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	Dosen Pembimbing 2
(Dr. Ir. Andrianingsih, S.Kom., MMSI.)	(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 Hindarto, S.Kom., M.Kom.
0317016905

Ketua Program Studi

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 karna 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 sayang. 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

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

DAFTAR TABEL

DAFTAR GAMBAR

DAFTAR LAMPIRAN