

**U-TAPIS KLITIK: DETEKSI PENYALAHGUNAAN TULISAN
MENGUNAKAN RANDOM FOREST DAN RULE-BASED**



LAPORAN RESEARCH TECHNOLOGY PROGRAM

**ALFONSUS VEGA HENTYA TANOJO
00000068825**

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

**U-TAPIS KLITIK: DETEKSI PENYALAHGUNAAN TULISAN
MENGUNAKAN RANDOM FOREST DAN RULE-BASED**



LAPORAN RESEARCH TECHNOLOGY PROGRAM

**ALFONSUS VEGA HENTYA TANOJO
00000068825**

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

HALAMAN PERNYATAAN TIDAK PLAGIAT

Dengan ini saya,

Nama : Alfonsus Vega Hentya Tanojo
Nomor Induk Mahasiswa : 00000068825
Program Studi : Informatika

Menyatakan dengan sesungguhnya bahwa Laporan saya yang berjudul:

U-TAPIS Klitik: Deteksi Penyalahgunaan Tulisan Menggunakan Random Forest dan Rule-Based

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, 2 Januari 2026



(Alfonsus Vega Hentya Tanojo)

UMM
UNIVERSITAS
MULTIMEDIA
NUSANTARA

HALAMAN PERNYATAAN PENGGUNAAN BANTUAN KECERDASAN ARTIFISIAL (AI)

Saya yang bertanda tangan di bawah ini:

Nama Lengkap : Alfonsus Vega Hentya Tanojo
NIM : 00000068825
Program Studi : Informatika
Judul Laporan : U-TAPIS Klitik: Deteksi Penyalahgunaan Tulisan
Menggunakan Random Forest dan Rule-Based

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, 7 Januari 2026



(Alfonsus Vega Hentya Tanojo)

HALAMAN PERNYATAAN KEABSAHAN PERUSAHAAN

Yang bertanda tangan di bawah ini:

Nama : Alfonsus Vega Hentya Tanojo
NIM : 00000068825
Program Studi : Informatika
Fakultas : Teknik & Informatika

menyatakan bahwa saya melaksanakan kegiatan di:

Nama Perusahaan/Organisasi : LPPM Universitas Multimedia
Nusantara

Alamat : Jalan Scientia Boulevard
Gading, Curug Sangereng,
Serpong, Kabupaten
Tangerang, Banten 15810

Email Perusahaan/Organisasi : lppm@umn.ac.id

1. Perusahaan/Organisasi tempat saya melakukan kegiatan dapat di validasi keberadaannya.
2. Jika dikemudian hari, terbukti ditemukan kecurangan/penyimpangan data yang tidak valid di perusahaan/organisasi tempat saya melakukan kegiatan, maka:
 - (a) Saya bersedia menerima konsekuensi dinyatakan **TIDAK LULUS** untuk mata kuliah yang telah saya tempuh.
 - (b) Saya bersedia menerima semua sanksi yang berlaku sebagaimana ditetapkan dalam peraturan yang berlaku di Universitas Multimedia Nusantara.

Pernyataan ini saya buat dengan sebenar-benarnya dan digunakan sebagaimana mestinya.

Mengetahui,
Dosen Pembimbing

Tangerang, 2 Januari 2026

Yang menyatakan,



Digitally signed
by Marlinda
Vasty Overbeek
Date: 2026.01.15
15:17:09 +07'00'



Marlinda Vasty Overbeek, S.Kom,
M.Kom

Alfonsus Vega Hentya Tanojo

UMN
UNIVERSITAS
MULTIMEDIA
NUSANTARA

**HALAMAN PERSETUJUAN PUBLIKASI KARYA ILMIAH UNTUK
KEPENTINGAN AKADEMIS**

Yang bertanda tangan di bawah ini:

Nama : Alfonsus Vega Hentya Tanojo
NIM : 00000068825
Program Studi : Informatika
Jenjang : S1
Judul Karya Ilmiah : U-TAPIS Klitik: Deteksi
Penyalahgunaan Tulisan
Menggunakan Random Forest dan
Rule-Based

Menyatakan dengan sesungguhnya bahwa saya bersedia:

- ☒ 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, 2 Januari 2026

Yang menyatakan



Alfonsus Vega Hentya Tanojo

**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

Puji syukur dipanjatkan ke hadirat Tuhan Yang Maha Esa, karena atas rahmat dan karunia-Nya, Tugas Akhir ini yang berjudul “U-TAPIS KLITIK: DETEKSI PENYALAHGUNAAN TULISAN MENGGUNAKAN RANDOM FOREST DAN RULE-BASED” dapat diselesaikan. Penelitian ini disusun dengan tujuan untuk memenuhi salah satu syarat dalam menyelesaikan pendidikan di program studi Informatika sekaligus untuk mengembangkan pemahaman dan kemampuan penelitian di bidang NLP.

Mengucapkan terima kasih

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 Vincentius Kurniawan, S.Kom., M.Eng.Sc., sebagai Pembimbing pertama yang telah memberikan bimbingan, arahan, dan motivasi atas terselesainya tugas akhir ini.
5. Ibu Marlinda Vasty Overbeek, S.Kom, M.Kom, sebagai Pembimbing lapangan yang telah memberikan bimbingan, arahan, dan motivasi atas terselesainya tugas akhir ini.
6. Ibu Dr. Niknik M. Kuntarto, M. Si.dan Bapak Samiaji Bintang Nusantara S.T., M.A., sebagai Pakar bahasa yang telah memberikan bimbingan, arahan, dan motivasi atas terselesainya tugas akhir ini.
7. Bapak Rudi Sutomo, S.Kom., M.Si., M.Kom., sebagai Pengembang *website* yang telah memberikan bimbingan, arahan, dan motivasi atas terselesainya tugas akhir ini.
8. Bapak Moeljono Widjaja, B.Sc., M.Sc., Ph.D., sebagai Pakar data analisis yang telah memberikan bimbingan, arahan, dan motivasi atas terselesainya tugas akhir ini.

9. 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 pengembangan ilmu pengetahuan, khususnya di bidang Natural Language Processing, serta menjadi referensi dan kontribusi bagi penelitian selanjutnya yang berkaitan dengan deteksi kesalahan kebahasaan dalam teks berbahasa Indonesia.

Tangerang, 2 Januari 2026



Alfonsus Vega Hentya Tanojo



U-TAPIS KLITIK: DETEKSI PENYALAHGUNAAN TULISAN MENGUNAKAN RANDOM FOREST DAN RULE-BASED

Alfonsus Vega Hentya Tanojo

ABSTRAK

Kesalahan penggunaan klitik dalam bahasa Indonesia masih sering ditemukan pada teks berita dan tulisan formal, serta sulit dideteksi karena kemiripan strukturnya dengan bentuk yang benar. Penelitian ini mengusulkan sistem deteksi kesalahan klitik berbasis pendekatan hibrida dengan mengombinasikan model klasifikasi *Random Forest* dan verifikasi *rule-based*. Dataset dibangun secara otomatis menggunakan *Python* dengan membentuk kalimat klitik benar dan salah yang bersumber dari kamus bahasa Indonesia, kamus bahasa Inggris, serta dataset akronim, sehingga menghasilkan total 1.028.975 data yang terbagi ke dalam tujuh kelas. Ekstraksi fitur dilakukan menggunakan metode *n-gram*, yaitu *bigram* dan *trigram*, untuk menangkap pola konteks dalam penggunaan klitik. Proses *hyperparameter tuning* diterapkan menggunakan *GridSearchCV* untuk memperoleh konfigurasi terbaik pada model *Random Forest*. Hasil eksperimen menunjukkan bahwa sistem yang diusulkan mencapai akurasi sebesar 0,99 dengan nilai *macro-average* dan *F1-score* sebesar 0,99 yang menandakan performa yang konsisten pada seluruh kelas. Komponen *rule-based* diterapkan sebagai mekanisme pascaproses untuk menangani kasus linguistik ambigu, seperti penggunaan klitik pada akronim dan kata asing. Evaluasi pada kalimat berita nyata menunjukkan bahwa pendekatan hibrida ini efektif dalam mendeteksi kesalahan penggunaan klitik.

Kata kunci: Deteksi Kesalahan Klitik; N-gram; Pemrosesan Bahasa Alami; Random Forest; Rule-Based

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

***U-TAPIS CLITICS: TYPING ERROR DETECTION USING RANDOM
FOREST AND RULE-BASED***

Alfonsus Vega Hentya Tanojo

ABSTRACT

Clitic usage errors frequently occur in Indonesian texts and are difficult to detect due to their structural similarity to correct forms. This study proposes a hybrid clitic error detection system that combines a Random Forest classification model with a rule-based verification approach. The dataset was automatically generated using Python by constructing correct and incorrect clitic sentences derived from Indonesian and English dictionaries as well as acronym datasets, resulting in a total of 1,028,975 data instances across seven classes. Feature extraction was performed using n-gram vectorization with bigram and trigram representations to capture contextual patterns in clitic usage. Hyperparameter tuning was conducted using GridSearchCV to obtain the optimal configuration for the Random Forest model. Experimental results show that the proposed system achieved an accuracy of 0.99 with macro-average and F1-scores of 0.99, indicating consistent performance across all classes. The rule-based component was applied as a post-processing mechanism to handle linguistically ambiguous cases such as acronyms and foreign words. Evaluation on real news sentences demonstrates that the proposed hybrid approach is effective in detecting clitic usage errors.

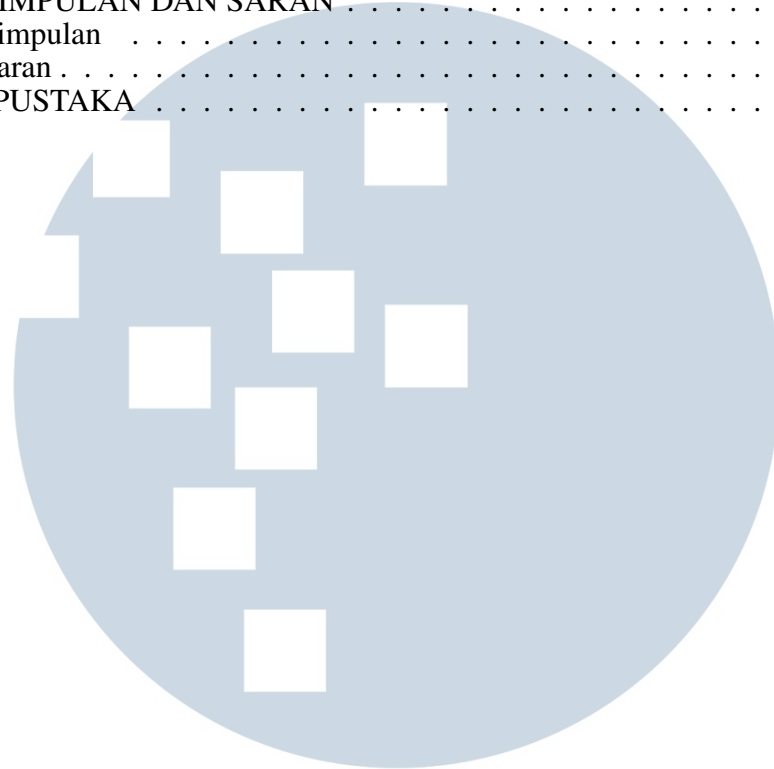
Keywords: Clitic Error Detection; N-gram; Natural Language Processing; Random Forest; Rule-Based System



DAFTAR ISI

HALAMAN JUDUL	i
PERNYATAAN TIDAK MELAKUKAN PLAGIAT	ii
PERNYATAAN PENGGUNAAN BANTUAN AI	iii
HALAMAN PERNYATAAN KEABSAHAN PERUSAHAAN	iv
HALAMAN PERSETUJUAN PUBLIKASI KARYA ILMIAH	vi
HALAMAN PERSEMBAHAN/MOTO	viii
KATA PENGANTAR	ix
ABSTRAK	xi
ABSTRACT	xii
DAFTAR ISI	xiii
DAFTAR TABEL	xv
DAFTAR GAMBAR	xvi
DAFTAR KODE	xvii
DAFTAR RUMUS	xviii
DAFTAR LAMPIRAN	xix
BAB 1 PENDAHULUAN	1
1.1 Latar Belakang Masalah	1
1.2 Rumusan Masalah	2
1.3 Tujuan Penelitian	3
1.4 Urgensi Penelitian	3
1.5 Luaran Penelitian	3
1.6 Manfaat Penelitian	3
BAB 2 LANDASAN TEORI	5
2.1 Kesalahan Penulisan	5
2.2 Aturan Penulisan Klitik	5
2.2.1 Ejaan yang Disempurnakan (EYD)	5
2.2.2 Pedoman Umum Ejaan Bahasa Indonesia (PUEBI)	6
2.3 N-Gram	7
2.4 Rule-based	7
2.5 Hypertuning Parameter	8
2.6 Decision Tree	9
2.7 Random Forest	10
2.8 Metrik Evaluasi	11
2.8.1 Accuracy	12
2.8.2 Precision	13
2.8.3 Recall	13
2.8.4 F1-score	13
BAB 3 METODOLOGI PENELITIAN	15
3.1 Tahapan Penelitian	15
3.1.1 Spesifikasi Sistem	16
3.1.2 Persiapan dan Pengumpulan Data	16
3.1.3 Rekayasa Fitur	18
3.2 Pembuatan Model	19
3.3 Penyusunan Rule-Based	21
BAB 4 HASIL DAN DISKUSI	23
4.1 Alur Deteksi Kesalahan Klitik	23
4.2 Hasil Pelatihan Model	24
4.3 Hasil Penyusunan Rule-based	24

4.3.1	Heatmap Rule dan Model	26
4.3.2	Hasil Model pada Kalimat Berita	27
BAB 5	SIMPULAN DAN SARAN	29
5.1	Simpulan	29
5.2	Saran	29
DAFTAR PUSTAKA		30



UMN
UNIVERSITAS
MULTIMEDIA
NUSANTARA

DAFTAR TABEL

Tabel 2.1	Tabel <i>Confusion Matrix</i> [1]	12
Tabel 3.1	Tabel Contoh Dataset	18
Tabel 3.2	Kombinasi Parameter <i>Random Forest</i>	20
Tabel 4.1	Tabel Contoh Deteksi Model	28



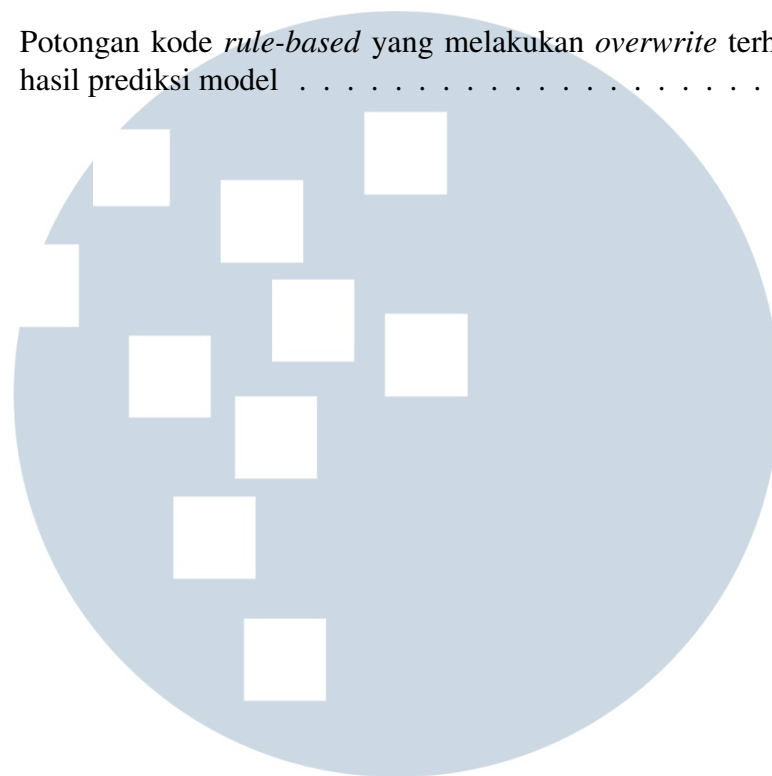
DAFTAR GAMBAR

Gambar 2.1	Diagram skematik <i>decision tree</i> [2]	10
Gambar 2.2	<i>Pseudocode random forest</i> [3]	11
Gambar 2.3	Diagram skematik algoritma <i>random forest</i> [4]	11
Gambar 3.1	<i>Pipeline</i> penelitian	15
Gambar 3.2	<i>Flowchart</i> pembuatan dataset	17
Gambar 3.3	Alur Rekayasa Fitur	19
Gambar 3.4	Alur Pembuatan Model	21
Gambar 4.1	Alur deteksi klitik	23
Gambar 4.2	<i>Heatmap</i> Model dan <i>Rule</i>	26



DAFTAR KODE

Kode 4.1	Potongan kode <i>rule-based</i> yang melakukan <i>overwrite</i> terhadap hasil prediksi model	24
----------	---	----



UMN
UNIVERSITAS
MULTIMEDIA
NUSANTARA

DAFTAR LAMPIRAN

Lampiran 1	PRO-STEP-01 Cover Letter	34
Lampiran 2	PRO-STEP-02 PRO-STEP Card	35
Lampiran 3	PRO-STEP-03 PRO-STEP Daily Task	36
Lampiran 4	PRO-STEP-04 Verification Form	44
Lampiran 5	PRO-STEP-05 Letter of Acceptance	45
Lampiran 6	Formulir Bimbingan	46
Lampiran 7	Hasil Persentase Turnitin	47
Lampiran 8	Draft Artikel Nasional	48
Lampiran 9	Surat Kerjasama	54

