

**IMPLEMENTASI ALGORITMA RANDOM FOREST PADA
MODEL DETEKSI KOMENTAR PELECEHAN SEKSUAL
VERBAL BERBAHASA INDONESIA DI MEDIA SOSIAL X**



SKRIPSI

**LEIDEOVICO YUDHISTI
00000055683**

**PROGRAM STUDI INFORMATIKA
FAKULTAS TEKNIK DAN INFORMATIKA
UNIVERSITAS MULTIMEDIA NUSANTARA
TANGERANG
2025**

**IMPLEMENTASI ALGORITMA RANDOM FOREST PADA
MODEL DETEKSI KOMENTAR PELECEHAN SEKSUAL
VERBAL BERBAHASA INDONESIA DI MEDIA SOSIAL X**



SKRIPSI

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

LEIDEOVICO YUDHISTI

00000055683

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

TANGERANG

2025

HALAMAN PERNYATAAN TIDAK PLAGIAT

Dengan ini saya,

Nama : Leideovico Yudhisti
Nomor Induk Mahasiswa : 00000055683
Program Studi : Informatika

Skripsi dengan judul:

Implementasi Algoritma Random Forest pada Model Deteksi Komentar Pelecehan Seksual Verbal Berbahasa Indonesia di Media Sosial X

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, 4 Juli 2025



(Leideovico Yudhisti)

HALAMAN PENGESAHAN

Skripsi dengan judul

IMPLEMENTASI ALGORITMA RANDOM FOREST PADA MODEL DETEKSI KOMENTAR PELECEHAN SEKSUAL VERBAL BERBAHASA INDONESIA DI MEDIA SOSIAL X

oleh

Nama : Leideovico Yudhisti
NIM : 00000055683
Program Studi : Informatika
Fakultas : Fakultas Teknik dan Informatika

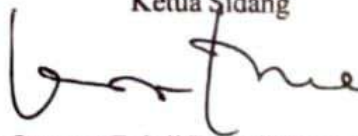
Telah diujikan pada hari Kamis, 10 Juli 2025

Pukul 15.00 s/s 17.00 dan dinyatakan

LULUS

Dengan susunan penguji sebagai berikut

Ketua Sidang



(Dr. Ivransa Zuhdi Pane, B.Eng., M.Eng.) (Eka Jaya Harsono, S.Kom., M.Eng.Sc.)

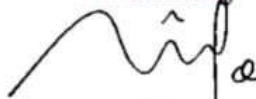
NIDN: 8812520016

Penguji



NIDN: 834377167230333

Pembimbing



(Suwito Pomalingo, S.Kom., M.Kom.)

NIDN: 0911098201

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 : Leidcovico Yudhisti
NIM : 00000055683
Program Studi : Informatika
Jenjang : S1
Judul Karya Ilmiah : Implementasi Algoritma Random
Forest pada Model Deteksi Komentar
Pelecehan Seksual Verbal Berbahasa
Indonesia di Media Sosial X

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, 4 Juli 2025

Yang menyatakan



Leideovico Yudhisti

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



UMN
UNIVERSITAS
MULTIMEDIA
NUSANTARA

HALAMAN PERSEMBAHAN / MOTTO

"A good name is to be more desired than great wealth, Favor is better than silver and gold."

Proverbs 22:1 (NASB)



KATA PENGANTAR

Dengan penuh rasa syukur ke hadirat Tuhan Yang Maha Esa, penulis dapat menyelesaikan Tugas Akhir yang berjudul "Implementasi Algoritma Random Forest pada Model Deteksi Komentar Pelecehan Seksual Verbal Berbahasa Indonesia di Media Sosial X". Penelitian ini bertujuan untuk menganalisis performa algoritma Random Forest dalam mendeteksi komentar pelecehan seksual berbasis teks. Penulis ingin mengucapkan terima kasih yang sebesar-besarnya 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 Suwito Pomalingo, S.Kom., M.Kom., sebagai Pembimbing yang telah memberikan bimbingan, arahan, dan motivasi atas terselesainya tugas akhir ini.
5. Tiga mahasiswa Program Studi Bahasa Indonesia yang telah mendukung proses skripsi ini dalam melakukan pelabelan dataset untuk mendukung penelitian ini.
6. Keluarga saya yang telah memberikan bantuan dukungan material dan moral, sehingga penulis dapat menyelesaikan tugas akhir ini.

Semoga karya ilmiah ini bermanfaat bagi pengembangan metode deteksi pelecehan seksual verbal di media sosial serta menjadi referensi bagi penelitian selanjutnya dalam menerapkan algoritma Random Forest untuk kasus serupa.

Tangerang, 4 Juli 2025



Leideovico Yudhisti

IMPLEMENTASI ALGORITMA RANDOM FOREST PADA MODEL DETEKSI KOMENTAR PELECEHAN SEKSUAL VERBAL BERBAHASA INDONESIA DI MEDIA SOSIAL X

Leideovico Yudhisti

ABSTRAK

Meningkatnya penyebaran komentar pelecehan seksual verbal di media sosial X menjadi salah satu bentuk Kekerasan Berbasis Gender *Online* (KBGO) yang mengancam remaja di Indonesia. Penelitian sebelumnya seringkali memiliki cakupan yang terlalu luas, seperti deteksi seksisme secara umum, dan belum secara spesifik mengukur potensi performa terbaik dari algoritma *Random Forest* pada komentar pelecehan seksual. Penelitian ini bertujuan untuk mengimplementasikan algoritma *Random Forest* dan menemukan performa terbaiknya untuk model deteksi komentar pelecehan seksual verbal berbahasa Indonesia. Metodologi penelitian dimulai dengan pengumpulan 1.000 data komentar dari media sosial X, yang kemudian dilabeli secara manual. Untuk mengatasi masalah kelas tidak seimbang, diterapkan teknik *random undersampling* yang menghasilkan 757 data. Data tersebut kemudian melalui tahap *preprocessing* dan *feature extraction* menggunakan TF-IDF. Eksperimen dilakukan dengan membandingkan model *baseline* dengan model yang telah melalui proses *hyperparameter tuning* menggunakan *GridSearchCV*. Hasil evaluasi menunjukkan bahwa model dengan performa terbaik berhasil mencapai akurasi sebesar 87,50%. Peningkatan paling signifikan terlihat pada metrik *Recall* untuk kelas *Sexual Harassment* yang mencapai 92,86%, menunjukkan kemampuan model yang andal dalam mengidentifikasi komentar pelecehan seksual. Temuan ini menunjukkan algoritma *Random Forest* efektif untuk diimplementasikan pada model deteksi otomatis komentar pelecehan seksual verbal berbahasa Indonesia di media sosial X.

Kata kunci: Algoritma *Random Forest*, Deteksi Pelecehan Seksual, *Hyperparameter Tuning*, KBGO, Klasifikasi Teks, TF-IDF

U N I V E R S I T A S
M U L T I M E D I A
N U S A N T A R A

**IMPLEMENTATION OF THE RANDOM FOREST ALGORITHM IN A
MODEL FOR DETECTING VERBAL SEXUAL HARASSMENT COMMENTS
IN INDONESIAN ON SOCIAL MEDIA X**

Leideovico Yudhisti

ABSTRACT

The increasing spread of verbal sexual harassment comments on social media X has become a form of Online Gender-Based Violence (OGBV) that threatens adolescents in Indonesia. Previous research has often been too broad in scope, such as detecting sexism in general, and has not specifically measured the optimal performance potential of effective algorithms for this particular case. This research aims to implement the Random Forest algorithm and find its best performance for a model that detects Indonesian-language verbal sexual harassment comments. The research methodology began with the collection of 1,000 comments from social media X, which were then manually labeled. To address the class imbalance problem, a random undersampling technique was applied, resulting in a dataset of 757 instances. This data then underwent text preprocessing and feature extraction using TF-IDF. The experiment was conducted by comparing a baseline model with a model optimized through a hyperparameter tuning process using GridSearchCV. Evaluation results show that the best-performing model achieved an accuracy of 87.50%. The most significant improvement was observed in the Recall metric for the Sexual Harassment class, which reached 92.86%, indicating the model's reliable ability to identify harassment cases. These findings demonstrate that a Random Forest algorithm is effective for implementing an automated detection model for verbal sexual harassment comments.

Keywords: *Gender-Based Online Violence, Hyperparameter Tuning, Random Forest Algorithm, Sexual Harassment Detection, Social Media X, Text Classification*

U N I V E R S I T A S
M U L T I M E D I A
N U S A N T A R A

DAFTAR ISI

HALAMAN JUDUL	i
PERNYATAAN TIDAK MELAKUKAN PLAGIAT	ii
HALAMAN PENGESAHAN	iii
HALAMAN PERSETUJUAN PUBLIKASI KARYA ILMIAH	iv
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	3
1.3 Batasan Permasalahan	3
1.4 Tujuan Penelitian	4
1.5 Manfaat Penelitian	4
1.6 Sistematika Penulisan	5
BAB 2 LANDASAN TEORI	7
2.1 Tinjauan Teori	7
2.1.1 Media Sosial dan Remaja	7
2.1.2 Kekerasan Berbasis Gender Online	7
2.1.3 Pelecehan Seksual Berbasis Online	8
2.1.4 Data Collection	9
2.1.5 Data Labelling	10
2.1.6 Resampling	10
2.1.7 Text Preprocessing	10
2.1.8 Feature Extraction	12
2.1.9 Term Frequency-Inverse Document Frequency (TF-IDF)	13
2.1.10 Data Splitting	14
2.1.11 Random Forest	14
2.1.12 Ensemble Learning	18
2.1.13 Hyperparameter Tuning	19
2.1.14 GridSearchCV	19
2.1.15 Evaluasi Model	20
2.2 Jabaran Penelitian Terdahulu	22
BAB 3 METODOLOGI PENELITIAN	30
3.1 Spesifikasi sistem yang Digunakan	32
3.2 Studi Literatur	33
3.3 Data Collection	34
3.4 Data Labelling	35
3.5 Random Undersampling	37
3.6 Text Preprocessing	38
3.7 Feature Extraction	40
3.8 Data Splitting	42

3.9	Random Forest dan Hyperparameter Tuning	43
3.10	Evaluasi Model	46
BAB 4	HASIL DAN DISKUSI	48
4.1	Data Collection	48
4.2	Data Labelling	50
4.3	Random Undersampling	52
4.4	Data Loading dan Persiapan Dataset	53
4.5	Text Preprocessing	55
4.6	Feature Extraction	60
4.7	Data Splitting	64
4.8	Random Forest dengan Default Parameter	66
4.9	Random Forest dengan Best Parameter	72
4.10	Evaluasi Model	74
4.11	Hasil dan Diskusi	82
BAB 5	SIMPULAN DAN SARAN	84
5.1	Simpulan	84
5.2	Saran	84
DAFTAR PUSTAKA		86



DAFTAR TABEL

Tabel 2.1	Tabel Daftar Penelitian Terdahulu	22
Tabel 3.1	Spesifikasi Sistem yang Digunakan	33
Tabel 3.2	Perbandingan Set Parameter Model	45
Tabel 4.1	Perbandingan Distribusi Data Sebelum dan Sesudah Undersampling	52
Tabel 4.2	Hasil Tahap Pembersihan Awal	56
Tabel 4.3	Hasil Tahap Penghapusan Karakter Berulang	57
Tabel 4.4	Hasil Tahap Case Folding & Penghapusan Tanda Baca	57
Tabel 4.5	Hasil Tahap Tokenisasi	58
Tabel 4.6	Hasil Tahap Normalisasi	58
Tabel 4.7	Hasil Tahap Penghapusan Stopwords	59
Tabel 4.8	Hasil Tahap Stemming	59
Tabel 4.9	Contoh Perbandingan Data Sebelum dan Sesudah Keseluruhan Proses Preprocessing	60
Tabel 4.10	Visualisasi Transformasi Komentar ke Vektor TF-IDF	62
Tabel 4.11	Parameter Default dalam Model Baseline Random Forest	67
Tabel 4.12	Ruang Pencarian Hyperparameter	73
Tabel 4.13	Tabulasi Confusion Matrix Model Baseline	76
Tabel 4.14	Tabulasi Confusion Matrix Model Hasil Tuning	77
Tabel 4.15	Perbandingan Metrik Evaluasi Model Baseline dan Tuned	80
Tabel 4.16	Perbandingan Parameter Random Forest	81
Tabel 4.17	Perbandingan Kinerja Model Penelitian dengan Acuan	83



DAFTAR GAMBAR

Gambar 2.1	Cara Kerja <i>Random Forest</i>	16
Gambar 2.2	Tabel Confusion Matrix	20
Gambar 3.1	Diagram tahapan Penelitian	30
Gambar 3.2	Flowchart Data Collection	34
Gambar 3.3	Flowchart Data Labelling	36
Gambar 3.4	Flowchart Random Undersampling	37
Gambar 3.5	Flowchart Text Preprocessing	38
Gambar 3.6	Flowchart Feature Extraction	40
Gambar 3.7	Flowchart Data Splitting	42
Gambar 3.8	Flowchart Modelling Random Forest	44
Gambar 3.9	Flowchart Evaluasi Model	46
Gambar 4.1	Tampilan Sebagian Dataset Hasil Crawling Data	50
Gambar 4.2	Tampilan Antarmuka Proses Pelabelan oleh Tiga Anotator	51
Gambar 4.3	Implementasi Rumus Modus pada Google Sheets	51
Gambar 4.4	Tampilan Dataset Setelah Kolom Tidak Relevan Dihapus	55
Gambar 4.5	Jumlah Total Komentar dalam Dataset	55
Gambar 4.6	Contoh Output Vektor TF-IDF dan Kata dengan Bobot Tertinggi	62
Gambar 4.7	Tabel Distribusi Hasil Pembagian Data Latih dan Uji	66
Gambar 4.8	Visualisasi Pohon Keputusan Pertama dari Model Random Forest	69
Gambar 4.9	Confusion Matrix pada Model Random Forest dengan Parameter Default (Baseline)	75
Gambar 4.10	Confusion Matrix pada Model Random Forest dengan Parameter Hasil Tuning	76



DAFTAR KODE

Kode 4.1	Autentikasi Token dan Instalasi Pustaka	48
Kode 4.2	Definisi Parameter untuk Proses Crawling Data	48
Kode 4.3	Eksekusi Crawling Data dengan Tweet Harvest	49
Kode 4.4	Impor Library Pandas dan NumPy	53
Kode 4.5	Membaca Dataset CSV Menggunakan Pandas	54
Kode 4.6	Menghapus Kolom yang Tidak Diperlukan	54
Kode 4.7	Menampilkan Data untuk Verifikasi	54
Kode 4.8	Menampilkan Jumlah Total Komentar	55
Kode 4.9	Fungsi untuk Pembersihan Awal Teks	56
Kode 4.10	Fungsi untuk Menghapus Karakter Berulang	56
Kode 4.11	Fungsi untuk Case Folding & Penghapusan Tanda Baca	57
Kode 4.12	Fungsi untuk Tokenisasi	57
Kode 4.13	Fungsi untuk Normalisasi Kata	58
Kode 4.14	Fungsi untuk Menghapus Stopwords	59
Kode 4.15	Fungsi untuk Stemming	59
Kode 4.16	Implementasi Ekstraksi Fitur dengan TF-IDF	60
Kode 4.17	Library untuk Data Splitting	64
Kode 4.18	Ekstraksi Fitur dan Label	65
Kode 4.19	Pemisahan Data Latih dan Uji	65
Kode 4.20	Pelatihan Model Random Forest dengan Parameter Default	67
Kode 4.21	Prediksi pada Data Uji	72
Kode 4.22	Set Parameter untuk GridSearchCV	73
Kode 4.23	Proses Hyperparameter Tuning	73
Kode 4.24	Prediksi dengan Model Tuned	74
Kode 4.25	Membuat dan Memvisualisasikan Confusion Matrix untuk Model Baseline	75
Kode 4.26	Membuat dan Memvisualisasikan Confusion Matrix untuk Model Hasil Tuning	76
Kode 4.27	Kode Perhitungan Metrik Evaluasi untuk Model Baseline	78
Kode 4.28	Kode Perhitungan Metrik Evaluasi untuk Model Hasil Tuning	79

UNIVERSITAS
MULTIMEDIA
NUSANTARA

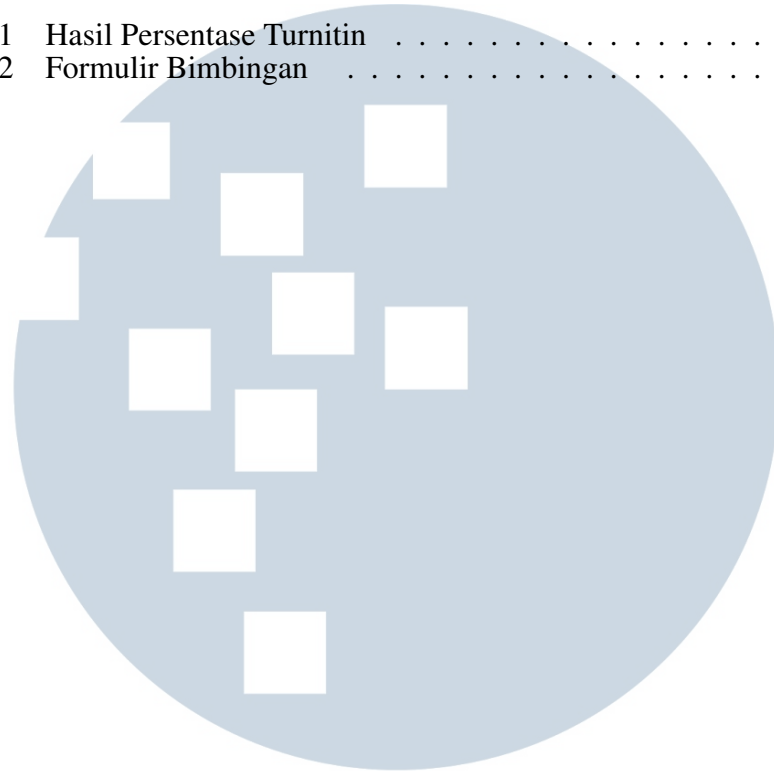
DAFTAR RUMUS

Rumus 2.1	<i>TF-IDF</i>	13
Rumus 2.2	<i>Gini Index</i>	16
Rumus 2.3	Majority Voting	17
Rumus 2.4	<i>Accuracy</i>	21
Rumus 2.5	<i>Precision</i>	21
Rumus 2.6	<i>Recall</i>	21
Rumus 2.7	<i>F1-score</i>	21



DAFTAR LAMPIRAN

Lampiran 1	Hasil Persentase Turnitin	92
Lampiran 2	Formulir Bimbingan	106



UMN
UNIVERSITAS
MULTIMEDIA
NUSANTARA