BAB I
PENDAHULUAN

1.1 Latar Belakang

Perkembangan internet dari tahun ke tahun membuat fasilitas yang ditawarkannya menjadi lebih banyak. Dari segi kemudahan akses hingga banyaknya fitur yang ditawarkan pastinya akan mudah membuat orang-orang tertarik untuk menggunakannya, sehingga tidak dipungkiri jika penggunanya tidaklah sedikit. Sebagai contohnya terdapat media sosial, internet atau mobile banking. Dua hal tersebut telah memberikan kemudahan bagi orang-orang untuk saling berhubungan meski dipisahkan oleh jarak dan membantu proses transaksi, yang mana bisa diakses dan dilakukan dari mana saja secara online dan *realtime*. Namun dibalik kemudahan tersebut, bisa menjadi celah keamanan bagi pengguna internet dalam bertransaksi secara online maupun mengakses media sosial seperti data pribadi, kata sandi, e-mail, informasi sensitif internet banking atau mobile banking. Hal tersebut biasanya dilakukan oleh pelaku dengan bantuan menggunakan situs phising. Situs Phising merupakan suatu metode untuk melakukan penipuan dengan mengelabui target dengan maksud untuk mencuri akun target. Istilah ini berasal dari kata "fishing" = "memancing" korban untuk terperangkap dijebakannya. Sedangkan situs phising adalah situs yang dibuat semirip mungkin dengan situs aslinya dan digunakan untuk mengelabui lalu menjebak korban yang mengaksesnya [1]. Phising juga terjadi pada media sosial seperti Facebook, Gmail, dan Twitter. Dimana situs phising yang dibuat akan memberikan informasi layaknya situs aslinya kemudian ketika informasi pengguna dimasukkan pada kolom *login*, situs tersebut akan mengarah ke situs phising

daripada ke situs aslinya [2]. Dari laporan APWG (*Anti-Phishing Working Group*) [3], tahun 2019 menjadi naik turunnya perbuatan phishing, meskipun begitu tetap saja terjadi peningkatan pesat pada bulan-bulan tertentu. Dari laporan APWG kuartal keempat 2019, pada bulan Oktober – Desember ada 162.555 situs web yang terdeteksi oleh APWG. Hasil tersebut turun dari 266.387 pada bulan Juli – September.

Hal tersebut bisa menimbulkan ketidakpercayaan pengguna internet pada situs media sosial maupun situs transaksi. Oleh sebab itu dibutuhan sistem yang mampu mendeteksi situs phishing untuk mencegah kriminalitas phishing sehingga mengurangi dan bisa menghindari kerugian situs phishing terhadap pengguna internet. Dari beberapa penelitian sebelumnya digunakan seleksi fitur dan metode klasifikasi untuk deteksi situs phishing dan non-phishing sehingga dapat dilakukan pencegahan terkena serangan phishing. Metode klasifikasi yang paling popular adalah Decision Tree, Naïve Bayes dan KNN. Metode klasifikasi menggunakan semua fitur yang terdapat pada sebuah data untuk membangun model, namun menurut [4] tidak semua fitur tersebut relevan terhadap hasil klasifikasi. Apabila hal tersebut pada data yang memiliki ukuran besar, maka kinerja klasifikasi akan menjadi tidak efisien dan efektif, misalkan beban yang dilakukan oleh komputer untuk melakukan komputasi menjadi lebih berat, waktu pemrosesan menjadi lebih lama. Kemudian diterapkanlah seleksi fitur berbasis Information Gain, seleksi fitur digunakan untuk menyeleksi fitur-fitur yang relevan pada dataset. Fitur yang memiliki relevansi rendah akan dibuang sehingga mengurangi beban pemrosesan pada kinerja metode klasifikasi.

Oleh karena itu pada penelitian ini akan diterapkan seleksi atribut yang berbasis Information Gain terhadap kinerja algoritma klasifikasi yang popular yaitu *decision tree* (CART), *naïve bayes* dan k-*nearest neighbor*. Dengan demikian kinerja klasifikasi yang dihasilkan dapat terukur secara sistematis setelah melalui seleksi atribut atau seleksi fitur, sehingga hasil penelitian ini dapat digunakan untuk menunjang penelitian lain yang sejenis dikemudian hari. Tentunya apabila model klasifikasi dari penelitian ini diimplementasikan, maka dapat mengurangi resiko pengguna internet terkena serangan phishing.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah disampaikan, di bawah ini adalah rumusan masalah atau sesuatu yang menjadi pertanyaan-pertanyaan yang harus diselesaikan oleh peneliti pada penelitian yang akan dibuat :

1 Apakah seleksi fitur mempengaruhi kinerja algoritma klasifikasi?
2 Fitur-fitur apa saja yang relevan terhadap hasil klasifikasi?
3 Adakah peningkatan kinerja klasifikasi setelah dilakukan seleksi fitur?
4 Adakah penurunan kinerja klasifikasi setelah dilakukan seleksi fitur?

1.3 Batasan Masalah

Untuk membatasi permasalahan yang lebih luas pada konteks objek, maka ruang lingkup dalam penelitian ini akan dibatasi sesuai kebutuhan dan kemampuan. Batasan masalah dalam penelitian ini antara lain:

1. Dataset untuk *training* dan *test* didapatkan dari repository UCI Machine Learning : Phishing Websites Data Set. Situs phishing didapatkan dari

PhishThank yang beralamatkan *http://phishtank.com*. Situs non-phishing didapatkan dari Alexa yang beralamatkan di *https://alexa.com*, dari Moz Top 500 yang beralamatkan di *https://moz.com/top500*.

2. Dalam penelitian ini akan digunakan Information Gain untuk seleksi atribut atau fitur pada dataset kemudian dilakukan klasifikasi menggunakan algoritma klasifikasi. Dengan batas hasil akurasi setelah seleksi atribut.
3. Metode klasifikasi yang digunakan dalam penelitian ini adalah *decision tree* (CART), *naïve bayes* dan k-*nearest neighbor*. Dengan batas hasil akurasi dan hasil penerapan.
4. Perhitungan menggunakan *library* dari scikit-learn.
5. Sistem web dibangun dengan *framework* Flask menggunakan Bahasa pemrograman Python.
6. Sistem yang digunakan sebagai pembukti bahwa seleksi fitur dan algoritma yang diimplementasikan berjalan untuk deteksi web phishing.

1.4 Maksud dan Tujuan Penelitian

1.4.1 Maksud Penelitian

Dibawah ini adalah maksud dari penelitian ini, yaitu:

1. Pemrosesan klasifikasi pada deteksi web phishing menjadi lebih efektif dan efisien setelah dilakukan seleksi atribut.
2. Mengetahui atribut atau fitur apa saja yang relevan pada data set.
3. Mengetahui pengaruh penerapan seleksi fitur pada kinerja klasifikasi.

1.4.2 Tujuan Penelitian

Di bawah ini adalah tujuan dari penelitian ini, yaitu:

1. Menerapkan seleksi fitur berbasis *information gain* untuk menguji pengaruhnya pada kinerja algoritma klasifikasi.
2. Menentukan fitur-fitur yang relevan berdasarkan hasil seleksi fitur berbasis *information gain*.
3. Mengetahui peningkatan kinerja klasifikasi setelah diterapkannya seleksi fitur.
4. Mengetahui penurunan kinerja klasifikasi setelah diterapkannya seleksi fitur.

1.5 Metode Penelitian

1.5.1 Metode Pengumpulan Data

Berikut adalah metode yang dilakukan dalam pengumpulan data:

1. Data *training* dan *test* dalam penelitian ini didapatkan dari repository UCI Machine Learning bagian Phishing Websites Data Set.
2. Data situs phishing didapatkan dari PhishTank.
3. Data situs non-phishing (situs otentik) didapatkan dari Alexa dan Moz Top 500.

1.5.2 Tahap-Tahap Penelitian

Dalam penelitian ini, peneliti menggunakan dataset yang berasal dari repository UCI Machine Learning bagian Phishing Websites Data Set. Seleksi fitur akan diterapkan pada dataset tersebut, kemudian akan diseleksi fitur yang memiliki

tingkat relevansi tinggi satu sama lain. Fitur yang memiliki relevansi rendah atau tidak memiliki relevansi tidak akan dipakai. Data hasil seleksi tersebut yang nantinya akan digunakan untuk menguji ketiga metode klasifikasi yaitu *decision tree* (CART), *naïve bayes*, dan k-*nearest neighbor*. Sehingga hasil uji tersebut bisa dibandingkan mana yang memiliki hasil akurasi tertinggi, kinerja terbaik dan juga berapa peningkatan atau penurunan yang terjadi pada ketiga klasifikasi tersebut. Kemudian dari ketiga algoritma klasifikasi tersebut akan dipilih lalu digunakan untuk melakukan deteksi web phishing dengan melakukan prediksi berdasarkan hasil latih yang telah dilakukan.

1.6 Sistematika Penulisan

Pada bagian ini dituliskan urutan-urutan dan sistematika penulisan yang dilakukan. Berikut ringkasan mengenai isi masing-masing bab:

- **BAB I PENDAHULUAN**

  Pada bab ini dijelaskan mengenai latar belakang, rumusan masalah, batasan masalah, maksud dan tujuan penelitian, manfaat penelitian, metode penelitan, dan sistematika penulisan dalam penelitan.

- **BAB II LANDASAN TEORI**

  Pada bab ini akan dijelaskan mengenai landasan teori-teori dan kajian pustaka dari berbagai penelitian yang memiliki keterkaitan dengan penelitian ini. Hal tersebut berguna untuk memperkuat dasar, analisa, penulisan dan alasan dilakukannya penelitian ini. Sumber dari landasan teori ini juga berasal dari buku, jurnal yang secara fisik maupun berasal dari internet.

- **BAB III METODOLOGI PENELITIAN**

  Pada bab ini akan dijelaskan mengenai langkah-langkah penelitian beserta metode yang digunakan.

- **BAB IV HASIL DAN PEMBAHASAN**

  Pada bab ini akan dilakukan perancangan sistem serta pembahasannya. Kemudian di bab ini juga hasil penelitian akan didapatkan dan dilakukan pembahasannya.

- **BAB V KESIMPULAN**

  Pada bab ini berisi kesimpulan dari penelitian ini dan juga saran bagi penelitian mendatang yang berasal dari kekurangan dari penelitian ini.