BAB II
TINJAUAN PUSTAKA

2.1 Teori Dalam Skripsi

Berikut merupakan beberapa teori yang diimplementasikan dan digunakan selama melakukan penelitian ini:

2.1.1 *Phishing*

Phishing adalah bentuk serangan rekayasa sosial di mana penyerang mencoba untuk secara menipu mendapatkan informasi sensitif pengguna dengan mengirimkan email yang mengklaim berasal dari organisasi yang sah. Mereka menipu pengguna untuk memberikan informasi rahasia yang dapat digunakan untuk pencurian identitas [10]. Bahaya ini terus berkembang karena peningkatan dalam penipuan, perusakan identitas, penipuan, dan serangan online ganda. Sebagian besar serangan disampaikan melalui email yang menarik pengguna untuk mengklik tautan yang tertanam dalam email dan membawa mereka ke situs web berbahaya. Para penyerang biasanya menargetkan informasi keuangan pengguna akhir dengan mengklaim sebagai bank mereka, perusahaan utilitas, *HM Revenue and Customs*, atau lembaga pemerintah lainnya untuk meyakinkan pengguna akhir membuka dokumen yang dilampirkan ke email, yang kemudian menargetkan informasi sensitif di sistem mereka.

2.1.2 Website Feature

Dalam penelitian yang dilakukan oleh Rami M. Mohammad dan rekan-rekannya yang berjudul "Penilaian Fitur Terkait Situs *Web Phishing* dengan Menggunakan Teknik Otomatis," mereka mengusulkan aturan atau kriteria untuk secara otomatis mengidentifikasi fitur-fitur tertentu. Struktur aturan ini terbagi menjadi empat kelas utama di mana fitur-fitur diklasifikasikan dan

ditempatkan sesuai dengan kelasnya. Keempat kelas tersebut mencakup fitur-fitur pada *address bar*, fitur yang dianggap abnormal, fitur yang diekstrak dari komponen HTML dan JavaScript, serta fitur berdasarkan domain. Dalam kelas fitur pada *address bar*, beberapa contohnya melibatkan penggunaan alamat IP, panjang URL, penggunaan simbol tertentu, dan penggunaan *subdomain*. Untuk kelas fitur yang dianggap abnormal, contoh-fiturnya mencakup URL permintaan, URI dari *anchor*, penangan formulir *server*, dan URL yang dianggap abnormal. Fitur yang diekstrak dari komponen HTML dan JavaScript mencakup halaman *redirect*, pemblokiran klik kanan, dan penggunaan *mouseover* untuk menyembunyikan tautan. Sementara itu, kelas fitur berdasarkan domain mencakup usia *domain*, catatan DNS, dan lalu lintas situs *web*.

2.2 Teori tentang *Framework* / Algoritma yang digunakan

Framework penelitian merujuk pada struktur atau kerangka kerja konseptual yang digunakan untuk merancang, mengembangkan, dan mengorganisasi suatu penelitian. Hal tersebut memberikan landasan teoritis dan metodologis yang memandu peneliti dalam merancang studi, mengumpulkan data menganalisis temuan, dan menyajikan hasil. Berikut merupakan *framework* atau algoritma yang digunakan dalam penelitian ini:

2.1.1 Algoritma *Random Forest*

Random Forest merupakan salah satu algoritma yang dapat digunakan untuk pengerjaan dengan metode klasifikasi. *Random Forest*, seperti namanya, berisi sejumlah besar pohon keputusan individual yang bertindak sebagai kelompok untuk menentukan *output*. Setiap pohon dalam *Random Forest* menentukan prediksi kelas, dan hasilnya akan menjadi kelas yang paling diprediksi di antara keputusan pohon-pohon tersebut. Alasan dari hasil yang mengagumkan dari *Random Forest* adalah karena pohon-pohon tersebut saling

melindungi dari kesalahan individu. Meskipun beberapa pohon mungkin memprediksi jawaban yang salah, banyak pohon lainnya akan memperbaiki prediksi akhir, sehingga sebagai kelompok pohon-pohon tersebut dapat bergerak ke arah yang benar. *Random Forest* mencapai pengurangan *overfitting* dengan menggabungkan banyak pembelajar lemah yang underfit karena mereka hanya menggunakan *subset* dari semua sampel pelatihan. *Random Forest* dapat mengatasi sejumlah besar variabel dalam suatu set data. Selain itu, selama proses konstruksi hutan, mereka membuat perkiraan tidak bias tentang kesalahan generalisasi. Selain itu, mereka dapat memperkirakan data yang hilang dengan baik. Kekurangan utama dari *Random Forest* adalah kurangnya reproduktivitas karena proses konstruksi hutan bersifat acak. Selain itu, sulit untuk menginterpretasi model akhir dan hasil yang berikutnya, karena melibatkan banyak pohon keputusan independen [6]. Untuk *Random Forest* yang terdiri dari *N trees* maka dapat dirumuskan sebagai berikut:

$$l(y) = argmax_c(\sum_{n=1}^{N} I_{h_n(y)=c})$$

Gambar 2. 1 Rumus *Random Forest*

Gambar 2.1 merupakan gambar dari penulisan rumus *Random Forest*. Dimana *I* merupakan suatu fungsi indikator dan h_n merupakan pohon ke-n dari *Random Forest*.

2.1.2 Algoritma *K-Nearest Neighbour*

K-Nearest Neighbour merupakan metode yang menggunakan algoritma *supervised* dimana hasil dari *query instance* yang baru diklasifikasikan berdasarkan mayoritas dari *label class* pada KNN. Tujuan dari algoritma KNN adalah untuk mengklasifikasikan objek baru berdasarkan atribut dan *training data*. Algoritma KNN bekerja berdasarkan jarak terpendek dari *query instance*

ke *training data* untuk menentukan *value* K. Dalam proses perhitungan jarak *value* tersebut dapat menggunakan beberapa rumus perhitungan jarak, seperti rumus *euclidean*. Rumus tersebut digunakan untuk menghitung jarak antar titik pada bidang Kartesian.

2.1.3 Algoritma *Support Vector Machine*

Support Vector Machine merupakan algoritma pembelajaran mesin yang digunakan untuk tugas klasfikasi dan regresi. Tujuan utama dari penggunaan SVM dalam konteks klasifikasi adalah menemukan *hyperlane* terbaik yang memisahkan dua kelas data. *Hyperlane* ini dipilih sedemikian rupa sehingga memiliki *margin* terbesar antara dua kelas, diukur sebagai jarak terdekat antara titik-titik data dari kedua kelas ke *hyperlane*.

2.1.4 Framework Flask

Flask merupakan *web microframework* yang berbasis dengan bahasa pemrograman *Python*. *Framework* ini memiliki beberapa fungsi yang cocok untuk diimplementasi dalam pengembangan program *website*. Meskipun termasuk dalam kategori *framework* yang ringan, *Flask* memiliki fungsi yang dapat dikembangkan sesuai dengan kebutuhan dari *developer* dan terintegrasi dengan *database* melalui *SQLAlchemy* dengan baik sesuai dengan prosedur yang berlaku [11].

2.3 Teori tentang *Tools* / *Software* yang digunakan

Terdapat beberapa *tools* atau *software* yang digunakan oleh peneliti demi keberlangsungan penelitian ini. Beberapa *tools* yang digunakan dalam penelitian adalah sebagai berikut:

2.3.1 Visual Studio Code

Visual Studio Code merupakan aplikasi teks *editor* untuk *source code* yang dikembangkan oleh perusahaan teknologi asal Amerika, Microsoft. Aplikasi ini dapat dioperasikan pada berbagai macam sistem operasi, seperti Windows,

MacOS, dan Linux. *Visual Studio Code* sendiri pada umumnya digunakan untuk kepentingan pengembangan *website*, tetapi juga mampu digunakan untuk kepentingan dengan bahasa pemrograman yang lebih luas dan memiliki beragam *extension* yang tersedia pada aplikasi.

2.3.2 Chrome Browser

Google Chrome merupakan aplikasi yang berfungsi untuk mencari, mengakses, dan menampilkan segala bentuk informasi yang tersedia di internet. *Google Chrome* merupakan salah satu *web browser* yang memiliki pengguna terbanyak di dunia. Dirilis pada tanggal 2 September 2008 oleh perusahaan teknologi Google kini *Google Chrome* sudah dapat diakses dengan menggunakan beragam sistem operasi, seperti Linux, MacOS, Microsoft, iOS, dan Android.

2.3.3 Python

Python merupakan salah satu bahasa pemrograman yang memiliki banyak kegunaan, seperti pengembangan *website*, pengembangan *software*, data sains, dan *machine learning*. Bahasa pemrograman ini memiliki keunggulan yang efisien dan mudah untuk dipelajari oleh para *developer* serta dapat dijalankan di berbagai macam *platform*. *Python* juga dapat terintegrasi dengan baik pada semua tipe sistem.

2.3.4 JavaScript

JavaScript merupakan bahasa pemrograman yang dapat digunakan untuk pengembangan *website*. Fitur-fitur yang tersedia pada bahasa pemrograman ini bersifat lebih dinamis dan interaktif, seperti *Object Oriented*, *client-side, high-level programming,* dan *loosely typed.* Tidak hanya digunakan untuk pengembangan *website,* JavaScript juga dapat digunakan untuk melakukan pengembangan aplikasi, *tools,* atau bahkan *game* pada *website*.

2.3.5 HTML dan CSS

Baik HTML dan CSS merupakan *tools* yang digunakan untuk membuat kerangka dasar pengembangan suatu *website*. Untuk HTML sendiri merupakan bahasa *markup* yang membangun kerangka dan mengatur sebuah *website* yang akan ditampilkan pada *browser*. Sedangkan untuk CSS sendiri merupakan aturan yang berfungsi mengendalikan berbagai macam *element* dan komponen yang ada pada *website* sehingga lebih terstruktur dan seragam.

2.3.6 API

Application Programming Interface (API) merupakan kumpulan aturan dan protokol yang memungkinkan berbagai perangkat lunak atau aplikasi untuk berkomunikasi antara satu dengan yang lainnya. API menyediakan cara bagi *developer* untuk mengakses fungsionalitas atau data dari suatu sistem tanpa perlu mengetahui detail internal dari sistem tersebut. API dapat berbentuk RESTFul API (*Representational State Transfer*), SOAP (*Simple Object Access Protocol*), atau jenis-jenis lainnya tergantung pada desain dan kebutuhan sistem yang sedang dikembangkan.

2.4 Penelitian Terdahulu

Dalam mengerjakan proyek ini, digunakan 4 penelitian terdahulu sebagai sumber literatur guna mendukung terselesaikannya penelitian ini. Pemilihan penelitian terdahulu didasarkan pada kemiripan isi pembahasan penelitian, yaitu tentang *phising* detection. Secara lebih spesifik mengarah ke implementasi algoritma klasifikasi sebagai metode untuk mendeteksi link *phising*. Penelitian-penelitian terdahulu dicari dengan sumber dari Google Scholar. Pemilihan dua sumber tersebut didasarkan pada kredibilitas dan legalitas penelitian-penelitian yang telah di-*upload* ke sumber tersebut. Berikut merupakan tabel beberapa penelitian terdahulu yang digunakan sebagai sumber literatur:

Tabel 2. 1 Tabel Penelitian Terdahulu

Tabel Penelitian Terdahulu	
Penulis	Mummad Adipa, Ahmad Turmudi Zy, M. Makmun Effendi
Nama Jurnal	Jurnal Restikom: Riset Teknik Informatika dan Komputer, ISSN 2686-4797 Vol. 5, No.2, Agustus 2023, hlm.148-157, https://restikom.nusaputra.ac.id
Judul	Klasifikasi Email Phishing Menggunakan Algoritma K-Nearest Neighbor
Permasalahan	Permasalahan pada penelitian ini, yaitu meneliti untuk menerapkan metode algoritma K-Nearest Neighbor untuk proses mengklasifikasi suatu email termasuk klasifikasi *phishing* atau aman untuk dikunjungi. Ini dilakukan guna mengantisipasi maraknya kejadian *phishing* yang terjadi di Indonesia pada tahun 2022.
Tabel Penelitian Terdahulu	
Penulis	Ankit Kumar Jaln, B. B. Gupta
Nama Jurnal	Journal of Ambient Intelligence and Humanized Computing, *Received*: 11 December 2017 / *Accepted*: 14 April 2018, https://doi.org/10.1007/s12652-018-0798-z
Judul	*A Machine Learning Based Approach for Phishing Detection Using Hyperlinks Information*
Permasalahan	Pertumbuhan berbagai teknik baru dalam deteksi phishing belakangan ini telah melibatkan penerapan berbagai teknik berbasis *machine learning*. Dalam teknik-teknik berbasis *machine learning*, sebuah algoritma klasifikasi dilatih menggunakan beberapa fitur yang dapat membedakan situs *web phishing* dari yang sah (Jain dan Gupta 2016a). Fitur-fitur ini diekstraksi dari berbagai sumber seperti URL, sumber halaman mesin pencari, lalu lintas situs web, mesin pencari, DNS, dan lain sebagainya. Metode berbasis *machine learning* yang sudah ada mengekstraksi fitur dari pihak ketiga, mesin pencari, dan sebagainya. Oleh karena itu, metode-metode tersebut cenderung rumit, lambat, dan

Tabel Penelitian Terdahulu	
	tidak sesuai untuk lingkungan waktu nyata. Situs *web phishing* bersifat singkat, dan ribuan situs *web* palsu dibuat setiap hari. Oleh karena itu, diperlukan solusi deteksi *phishing* yang *real-time*, cepat, dan cerdas.
Tabel Penelitian Terdahulu	
Penulis	Mohammed Almseidin, AlMaha Abu Zuraiq, Mouhammad Al-Kasassbeh, Nidal Alniadami
Nama Jurnal	*International Journal of Interactive Mobile Technologies* Vol. 13, No.12 (Dec 18, 2019), pp.171-183, https://www.learnteclib.org/p/216410/
Judul	*Phishing Detection Based on Machine Learning and Feature Selection Methods*
Permasalahan	Permasalahan pada penelitian ini *cyber attack* yang paling banyak ditemukan pada tahun 2018 adalah *phishing attack*. Dalam proses mendeteksi serangan *phishing,* yang menjadi tantangan terbesarnya adalah menemukan teknik yang tepat dan efektif untuk mendeteksi serangan tersebut. Para oknum yang melakukan serangan *phishing* juga tidak tinggal diam dengan terus melakukan peningkatan strategi yang dapat membuat *web page* yang sekaligus mampu melindungi diri mereka sendiri dari banyak bentuk deteksi *phishing*. Metode deteksi yang efektif dan tangguh sangat diperlukan untuk melawan teknik adaptif yang digunakan oleh para oknum serangan tersebut.
Tabel Penelitian Terdahulu	
Penulis	Ozgur Koray Sahingoz, Ebubekir Buber, Onder Demir, Banu Diri
Nama Jurnal	*Expert Systems With Applications,* Vol.117, 1 March 2019, *pages* 345-357, https://doi.org/10.1016/j.eswa.2018.09.029
Judul	*Machine Learning Based Phishing Detection from URLs*
Permasalahan	Penelitian ini berfokus proses deteksi *phishing web page* secara *real-time* dengan cara menyelidiki URL dari halaman *web* tersebut menggunakan algoritma *machine learning* dan

Tabel Penelitian Terdahulu	
	berbagai set *feature.* Dalam proses eksekusi dari algoritma *machine learning*, bukan hanya *dataset* tetapi ekstraksi dari fitur-fitur yang ada di *dataset* juga sangat penting. Oleh karena itu langkah pertama adalah dengan mengumpulkan beragam URL, baik yang *phishing* atau aman untuk keperluan *dataset*. Kemudian mendefinisikan tiga jenis set fitur yang berbeda, yaitu *Word Vector,* berbasis NLP, dan fitur *Hybrid* untuk mengukur efisiensi dari sistem yang diusulkan.

Tabel 2.1 merupakan tabel yang berisikan informasi tentang gambaran besar dari beberapa penelitian terdahulu yang digunakan sebagai referensi dalam mengerjakan penelitian ini. Penelitian-penelitian terdahulu memiliki peran yang signifikan dalam mengembangkan ilmu terkait pembuatan *website* untuk deteksi *phishing*. Riset-riset sebelumnya telah menyumbangkan pemahaman mendalam tentang metode-metode serangan *phishing*, pola perilaku penipuan *online*, serta identifikasi ciri-ciri halaman web yang mencurigakan. Informasi yang diperoleh dari penelitian-penelitian tersebut menjadi dasar pengetahuan yang berharga bagi pengembang alat deteksi *phishing*. Kontribusi-kontribusi ini meliputi pengembangan model kecerdasan buatan yang mampu mengenali pola *phishing*, analisis data historis untuk mengidentifikasi *tren* serangan, dan perbaikan terus-menerus terhadap teknik deteksi. Dengan memanfaatkan temuan-temuan sebelumnya, dalam penelitian dapat merancang solusi yang lebih efektif dan tangguh dalam melawan ancaman *phishing* di dunia daring.