KLASIFIKASI HALAMAN WEB BERBASIS *MACHINE LEARNING* UNTUK OPTIMASI SEO MENGGUNAKAN FITUR NUMERIK DAN SEMANTIK BERBASIS INDOBERT

SITI NURADILLA

PROGRAM STUDI MAGISTER STATISTIKA DAN SAINS DATA
SEKOLAH SAINS DATA, MATEMATIKA, DAN INFORMATIKA
INSTITUT PERTANIAN BOGOR
BOGOR
2026

KLASIFIKASI HALAMAN WEB BERBASIS *MACHINE LEARNING* UNTUK OPTIMASI SEO MENGGUNAKAN FITUR NUMERIK DAN SEMANTIK BERBASIS INDOBERT

SITI NURADILLA

Tesis
sebagai salah satu syarat untuk memperoleh gelar
Magister pada
Program Studi Statistika dan Sains Data

PROGRAM STUDI MAGISTER STATISTIKA DAN SAINS DATA
SEKOLAH SAINS DATA, MATEMATIKA, DAN INFORMATIKA
INSTITUT PERTANIAN BOGOR
BOGOR
2026

PERNYATAAN MENGENAI TESIS DAN SUMBER INFORMASI SERTA PELIMPAHAN HAK CIPTA

Dengan ini saya menyatakan bahwa tesis dengan judul "Klasifikasi Halaman Web Berbasis *Machine Learning* untuk Optimasi SEO Menggunakan Fitur Numerik dan Semantik Berbasis IndoBERT" adalah karya saya dengan arahan dari dosen pembimbing dan belum diajukan dalam bentuk apa pun kepada perguruan tinggi mana pun. Sumber informasi yang berasal atau dikutip dari karya yang diterbitkan maupun tidak diterbitkan dari penulis lain telah disebutkan dalam teks dan dicantumkan dalam Daftar Pustaka di bagian akhir tesis ini.

Dengan ini saya melimpahkan hak cipta dari karya tulis saya kepada Institut Pertanian Bogor.

Bogor, Mei 2026

Siti Nuradilla
M0501241010

RINGKASAN

SITI NURADILLA. Klasifikasi Halaman Web Berbasis *Machine Learning* untuk Optimasi SEO Menggunakan Fitur Numerik dan Semantik Berbasis IndoBERT. Dibimbing oleh BUDI SUSETYO dan CICI SUHAENI.

Meningkatnya volume data teks telah memperkuat kebutuhan akan metode yang efektif untuk mengekstraksi informasi yang bermakna, khususnya pada *search engine optimization* (SEO). Pada proses optimasi SEO, ekstraksi makna semantik menjadi krusial karena relevansi halaman tidak hanya dipengaruhi oleh panjang karakter komponen *on-page*, namun juga koherensi antar komponennya, meliputi *title*, *meta description*, H1, dan *address*. Namun, evaluasi SEO masih dilakukan secara manual, sehingga kurang efisien dan rentan terhadap bias subjektif. Penelitian sebelumnya hanya berfokus pada indikator permukaan seperti kepadatan kata kunci dan bergantung pada dataset empiris, sehingga belum menjamin keandalan model pada kondisi data yang berbeda. Oleh karena itu, diperlukan pembangkitan data sintetis yang dapat merepresentasikan beragam skenario distribusi data, sehingga ketahanan dan konsistensi performa model dapat diuji secara lebih komprehensif.

Untuk mengatasi keterbatasan tersebut**,** penelitian ini bertujuan untuk mengevaluasi kemampuan model GPT yang di-*fine-tune* dalam membangkitkan data sintetis yang merepresentasikan karakteristik SEO *on-page*. Selanjutnya, penelitian ini juga menganalisis karakteristik model klasifikasi halaman web dengan memanfaatkan fitur numerik berupa panjang karakter komponen SEO *on-page*, serta fitur semantik berupa tingkat koherensi antar komponen yang diperoleh melalui proses *embedding* menggunakan IndoBERT. Koherensi semantik diukur menggunakan *cosine similarity* antar komponen SEO *on-page* untuk merepresentasikan keselarasan konteks antar komponen. Berdasarkan fitur tersebut, halaman web diklasifikasikan menjadi SEO-*friendly* dan *non*-SEO-*friendly* menggunakan Random Forest, XGBoost, LightGBM, dan TabNet. Penelitian menggunakan delapan dataset, terdiri atas satu dataset empiris berisi 10.791 halaman web dan tujuh dataset sintetis dengan variasi tingkat koherensi (20%–80%). Data empiris dikumpulkan menggunakan Screaming Frog SEO Spider untuk memperoleh komponen SEO *on-page* seperti *title*, *meta description*, *address*, dan *heading*, sedangkan Semrush digunakan untuk memperoleh data peringkat halaman pada SERP yang dimanfaatkan dalam proses pelabelan kelas SEO-*friendly* dan *non*-SEO-*friendly*. Proses penelitian meliputi prapemrosesan, representasi teks, penghitungan koherensi, serta pelatihan dan evaluasi model, sementara kualitas data sintetis dievaluasi menggunakan metrik *diversity*, *novelty*, dan *duplication*.

Hasil eksplorasi menunjukkan bahwa halaman SEO-*friendly* memiliki pola yang lebih konsisten dalam panjang teks dan koherensi semantik dibandingkan halaman *non*-SEO-*friendly*. Selain itu, model GPT mampu menghasilkan data sintetis dengan kualitas yang baik, ditunjukkan oleh nilai *diversity* yang tinggi (0,9–1), *novelty* pada rentang 0,7–0,85, serta tingkat duplikasi yang sangat rendah (<0,06%). Temuan ini menunjukkan bahwa data sintetis yang dihasilkan mampu merepresentasikan variasi kondisi data dan mendukung proses klasifikasi dengan lebih komprehensif.

Pada tahap pemodelan, dilakukan perbandingan antara Random Forest, XGBoost, LightGBM, dan TabNet pada data empiris dan data sintetis. Pada data sintetis, performa model menunjukkan pola yang berbeda pada setiap tingkat koherensi. XGBoost dan LightGBM cenderung lebih kompetitif pada koherensi rendah hingga menengah, sedangkan TabNet mulai menunjukkan performa yang lebih tinggi pada koherensi tinggi. Pada data empiris, Random Forest memperoleh performa yang tinggi dengan *balanced accuracy* sebesar 0,8677, diikuti oleh XGBoost (0,8654) dan LightGBM (0,8605). Temuan ini sejalan dengan kajian awal (*baseline*) yang menunjukkan bahwa Random Forest sangat efektif dalam menangkap pola dominan pada distribusi data nyata yang relatif stabil. Namun, Random Forest cenderung mengalami penurunan performa pada struktur semantik antar kelas yang kurang tegas. Sebaliknya, model boosting yaitu XGBoost dan LightGBM menunjukkan performa yang lebih konsisten di berbagai tingkat koherensi, dengan nilai *balanced accuracy* yang kompetitif serta variasi performa yang relatif kecil. Performa terbaik secara umum dicapai pada skenario koherensi menengah (40%), di mana perbedaan karakteristik antar kelas menjadi lebih jelas. Sementara itu, pada koherensi rendah (20%–30%) dan tinggi (70%–80%), performa model cenderung menurun akibat meningkatnya ambiguitas atau homogenitas distribusi data.

Hasil uji statistik menunjukkan bahwa performa klasifikasi berbeda pada setiap tingkat koherensi dan bergantung pada jenis model yang digunakan. Uji lanjut perbandingan nilai tengah berganda dengan penyesuaian Holm memperlihatkan bahwa keunggulan model bersifat kontekstual, di mana Random Forest unggul pada data empiris, XGBoost dan LightGBM menunjukkan performa yang kompetitif pada koherensi rendah hingga menengah, sedangkan TabNet unggul pada koherensi tinggi. Dalam konteks implementasi sistem evaluasi SEO otomatis pada lingkungan data yang dinamis, model berbasis *boosting*, khususnya XGBoost, menunjukkan performa yang relatif stabil baik pada data empiris maupun berbagai tingkat koherensi data sintetis. Temuan ini menunjukkan bahwa model berbasis *boosting* memiliki potensi yang baik untuk digunakan pada kondisi data SEO *on-page* yang bervariasi.

Kata kunci: *IndoBERT, koherensi semantik, machine learning, search engine optimization.*

SUMMARY

SITI NURADILLA. Machine Learning-Based Web Page Classification for SEO Optimization Using Numerical and Semantic Features with IndoBERT. Supervised by BUDI SUSETYO and CICI SUHAENI.

The increasing volume of textual data has strengthened the need for effective methods to extract meaningful information, particularly in the context of search engine optimization (SEO). In SEO optimization, semantic information extraction is crucial because page relevance is influenced not only by the character length of on-page components, but also by the coherence among components, including title, meta description, H1, and address. However, SEO evaluation is still commonly performed manually, making it less efficient and prone to subjective bias. Previous studies have primarily focused on surface-level indicators such as keyword density and relied heavily on empirical datasets, which limits the reliability of models under varying data conditions. Therefore, synthetic data generation is required to represent diverse data distribution scenarios, enabling a more comprehensive evaluation of model robustness and performance consistency.

To address these limitations, this study aims to evaluate the capability of a fine-tuned GPT model in generating synthetic data that represent the characteristics of SEO on-page components. Furthermore, this study analyzes the characteristics of web page classification models by utilizing numerical features in the form of character lengths of SEO on-page components, as well as semantic features represented by the coherence level among components obtained through an embedding process using IndoBERT. Semantic coherence is measured using cosine similarity among SEO on-page components to represent contextual alignment between components. Based on these features, web pages are classified into SEO-friendly and non-SEO-friendly categories using Random Forest, XGBoost, LightGBM, and TabNet.

This study utilizes eight datasets, consisting of one empirical dataset containing 10,791 web pages and seven synthetic datasets with varying coherence levels (20%–80%). The empirical data were collected using Screaming Frog SEO Spider to obtain SEO on-page components such as title, meta description, address, and heading, while Semrush was used to extract SERP ranking data utilized in the labeling process for SEO-friendly and non-SEO-friendly classes. The research process includes preprocessing, text representation, coherence calculation, as well as model training and evaluation, while the quality of synthetic data is evaluated using diversity, novelty, and duplication metrics.

Exploratory results indicate that SEO-friendly pages exhibit more consistent patterns in text length and semantic coherence compared to non-SEO-friendly pages. Additionally, the GPT model successfully generates high-quality synthetic data, as indicated by high diversity scores (0.9–1.0), novelty values ranging from 0.7 to 0.85, and extremely low duplication rates (<0.06%). These findings suggest that the generated synthetic data effectively represent diverse data conditions and support more comprehensive classification experiments.

In the modeling stage, Random Forest, XGBoost, LightGBM, and TabNet were compared using both empirical and synthetic datasets. On synthetic datasets, model performance patterns varied across different coherence levels. XGBoost and

LightGBM tended to perform more competitively at low to medium coherence levels, whereas TabNet achieved higher performance at high coherence levels. On the empirical dataset, Random Forest achieved a high balanced accuracy of 0.8677, followed by XGBoost (0.8654) and LightGBM (0.8605). These findings are consistent with the baseline analysis, which showed that Random Forest is highly effective in capturing dominant patterns within relatively stable real-world data distributions. However, Random Forest tended to experience performance degradation when semantic boundaries between classes became less distinct. In contrast, boosting-based models, namely XGBoost and LightGBM, demonstrated more consistent performance across various coherence levels, achieving competitive balanced accuracy values with relatively low performance variation. Overall, the best performance was generally observed at medium coherence levels (40%), where semantic relationships among classes became more distinguishable. Meanwhile, at low coherence levels (20%–30%) and high coherence levels (70%–80%), model performance tended to decline due to increasing ambiguity or homogeneity in data distributions.

Statistical analysis results showed that classification performance differed across coherence levels and depended on the type of model used. Holm-adjusted multiple comparison tests further revealed that model superiority was contextual, where Random Forest performed best on empirical data, XGBoost and LightGBM showed competitive performance at low to medium coherence levels, and TabNet achieved the best performance at high coherence levels. In the context of implementing automated SEO evaluation systems in dynamic data environments, boosting-based models, particularly XGBoost, demonstrated relatively stable performance across both empirical data and multiple synthetic coherence scenarios. These findings suggest that boosting-based models have strong potential for implementation in on-page SEO evaluation tasks involving diverse data characteristics.

Keywords: *IndoBERT, machine learning, search engine optimization, semantic coherence.*

KLASIFIKASI HALAMAN WEB BERBASIS *MACHINE LEARNING* UNTUK OPTIMASI SEO MENGGUNAKAN FITUR NUMERIK DAN SEMANTIK BERBASIS INDOBERT

SITI NURADILLA

Tesis
sebagai salah satu syarat untuk memperoleh gelar
Magister pada
Program Studi Statistika dan Sains Data

PROGRAM STUDI MAGISTER STATISTIKA DAN SAINS DATA
SEKOLAH SAINS DATA, MATEMATIKA, DAN INFORMATIKA
INSTITUT PERTANIAN BOGOR
BOGOR
2026

Penguji Luar Komisi Pembimbing pada Ujian Tesis:
Prof. Dr. Ir. Hari Wijayanto, M.Si.

Judul Tesis : Klasifikasi Halaman Web Berbasis *Machine Learning* untuk Optimasi SEO Menggunakan Fitur Numerik dan Semantik Berbasis IndoBERT
Nama : Siti Nuradilla
NIM : M0501241010

Disetujui oleh

Pembimbing 1:
Dr. Ir. Budi Susetyo, M.S.

Pembimbing 2:
Cici Suhaeni, S.Si., M.Si., Ph.D.

Diketahui oleh

Ketua Program Studi:
Dr. Agus Mohamad Soleh, S.Si., M.T.
NIP. 19750315 199903 1 004

Dekan Sekolah Sains Data, Matematika, dan Informatika:
Prof. Dr. Ir. Agus Buono, M.Si., M.Kom.
NIP. 19660702 199302 1 001

Tanggal Ujian:
7 Mei 2026

Tanggal Pengesahan:

PRAKATA

Puji syukur penulis panjatkan ke hadirat Allah Subhanahu wa Ta'ala atas limpahan rahmat, taufik, dan karunia-Nya sehingga penulis dapat menyelesaikan tesis ini dengan baik. Karya ini disusun sebagai bagian dari pemenuhan tugas akhir dengan judul: **"**Klasifikasi Halaman Web Berbasis *Machine Learning* untuk Optimasi SEO Menggunakan Fitur Numerik dan Semantik Berbasis IndoBERT**"**.

Pada proses penyusunan tesis ini, sangat banyak bantuan dari berbagai pihak. Penulis menyampaikan terima kasih dan penghargaan setinggi-tingginya kepada:

1. Bapak Dr. Ir. Budi Susetyo, M.S. dan Ibu Cici Suhaeni, S.Si., M.Si., Ph.D selaku dosen pembimbing yang dengan penuh dedikasi telah memberikan arahan, bimbingan, motivasi, serta berbagai masukan yang membangun selama proses penelitian hingga penyusunan karya ilmiah ini.
2. Bapak Prof. Dr. Ir. Hari Wijayanto, M.Si. selaku penguji luar komisi pembimbing yang telah memberikan saran, masukan, dan wawasan yang sangat berharga dalam menyempurnakan karya ilmiah ini.
3. Bapak Dr. Agus Mohamad Soleh, S.Si., M.T. selaku pimpinan sidang tesis sekaligus Ketua Program Studi Statistika dan Sains Data atas arahan, saran, dan masukan yang bermanfaat bagi penyempurnaan karya ilmiah in
4. Seluruh dosen Program Studi Statistika dan Sains Data IPB University yang telah memberikan ilmu pengetahuan, pengalaman, serta pembelajaran yang bermanfaat selama penulis menempuh pendidikan.
5. Seluruh tenaga kependidikan Program Studi Statistika dan Sains Data IPB University yang telah membantu penulis dalam proses administrasi dan pelayanan akademik.
6. Kedua orang tua serta keluarga penulis yang senantiasa memberikan doa, kasih sayang, dukungan moral maupun material, serta semangat selama proses akademik dan penyusunan karya ilmiah ini.
7. Teman dekat penulis yang selalu hadir memberikan dukungan, bantuan, motivasi, serta menemani perjalanan penulis sejak awal perkuliahan hingga terselesaikannya karya ilmiah ini.
8. Penulis juga mengucapkan terima kasih kepada pihak Beasiswa Pendidikan Indonesia yang telah memfasilitasi dalam pelaksanaan Pendidikan dan penelitian.

Penulis menyadari bahwa karya ilmiah ini masih memiliki keterbatasan. Oleh karena itu, penulis membuka diri terhadap saran dan kritik yang membangun demi kesempurnaan karya ini di masa mendatang.

Akhir kata, semoga karya ini dapat memberikan manfaat bagi semua pihak yang membutuhkan dan berkontribusi nyata bagi pengembangan ilmu pengetahuan dan perumusan kebijakan berbasis data di Indonesia.

Bogor, Mei 2026

Siti Nuradilla

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

DAFTAR LAMPIRAN