BAB 2
LANDASAN TEORI

2.1 Tinjauan Teori

Bab ini menguraikan landasan teori yang menjadi dasar dalam penelitian mengenai deteksi smishing berbasis *machine learning* dan *Explainable Artificial Intelligence (XAI*). Tinjauan pustaka ini mencakup konsep dasar smishing sebagai bentuk ancaman keamanan siber pada komunikasi mobile, penerapan *machine learning* dan *deep learning* dalam klasifikasi teks, serta pentingnya pendekatan *XAI* untuk meningkatkan transparansi dan interpretabilitas model.

2.1.1 Smishing

Smishing merupakan salah satu bentuk serangan *phishing* yang memanfaatkan layanan pesan singkat (*Short Message Service/SMS*) sebagai media utama untuk menipu korban. Istilah *smishing* berasal dari gabungan kata *SMS* dan *phishing*, yang merujuk pada upaya manipulasi pengguna agar memberikan informasi sensitif seperti kredensial akun atau data keuangan [12].

Berbeda dengan *phishing* berbasis *email*, *smishing* memiliki tingkat keterbacaan yang lebih tinggi karena *SMS* bersifat langsung dan sering kali dipercaya oleh pengguna. Serangan ini umumnya disamarkan sebagai pesan resmi dari institusi terpercaya dan sering mengandung unsur urgensi atau ancaman untuk mendorong korban bertindak cepat [4].

Perkembangan terbaru menunjukkan bahwa *smishing* memanfaatkan teknik penyamaran *URL*, variasi bahasa, dan konten berbasis kecerdasan buatan, sehingga menjadi ancaman serius dalam keamanan komunikasi *mobile*.

2.1.2 Machine Learning untuk Deteksi Smishing

Machine learning merupakan pendekatan komputasional yang memungkinkan sistem belajar dari data tanpa diprogram secara eksplisit. Dalam deteksi *smishing*, pendekatan ini digunakan untuk melakukan klasifikasi teks berdasarkan pola linguistik [2].

Secara umum, proses klasifikasi dapat dirumuskan sebagai berikut:

$$y = f(x) \tag{2.1}$$

di mana x merupakan representasi fitur dari pesan *SMS* dan $y \in \{0, 1\}$ adalah label kelas (0 = *non-smishing*, 1 = *smishing*).

2.1.3 TF-IDF (*Term Frequency-Inverse Document Frequency*)

TF-IDF merupakan salah satu metode ekstraksi fitur yang paling umum digunakan dalam pengolahan teks dan *text mining*. Metode ini digunakan untuk mengukur tingkat kepentingan suatu kata dalam sebuah dokumen terhadap keseluruhan kumpulan dokumen (*corpus*). *TF-IDF* bekerja dengan memberikan bobot tinggi pada kata yang sering muncul dalam suatu dokumen tetapi jarang muncul pada dokumen lain. Dengan demikian, kata-kata yang memiliki karakteristik khusus terhadap suatu dokumen akan memiliki nilai yang lebih besar dibandingkan kata umum yang muncul di hampir semua dokumen.

Pada penelitian deteksi *smishing*, *TF-IDF* digunakan untuk mengubah data *SMS* yang berbentuk teks menjadi representasi numerik sehingga dapat diproses oleh algoritma *machine learning* seperti *Logistic Regression* dan *Random Forest*. Metode ini efektif karena mampu merepresentasikan kata-kata penting yang sering muncul pada pesan *smishing*, seperti *"verify", "urgent", "bank", atau "click"*.

Secara matematis, *TF-IDF* terdiri dari dua komponen utama, yaitu *Term Frequency* (TF) dan *Inverse Document Frequency* (IDF).

Rumus 2.2 menunjukkan cara perhitungan *Term Frequency*.

$$TF(t,d) = \frac{f_{t,d}}{\sum_k f_{k,d}} \tag{2.2}$$

Rumus 2.3 menunjukkan cara perhitungan *Inverse Document Frequency*.

$$IDF(t) = \log\left(\frac{N}{df_t}\right) \tag{2.3}$$

Rumus 2.4 menunjukkan cara perhitungan nilai *TF-IDF* secara keseluruhan.

$$TFIDF(t,d) = TF(t,d) \times IDF(t) \tag{2.4}$$

di mana:

- $f_{t,d}$ = frekuensi term t dalam dokumen d
- N = jumlah total dokumen
- df_t = jumlah dokumen yang mengandung term t

Metode *TF-IDF* pertama kali diperkenalkan dalam bidang *information retrieval* dan masih menjadi pendekatan populer dalam klasifikasi teks karena sederhana namun efektif.

2.1.4 Logistic Regression

Logistic Regression merupakan algoritma klasifikasi *supervised learning* yang digunakan untuk memprediksi probabilitas suatu data termasuk ke dalam kelas tertentu. Pada kasus deteksi *smishing*, model ini digunakan untuk menentukan apakah suatu *SMS* termasuk kategori *smishing* atau bukan. *Logistic Regression* banyak digunakan karena memiliki interpretabilitas yang baik, proses pelatihan yang cepat, dan performa yang cukup tinggi pada data teks berdimensi besar.

Berbeda dengan regresi linear yang menghasilkan *output* kontinu, *Logistic Regression* menggunakan fungsi *sigmoid* untuk mengubah hasil perhitungan linear menjadi probabilitas bernilai antara 0 dan 1. Jika probabilitas melebihi ambang tertentu (biasanya 0.5), maka data diklasifikasikan ke kelas positif.

Fungsi *sigmoid* pada *Logistic Regression* dirumuskan sebagai berikut:

Rumus 2.5 menunjukkan cara perhitungan fungsi *sigmoid*.

$$P(y=1|x) = \frac{1}{1+e^{-(w^T x+b)}} \tag{2.5}$$

di mana:

- x = vektor fitur input
- w = bobot model
- b = bias
- $P(y=1|x)$ = probabilitas data termasuk ke kelas positif

Logistic Regression menjadi salah satu metode dasar yang sering digunakan dalam klasifikasi teks dan keamanan siber karena kemampuannya dalam menangani data berdimensi tinggi.

2.1.5 Random Forest

Random Forest merupakan algoritma *ensemble learning* yang dibangun dari kumpulan *decision tree*. Metode ini bekerja dengan membuat banyak pohon keputusan menggunakan subset data dan subset fitur yang berbeda, kemudian menggabungkan hasil prediksi seluruh pohon untuk menghasilkan keputusan akhir. Pendekatan ini mampu mengurangi risiko *overfitting* yang sering terjadi pada *single decision tree*.

Dalam deteksi smishing, *Random Forest* mampu mengenali pola kompleks dari kombinasi kata-kata tertentu yang sering muncul pada pesan penipuan. Selain itu, algoritma ini memiliki kemampuan yang baik dalam menangani data berdimensi tinggi dan fitur yang banyak, seperti hasil ekstraksi *TF-IDF*.

Prediksi akhir pada *Random Forest* diperoleh melalui mekanisme *majority voting*, yang dirumuskan sebagai berikut:

$$\hat{y} = \text{mode}(T_1(x), T_2(x), ..., T_n(x)) \tag{2.6}$$

di mana:

- $T_i(x)$ = prediksi dari pohon keputusan ke-i
- $\hat{y}$ = hasil prediksi akhir

Random Forest diperkenalkan oleh Breiman dan dikenal memiliki performa tinggi serta stabilitas yang baik pada berbagai permasalahan klasifikasi.

2.1.6 Deep Learning dalam Keamanan Siber

Deep Learning adalah salah satu cabang dari *machine learning* yang menggunakan jaringan saraf tiruan (*artificial neural network*) dengan lapisan yang banyak guna mempelajari representasi dari data yang kompleks. Dalam bidang keamanan siber, *deep learning* banyak digunakan untuk mendeteksi serangan *phishing*, *malware*, *spam*, dan *smishing* karena mampu mengenali pola tersembunyi dari data dalam jumlah besar.

Long Short-Term Memory (LSTM) merupakan salah satu dari aristektur *deep learning* yang populer untuk data tipe teks. *LSTM* dirancang untuk mengatasi *vanishing gradient* pada data berurutan. *LSTM* memiliki kemampuan untuk mengingat informasi jangka panjang melalui mekanisme gerbang (*gates*) yang mengatur aliran informasi.

2.1.7 Explainable Artificial Intelligence (XAI)

Explainable Artificial Intelligence (XAI) merupakan suatu pendekatan yang dirancang untuk meningkatkan keterbukaan serta kemampuan interpretasi pada model *Artificial Intelligence (AI)*, sehingga proses dan dasar pengambilan keputusan yang dilakukan oleh model dapat dipahami oleh pengguna atau manusia. Dalam bidang keamanan siber, *XAI* sangat penting karena membantu analis memahami alasan di balik prediksi model, terutama dalam sistem deteksi ancaman seperti *smishing*.

Salah satu metode *XAI* yang populer adalah *SHAP* (*Shapley Additive Explanations*). *SHAP* bekerja berdasarkan teori permainan (*game theory*) dengan menghitung kontribusi setiap fitur terhadap hasil prediksi model. Dengan *SHAP*, pengguna dapat mengetahui fitur mana yang paling memengaruhi keputusan model.

Nilai *Shapley* pada *SHAP* dirumuskan sebagai berikut:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f(S \cup \{i\}) - f(S)] \tag{2.7}$$

di mana:

- F = himpunan seluruh fitur
- S = subset fitur
- ϕ_i = kontribusi fitur ke-i

Selain *SHAP*, metode *LIME* (*Local Interpretable Model-Agnostic Explanations*) juga sering digunakan untuk menjelaskan prediksi model secara lokal dengan membangun model *linear* sederhana di sekitar sampel yang dianalisis. Dengan integrasi *XAI*, sistem deteksi smishing tidak hanya berfokus pada akurasi, tetapi juga pada transparansi dan akuntabilitas model [13, 14].