MULTINOMIAL NAÏVE BAYES UNTUK DETEKSI *PHISHING* URL DENGAN PERBANDINGAN REPRESENTASI FITUR *BAG OF WORDS* DAN TF-IDF

SKRIPSI

Diajukan untuk memenuhi salah satu syarat mencapai derajat Sarjana

Program Studi S1 Teknik Komputer

disusun oleh

GEOREL JEFERSON FRANSISKUS BONAI

22.83.0833

Kepada

FAKULTAS ILMU KOMPUTER

UNIVERSITAS AMIKOM YOGYAKARTA

YOGYAKARTA

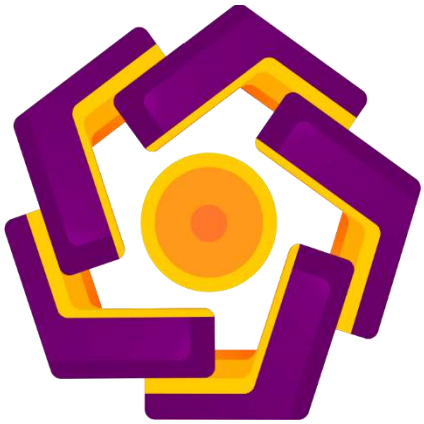
2026

MULTINOMIAL NAÏVE BAYES UNTUK DETEKSI *PHISHING* URL DENGAN PERBANDINGAN REPRESENTASI FITUR *BAG OF WORDS* DAN TF-IDF

SKRIPSI

untuk memenuhi salah satu syarat mencapai derajat Sarjana

Program Studi S1 Teknik Komputer



disusun oleh

GEOREL JEFERSON FRANSISKUS BONAI

22.83.0833

Kepada

FAKULTAS ILMU KOMPUTER

UNIVERSITAS AMIKOM YOGYAKARTA

YOGYAKARTA

2026

HALAMAN PERSETUJUAN

SKRIPSI

MULTINOMIAL NAÏVE BAYES UNTUK DETEKSI *PHISHING* URL DENGAN PERBANDINGAN REPRESENTASI FITUR *BAG OF WORDS* DAN TF-IDF

yang disusun dan diajukan oleh

Georel Jeferson Fransiskus Bonai

22.83.0833

telah disetujui oleh Dosen Pembimbing Skripsi
pada tanggal 23 Januari 2026

Dosen Pembimbing,

Tonny Hidayat, S.Kom., M.Kom., Ph.D
NIK. 190302182

HALAMAN PENGESAHAN

SKRIPSI

MULTINOMIAL NAÏVE BAYES UNTUK DETEKSI *PHISHING* URL DENGAN PERBANDINGAN REPRESENTASI FITUR *BAG OF WORDS* DAN TF-IDF

yang disusun dan diajukan oleh

Georel Jeferson Fransiskus Bonai

22.83.0833

Telah dipertahankan di depan Dewan Penguji
pada tanggal 23 Januari 2026

Susunan Dewan Penguji

Nama Penguji	Tanda Tangan
Melwin Syafrizal, S.Kom., M.Eng., Ph.D **NIK. 190302105**	
Senie Destya, M.Kom **NIK. 190302312**	
Tonny Hidayat, S.Kom., M.Kom., Ph.D **NIK. 190302182**	

Skripsi ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Sarjana Komputer
Tanggal 23 Januari 2026

DEKAN FAKULTAS ILMU KOMPUTER

Prof. Dr. Kusrini, M.Kom.
NIK. 190302106

HALAMAN PERNYATAAN KEASLIAN SKRIPSI

Yang bertandatangan di bawah ini,

Nama mahasiswa : Georel Jeferson Fransiskus Bonai
NIM : 22.83.0833

Menyatakan bahwa Skripsi dengan judul berikut:

***MULTINOMIAL NAÏVE BAYES* UNTUK DETEKSI *PHISHING* URL DENGAN PERBANDINGAN REPRESENTASI FITUR *BAG OF WORDS* DAN TF-IDF**

Dosen Pembimbing : Tonny Hidayat, S.Kom., M.Kom., Ph.D

1. Karya tulis ini adalah benar-benar ASLI dan BELUM PERNAH diajukan untuk mendapatkan gelar akademik, baik di Universitas AMIKOM Yogyakarta maupun di Perguruan Tinggi lainnya.
2. Karya tulis ini merupakan gagasan, rumusan dan penelitian SAYA sendiri, tanpa bantuan pihak lain kecuali arahan dari Dosen Pembimbing.
3. Dalam karya tulis ini tidak terdapat karya atau pendapat orang lain, kecuali secara tertulis dengan jelas dicantumkan sebagai acuan dalam naskah dengan disebutkan nama pengarang dan disebutkan dalam Daftar Pustaka pada karya tulis ini.
4. Perangkat lunak yang digunakan dalam penelitian ini sepenuhnya menjadi tanggung jawab SAYA, bukan tanggung jawab Universitas AMIKOM Yogyakarta.
5. Pernyataan ini SAYA buat dengan sesungguhnya, apabila di kemudian hari terdapat penyimpangan dan ketidakbenaran dalam pernyataan ini, maka SAYA bersedia menerima SANKSI AKADEMIK dengan pencabutan gelar yang sudah diperoleh, serta sanksi lainnya sesuai dengan norma yang berlaku di Perguruan Tinggi.

Yogyakarta, 23 Januari 2026

Yang Menyatakan,

Georel Jeferson Fransiskus Bonai

HALAMAN PERSEMBAHAN

Skripsi ini penulis persembahkan kepada:

Orang tua tercinta dan wali serta keluarga, yang senantiasa memberikan doa, kasih sayang, dukungan moral, serta pengorbanan yang tak ternilai selama penulis menempuh pendidikan.

Dosen pembimbing, yang telah dengan sabar memberikan bimbingan, arahan, dan motivasi sehingga skripsi ini dapat diselesaikan dengan baik.

Para dosen penguji, atas saran, kritik, dan masukan yang membangun demi penyempurnaan skripsi ini.

Semoga karya ini dapat menjadi bentuk rasa syukur dan tanggung jawab penulis atas ilmu yang telah diperoleh.

KATA PENGANTAR

Puji dan syukur penulis panjatkan ke hadirat Tuhan Yesus Kristus, karena atas rahmat dan karunia-Nya penulis dapat menyelesaikan skripsi ini dengan judul ”*Multinomial Naïve Bayes* untuk Deteksi *Phishing* URL dengan Perbandingan Representasi Fitur *Bag of Words* dan TF-IDF” Skripsi ini disusun sebagai salah satu syarat untuk memperoleh gelar Sarjana pada Program Studi Teknik Komputer, Fakultas Ilmu Komputer, Universitas Amikom Yogyakarta.

Penulis menyadari bahwa dalam proses penyusunan skripsi ini tidak terlepas dari bantuan, bimbingan, dan dukungan dari berbagai pihak. Oleh karena itu, pada kesempatan ini penulis ingin menyampaikan ucapan terima kasih yang sebesar-besarnya kepada:

1. Bapak Tonny Hidayat, S.Kom., M.Kom., Ph.D selaku Dosen Pembimbing, yang telah meluangkan waktu, tenaga, dan pikiran untuk memberikan bimbingan, arahan, serta masukan yang sangat berarti selama proses penyusunan skripsi ini hingga dapat diselesaikan dengan baik.
2. Bapak dan Ibu Dosen Penguji, yang telah memberikan saran, kritik, dan masukan yang membangun dalam proses pendadaran, sehingga skripsi ini dapat disempurnakan.
3. Orang Tua dan Keluarga tercinta, yang telah memberikan dukungan moral, doa, serta bantuan finansial selama penulis menempuh pendidikan hingga tahap penyelesaian skripsi ini.
4. Yoga Pangestu teman saya, yang telah banyak membantu penulis, baik melalui diskusi, pemberian contoh kerangka presentasi, contoh skripsi, naskah, maupun berbagai informasi yang berkaitan dengan isi dan penyusunan skripsi ini.

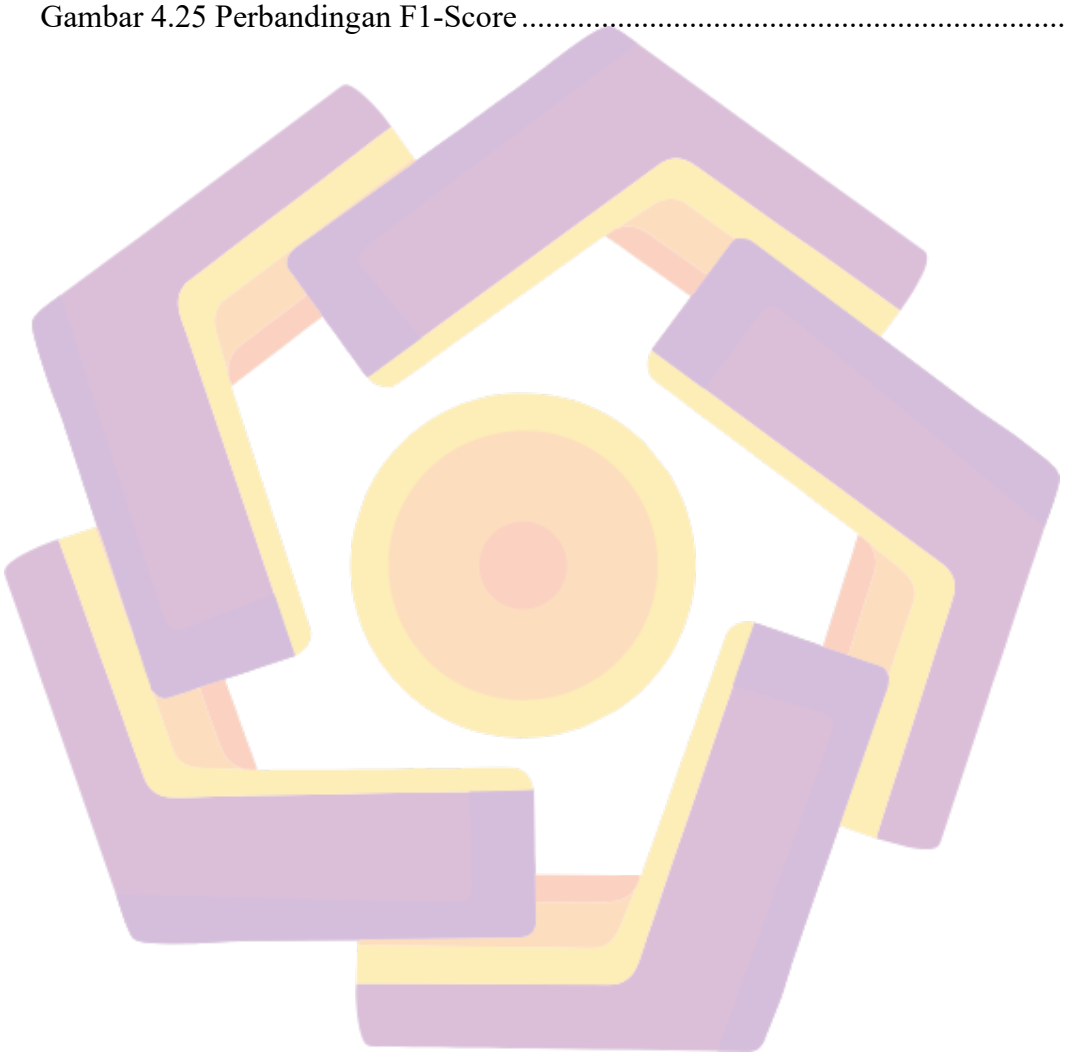
Yogyakarta, 23 Januari 2026

Penulis

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR



DAFTAR LAMPIRAN

DAFTAR LAMBANG DAN SINGKATAN

SVM	Support Vector Machine
BoW	Bag of Words
TF-IDF	Term Frequency-Inverse Document Frequency
NLP	Natural Language Processing
URL	Uniform Resource Locator
CV	Cross Validation
SMOTE	Synthetic Minority Over-sampling Technique
CSV	Comma Separated Values
HTTPS	Hypertext Transfer Protocol Secure

DAFTAR ISTILAH

Phishing	Teknik penipuan untuk memperoleh informasi sensitif melalui URL palsu
Klasifikasi	Proses pengelompokan data ke dalam kelas tertentu
Dataset	Kumpulan data yang digunakan dalam penelitian
Preprocessing	Tahapan pembersihan dan normalisasi data
Feature Extraction	Proses mengubah data mentah menjadi fitur numerik
Text Vectorization	Proses mengubah teks menjadi vektor numerik
Imbalanced Dataset	Dataset dengan distribusi kelas tidak seimbang
Confusion Matrix	Tabel evaluasi hasil klasifikasi
Accuracy	Ukuran ketepatan prediksi model
Precision	Ketepatan prediksi positif
Recall	Kemampuan model menemukan data positif
F1-Score	Rata-rata harmonik precision dan recall
Character N-Gram	Inti dari pendekatan URL-Based Phishing
Model Training	Tahapan Pembelajaran Model
Model Testing	Tahapan Evaluasi Model
Overfitting	Resiko Model terlalu cocok pada data latih
Real-time Detection	Konteks Aplikasi Phishing

INTISARI

Phishing merupakan salah satu bentuk serangan siber yang memanfaatkan URL palsu untuk menipu pengguna agar memberikan informasi sensitif. Seiring meningkatnya jumlah serangan phishing, diperlukan metode deteksi otomatis yang efektif dan efisien. Penelitian ini bertujuan untuk membangun dan mengevaluasi model deteksi phishing URL menggunakan algoritma Multinomial Naïve Bayes dengan dua metode representasi teks, yaitu Bag of Words (BoW) dan Term Frequency–Inverse Document Frequency (TF-IDF).

Dataset yang digunakan terdiri dari beberapa kelas URL, yaitu benign, phishing, malware, dan defacement, yang memiliki distribusi data tidak seimbang. Untuk mengatasi permasalahan tersebut, diterapkan teknik Synthetic Minority Over-sampling Technique (SMOTE) pada data latih. Proses penelitian meliputi tahapan preprocessing teks, representasi fitur menggunakan character n-gram, pelatihan model, serta evaluasi menggunakan metrik accuracy, precision, recall, dan F1-score. Selain itu, dilakukan 5-Fold Cross Validation untuk mengukur stabilitas dan kemampuan generalisasi model.

Hasil pengujian menunjukkan bahwa kedua metode representasi teks mampu menghasilkan performa klasifikasi yang baik. Representasi BoW menghasilkan accuracy sebesar (78,20%) dan F1-score sebesar (78,59%), sedangkan TF-IDF memperoleh accuracy sebesar (77,90%) dan F1-score sebesar (79,16%). Dari sisi stabilitas, hasil validasi menunjukkan nilai CV-Mean yang konsisten pada kedua metode. Selain itu, BoW memiliki waktu komputasi vektorisasi yang lebih rendah dibandingkan TF-IDF, sehingga lebih efisien dalam pemrosesan data skala besar. Berdasarkan evaluasi kinerja, stabilitas model, dan efisiensi waktu komputasi, model Multinomial Naïve Bayes dengan representasi BoW dan penerapan SMOTE dipilih sebagai model terbaik dalam penelitian ini.

Kata kunci: phishing URL, Multinomial Naïve Bayes, Bag of Words, TF-IDF, SMOTE

ABSTRACT

Phishing is one of the most common cyber threats that exploits fraudulent URLs to deceive users into disclosing sensitive information. As phishing attacks continue to increase, effective and efficient automated detection methods are required. This study aims to develop and evaluate a phishing URL detection model using the Multinomial Naïve Bayes algorithm with two text representation approaches, namely Bag of Words (BoW) and Term Frequency–Inverse Document Frequency (TF-IDF).

The dataset used in this study consists of several URL classes, including benign, phishing, malware, and defacement, which exhibit an imbalanced class distribution. To address this issue, the Synthetic Minority Over-sampling Technique (SMOTE) was applied to the training data. The research process includes text preprocessing, feature representation using character n-grams, model training, and performance evaluation using accuracy, precision, recall, and F1-score metrics. In addition, 5-Fold Cross Validation was conducted to assess model stability and generalization capability.

The experimental results show that both text representation methods achieve good classification performance. The BoW representation obtained an accuracy of (78.20%) and an F1-score of (78.59%), while TF-IDF achieved an accuracy of (77.90%) and an F1-score of (79.16%). Cross-validation results indicate consistent performance across folds for both approaches. In terms of computational efficiency, BoW requires lower vectorization time compared to TF-IDF, making it more suitable for large-scale phishing URL detection. Based on overall classification performance, model stability, and computational efficiency, the Multinomial Naïve Bayes model with BoW representation and SMOTE is selected as the best-performing model in this study.

Keyword**:** *phishing URL*, *Multinomial Naïve Bayes*, *Bag of Words*, TF-IDF, SMOTE