JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 4, No. 6 Tahun 2026
ISSN 3025-0919 (media online)
Hal 1473-1479

Penerapan Algoritma Naive Bayes untuk Deteksi *Website Phishing* Berbasis Fitur URL dan Konten

Andika Arya Pratama[1], Aqidatul Nur Robby[2], Fready Anggara[3], Jordi Ricaldo[4], Mohamad Ryan Syekhan Ramadan[5], Rahmawati [6]

[1-6] Program Studi Teknik Informatika, Universitas Pamulang, Tangerang Selatan, Indonesia
Email: [1]andikaaryapratama0805@gmail.com, [2]aqidrobby9@gmail.com, [3]freadyanggara@gmail.com, [4]jordiricaldo555@gmail.com, [5]ryansyekhanx@gmail.com, [6]dosen02394@unpam.ac.id

Abstrak−Ancaman siber berupa *phishing* terus berkembang dan menjadi salah satu metode kejahatan digital paling merugikan di era internet. *Website phishing* dirancang untuk menipu pengguna agar menyerahkan informasi sensitif seperti kredensial akun, data perbankan, maupun informasi pribadi. Penelitian ini mengimplementasikan algoritma *Categorical Naive Bayes* untuk mengklasifikasikan *website phishing* berbasis fitur URL dan konten halaman web. *Dataset* yang digunakan adalah *Phishing Website*s *Dataset* dari UCI *Machine learning* Repository yang memuat 11.055 *instance* dengan 30 fitur kategorikal. Data dibagi dengan rasio 80:20 untuk data latih dan data uji. Hasil penelitian menunjukkan bahwa model *Categorical Naive Bayes* mencapai akurasi 92,94% pada data uji, dengan *Precision* rata-rata 0,93, *Recall* rata-rata 0,93, dan *F1-Score* rata-rata 0,93. Selisih akurasi antara data latih (92,99%) dan data uji (92,94%) hanya sebesar 0,05%, yang mengindikasikan bahwa model tidak mengalami *overfitting* dan memiliki kemampuan generalisasi yang sangat baik. Algoritma Naive Bayes terbukti merupakan pilihan yang efektif dan efisien untuk deteksi *phishing* pada data fitur bertipe kategorikal dengan biaya komputasi yang rendah.

Kata Kunci: *Naive Bayes; deteksi phishing; keamanan siber; machine learning; klasifikasi URL*

Abstract−*Cyber threats in the form of phishing continue to grow and become one of the most harmful digital crime methods in the internet era. Phishing websites are designed to deceive users into submitting sensitive information such as account credentials, banking data, and personal information. This study implements the Categorical Naive Bayes algorithm to classify phishing websites based on URL features and web page content. The dataset used is the Phishing Websites Dataset from the UCI Machine learning Repository containing 11,055 instances with 30 categorical features. The data is split with an 80:20 ratio for Training and testing data. Results show that the Categorical Naive Bayes model achieved 92.94% accuracy on test data, with an average Precision of 0.93, average Recall of 0.93, and average F1-Score of 0.93. The accuracy difference between Training data (92.99%) and test data (92.94%) is only 0.05%, indicating that the model does not overfit and has excellent generalization ability. The Naive Bayes algorithm proves to be an effective and efficient choice for phishing detection on categorical feature data with low computational cost.*

Keywords: *Naive Bayes; phishing detection; cybersecurity; machine learning; URL classification*

1. PENDAHULUAN

Perkembangan teknologi informasi yang pesat membawa dampak signifikan terhadap perubahan pola kehidupan masyarakat, khususnya dalam aktivitas berbasis internet. Namun di balik manfaat tersebut, ancaman keamanan siber juga turut meningkat secara masif. Salah satu bentuk ancaman yang paling umum dan berbahaya adalah serangan *phishing* (Jain & Gupta, 2017).

Phishing merupakan teknik rekayasa sosial (*social engineering*) yang digunakan oleh pelaku kejahatan siber untuk mencuri informasi sensitif pengguna, seperti nama pengguna, kata sandi, nomor kartu kredit, dan data perbankan lainnya, dengan cara menyamar sebagai entitas terpercaya melalui media digital seperti *email*, pesan teks, atau *website* palsu (Khonji et al., 2013).

Menurut laporan *Anti-Phishing Working Group (APWG)*, jumlah serangan *phishing* terus meningkat dari tahun ke tahun. Pada tahun 2023, lebih dari 4,7 juta serangaunik tercatat, menjadikannya salah satu ancaman siber terbesar secara global (Anti-Phishing Working Group, 2024). Di Indonesia, Badan Siber dan Sandi Negara (BSSN) juga mencatat peningkatan insiden siber yang signifikan, di mana *phishing* menempati peringkat atas sebagai vektor serangan utama.

Pendekatan konvensional seperti daftar hitam (*blacklist*) URL terbukti tidak cukup efektif karena penyerang dapat dengan mudah membuat *domain* baru dalam hitungan menit (Afroz & Greenstadt, 2011). Oleh karena itu, diperlukan pendekatan yang lebih cerdas dan adaptif, salah satunya adalah penerapan teknik *machine learning* untuk mendeteksi pola-pola *phishing* secara otomatis.

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 4, No. 6 Tahun 2026
ISSN 3025-0919 (media online)
Hal 1473-1479

Website phishing adalah halaman web yang sengaja dibuat untuk meniru tampilan situs web yang sah guna menipu pengguna (Mohammad et al., 2014a) . Karakteristik umum *website phishing* meliputi penggunaan alamat IP langsung sebagai URL, subdomain yang mencurigakan, nama *domain* yang menyerupai nama *brand* terkenal dengan sedikit modifikasi, serta penggunaan sertifikat SSL palsu.

Penelitian terdahulu mengidentifikasi bahwa fitur-fitur *website* dapat dibagi menjadi tiga kategori utama (Fette et al., 2007):

- Fitur berbasis URL: Meliputi panjang URL, penggunaan alamat IP, simbol khusus seperti "@" dan tanda hubung, penggunaan layanan pemendekan URL (shortening service), dan lain-lain.
- Fitur berbasis konten halaman: Meliputi analisis *anchor tag*, *Server Form Handler (SFH)*, keberadaan iframe, *popup*, dan teknik penyembunyian menu konteks (*right-click disabled*).
- Fitur berbasis *domain*: Meliputi usia *domain*, rekam jejak DNS, dan informasi WHOIS.

Naive Bayes adalah algoritma klasifikasi probabilistik yang didasarkan pada Teorema Bayes dengan asumsi independensi kondisional (*conditional independence*) yang kuat di antara setiap pasang fitur (Mitchell, 1997). Secara matematis, Teorema Bayes diformulasikan sebagai:

$$P(C|X) = P(X|C) \times P(C) / P(X)$$

Di mana:

- P(C|X) adalah probabilitas *posterior* dari kelas C diberikan fitur X
- P(C) adalah probabilitas *prior* dari kelas C
- P(X|C) adalah *likelihood* dari fitur X diberikan kelas C
- P(X) adalah probabilitas bukti (konstanta normalisasi)

Dengan asumsi independensi, maka: $P(X|C) = \prod P(xi|C)$. *Categorical Naive Bayes* (*Categoricalnb*) adalah varian Naive Bayes yang dirancang khusus untuk data fitur bertipe kategorikal, di mana setiap fitur diasumsikan mengikuti distribusi kategoris (Pedregosa et al., 2011). Model ini sangat sesuai digunakan pada *dataset phishing* yang memiliki nilai fitur kategorikal diskret.

Tabel berikut merangkum beberapa penelitian terdahulu yang relevan:

Tabel 1. Perbandingan Penelitian Terkait

Peneliti	Tahun	Metode	*Dataset*	Akurasi
Mohammad et al.	2014	*Hybrid* (Rule+ML)	UCI *Phishing*	93,44%
Sahingoz et al.	2019	*Random forest*	80K URL	97,98%
Rao & Pais	2019	*Logistic Regression*	10K URL	93,28%
Zhu et al.	2020	*XGBoost*	UCI *Phishing*	97,10%
Penelitian ini	2022	Naive Bayes (*Categoricalnb*)	UCI *Phishing*	92,94%

Dibandingkan dengan penelitian terdahulu, pendekatan yang diusulkan dalam penelitian ini menggunakan algoritma yang lebih sederhana dengan kompleksitas komputasi yang rendah, namun tetap menghasilkan performa yang kompetitif. Keunggulan utama Naive Bayes adalah waktu pelatihan yang sangat cepat dan kebutuhan memori yang minimal.

Berdasarkan latar belakang tersebut penelitian ini bertujuan untuk, mengimplementasikan algoritma *Categorical Naive Bayes* untuk klasifikasi *website phishing*, mengevaluasi performa model menggunakan metrik akurasi, *Precision*, *Recall*, dan *F1-Score*, dan menganalisis kemampuan generalisasi model dengan membandingkan performa pada data latih dan data uji.

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 4, No. 6 Tahun 2026
ISSN 3025-0919 (media online)
Hal 1473-1479

Penelitian ini diharapkan dapat memberikan kontribusi berupa model deteksi *phishing* yang dapat diintegrasikan ke dalam sistem keamanan jaringan, peramban web (*browser extension*), atau aplikasi keamanan siber lainnya sebagai lapisan perlindungan tambahan bagi pengguna internet.

2. METODE

2.1 Alur Penelitian

Penelitian ini mengikuti metodologi *Knowledge Discovery in Databases (KDD)* yang terdiri atas tahapan berikut:

[Pengumpulan Data] → [Pra-pemrosesan Data] → [Ekstraksi Fitur] → [Pembagian Data] → [Pelatihan Model] → [Evaluasi Model] → [Analisis Hasil]

2.2 Dataset

Dataset yang digunakan dalam penelitian ini adalah "*Phishing Website*s *Dataset*" yang bersumber dari UCI *Machine learning* Repository (Mohammad et al., 2015). *Dataset* ini dikompilasi oleh Mohammad, Thabtah, dan McCluskey (2014) dan telah banyak digunakan sebagai *Benchmark* dalam penelitian deteksi *phishing*.

a. Karakteristik *Dataset*:

Tabel 2. Karakteristik Dataset

Properti	Nilai
Total *Instance*	11.055
Jumlah Fitur	30
Tipe Fitur	Kategorikal (-1, 0, 1)
Label Kelas	*Result* (-1 = *Phishing*, 1 = *Legitimate*)
Format File	ARFF (*Attribute-Relation File Format*)

b. Distribusi Kelas:

Tabel 3. Distribusi Kelas Dataset

Kelas	Label	Jumlah	Persentase
Phishing	-1	4.898	44,31%
Legitimate	1	6.157	55,69%

2.3 Deskripsi Fitur

Dataset memuat 30 fitur yang terbagi dalam kategori berikut:

Tabel 4. Daftar Fitur Dataset

No.	Nama Fitur	Kategori	Nilai
1	*having_IP_Address*	Berbasis URL	{-1, 1}
2	*URL_Length*	Berbasis URL	{-1, 0, 1}
3	*Shortining_Service*	Berbasis URL	{-1, 1}
4	*having_At_Symbol*	Berbasis URL	{-1, 1}

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 4, No. 6 Tahun 2026
ISSN 3025-0919 (media online)
Hal 1473-1479

5	*double_slash_redirecting*	Berbasis URL	{-1, 1}
6	*Prefix_Suffix*	Berbasis URL	{-1, 1}
7	*having_Sub_Domain*	Berbasis URL	{-1, 0, 1}
8	*SSLfinal_State*	Berbasis Keamanan	{-1, 0, 1}
9	*Domain_registeration_length*	Berbasis *Domain*	{-1, 1}
10	*Favicon*	Berbasis Konten	{-1, 1}
11	*Port*	Berbasis URL	{-1, 1}
12	*HTTPS_token*	Berbasis Keamanan	{-1, 1}
13	*Request_URL*	Berbasis Konten	{-1, 1}
14	*URL_of_Anchor*	Berbasis Konten	{-1, 0, 1}
15	*Links_in_tags*	Berbasis Konten	{-1, 0, 1}
16	*SFH*	Berbasis Konten	{-1, 0, 1}
17	*Submitting_to_email*	Berbasis Konten	{-1, 1}
18	*Abnormal_URL*	Berbasis *Domain*	{-1, 1}
19	*Redirect*	Berbasis URL	{0, 1}
20	*on_mouseover*	Berbasis Konten	{-1, 1}
21	*RightClick*	Berbasis Konten	{-1, 1}
22	*popUpWidnow*	Berbasis Konten	{-1, 1}
23	*Iframe*	Berbasis Konten	{-1, 1}
24	*age_of_domain*	Berbasis *Domain*	{-1, 1}
25	*DNSRecord*	Berbasis *Domain*	{-1, 1}
26	*web_traffic*	Berbasis *Domain*	{-1, 0, 1}
27	*Page_Rank*	Berbasis *Domain*	{-1, 1}
28	*Google_Index*	Berbasis *Domain*	{-1, 1}
29	*Links_pointing_to_page*	Berbasis *Domain*	{-1, 0, 1}
30	*Statistical_report*	Berbasis *Domain*	{-1, 1}

2.4 Pra-Pemrosesan Data

Proses pra-pemrosesan yang dilakukan meliputi:

1. Konversi Format: Data dibaca dari format ARFF menggunakan pustaka liac-arff dan dikonversi menjadi DataFrame Pandas.
2. Konversi Tipe Data: Semua nilai fitur dikonversi ke tipe integer karena seluruh fitur bersifat numerik kategorikal.

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 4, No. 6 Tahun 2026
ISSN 3025-0919 (media online)
Hal 1473-1479

3. Penanganan Nilai Negatif: *Categoricalnb* dari scikit-learn mensyaratkan nilai fitur non-negatif (≥ 0). Karena *dataset* mengandung nilai -1, dilakukan *Shifting* nilai dengan menambahkan nilai absolut dari nilai minimum (shift = 1), sehingga nilai {-1, 0, 1} menjadi {0, 1, 2}. Operasi ini dilakukan secara konsisten pada data latih dan data uji.

2.5 Pembagian Data

Data dibagi menggunakan teknik *Holdout Validation* dengan rasio:

- Data Latih (*Training* Set): 80% → 8.844 *instance*
- Data Uji (*Testing Set*): 20% → 2.211 *instance*

Pembagian dilakukan secara acak dengan random_state=42 untuk menjamin reprodusibilitas eksperimen.

2.6 Pembangunan Model

Model dibangun menggunakan kelas *Categoricalnb* dari pustaka scikit-learn versi 1.x. Seluruh parameter model menggunakan nilai *Default* (*Smoothing parameter* α = 1.0, yang setara dengan *Laplace smoothing*). Proses pelatihan dilakukan dengan metode fit(X_train, y_train).

2.7 Metrik Evaluasi

Model dievaluasi menggunakan metrik-metrik standar klasifikasi biner:

Akurasi: (TP + TN) / (TP + TN + FP + FN)
***Precision*:** TP / (TP + FP)
***Recall*:** TP / (TP + FN)
***F1-Score*:** 2 × (*Precision* × *Recall*) / (*Precision* + *Recall*)

Di mana TP = True Positive, TN = True Negative, FP = False Positive, FN = False Negative.

3. HASIL DAN PEMBAHASAN

3.1 Hasil Evaluasi Model

Setelah proses pelatihan dan pengujian, diperoleh hasil evaluasi sebagai berikut:

Tabel 5. Perbandingan Akurasi Data Latih dan Data Uji

Dataset	**Jumlah *Instance***	**Akurasi**
Data Latih (*Training*)	8.844	92,99%
Data Uji (*Testing*)	2.211	92,94%
Selisih	—	0,05%

Tabel 6. Confusion Matrix (Data Uji)

	Prediksi: *Phishing* (-1)	**Prediksi: *Legitimate* (1)**
Aktual: *Phishing* (-1)	861 (TP)	95 (FN)
Aktual: *Legitimate* (1)	61 (FP)	1.194 (TN)

Tabel 7. Classification Report (Data Uji)

Kelas	***Precision***	***Recall***	***F1-Score***	***Support***
Phishing (-1)	0,93	0,90	0,92	956
Legitimate (1)	0,93	0,95	0,94	1.255

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 4, No. 6 Tahun 2026
ISSN 3025-0919 (media online)
Hal 1473-1479

Macro Avg	0,93	0,93	0,93	2.211
Weighted Avg	0,93	0,93	0,93	2.211
Accuracy			0,93	2.211

3.2 Analisis Overfitting

Analisis *overfitting* dilakukan dengan membandingkan performa model pada data latih dan data uji. Hasil menunjukkan selisih akurasi yang sangat kecil, yaitu hanya 0,05% (92,99% vs 92,94%). Hal ini mengindikasikan bahwa model Naive Bayes yang dibangun tidak mengalami *overfitting* dan memiliki kemampuan generalisasi yang sangat baik.

Kondisi ini dipengaruhi oleh dua faktor utama:

1. Karakteristik Naive Bayes: Algoritma Naive Bayes secara inheren merupakan model dengan bias tinggi dan varians rendah (*High bias, low variance*). Asumsi independensi antar-fitur yang kuat membuat model lebih resisten terhadap *overfitting* dibandingkan model kompleks seperti *Decision Tree* tanpa *Pruning*.
2. Ukuran *Dataset* yang Memadai: Dengan 8.844 sampel data latih, distribusi probabilitas kelas dan fitur dapat diestimasi secara stabil, meminimalkan risiko model menghafal *Noise* dalam data latih.

3.3 Analisis Kesalahan Klasifikasi

Dari *Confusion Matrix* (Tabel 6), terdapat dua jenis kesalahan:

- *False Negative (*FN = 95): 95 *website phishing* yang salah diklasifikasikan sebagai *Legitimate*. Ini merupakan jenis kesalahan yang paling kritis dalam konteks keamanan siber karena pengguna tidak mendapat peringatan terhadap ancaman nyata.
- *False Positive* (FP = 61): 61 *website Legitimate* yang salah diklasifikasikan sebagai *phishing*. Kesalahan ini bersifat lebih dapat ditoleransi karena hanya menyebabkan ketidaknyamanan pengguna.

Nilai *Recall* untuk kelas *phishing* sebesar 0,90 (90%) berarti model berhasil mendeteksi 90% dari seluruh *website phishing* yang ada, yang merupakan performa yang cukup baik untuk skenario deteksi ancaman.

3.4 Pembahasan

Hasil penelitian ini menunjukkan bahwa *Categorical Naive Bayes* merupakan pilihan algoritma yang tepat dan efisien untuk klasifikasi *website phishing* berbasis fitur kategorikal. Keunggulan utamanya adalah:

1. Efisiensi Komputasi: Proses pelatihan berjalan sangat cepat (di bawah 1 detik pada *dataset* ini) karena Naive Bayes hanya memerlukan satu kali pemindaian data untuk mengestimasi probabilitas.
2. Interpretabilitas: Model probabilistik mudah diinterpretasikan. Kontribusi setiap fitur terhadap keputusan klasifikasi dapat dianalisis secara eksplisit.
3. Ketahanan terhadap Fitur Tidak Relevan: Jika terdapat fitur yang kurang informatif, dampaknya terhadap performa keseluruhan minimal.
4. Tidak Memerlukan *Tuning hyperparameter* yang Intensif: Performa baik dapat dicapai dengan parameter *Default*.

Namun, terdapat beberapa keterbatasan yang perlu diperhatikan:

1. Asumsi Independensi: Asumsi independensi antar-fitur seringkali tidak realistis dalam praktek. Beberapa fitur URL mungkin berkorelasi satu sama lain.
2. Keterbatasan pada Fitur Baru: Model bergantung pada fitur-fitur yang telah didefinisikan secara statis. *Website phishing* yang menggunakan teknik-teknik baru mungkin tidak tertangkap.

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 4, No. 6 Tahun 2026
ISSN 3025-0919 (media online)
Hal 1473-1479

4. KESIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan, dapat ditarik beberapa kesimpulan: (1) Algoritma *Categorical Naive Bayes* berhasil diimplementasikan untuk mendeteksi *website phishing* menggunakan 30 fitur URL dan konten halaman web dari *dataset* UCI *Phishing Website*s. (2) Model yang dihasilkan mencapai akurasi 92,94% pada data uji, dengan nilai *Precision* rata-rata 0,93, *Recall* rata-rata 0,93, dan *F1-Score* rata-rata 0,93. (3) Model tidak mengalami *overfitting*, dibuktikan dengan selisih akurasi data latih (92,99%) dan data uji (92,94%) yang hanya sebesar 0,05%, menunjukkan kemampuan generalisasi model yang sangat baik. (4) Algoritma Naive Bayes terbukti merupakan pilihan yang efektif dan efisien untuk deteksi *phishing*, khususnya pada data fitur bertipe kategorikal, dengan biaya komputasi yang rendah.

Untuk penelitian selanjutnya, disarankan: (1) Melakukan perbandingan performa dengan algoritma lain seperti *Random forest*, *XGBoost*, dan *Support Vector Machine (SVM)* untuk mendapatkan gambaran yang lebih komprehensif. (2) Mengintegrasikan fitur *real-time* seperti konten HTML yang di-*scraping* secara langsung agar sistem dapat mendeteksi *phishing* dinamis yang belum masuk ke dalam *database*. (3) Mengeksplorasi teknik seleksi fitur (*Feature selection*) untuk mengidentifikasi fitur-fitur paling diskriminatif guna meningkatkan efisiensi model. (4) Menerapkan teknik *Ensemble learning* atau *stacking* untuk meningkatkan nilai Recall pada kelas *phishing*, guna meminimalkan kasus *false negative.*

REFERENCES

Afroz, S., & Greenstadt, R. (2011). PhishZoo: Detecting phishing websites by looking at them. In *Proceedings of the 5th IEEE International Conference on Semantic Computing* (pp. 368–375). IEEE.

Anti-Phishing Working Group. (2024). *Phishing activity trends report: Q4 2023*.

Fette, I., Sadeh, N., & Tomasic, A. (2007). Learning to detect phishing emails. In *Proceedings of the 16th International World Wide Web Conference* (pp. 649–656).

Jain, A. K., & Gupta, B. B. (2017). Phishing detection: Analysis of visual similarity based approaches. *Security and Communication Networks, 2017*, Article 5421046.

Khonji, M., Iraqi, Y., & Jones, A. (2013). Phishing detection: A literature survey. *IEEE Communications Surveys & Tutorials, 15*(4), 2091–2121.

Mitchell, T. M. (1997). *Machine learning*. McGraw-Hill.

Mohammad, R. M., Thabtah, F., & McCluskey, T. (2014a). Intelligent rule-based phishing websites classification. *IET Information Security, 8*(3), 153–160.

Mohammad, R. M., Thabtah, F., & McCluskey, T. (2014b). Predicting phishing websites based on self-structuring neural network. *Neural Computing and Applications, 25*(2), 443–458.

Mohammad, R. M., Thabtah, F., & McCluskey, T. (2015). *Phishing websites dataset*. UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets/phishing+websites

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Rao, R. S., & Pais, A. R. (2019). Detection of phishing websites using an efficient feature-based machine learning framework. *Neural Computing and Applications, 31*(8), 3851–3873.

Sahingoz, O. K., Buber, E., Demir, O., & Diri, B. (2019). Machine learning based phishing detection from URLs. *Expert Systems with Applications, 117*, 345–357.

Zhu, E., Ju, Y., Chen, Z., Liu, F., & Fang, X. (2020). DTOF-ANN: An artificial neural network phishing detection model based on decision tree and optimal features. *Applied Soft Computing, 95*, Article 106505.