

**OPTIMASI AKURASI KATEGORISASI KONTEN KESEHATAN
BERBASIS PENYAKIT MENGGUNAKAN MODEL *XGBOOST* DAN
*LIGHTGBM***

SKRIPSI SARJANA SISTEM INFORMASI

Oleh

Nanda Oktaviana

227006516076



**FAKULTAS TEKNOLOGI KOMUNIKASI DAN INFORMATIKA
PROGRAM STUDI SISTEM INFORMASI
UNIVERSITAS NASIONAL**

2026

**OPTIMASI AKURASI KATEGORISASI KONTEN
KESEHATAN BERBASIS PENYAKIT MENGGUNAKAN
MODEL XGBOOST DAN LIGHTGBM**

SKRIPSI SARJANA SISTEM INFORMASI

Karya ilmiah sebagai salah satu syarat untuk memperoleh gelar
Sarjana Sistem Informasi dari Fakultas Teknologi Komunikasi dan Informatika

Oleh

Nanda Oktaviana

227006516076



**FAKULTAS TEKNOLOGI KOMUNIKASI DAN INFORMATIKA
PROGRAM STUDI SISTEM INFORMASI
UNIVERSITAS NASIONAL**

2026

HALAMAN PENGESAHAN
TUGAS AKHIR

OPTIMASI AKURASI KATEGORISASI KONTEN KESEHATAN
BERBASIS PENYAKIT MENGGUNAKAN MODEL
XGBOOST DAN *LIGHTGBM*



Nanda Oktaviana
227006516076

Dosen Pembimbing 1

A handwritten signature in blue ink, consisting of a large, stylized 'A' followed by several vertical strokes.

(Dr. Ir. Andrianingsih, S.Kom., MMSI.)

Dosen Pembimbing 2

A handwritten signature in blue ink, featuring a large, stylized 'D' followed by several vertical strokes.

(Djarot Hindarto, S.Kom., M.Kom.)

PERNYATAAN KEASLIAN TUGAS AKHIR

Saya menyatakan dengan sesungguhnya bahwa Tugas Akhir dengan judul :

OPTIMASI AKURASI KATEGORISASI KONTEN KESEHATAN BERBASIS PENYAKIT MENGGUNAKAN MODEL *XGBOOST* DAN *LIGHTGBM*

Yang dibuat untuk melengkapi salah satu persyaratan menjadi Sarjana Komputer pada Program Studi Sistem Informasi Fakultas Teknologi Komunikasi dan Informatika Universitas Nasional, sebagaimana yang saya ketahui adalah bukan merupakan tiruan atau publikasi dari Tugas Akhir yang pernah diajukan atau dipakai untuk mendapatkan gelar di lingkungan Universitas Nasional maupun perguruan tinggi atau instansi lainnya, kecuali pada bagian – bagian tertentu yang menjadi sumber informasi atau acuan yang dicantumkan sebagaimana mestinya.



Jakarta, 2 Maret 2026



Nanda Oktaviana

227006516076

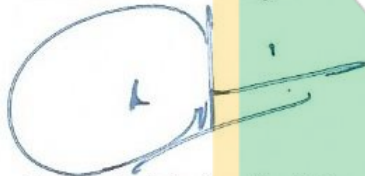
LEMBAR PERSETUJUAN REVIEW AKHIR

Tugas Akhir dengan judul :

OPTIMASI AKURASI KATEGORISASI KONTEN KESEHATAN BERBASIS PENYAKIT MENGGUNAKAN MODEL *XGBOOST* DAN *LIGHTGBM*

Dibuat untuk melengkapi salah satu persyaratan menjadi Sarjana Komputer pada Program Studi Sistem Informasi, Fakultas Teknologi Komunikasi dan Informatika Universitas Nasional. Tugas Akhir ini diujikan pada Sidang Review Akhir Semester Ganjil 2025-2026 pada tanggal 24 Februari Tahun 2026

Dosen Pembimbing I



Dr. Ir. Andrianingsih, S.Kom., MMSI.

0303097902

Dosen Pembimbing 2



Djarot Hingarto, S.Kom., M.Kom.

0317016905

Ketua Program Studi



UNIVERSITAS NASIONAL

Dr. Ir. Andrianingsih, S.Kom., MMSI.

0303097902

LEMBAR PERSETUJUAN JUDUL YANG TIDAK ATAU YANG DIREVISI


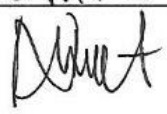
Nama : Nanda Oktaviana
NPM : 227006516076
Fakultas/Akademi : Fakultas Teknologi Komunikasi dan Informatika
Program Studi : Sistem Informasi
Tanggal Sidang : 24 Februari 2026

JUDUL DALAM BAHASA INDONESIA :

OPTIMASI AKURASI KATEGORISASI KONTEN KESEHATAN BERBASIS PENYAKIT MENGGUNAKAN MODEL XGBOOST DAN LIGHTGBM

JUDUL DALAM BAHASA INGGRIS :

OPTIMIZING THE ACCURACY OF DISEASE-BASED HEALTH CONTENT CATEGORIZATION USING XGBOOST AND LIGHTGBM MODELS

TANDA TANGAN DAN TANGGAL		
Pembimbing 1	Ka. Prodi	Mahasiswa
TGL : 2/3 2026	TGL : 2/3 2026	TGL : 2 Maret 2026
		

LEMBAR PERSETUJUAN JUDUL YANG TIDAK ATAU YANG DIREVISI

Nama : Nanda Oktaviana
NPM : 227006516076
Fakultas/Akademi : Fakultas Teknologi Komunikasi dan Informatika
Program Studi : Sistem Informasi
Tanggal Sidang : 24 Februari 2026


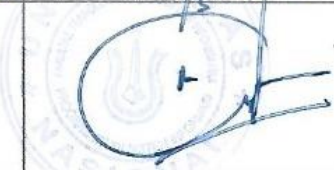

JUDUL DALAM BAHASA INDONESIA :

**OPTIMASI AKURASI KATEGORISASI KONTEN
KESEHATAN BERBASIS PENYAKIT MENGGUNAKAN
MODEL XGBOOST DAN LIGHTGBM**

JUDUL DALAM BAHASA INGGRIS :

**OPTIMIZING THE ACCURACY OF DISEASE-BASED
HEALTH CONTENT CATEGORIZATION USING
XGBOOST AND LIGHTGBM MODELS**

TANDA TANGAN DAN TANGGAL

Pembimbing 2	Ka, Prodi	Mahasiswa
TGL: 2 Maret 2026	TGL: 2/3 2026	TGL: 2 Maret 2026
		

KATA PENGANTAR

Segala puji Segala puji dan syukur penulis panjatkan ke hadirat Allah SWT atas limpahan rahmat, taufik, serta hidayah-Nya, sehingga penulis dapat menyelesaikan skripsi dengan judul “Forest”. Skripsi ini disusun sebagai salah satu syarat untuk memperoleh gelar Sarjana Komputer pada Program Studi Sistem Informasi, Fakultas Teknologi Komunikasi dan Informatika, Universitas Nasional.

Skripsi ini disusun sebagai salah satu syarat untuk memperoleh gelar Sarjana Komputer pada Program Studi Sistem Informasi, Fakultas Teknologi Komunikasi dan Informatika, Universitas Nasional. Penelitian dan penulisan skripsi ini tidak terlepas dari bantuan berbagai pihak, oleh karena itu peneliti menyampaikan banyak terima kasih terutama kepada dosen pembimbing Tugas Akhir yaitu Ibu Dr. Andrianingsih, S.Kom., MMSI. yang telah meluangkan banyak waktu, tenaga, pikiran, bimbingan, arahan, motivasi serta memaklumi segala kekurangan penulis selama penelitian tugas akhir dan penyusunan skripsi. Peneliti juga mengucapkan banyak terima kasih kepada:

1. Bapak Dr. Agung Triayudi, S.Kom., M.Kom. Selaku Dekan Fakultas Teknologi Komunikasi dan Informatika Universitas Nasional.
2. Ibu Ir. Endah Tri Esti Handayani, MMSI. Selaku Wakil Dekan Fakultas Teknologi Komunikasi dan Informatika Universitas Nasional.
3. Ibu Dr. Andrianingsih, S.Kom., MMSI., selaku Dosen Pembimbing I sekaligus Ketua Program Studi Sistem Informasi, yang telah memberikan arahan, motivasi, serta bimbingan selama proses penyusunan skripsi ini.
4. Bapak Djarot Hindarto, S.Kom., M.Kom., selaku Dosen Pembimbing II, yang telah memberikan bimbingan intensif, saran teknis, serta masukan mendalam dalam proses analisis data dan penyusunan jurnal ilmiah.
5. Kedua orang tua yang sangat ku cintai dan ku sayangi. Terima kasih telah memberikan doa, motivasi, dukungan, dan pengorbanan tanpa

batas yang bapak dan mama berikan tidak akan bisa terbalas, tapi insyaallah saya akan berusaha dan kerja keras untuk memberikan yang terbaik.

Akhir kata, semoga Allah SWT membalas kebaikan dan bantuan yang telah diberikan dengan hal yang lebih baik. Penulis mengharapkan kritik dan saran yang bersifat membangun dan semoga skripsi ini dapat memberikan manfaat di bidang Teknologi Informatika.



Jakarta, 8 Oktober 2025

A handwritten signature in black ink, appearing to read "Nanda Oktaviana".

Nanda Oktaviana

ABSTRAK

Konten kesehatan berbahasa Indonesia pada platform X (Twitter) berkembang sangat cepat dan memuat berbagai informasi serta pengalaman masyarakat terkait penyakit, sehingga diperlukan sistem yang mampu melakukan klasifikasi konten kesehatan secara otomatis dan akurat. Penelitian ini bertujuan untuk mengoptimalkan akurasi klasifikasi konten kesehatan berbasis penyakit dengan memanfaatkan IndoBERT embedding sebagai representasi semantik teks serta membandingkan performa algoritma XGBoost dan Light Gradient Boosting Machine (LightGBM) sebagai model klasifikasi. Data penelitian diperoleh melalui proses data scraping dari platform X dengan fokus pada tiga kategori penyakit, yaitu jantung, hipertensi, dan diabetes, kemudian dilakukan prapemrosesan teks dan pembentukan embedding IndoBERT. Selanjutnya, embedding digunakan sebagai masukan untuk pelatihan model XGBoost dan LightGBM, dengan evaluasi kinerja menggunakan metrik accuracy, precision macro, recall macro, dan F1-score macro melalui pendekatan K-Fold Cross Validation. Hasil penelitian menunjukkan bahwa kombinasi IndoBERT + LightGBM menghasilkan performa terbaik dengan nilai accuracy sebesar 85,26%, precision macro 85,29%, recall macro 85,26%, dan F1-score macro 85,27%, lebih tinggi dibandingkan kombinasi IndoBERT + XGBoost yang memperoleh akurasi 83,25%. Penelitian ini diharapkan dapat memberikan kontribusi dalam pengembangan sistem klasifikasi konten kesehatan digital berbahasa Indonesia yang lebih akurat dan kontekstual, serta mendukung peningkatan literasi kesehatan digital di Indonesia.

Kata Kunci: IndoBERT Embedding, XGBoost, LightGBM, Klasifikasi Teks, Konten Kesehatan

ABSTRACT

Indonesian-language health content on the X (Twitter) platform is growing rapidly and contains various types of information and public experiences related to diseases, creating the need for a system capable of automatically and accurately classifying health content. This study aims to optimize the accuracy of disease-based health content classification by utilizing IndoBERT embeddings as semantic text representations and by comparing the performance of the XGBoost and Light Gradient Boosting Machine (LightGBM) algorithms as classification models. The research data were obtained through a data scraping process from the X platform, focusing on three disease categories, namely heart disease, hypertension, and diabetes, followed by text preprocessing and the generation of IndoBERT embeddings. The embeddings were then used as input for training the XGBoost and LightGBM models, with performance evaluation conducted using accuracy, macro precision, macro recall, and macro F1-score metrics through a K-Fold Cross Validation approach. The results show that the IndoBERT + LightGBM combination achieved the best performance, with an accuracy of 85.26%, macro precision of 85.29%, macro recall of 85.26%, and macro F1-score of 85.27%, outperforming the IndoBERT + XGBoost combination, which achieved an accuracy of 83.25%. This study is expected to contribute to the development of more accurate and context-aware Indonesian-language digital health content classification systems and to support the improvement of digital health literacy in Indonesia.

Keywords: *IndoBERT Embedding, XGBoost, LightGBM, Text Classification, Health Content*



DAFTAR ISI

OPTIMASI AKURASI KATEGORISASI KONTEN KESEHATAN BERBASIS PENYAKIT MENGGUNAKAN MODEL <i>XGBOOST</i> DAN <i>LIGHTGBM</i>	1
HALAMAN PENGESAHAN	iii
PERNYATAAN KEASLIAN TUGAS AKHIR	iv
LEMBAR PERSETUJUAN REVIEW AKHIR	v
LEMBAR PERSETUJUAN JUDUL YANG TIDAK ATAU YANG DIREVISI	vi
KATA PENGANTAR	viii
ABSTRAK	x
ABSTRACT	xi
DAFTAR ISI	xii
DAFTAR TABEL	xv
DAFTAR GAMBAR	xvi
DAFTAR LAMPIRAN	xvii
BAB I PENDAHULUAN	1
1.1 Latar Belakang Masalah	1
1.2 Rumusan Masalah	5
1.3 Tujuan Penelitian	6
1.4 Batasan Masalah	7
1.5 Manfaat Penelitian	8
1.6 Kontribusi Penelitian	9
BAB II TINJAUAN PUSTAKA	11
2.1 Studi Literatur	11
2.2 Research Positioning	17
2.3 Fishbone Diagram	21
2.4 Landasan Teori	23
2.4.1 Konten Kesehatan Digital Berbasis Penyakit	23

2.4.2	Twitter (Platform X)	24
2.4.3	<i>Data Scraping</i>	25
2.4.4	<i>Machine Learning</i>	26
2.4.5	<i>Natural Language Processing (NLP)</i>	26
2.4.6	Model <i>Transformer</i> dan <i>IndoBERT Embedding</i>	28
2.4.7	Algoritma <i>XGBoost (Extreme Gradient Boosting)</i>	30
2.4.8	Algoritma <i>Light Gradient Boosting Machine (LightGBM)</i>	31
2.4.9	Evaluasi Model (<i>Accuracy, Precision, Recall, F1-Score</i>)....	32
2.4.10	Python	34
2.4.11	<i>Framework Django</i>	35
BAB III METODOLOGI PENELITIAN		37
3.1	Desain Penelitian	37
3.1.1	Pendekatan Penelitian	37
3.1.2	Tahapan Alur Penelitian	38
3.2	Fokus Penelitian	42
3.3	Objek Penelitian	43
3.4	Variabel Penelitian dan Definisi Operasional	44
3.5	Hipotesis Penelitian	44
3.6	Lokasi dan Waktu Penelitian	45
3.7	Teknik Pengumpulan Data	46
3.8	Teknik Analisis Data	47
BAB IV HASIL DAN PEMBAHASAN		50
4.1	Gambaran Umum <i>Dataset</i>	50
4.1.1	Karakteristik <i>Dataset</i>	50
4.2	Tahapan Prapemrosesan Teks	52
4.2.1	Pembersihan Teks	52
4.2.2	<i>Case Folding</i>	54
4.2.3	<i>Normalization</i>	54
4.2.4	Hasil Pelabelan Data	55
4.2.5	<i>Tokenization IndoBERT</i>	56
4.2.6	<i>IndoBERT Embedding</i>	56

4.3	Hasil Pemodelan dan Evaluasi.....	57
4.3.1	Hasil Klasifikasi Menggunakan XGBoost.....	57
4.3.2	Hasil Klasifikasi Menggunakan <i>LightGBM</i>	58
4.3.3	Perbandingan Kinerja <i>XGBoost</i> dan <i>LightGBM</i>	59
4.3.4	<i>Evaluation</i>	60
4.3.5	Visualisasi <i>Wordcloud</i> Tweet Kesehatan.....	64
4.3.6	Hasil dan Analisis <i>Interpretabilitas</i> Model Menggunakan SHAP	67
4.4	Perhitungan Manual Algoritma	70
4.3.1	Ruang Lingkup Perhitungan Manual	70
4.3.2	Penentuan Kode Kelas	71
4.3.3	Perhitungan Manual Pohon Keputusan.....	71
4.3.4	Langkah 1 – Menghitung Gini Node Akar	72
4.3.5	Langkah 2 – Mendefinisikan Fitur F1 dan Split	72
4.3.6	Langkah 3 – Menghitung Gini Node Kiri dan Kanan.....	73
4.3.7	Langkah 4 – Gini Setelah Split dan <i>Gini Gain</i>	74
4.3.8	Keterkaitan Perhitungan Manual dengan Algoritma <i>LightGBM</i>	74
4.3.9	Keterkaitan Perhitungan Manual dengan Algoritma <i>LightGBM</i>	75
4.3.10	Validasi Perhitungan Manual dengan Implementasi <i>Python</i>	79
4.4	<i>Deployment</i>	80
BAB V PENUTUP		83
5.1	Kesimpulan	83
5.2	Saran	83
DAFTAR PUSTAKA		85
LAMPIRAN.....		88

DAFTAR TABEL

Tabel 2.1 Penelitian Terdahulu	11
Tabel 2.2 Penelitian Sebelumnya dan Penelitian Ini.....	18
Tabel 3.1 Timeline Penelitian	46
Tabel 4.1 Perbandingan Sebelum dan Sesudah Pembersihan Teks	53
Tabel 4.2 Perbandingan Sebelum dan Sesudah Case Folding	54
Tabel 4.3 Perbandingan Sebelum dan Sesudah Normalization	55
Tabel 4.4 Hasil Tokenization Setelah Normalization	56
Tabel 4.5 Akurasi <i>LightGBM</i> pada <i>5-Fold Cross Validation</i>	57
Tabel 4.6 Akurasi <i>LightGBM</i> pada <i>5-Fold Cross Validation</i>	58
Tabel 4.7 Perbandingan Kinerja Rata-Rata Model	59
Tabel 4.8 Sampel Data <i>Tweet</i> Kesehatan untuk Perhitungan Manual.....	71
Tabel 4.9 Nilai Fitur F1 (kata “gula/diabetes”) pada Sampel Data <i>Tweet</i>	73
Tabel 4.10 Pemetaan Kelas Penyakit ke Kode Kelas dan Label.....	75
Tabel 4.11 Nilai Gradien dan Hessian untuk Setiap Sampel pada Satu Iterasi XGBoost.....	76
Tabel 4.12 Perubahan Probabilitas Prediksi Diabetes Setelah Satu Iterasi XGBoost	79

DAFTAR GAMBAR

Gambar 1.1 Sepuluh Penyakit Penyebab Kematian Tertinggi di Indonesia Tahun 2024 <i>Sumber: Laporan WHO, dikutip dari Inilah.com (Desember 2024)</i>	2
Gambar 2.1 <i>Research Positioning</i>	19
Gambar 2.2 <i>Fishbone Diagram</i>	21
Gambar 2.3 Platform X	24
Gambar 2.4 Cakupan dan Komponen <i>Natural Language Processing (NLP)</i>	27
Gambar 2.5 Phyton	34
Gambar 3.1 Tahap Alur Penelitian	39
Gambar 3.2 Alur Proses Pengolahan Data dan Pemodelan	48
Gambar 4.1 Hasil Data Tweet Yang Telah Dikumpulkan	51
Gambar 4.2 Hasil Pelabelan Data	55
Gambar 4.3 Hasil <i>embedding IndoBERT</i> untuk satu tweet Kesehatan	57
Gambar 4.4 Classification Report model <i>IndoBERT + XGBoost</i>	60
Gambar 4.5 <i>Normalized Confusion Matrix</i> model <i>IndoBERT + XGBoost</i>	61
Gambar 4.6 Classification Report model <i>LightGBM + XGBoost</i>	62
Gambar 4.7 <i>Normalized Confusion Matrix</i> model <i>IndoBERT + LightGBM</i>	63
Gambar 4.8 <i>Word Cloud Jantung</i>	65
Gambar 4.9 <i>Word Cloud Hipertensi</i>	66
Gambar 4.10 <i>Word Cloud – Diabetes</i>	67
Gambar 4.11 SHAP Bart Plot – <i>IndoBERT × XGBoost</i>	68
Gambar 4.12 SHAP Bar Plot – <i>IndoBERT × LightGBM</i>	69
Gambar 4.13 Deployment	81

DAFTAR LAMPIRAN

Lampiran 1. <i>Source Code Source Code Scraping</i> untuk Pengumpulan Data	88
Lampiran 2. <i>Letter of Acceptance</i>	89
Lampiran 3. Turnitin	90
Lampiran 4. <i>AI Checker</i>	91

