

**KLASIFIKASI BERITA HOAKS DAN FAKTA MENGGUNAKAN
MODEL INDOBERT BERBASIS NATURAL LANGUAGE
PROCESSING (NLP)**

SKRIPSI

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



disusun oleh
SALMAN AL FARISY
22.11.4707

Kepada

**FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA
2026**

**KLASIFIKASI BERITA HOAKS DAN FAKTA MENGGUNAKAN
MODEL INDOBERT BERBASIS NATURAL LANGUAGE
PROCESSING (NLP)**

SKRIPSI

untuk memenuhi salah satu syarat mencapai derajat Sarjana
Program Studi Informatika



disusun oleh

SALMAN AL FARISY

22.11.4707

Kepada

**FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA**

2026

HALAMAN PERSETUJUAN

SKRIPSI

**KLASIFIKASI BERITA HOAKS DAN FAKTA MENGGUNAKAN
MODEL INDOBERT BERBASIS NATURAL LANGUAGE
PROCESSING(NLP)**

yang disusun dan diajukan oleh

Salman Al Farisy

22.11.4707

telah disetujui oleh Dosen Pembimbing Skripsi
pada tanggal 26 Januari 2026

Dosen Pembimbing



Feriati Fauzi Abdulloh, M.Kom
NIK. 190302276

HALAMAN PENGESAHAN

SKRIPSI

**KLASIFIKASI BERITA HOAKS DAN FAKTA MENGGUNAKAN
MODEL INDOBERT BERBASIS NATURAL LANGUAGE
PROCESSING (NLP)**

yang disusun dan diajukan oleh

Salman Al Farisy

22.11.4707

Telah dipertahankan di depan Dewan Penguji
pada tanggal 26 Januari 2026

Susunan Dewan Penguji

Nama Penguji

Tanda Tangan

Rizqi Sukma Kharisma, S.Kom., M.Kom.
NIK. 190302215

Ali Mustopa, S.Kom., M.Kom.
NIK. 190302192

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

Skrripsi ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Sarjana Komputer
Tanggal 26 Januari 2026

DEKAN FAKULTAS ILMU KOMPUTER



Prof. Dr. Kusriani, M.Kom.
NIK. 190302106

HALAMAN PERNYATAAN KEASLIAN SKRIPSI

Yang bertandatangan di bawah ini,

Nama mahasiswa : Salman Al Farisy
NIM : 22.11.4707

Menyatakan bahwa Skripsi dengan judul berikut:

Klasifikasi Berita Hoaks dan Fakta Menggunakan Model IndoBERT Berbasis Natural Language Processing (NLP)

Dosen Pembimbing : Ferian Fauzi Abdulloh, M.Kom

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, 26 Januari 2026

Yang Menyatakan,



Salman Al Farisy

HALAMAN PERSEMBAHAN

Dengan rasa syukur dan hormat, karya ilmiah ini penulis persembahkan kepada:

1. Kedua orang tua penulis, Bapak Darman dan Ibu Suprihatun, atas doa, kasih sayang, dukungan, serta pengorbanan yang sangat tulus sejak awal hingga skripsi ini diselesaikan.
2. Bapak Ferian Fauzi Abdulloh, M.Kom, selaku dosen pembimbing, atas bimbingan, arahan, serta kesabaran yang sudah diberikan selama penyusunan skripsi ini.
3. Seluruh dosen dan staf pengajar Program Studi SI Informatika di Universitas Amikom Yogyakarta, atas ilmu dan pengalaman selama akademik yang telah diberikan selama masa studi.

KATA PENGANTAR

Puji dan Syukur penulis panjatkan ke hadirat Allah SWT atas rahmat dan karunia-Nya sehingga penulis dapat menyelesaikan skripsi yang berjudul **“Klasifikasi Berita Hoaks dan Fakta Menggunakan Model IndoBERT Berbasis Natural Language Processing (NLP)”**. Skripsi ini disusun sebagai salah satu syarat untuk memperoleh gelar sarjana pada Program Studi S1 Informatika, Universitas Amikom Yogyakarta.

Penulis menyadari bahwa proses penyusunan skripsi ini tidak lepas dari bantuan, arahan, serta dukungan. Maka dari itu, pada kesempatan ini penulis menyampaikan ucapan terima kasih kepada:

1. Allah SWT atas rahmat dan petunjuk yang diberikan kepada penulis selama proses penyusunan skripsi ini.
2. Seluruh dosen Fakultas Ilmu Komputer Universitas Amikom Yogyakarta atas ilmu yang diberikan selama penulis selama masa studi.
3. Bapak Ferian Fauzi Abdullah, M.Kom, selaku dosen pembimbing yang telah memberikan arahan, bimbingan, serta masukan yang berharga selama proses penyusunan skripsi ini.
4. Kedua orang tua penulis, Bapak Darman dan Ibu Suprihatun, atas doa, perhatian, serta dukungan yang senantiasa mengiringi langkah penulis.

Penulis menyadari bahwa skripsi ini masih memiliki keterbatasan. Oleh karena itu, penulis selalu terbuka terhadap kritik dan saran sebagai bahan evaluasi serta perbaikan di masa yang akan datang. Semoga skripsi ini dapat memberikan manfaat bagi pembaca dan turut berkontribusi bagi perkembangan ilmu pengetahuan, khususnya di bidang klasifikasi teks.

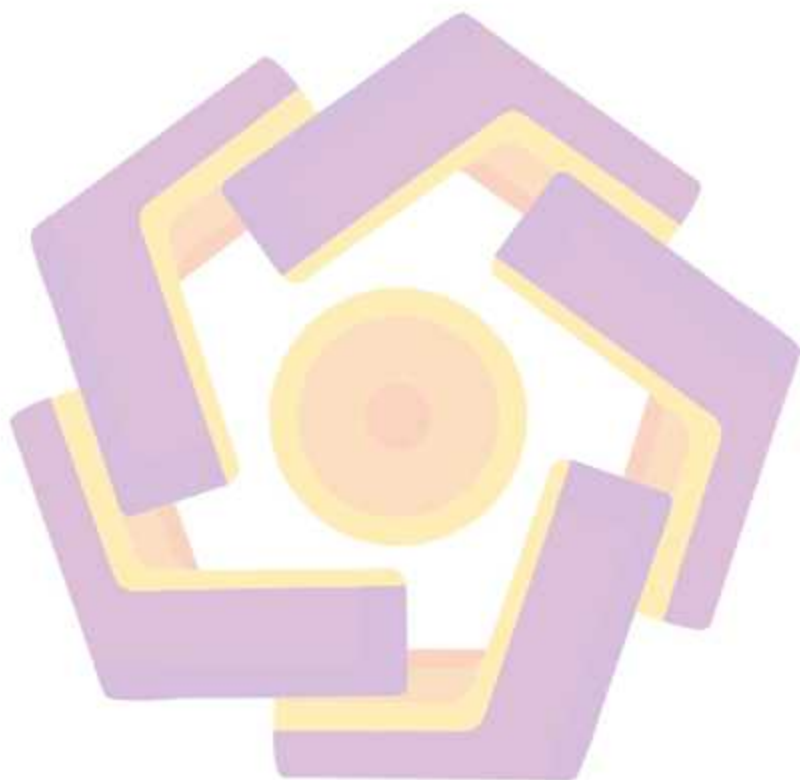
Yogyakarta, 7 Januari 2026

Salman Al Farisy

DAFTAR ISI

HALAMAN JUDUL.....	i
HALAMAN PERSETUJUAN.....	ii
HALAMAN PENGESAHAN.....	iii
HALAMAN PERNYATAAN KEASLIAN SKRIPSI.....	iv
HALAMAN PERSEMBAHAN.....	v
KATA PENGANTAR.....	vi
DAFTAR ISI.....	vii
DAFTAR TABEL.....	x
DAFTAR GAMBAR.....	xi
DAFTAR LAMBANG DAN SINGKATAN.....	xii
DAFTAR ISTILAH.....	xiii
INTISARI.....	xiv
ABSTRACT.....	xv
BAB I PENDAHULUAN.....	1
1.1 Latar Belakang.....	1
1.2 Rumusan Masalah.....	3
1.3 Batasan Masalah.....	3
1.4 Tujuan Penelitian.....	3
1.5 Manfaat Penelitian.....	3
1.6 Sistematika Penulisan.....	4
BAB II TINJAUAN PUSTAKA.....	5
2.1 Studi Literatur.....	5
2.2 Dasar Teori.....	11
2.2.1 Berita Hoaks dan Fakta.....	11
2.2.2 Klasifikasi.....	11
2.2.3 Hubungan Natural Language Processing (NLP) dengan BERT.....	12
2.2.4 IndoBERT.....	14
2.2.5 Arsitektur Transformer.....	15
2.2.6 Fine-Tuning.....	17

2.2.7	Hyperparameter.....	18
2.2.8	K-Fold Cross Validation	19
2.2.9	Evaluasi	19
BAB III METODE PENELITIAN		23
3.1	Alur Penelitian	23
3.1.1	Pengumpulan Data	24
3.1.2	Eksplorasi Data	24
3.1.3	Pra-pemrosesan	24
3.1.4	Pelatihan.....	25
3.1.5	Hyperparameter.....	25
3.1.6	Evaluasi.....	26
3.2	Alat dan Bahan.....	26
BAB IV HASIL DAN PEMBAHASAN		27
4.1	Pengumpulan Data	27
4.2	Eksplorasi Data	28
4.3	Pra-pemrosesan Data.....	30
4.3.1	Preprocessing Teks	31
4.3.2	Tokenisasi	31
4.3.3	Pembuatan Data Tensor	34
4.3.4	K-Fold Cross Validation	35
4.4	Pelatihan.....	36
4.5	Hyperparameter.....	38
4.5.1	Learning Rate.....	39
4.5.2	Batch Size	39
4.5.3	Pengaturan Eksperimen Hyperparameter.....	40
4.6	Evaluasi.....	41
4.6.1	Analisis Akurasi Model	41
4.6.2	Analisis Matrix.....	43
BAB V PENUTUP.....		46
5.1	Kesimpulan	46
5.2	Saran.....	47
REFERENSI		48



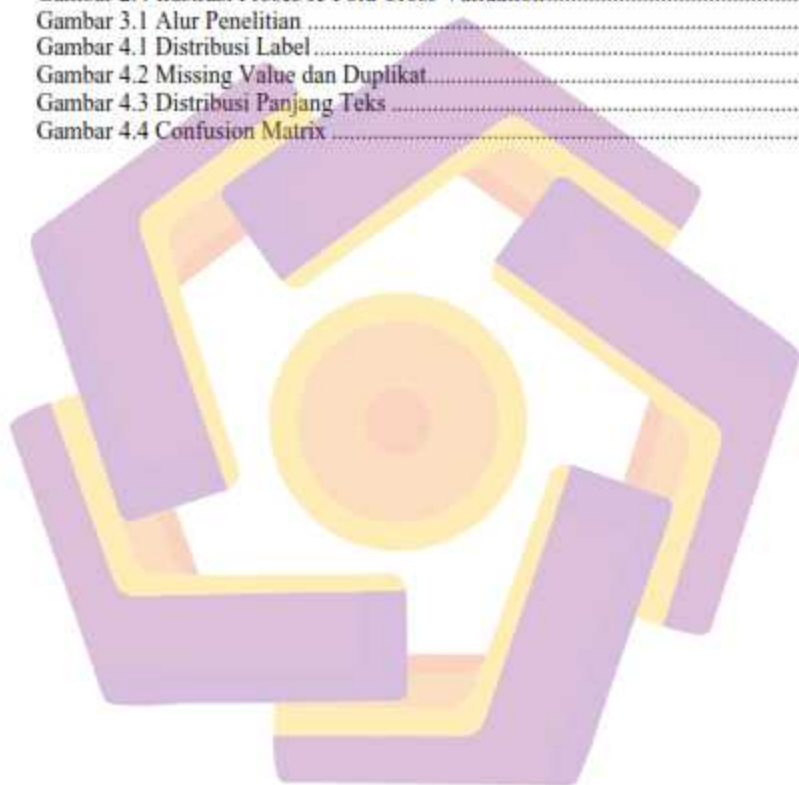
DAFTAR TABEL

Tabel 2.1. Keaslian Penelitian.....	7
Tabel 2.2 Confusion Matrix	20
Tabel 4.1 Contoh Dataset.....	27
Tabel 4.2 Hasil Token.....	32
Tabel 4.3 Hasil vocabulary IndoBERT.....	33
Tabel 4.4 Hasil Attention Mask	35
Tabel 4.5 Proses Pembentukan Embedding.....	36
Tabel 4.6 Representasi Dimensi Vektor Pada Encoder.....	37
Tabel 4.7 Dimensi Representasi Vektor pada Lapisan Klasifikasi.....	38
Tabel 4.8 Kombinasi Hyperparameter.....	40
Tabel 4.9 Urutan Akurasi Teratas	42
Tabel 4.10 <i>Classification Report</i>	44



DAFTAR GAMBAR

Gambar 2.1. Embedding dalam BERT	13
Gambar 2.2. Arsitektur BERT pada Lapisan Encoder.....	16
Gambar 2.3 Ilustrasi Fine-tuning pada Pre-training BERT	18
Gambar 2.4 Ilustrasi Proses K-Fold Cross Validation.....	19
Gambar 3.1 Alur Penelitian	23
Gambar 4.1 Distribusi Label	29
Gambar 4.2 Missing Value dan Duplikat.....	29
Gambar 4.3 Distribusi Panjang Teks	30
Gambar 4.4 Confusion Matrix	43



DAFTAR LAMBANG DAN SINGKATAN



BERT	Bidirectional Encoder Representations from Transformers
b_1	Bias Layer Pertama
b_2	Bias Layer Kedua
d_k	Dimensi Key
FN	False Negative
FP	False Positif
h	Jumlah Head
IndoBERT	Indonesian Bidirectional Encoder Representations from Transformers
K	Key
MLM	Masked Language Modeling
NLP	Natural Language Processing
NSP	Next Sentence Prediction
Q	Query
TN	True Negative
TP	True Positive
V	Value
W_1	Weight Layer Pertama
W_2	Weight Layer Kedua
W_l^k	Weight Key
W_l^v	Weight Value
W_l^o	Weight Output
W_l^q	Weight Query
X	Vektor Input

DAFTAR ISTILAH

Berita Fakta	Informasi asli yang dapat dibuktikan kebenarannya dengan data, bukti, serta sumber resmi yang dapat dipertanggungjawabkan.
Berita Hoaks	Suatu informasi dengan konteks menyesatkan yang bersifat di sengaja atau tidak sengaja yang dapat mengubah cara pandang, tindakan seseorang, persepsi, bahkan emosi.
Dataset	Kumpulan data berita yang digunakan dalam proses pelatihan serta pengujian model.
Embedding	Representasi numerik dari teks sebagai masukan model.
Encoder	Komponen model IndoBERT untuk memproses serta membentuk representasi kontekstual.
Fine-Tuning	Proses menyesuaikan model yang sudah melakukan tahap pre-training untuk menyelesaikan tugas spesifik.
Hyperparameter	Parameter yang ditentukan sebelum proses pelatihan model.
Klasifikasi Teks	Pengelompokan atau mengategorikan data ke kelas tertentu berdasarkan pola dan karakteristik data.
Overfitting	Model menghafal data latih dan buruk menghadapi data uji.
Preprocessing	Proses merubah data mentah menjadi data yang siap guna untuk melakukan pelatihan.
Tensor	Bentuk data yang dihasilkan dari pengolahan dataset teks melalui proses tokenisasi dan pembentukan attention mask sehingga data siap digunakan sebagai masukan model.
Tokenisasi	Tahapan untuk mengubah teks menjadi representasi numerik berupa urutan token ID yang dapat diproses oleh model IndoBERT.
Transformer	Arsitektur model yang dirancang dengan menggunakan mekanisme attention sehingga dapat menangkap hubungan antar kata.

INTISARI

Meningkatnya berita hoaks tersebar di Indonesia mengakibatkan terjadinya kesalahpahaman serta membuat kepercayaan masyarakat menurun pada informasi yang telah diterima, karena di era modernisasi ini perkembangan teknologi termasuk internet yang dapat memudahkan akses untuk menangkap informasi melalui media digital yang berakibatkan masyarakat mengalami kesulitan dalam membedakan antara berita hoaks dan berita fakta. Untuk mengatasi masalah tersebut, penelitian ini mengembangkan sistem klasifikasi yang dapat membedakan antara berita hoaks dan fakta secara akurat.

Penelitian ini menggunakan pendekatan Natural Language Processing (NLP) dengan model IndoBERT. Tahapan dalam penelitian ini meliputi prapemrosesan data yang terdiri preprocessing teks, tokenisasi, serta pembuatan dataset dalam bentuk tensor untuk diproses oleh arsitektur transformer. Proses pelatihan dilakukan melalui fine-tuning IndoBERT dengan melakukan eksperimen dari beberapa kombinasi hyperparameter learning rate ($1e-5$, $2e-5$, $3e-5$, $4e-5$, $5e-5$) dan batch size (8, 16, 32). Evaluasi dilakukan menggunakan metode K-Fold Cross Validation untuk mendapatkan hasil yang stabil serta mengurangi bias, serta metrik berupa akurasi, confusion matrix, dan classification report.

Hasil pada penelitian ini menunjukkan bahwa kombinasi learning rate $2e-5$ dengan batch size 8 menghasilkan performa yang paling optimal dengan akurasi sebesar 0,9887. Temuan ini menandakan bahwa pemilihan hyperparameter yang tepat dapat berpengaruh terhadap kinerja IndoBERT dalam klasifikasi berita hoaks dan fakta. Penelitian ini memberikan kontribusi dalam menentukan konfigurasi hyperparameter yang optimal pada proses fine-tuning IndoBERT serta dapat dimanfaatkan oleh penelitian selanjutnya dan masyarakat sebagai dasar pengembangan sistem deteksi berita hoaks dan fakta yang lebih akurat.

Kata kunci: Klasifikasi, Hoaks, IndoBERT, Fine-tuning, Hyperparameter

ABSTRACT

The increase in hoax news spreading in Indonesia has led to misunderstandings and a decline in public trust in the information received. In this era of modernization, technological developments, including the internet, have made it easier to access information through digital media, resulting in the public finding it difficult to distinguish between hoax news and factual news. To overcome this problem, this study developed a classification system that can accurately distinguish between hoax news and factual news.

This study uses a Natural Language Processing (NLP) approach with the IndoBERT model. The stages in this study include data preprocessing, which consists of text preprocessing, tokenization, and dataset creation in the form of tensors to be processed by the transformer architecture. The training process was carried out through fine-tuning IndoBERT by conducting experiments with several combinations of hyperparameters: learning rate (1e-5, 2e-5, 3e-5, 4e-5, 5e-5) and batch size (8, 16, 32). The evaluation was conducted using the K-Fold Cross Validation method to obtain stable results and reduce bias, as well as metrics such as accuracy, confusion matrix, and classification report.

The results of this study show that the combination of a learning rate of 2e-5 with a batch size of 8 produces the most optimal performance with an accuracy of 0.9887. This finding indicates that the selection of appropriate hyperparameters can affect the performance of IndoBERT in classifying hoax and factual news. This study contributes to determining the optimal hyperparameter configuration in the IndoBERT fine-tuning process and can be utilized by future research and the public as a basis for developing a more accurate hoax and factual news detection system.

Keyword: Classification, Hoax, IndoBERT, Fine-tuning, Hyperparameters