat menyelesaikan skripsi ini.
2. Bapak Theopilus Bayu Sasongko, S.Kom, M.Eng selaku dosen pembimbing yang telah memberikan arahan dan bimbingan kepada penulis dalam pengerjaan skripsi.
3. Kedua orang tua penulis yang selalu memberikan doa dalam menyelesaikan skripsi ini.

Penulis menyadari bahwa skripsi ini masih jauh dari kesempurnaan, sehingga penulis mengharapkan kritik dari pembaca. Semoga skripsi ini memberikan manfaat bagi pembaca.

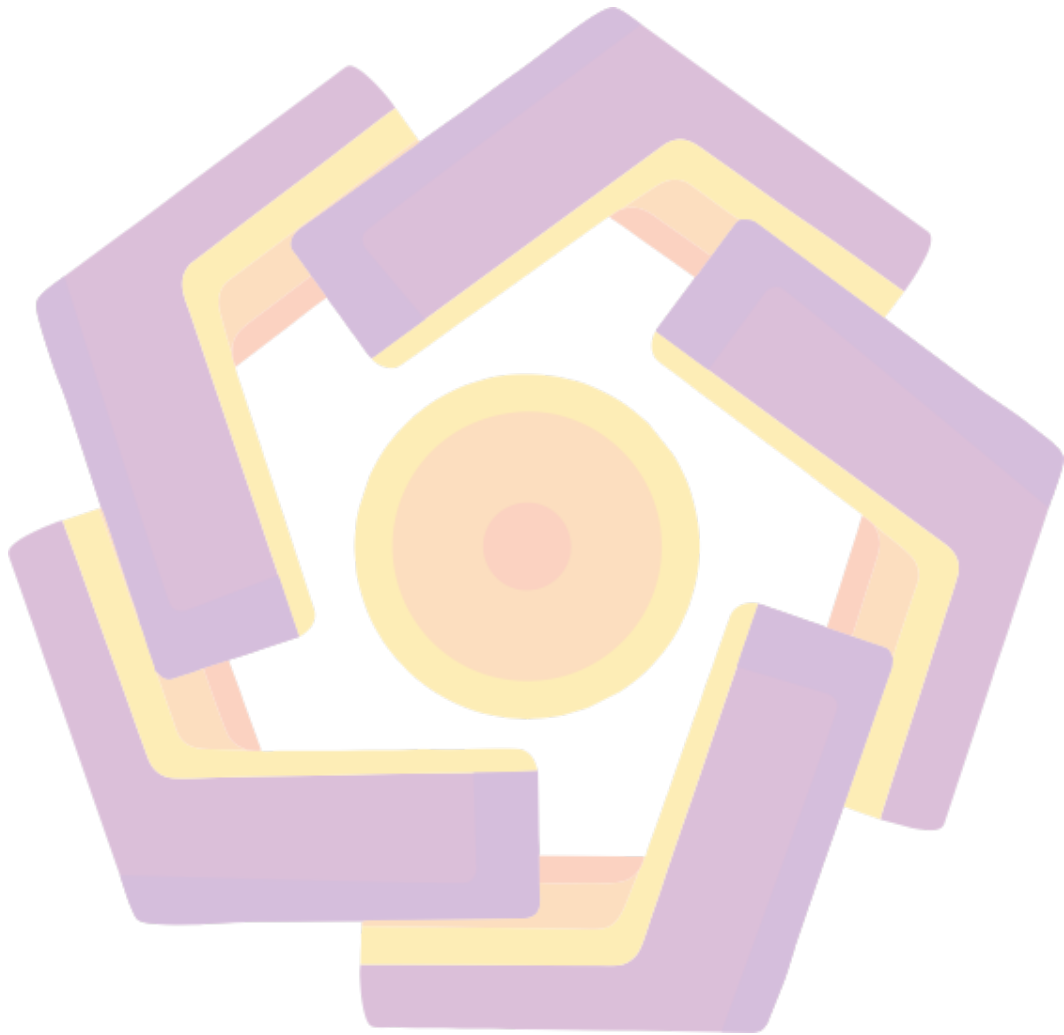
Yogyakarta, 20 November 2023

Penulis

## DAFTAR ISI

HALAMAN JUDUL .....	i
HALAMAN PERSETUJUAN.....	ii
HALAMAN PENGESAHAN .....	iii
HALAMAN PERNYATAAN KEASLIAN SKRIPSI .....	iv
KATA PENGANTAR .....	v
DAFTAR ISI.....	vi
DAFTAR TABEL.....	viii
DAFTAR GAMBAR .....	ix
DAFTAR LAMPIRAN.....	x
INTISARI .....	xi
ABSTRACT.....	xii
<b>BAB I PENDAHULUAN .....</b>	<b>1</b>
1.1 Latar Belakang .....	1
1.2 Rumusan Masalah.....	2
1.3 Batasan Masalah .....	3
1.4 Tujuan Penelitian .....	3
1.5 Manfaat Penelitian .....	3
1.6 Sistematika Penulisan .....	3
<b>BAB II TINJAUAN PUSTAKA .....</b>	<b>5</b>
2.1 Studi Literatur .....	5
2.2 Dasar Teori .....	10
<b>BAB III METODE PENELITIAN .....</b>	<b>27</b>
3.1 Alur Penelitian .....	27
3.2 Alat dan Bahan.....	31
<b>BAB IV HASIL DAN PEMBAHASAN .....</b>	<b>33</b>
4.1 Pengumpulan Data.....	33
4.2 Eksplorasi Data .....	36
4.3 Pra-Pemrosesan.....	36
4.4 Pelatihan.....	44
4.5 Evaluasi.....	47
<b>BAB V PENUTUP .....</b>	<b>53</b>
5.1 Kesimpulan .....	53
5.2 Saran .....	53
<b>REFERENSI .....</b>	<b>54</b>

LAMPIRAN.....57





## DAFTAR TABEL

Tabel 2.1 Keaslian Penelitian .....	7
Tabel 4.1 Contoh Dataset A .....	33
Tabel 4.2 Contoh Dataset B .....	35
Tabel 4.3 Hasil Pengodean Label Dataset A.....	37
Tabel 4.4 Hasil Tokenisasi.....	37
Tabel 4.5 Lanjutan .....	38
Tabel 4.6 <i>Vocabulary</i> IndoBERT <sub>BASE</sub> .....	39
Tabel 4.7 Proses <i>Embedding</i> .....	41
Tabel 4.8 Hasil <i>embedding</i> yang telah dilakukan <i>padding</i> .....	42
Tabel 4.9 Arsitektur lapisan <i>embedding</i> .....	44
Tabel 4.10 Arsitektur blok <i>encoder</i> .....	45
Tabel 4.11 Arsitektur lapisan klasifikasi.....	46
Tabel 4.12 Hasil performa IndoBERT <sub>BASE</sub> pada Dataset A.....	47
Tabel 4.13 Hasil performa IndoBERT <sub>BASE</sub> pada Dataset B.....	47
Tabel 4.14 Hasil <i>confusion matrix</i> pada Dataset A.....	48
Tabel 4.15 Hasil <i>confusion matrix</i> pada Dataset B.....	48
Tabel 4.16 Waktu pelatihan .....	49
Tabel 4.17 Perbandingan performa model pada Dataset A .....	50
Tabel 4.18 Perbandingan performa model pada Dataset B.....	52

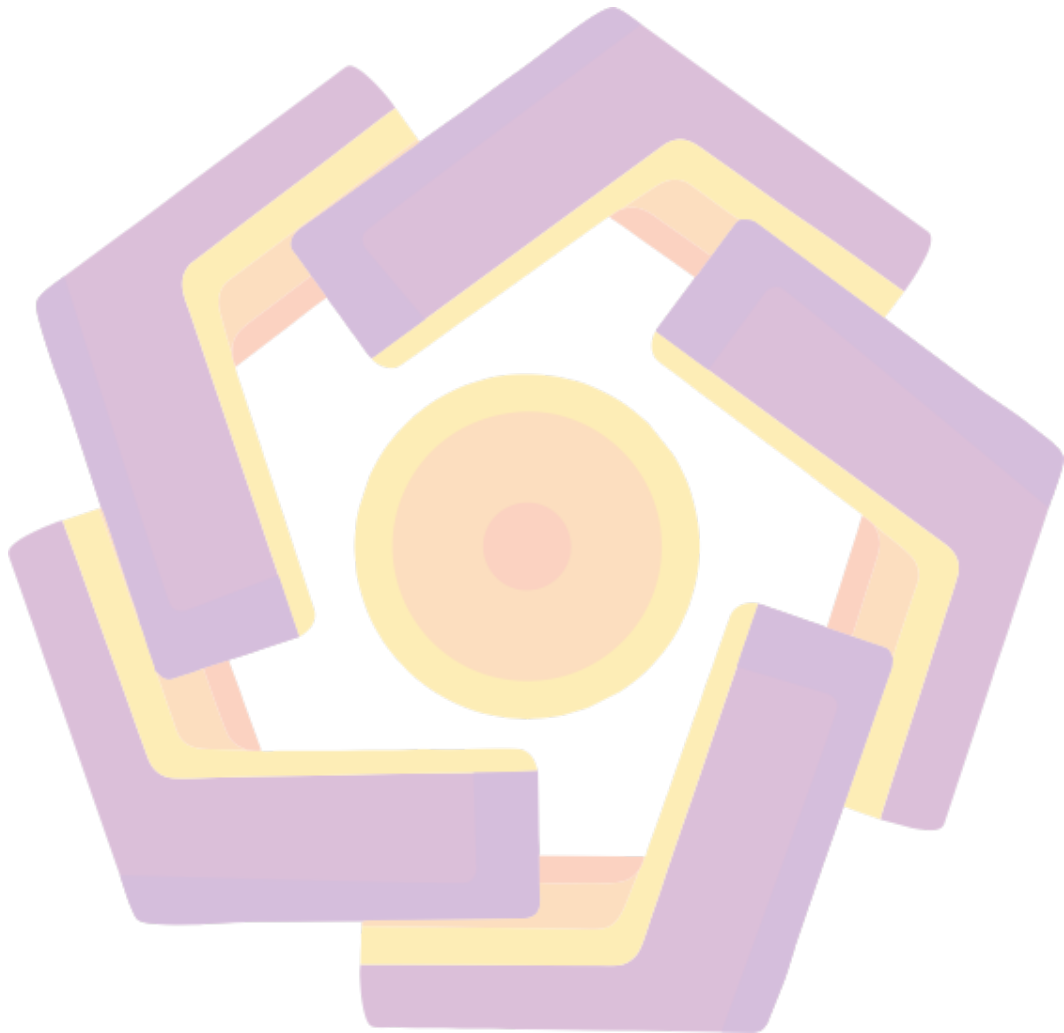
## DAFTAR GAMBAR

Gambar 2.1 Ilustrasi dari sebuah LSTM <i>cell</i> [30].	15
Gambar 2.2 Perbandingan komponen pada LSTM dan GRU[33].	16
Gambar 2.3 Arsitektur <i>Transformer</i> [15].	17
Gambar 2.4 Ilustrasi mekanisme <i>attention</i> [15].	19
Gambar 2.5 Representasi masukan pada BERT[18].	21
Gambar 2.6 Ilustrasi <i>fine-tuning</i> untuk tugas klasifikasi[18].	23
Gambar 2.7 Contoh format <i>confusion matrix</i> [35].	24
Gambar 3.1 Alur penelitian.	28
Gambar 3.2 Alur <i>embedding</i> .	30
Gambar 3.3 Alur pelatihan.	31
Gambar 4.1 Distribusi Label.	36
Gambar 4.2 <i>Loss Head-Only</i> Dataset A.	50
Gambar 4.3 <i>Loss Tail-Only</i> Dataset A.	51
Gambar 4.4 <i>Loss Head+Tail</i> Dataset A.	52



## DAFTAR LAMPIRAN

Lampiran 1 Kode sumber untuk pelatihan.....	57
---	----



## INTISARI

Berita palsu menyebar dengan cepat karena akses internet yang mudah. Klasifikasi menggunakan *deep learning* dapat digunakan sebagai cara untuk mengidentifikasi sebuah berita. Arsitektur *deep learning* yang bisa digunakan untuk tugas klasifikasi adalah arsitektur *recurrent*. Namun, arsitektur *recurrent* memiliki keterbatasan yaitu pemahaman konteks yang kurang baik. Arsitektur *Transformer* dibuat untuk menangani keterbatasan arsitektur *recurrent* tersebut. Salah satu algoritma yang didasarkan pada arsitektur *Transformer* adalah BERT. Penelitian ini bertujuan untuk menggunakan BERT dalam masalah klasifikasi berita palsu berbahasa Indonesia untuk melihat hasil performanya serta perbandingan performa BERT terhadap arsitektur *recurrent*.

Dataset yang digunakan berjumlah dua dataset yang berasal dari penelitian sebelumnya. Dataset A merepresentasikan dataset kecil sedangkan Dataset B merepresentasikan dataset besar. Token yang panjangnya melebihi ketentuan dipotong menggunakan tiga jenis *truncation* yaitu *head-only*, *tail-only*, dan *head+tail*. *K-fold Cross Validation* digunakan sebagai validasi hasil performa BERT. Model *pre-trained* BERT yang digunakan untuk pelatihan adalah IndoBERT<sub>BASE</sub>.

Hasil penelitian menunjukkan bahwa BERT menghasilkan performa terbaik pada *head-only truncation* di Dataset A serta Dataset B. BERT dengan *head-only truncation* menghasilkan akurasi sebesar 67%, *precision* sebesar 53%, *recall* sebesar 48%, dan *f-score* sebesar 45% pada Dataset A serta menghasilkan akurasi sebesar 93%, *precision* sebesar 90%, *recall* sebesar 90%, dan *f-score* sebesar 89% pada Dataset B. Perbandingan BERT dengan model *recurrent* menunjukkan bahwa LSTM dan GRU dapat mengungguli BERT dalam dataset yang kecil. Sedangkan pada dataset yang besar, BERT dapat meningkatkan akurasi sebesar 8%, *precision* 6%, dan *recall* sebesar 7% daripada LSTM.

**Kata kunci:** berita palsu, BERT, penambahan teks.

## ABSTRACT

Fake news spreads quickly due to easy internet access. Classification using deep learning can be used to identify news. The deep learning architecture that can be used for classification tasks is recurrent architecture. However, recurrent architectures have limitations, namely poor context understanding. Transformer architecture was created to address the limitations of the recurrent architecture. One of the algorithms based on Transformer architecture is BERT. This study aims to use BERT in the problem of classification of fake news in Indonesian to see the results of its performance and comparison of BERT performance against recurrent architecture.

The datasets used amounted to two datasets derived from previous research. Dataset A represents a small dataset while Dataset B represents a large dataset. Tokens that exceeded the required length were truncated using three types of truncation: head-only, tail-only, and head+tail. K-fold Cross Validation was used to validate the BERT performance results. The pre-trained BERT model used for training is IndoBERT<sub>BASE</sub>.

The results show that BERT produces the best performance on head-only truncation on Dataset A and Dataset B. BERT with head-only truncation produces an accuracy of 67%, precision of 53%, recall of 48%, and f-score of 45% on Dataset A and produces an accuracy of 93%, precision of 90%, recall of 90%, and f-score of 89% on Dataset B. Comparison of BERT with recurrent models shows that LSTM and GRU can outperform BERT in small datasets. While on large datasets, BERT can improve accuracy by 8%, precision by 6%, and recall by 7% than LSTM.

**Keyword:** fake news, BERT, text mining.