BAB I

PENDAHULUAN

1.1 Latar Belakang Masalah

Era modern saat ini hampir setiap perusahaan atau lembaga memiliki situs website sendiri untuk memenuhi dan menunjang kebutuhan dalam menjalankan tujuan perusahaan. Website sangat bermanfaat bagi perusahaan atau lembaga karena dapat memberi informasi profil atau membantu mengiklankan produk atau jasa perusahaan/lembaga. Website juga memudahkan masyarakat karena hanya dengan *online* melalui gadget bisa mendapat informasi atau keperluan lain tanpa harus keluar mendatangi perusahaan/lembaga tersebut.

Seiring perkembangan teknologi website semakin banyak tersebar di internet kejahatan dalam internet seperti yang dilaporkan oleh APWG (*Anti-Phishing Working Group*) [1] berbanding lurus dengan perkembangan teknologi termasuk internet. Dari berbagai banyak situs website saat ini terdapat oknum atau pihak yang ingin mengambil data-data pengguna untuk kepentingan pribadi mereka dengan cara memancing/mengundang mereka masuk ke situs website palsu (*phishing*) mereka yang dibuat hampir sama dengan situs website aslinya. Pengunjung yang masuk ke website palsu tersebut diminta data-data pribadi mereka untuk kepentingan dan bisa merugikan pengunjung website tersebut. Dari masalah ini maka website phising bisa dikatakan adalah tindak penipuan terhadap masyarakat.

Karena itu dibutuhkan cara untuk dapat mengetahui bagaimana membedakan antara situs website legal atau phising yaitu dengan *data mining*.

Data mining adalah proses penggalian atau ekstraksi data mulai dari pengumpulan, pemilahan, teknik statistik, *machine learning*, matematika, dan kecerdasan buatan untuk mencari pengetahuan, pola, atau informasi dari sekumpulan data yang sangat besar [2].

Data mining memiliki banyak fungsi seperti yang dijabarkan oleh [3] yakni deskripsi, prediksi, estimasi, klasifikasi, klastering, dan asosiasi. Dalam penelitian ini akan menggunakan fungsi klasifikasi untuk mengelompokkan website apakah termasuk legal atau phising. Klasifikasi sendiri memiliki banyak algoritma seperti C4.5, ID3, K-Nearest Neighbor, Naïve Bayes, dll. Algoritma yang digunakan dalam penelitian ini adalah Naïve Bayes dengan cara memasukan variabel – variabel seperti jumlah *link*, sub domain, umur domain, *traffic* website, dll kemudian menghitung dengan metode Naïve Bayes untuk memutuskan apakah suatu website termasuk phising atau legal. Naïve Bayes dipilih karena berdasarkan penelitian [4] algoritma ini memiliki kecepatan *training time* tercepat yakni 0,04 detik dan tingkat akurasi yang tidak terlalu jauh berbeda dengan algoritma klasifikasi lain. Faktor *training time* yang cepat ini penting karena data yang digunakan dalam penelitian ini sebanyak 11.055.

Hasil penelitian ini diharapkan adalah informasi/pola/pengetahuan yang didapat dari proses *data mining* akan bermanfaat dan dapat digunakan untuk mengklasifikasikan apakah suatu website termasuk phising atau legal. Dari hasil yang didapat akan diimplementasikan menggunakan bahasa python dengan memasukkan URL (*Uniform Resource Locator*) dari sebuah website yang akan dicek. Pengguna tinggal menekan tombol dan kemudian menunggu

proses pengolahan dari pola/pengetahuan yang sudah didapat dan akan diklasifikasikan apakah website yang dicek tersebut website phising atau legal.

1.2 Rumusan Masalah

Adapun yang menjadi rumusan masalah berdasarkan latar belakang yang telah dikemukakan adalah sebagai berikut:

1. Bagaimana cara mengklasifikasikan website termasuk phising atau legal?
2. Seberapa tinggi tingkat akurasi metode naïve bayes dalam memecahkan masalah ini?

1.3 Batasan Masalah

1. Dataset yang diambil untuk mencari pola/pengetahuan diambil dari situs UCI Machine Learning Repository.
2. Data sampel website phising diambil dari situs PhishTank.
3. Penelitian ini hanya membahas website phising tidak meliputi *phishing email*.
4. Variabel-variabel yang digunakan untuk mencari pola/pengetahuan dari dataset website phising yang didapat adalah 11 variabel dengan rincian 10 atribut dan 1 target kelas.
5. Evaluasi hasil yang didapat menggunakan *confusion matrix* dan ROC (*Receiver Operating Characteristic*) *curve*.
6. Proses implementasi/pengujian klasifikasi website dengan cara memasukkan URL (*Uniform Resource Locator*) website yang diuji.

1.4 Tujuan Penelitian

Tujuan penelitian ini berdasarkan rumusan masalah yang telah ditulis diatas adalah :

1. Mengklasifikasikan website phising atau legal.
2. Mengetahui tingkat akurasi metode naïve bayes dalam mengklasifikasi website phising atau legal.

1.5 Metode Penelitian

Penelitian ini masuk dalam penelitian kuantitatif yaitu proses menemukan pengetahuan menggunakan data berupa angka sebagai alat menganalisis keterangan mengenai apa yang ingin diketahui yakni data website.

1.5.1 Model Penelitian

Alur dalam penelitian ini menggunakan standar yang digunakan dalam *data mining* sesuai buku karya Larose, Daniel T. berjudul "Data Mining Methods and Models" [5] yaitu CRISP-DM (*Cross-Industry Standard Process Model for Data Mining*) yang memiliki tahapan sebagai berikut:

1. Business Understanding

   Tahap pertama dalam CRISP-DM yakni mendefinisikan proyek atau tujuan apa yang akan dilakukan dalam penelitian.

2. Data Understanding

   Secara garis besar tahap ini untuk mengumpulkan data dan memeriksa data sehingga dapat mengidentifikasi masalah dalam data.

3. Data Preparation

Memperbaiki masalah dalam data dan mempersiapkan data untuk memastikan data tepat dan siap diolah untuk algoritma yang digunakan. Dalam tahap ini dataset dibagi menjadi dua, data training yaitu data yang digunakan untuk mencari pengetahuan/pola dan data testing yaitu data untuk menguji pola yang sudah didapat dari data training.

4. Modelling

Teknik modelling yang bervariasi seperti *classification*, *scoring*, *ranking*, *clustering*, dsb dipilih dan diaplikasikan.

5. Evaluation

Melakukan evaluasi terhadap model untuk mengambil keputusan mengenai penggunaan *data mining*.

6. Deployment

Penerapan model yang sudah dibuat, pengetahuan yang didapat butuh untuk diorganisasikan dan dipresentasikan dengan cara dimana pengguna mudah untuk memahami.

1.5.2 Pengumpulan Data

Sekumpulan data yang digunakan dalam penelitian ini untuk diolah dan dicari pengetahuan atau dalam istilah *data mining* disebut dataset didapat dari situs UCI Machine Learning Repository. Sedangkan untuk sampel website phising yang akan digunakan untuk menguji pengetahuan/pola yang diperoleh setelah proses *data mining* didapat dari situs PhishTank.

1.5.3 Alat Penelitian

Perangkat lunak yang digunakan:

1. Microsoft Excel
2. Rapidminer Studio
3. PyCharm Community Edition (Python IDE)
4. Web browser: Mozilla Firefox

1.6 Sistematika Penulisan

1.6.1 Bab I - Pendahuluan

Dalam bab ini akan menguraikan latar belakang, rumusan masalah, batasan masalah, tujuan penelitian, metode penelitian, dan sistematika penulisan.

1.6.2 Bab II - Landasan Teori

Pada bab ini akan dibahas mengenai teori-teori yang mendukung dalam proses penyusunan penelitian ini. Berisi mengenai definisi-definisi dan teori-teori yang menjadi dasar dalam penulisan penelitian yang diambil dari berbagai sumber.

1.6.3 Bab III - Metode Penelitian

Menjelaskan bagaimana proses penelitian dilakukan mulai dari p dari awal proses pengolahan dataset, pengolahan data, persiapan data, evaluasi, dan terakhir yakni implementasi.

1.6.4 Bab IV - Hasil dan Pembahasan

Penjelasan mengenai bagaimana mendapatkan pengetahuan/pola yang sudah diolah dengan *data mining* yang kemudian diuji untuk mencari tahu tingkat keakuratan hasil. Kemudian dibahas dan dievaluasi sebelum diimplementasikan menjadi website dalam bahasa python.

1.6.5 Bab V - Penutup

Bab ini berisi kesimpulan hasil penelitian dari penjelasan bab-bab sebelumnya dan saran yang didapat dari penelitian yang telah dilakukan.