BAB I
PENDAHULUAN

1.1 Latar Belakang

Berkembang pesatnya teknologi internet di dunia saat ini tidak dapat dihindari, pengaruh dan manfaat yang didapat sangat besar dalam kehidupan manusia. Akses yang mudah dan *user friendly* membuat seseorang dapat dengan mudahnya menerima sebuah informasi dari berbagai belahan dunia melalui internet. *Email* merupakan salah satu contoh produk teknologi yang dapat mengirim dan menerima sebuah informasi baru yang menggantikan media berkomunikasi melalui surat konvensional [1].

Keterbatasan jarak pengiriman ke penerima yang jauh dimiliki oleh surat konvensional akan mengakibatkan lebih lama seseorang dalam menerima sebuah surat, kendala tersebut akan mudah teratasi jika seseorang menggunakan *email*. Selain memiliki kelebihan waktu pengiriman yang lebih cepat dan hemat, *email* dapat memuat informasi selain tulisan seperti file dokumen, gambar, audio, dan video[2]. Dengan segala kelebihan yang tidak dimiliki oleh surat konvensional, kini *email* menjadi sebuah jaringan informasi yang paling sering digunakan.

Penggunaan email di Indonesia terbilang cukup besar, Indonesia menduduki peringkat ketiga dalam penggunaan email dengan presentase 38,5%[3]. Pengguna *email* yang banyak dikarenakan berbagai kelebihannya tak serta merta membuat *email* tidak memiliki kekurangan. Seiring bertumbuhnya penggunaan internet terutama pada email, banyak orang menyalahgunakan manfaat utama pada email sebagai ajang promosi iklan, *phising*, bahkan mengirimkan sebuah virus. Kategori email yang tidak penting bagi pengguna email dapat disebut dengan *email SPAM* (*Stupid Pointless Annoying Message*).

Salah satu solusi untuk mengatasi permasalahan *email SPAM* tersebut adalah dengan teknik penyaringan *email SPAM* yaitu berdasarkan kandungan isi atau konten dari *email* sendiri[1]. Para penyedia layanan *email* seperti *Google*, *Yahoo*, dan *Outlook* sudah mempersiapkan layanan untuk menyaring suatu *email*. Namun, Tidak sedikit ditemukan kesalahan seperti *email* yang seharusnya asli dianggap *SPAM* oleh penyedia layanan atau sebaliknya[4]. Dalam hal ini pengguna juga tidak dapat menghindari masalah serius dalam menangani *email SPAM* yang didapat, sehingga pengguna terus mengalami berbagai kerugian.

Teknik penyaringan *email* adalah suatu proses yang memisahkan *email* berdasarkan kategorinya baik *SPAM* maupun bukan *SPAM* atau biasa disebut dengan *ham*. Dalam pengklasifikasian *email SPAM*, dibutuhkan suatu sistem cerdas yang dapat menyortir/mengklasifikasikan *email SPAM* dan bukan *SPAM (ham)* secara baik dan benar[5]. Terdapat banyak algoritma yang dapat membantu dalam melakukan klasifikasi *email SPAM*, diantaranya adalah *Naïve Bayes*, *Support Vector Machine (SVM)*, *Artifial Neural Networking (ANN)*, *Logistic Regression*, *K-Nearest Neigbors (KNN)*, *Random Forest*, dan *Decission Tree*[6].

Pada penelitian terdahulu yang berkaitan dengan klasifikasi *email SPAM*, metode yang paling sering digunakan adalah metode *Naïve Bayes*, hal ini dikarenakan metode *Naïve Bayes* memiliki hasil akurasi yang tinggi meskipun dataset yang digunakan tidak banyak. Oleh karena itu, peneliti akan melalukan perbandingan metode *Naïve Bayes* dengan metode *Random Forest* dan *K-Nearest Neighboor (KNN)* untuk melakukan proses klasifikasi *email SPAM*, peneliti akan membandingkan tingkat akurasi pada metode yang dipilih agar mendapatkan tingkat akurasi tertinggi untuk proses klasifikasi *email SPAM* pada penelitian ini.

Metode penelitian *Naïve Bayes* sering digunakan peneliti dalam melakukan sebuah klasifikasi model. Penggunaan metode *Naïve Bayes* sering digunakan peneliti dalam melakukan klasifikasi *email SPAM* karena metode ini memiliki kesederhanaan dalam konsepnya. *Naïve Bayes* bekerja berdasarkan kemunculan fitur atau ciri terhadap suatu kelas yang kemudian dihitung kemungkinannya,

sehingga sesuai dengan karakteristik fitur atau ciri *email* yang dijadikan acuan untuk membedakan mana yang termasuk *email SPAM* dan mana yang bukan (*ham*)[5]. Penelitian tentang metode *Naïve Bayes* telah dilakukan sebelumnya yaitu klasifikasi dalam memprediksi besarnya penggunaan listrik rumah tangga mendapatkan hasil akurasi sebesar 78,33%[7].

Random Forest merupakan metode yang digunakan sebagai pengklasifikasian untuk data yang jumlahnya besar. Klasifikasi ini dilakukan melalui penyatuan beberapa pohon (*tree*) dengan *training* pada sampel data yang akan digunakan. Sebagai metode klasifikasi, metode *Random Forest* ini telah teruji di beberapa penelitian sebelumnya, dimana metode ini mampu menghasilkan kinerja yang baik dengan akurasi yang tinggi[13]. Hal ini dikarenakan semakin banyak penggunaan pohon (*tree*) akan semakin mempengaruhi hasil akurasi. Berdasarkan pengujian yang telah dilakukan tentang klasifikasi data bank *marketing* menggunakan metode *Random Forest*, penelitian tersebut mendapatkan tingkat akurasi sebesar 88,30%[14].

KNN adalah metode yang menggunakan algoritma *supervised*, dimana metode ini bertujuan untuk menemukan pola baru dalam data dengan menghubungkan pola data yang sudah ada dengan data yang baru[15]. Metode ini bekerja dengan mencari pola latih terdekat yang kemudian menentukan keputusannya berdasarkan jumlah pola terbanyak. Hal ini menyebabkan akurasi yang didapat oleh *KNN* tinggi dalam mengklasifikasikan sebuah data. Penelitian sebelumnya tentang metode *KNN*, diperoleh hasil dari akurasi sistem pendeteksi citra tanaman dengan hasil akurasi 92%[16].

Sebelum melakukan klasifikasi *email SPAM*, peneliti akan melakukan praproses data menggunakan metode *stop-word removal, stemming* data, dan tokenisasi. Untuk tambahan kelengkapan saat melakukan praproses data, peneliti akan menambahkan metode penghilangan tanda baca/normalisasi dan *casefolding*. Dengan melakukan tambahan pada praproses data, diharapkan akan semakin meningkatkan akurasi dalam upaya klasifikasi *email SPAM*. Klasifikasi data *email*

SPAM dengan metode *Naïve Bayes* ini akan menggunakan data set yang diambil dari github.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, rumusan masalah pada penelitian ini adalah "Bagaimana hasil perbandingan implementasi antara metode *Naïve Bayes, Random Forest,* dan *K-Nearest Neighboor* terhadap sistem pengklasifikasian *spam Email*?"

1.3 Batasan Masalah

Agar pembahasan tidak menyimpang, beberapa batasan masalah yang digunakan dalam penelitian ini adalah sebagai berikut.

Metode yang digunakan untuk klasifikasi adalah *Naïve Bayes, Random Forest,* dan *K-Nearest Neighboor.*

Dataset yang digunakan yaitu data *email SPAM* dan *ham* yang didapat dari github.

Dataset yang digunakan berbahasa inggris.

Hasil klasifikasi berguna untuk membuktikan bahwa *email* merupakan sebuah *spam/ham* dan nilai akurasi sistem.

Sistem digunakan sebagai pembuktian bahwa algoritma yang diimplementasikan berjalan.

1.4 Maksud dan Tujuan Penelitian

Berdasarkan rumusan masalah yang dijabarkan, tujuan dari penelitian klasifikasi *email SPAM* adalah untuk mengetahui perbandingan tingkat akurasi dari metode *Naïve Bayes, Random Forest,* dan *K-Nearest Neighboor* terhadap klasifikasi *email SPAM.*

1.5 Metode Penelitian

Metode penelitian dapat disebut juga tahapan atau langkah dalam melakukan suatu penelitian. Berikut merupakan tahapan – tahapan metode yang dipakai dalam penelitian ini.

Studi Pustaka

Studi pustaka dilakukan oleh peneliti untuk mengumpulkan informasi dari penelitian sebelumnya yang digunakan sebagai acuan peneliti dalam melakukan penelitian.

Pengumpulan Data

Data input yang digunakan berupa *email SPAM* dan *ham*. Pengumpulan data diambil dari github. Data yang digunakan menggunakan format .csv.

Analisis dan Perencanaan

Pada tahap metodologi penelitian yaitu melakukan perencanaan sistem yang dapat mengklasifikasikan data. Tahapan dalam proses perancangan sistem ini antara lain perancangan *pre-processing*, perancangan pembobotan, perancangan algoritma klasifikasi metode *Naïve Bayes*, *Random Forest*, dan *K-Nearest Neighboor*.

Implementasi

Implementasi dilakukan dengan menggunakan rancangan sistem yang telah dibuat pada tahapan sebelumnya.

Pengujian

Pengujian dilakukan ke sistem yang telah diimplementasikan. Pengujian dilakukan untuk evaluasi model dan mengetahui nilai akurasi dari algoritma yang digunakan dalam penelitian ini.

Penulisan Laporan

Hasil dari penelitian kemudian dituliskan dalam sebuah laporan. Pada penulisan laporan penelitian akan ditarik sebuah kesimpulan berdasarkan hasil pembahasan dan pengujian, serta memberikan saran untuk penelitian selanjutnya.

1.6 Sistematika Penulisan

BAB I PENDAHULUAN

Bagian pendahuluan menjelaskan mengenai usulan peneliti tentang latar belakang, rumusan masalah, maksud tujuan penelitian, dan sistematika penulisan skripsi

BAB II LANDASAN TEORI

Bagian ini membahas mengenai landasan teori dari berbagai penelitian yang terkait sebelumnya untuk digunakan sebagai dasar melakukan penelitian agar dapat memahami konsep dan teori terhadap permasalahan yang akan diteliti. Bagian ini meliputi kajian pustakan dan dasar – dasar teori dari penelitian yang dilakukan.

BAB III METODOLOGI PENELITIAN

Bagian ini membahas komponen alat dan bahan apa saja yang dibutuhkan dalam penelitian serta langkah – langkah yang dijelaskan pada alur penelitian.

BAB IV HASIL DAN PEMBAHASAN

Bagian ini membahas mengenai hasil uji coba terhadap metode yang diimplementasikan serta pembahasan mengenai analisis hasil dari temuan penelitian setelah melakukan uji coba data.

BAB V KESIMPULAN DAN SARAN

Bagian ini membahas mengenai hasil dari penafsiran seluruh penelitian dan juga membahas mengenai saran bagi penelitian yang akan datang tentang kekurangan dari penelitian ini yang nantinya bisa dikembangkan pada penelitian selanjutnya.

DAFTAR PUSTAKA

Bagian ini berisikan daftar pustaka yang dipakai sebagai acuan referensi literatur dalam penelitian ini.