BAB I

`PENDAHULUAN

1.1 LATAR BELAKANG

Perkembangan teknologi saat ini telah mengalami kemajuan yang signifikan, terutama dalam bidang internet. Melalui internet, informasi dan berita dapat diakses serta diterima oleh masyarakat secara luas. Bahkan, internet memungkinkan individu untuk saling mengirim dan menerima pesan secara digital melalui fasilitas yang dikenal sebagai *email*. Namun, tidak semua orang menggunakan *email* dengan cara yang bijaksana dan tepat, yang terkadang dapat merugikan pihak lain. Spam *email* kini menjadi salah satu masalah utama yang dihadapi pengguna di dunia internet.

Email merupakan entitas penting dalam komunikasi digital melalui internet, meskipun perannya yang signifikan sebagai alat komunikasi daring, penggunaannya turut menghadirkan tantangan, terutama terkait kelebihan informasi dan pengelolaan pesan[1]. Email juga digunakan untuk beriklan dan mengirimkan file data. Saat ini, cara terbaik untuk berkomunikasi adalah melalui media elektronik. Untuk mengirim pesan elektronik, semua yang Anda butuhkan adalah koneksi internet [2].

Klasifikasi adalah cara yang dilakukan sebagai teknik untuk membentuk model klasifikasi dari contoh data pelatihan. Klasifikasi akan menganalisis input data dan membentuk model dengan menggambarkan kelas data[3]. Dalam prosesnya, klasifikasi dapat dilakukan dengan banyak cara baik secara manual

ataupun dengan bantuan teknologi. Klasifikasi yang dilakukan secara manual adalah klasifikasi yang dilakukan oleh manusia tanpa adanya bantuan dari algortima cerdas komputer. Sedangkan klasifikasi yang dilakukan dengan bantuan teknologi, memiliki beberapa algoritma, diantaranya *Naïve Bayes*, *Support Vector Machine*, *Decission Tree*[4].

Spam email dapat diklasifikasikan dengan menggunakan berbagai macam teknik, termasuk *Decision Tree*, *K-Nearest Neighbor* (KNN), *Naïve Bayes*, ID3, dan C4.5. Di antara teknik-teknik tersebut, naïve bayes merupakan pendekatan statistik langsung yang memiliki tingkat kesalahan yang rendah dan akurasi yang tinggi dalam proses klasifikasi, juga dikenal sebagai multinomial Model yang disederhanakan dari algoritma *Bayes, naïve bayes* merupakan teknik klasifikasi yang menggunakan metode statistik dan probabilitas untuk mengklasifikasikan teks atau dokumen. Pada klasifikasi *naïve bayes*, nilai kategori dari sebuah dokumen ditentukan oleh fitur-fitur atau kata-kata yang muncul pada dokumen yang akan diklasifikasikan[5].

K-Nearest Neighbors (KNN) merupakan salah satu algoritma pembelajaran mesin yang sering digunakan untuk klasifikasi teks, algoritma ini bekerja dengan mengidentifikasi *k* tetangga terdekat dalam ruang fitur dari data yang akan diklasifikasikan, salah satu keunggulan KNN adalah kesederhanaannya serta kemampuannya menghasilkan akurasi yang tinggi. Dalam aplikasi klasifikasi, seperti identifikasi *email* spam, KNN dapat digunakan untuk menentukan apakah sebuah *email* tergolong spam atau bukan berdasarkan fitur tertentu, seperti frekuensi kemunculan kata kunci, meskipun sederhana,

keberhasilan KNN sangat bergantung pada pemilihan nilai *k* yang tepat dan kualitas praproses data untuk memastikan performa yang optimal[6].

Term Frequency inverse Document Frequncy (TFIDF) adalah metode yang populer dalam reprentasi teks untuk analisis sentimen dan klasifikasi dokumen[7]. Metode ini mengukur pentingnya kata dalam dokumen relatif terhadap frekuensi kata tersebut dalam seluruh koleksi dokumen, sehingga membantu dalam menentukan kata-kata yang paling representasif[8]. Penelitian terbaru menunjukkan bahwa TFIDF dapat meningkatkan akurasi klasifikasi dalam teks, terutama ketika digunakan bersama dengan teknik lain seperti *Naive Bayes* dan *Latent Dirichlet Allocation*(LDA)[8]. Misalnya, Kim dan Gil (2019) mengembangkan sistem klasifikasi artikel penelitian berdasarkan TFIDF dan LDA, yang menunjukkan peningkatan signifikan dalam pengelompokan artiker berdasarkan topik yang serupa, Selain itu, penelitan Danya et al. (2024) menunjukkan bahwa penggunaan TFIDF dalam analisis sentimen ulasan film dapat meningkatkan akurasi klasifikasi ulasan menjadi positif atau negatif[7].

Dari uraian di atas, maka penelitian akan berfokus pada perbandigan algoritma *Naive Bayes* dan *K-Nearest Neighbor* pada klasifikasi spam *email* . Dengan judul penelitian **"PERBANDINGAN ALGORITMA *NAIVE BAYES* DAN *K-NEAREST NEIGHBOR* PADA KLASIFIKASI *EMAIL* ".**

1.2 RUMUSAN MASALAH

Bagaimana performa algoritma *Naive Bayes* dan *K-Nearest Neighbor* dalam klasifikasi *email*?

Bagaimana penerapan fitur TF-IDF dalam algoritma *Naive Bayes*dan *K-Nearest Neighbor*?

Algoritma manakah yang lebih efektif dan efisien dalam klasifikasi *email* berdasarkan hasil pengujian?

1.3 BATASAN MASALAH

Batasan-batasan dalam penelitian yang diambil oleh peneliti adalah:

1. Bahasa pemrograman yang digunakan adalah *Pyhton* dengan menggunakan *Google Colab.*
2. Data yang digunakan dalam penelitian ini adalah data *email*

   Menggunakan metode *Naive Bayes* dan *K-Nearest Neighbor*.
3. Fitur yang digunakan TF-IDF.

1.4 TUJUAN DAN MANFAAT PENELITIAN

1.4.1 Tujuan Penelitian

Berdasarkan latar belakang di atas, tujuan penelitian ini adalah:

1. Membandingkan performa Algoritma *Naive Bayes* dan *K-Nearest Neighbor* dalam klasifikasi *email*.
2. Memberikan rekomendasi metode yang lebih efektif digunakan untuk klasifikasi *email* berdasarkan hasil penelitian.

1.4.2 Manfaat Penelitian

Manfaat penelitian ini adalah :

1. Memberikan wawasan lebih mendalam mengenai perbandingan Algoritma *Naive Bayes* dan *K-Nearest Neighbor*.
2. Meningkatkan pemahaman tentang perbandigan klasifikasi *Naive Bayes* dan *K-Nearest Neighbor* pada platform cloud seperti *Google Colab*.
3. Dengan klasifikasi *email* yang lebih akurat, pengguna internet dapat terhindar dari *email* phishing, penipuan, atau serangan siber yang dapat mencuri data pribadi.

1.5 SISTEMATIKA PENULISAN

Sistematika penulisan laporan penelitian Perbandingan algoritma *Naive Bayes* dan *K-Nearest Neighbor* Klasifikasi *Email* , adalah sebagai berikut:

BAB I PENDAHULUAN

Bab ini membahas latar belakang penelitian, yaitu perkembangan teknologi internet dan tantangan yang dihadapi seperti *email*. Selain itu, diuraikan rumusan masalah berupa pertanyaan penelitian yang ingin dijawab, batasan penelitian untuk menjaga fokus, tujuan penelitian untuk membandingkan algoritma *Naive Bayes* dan KNN, serta manfaat penelitian dalam memberikan rekomendasi metode terbaik. Bagian ini juga mencakup sistematika penulisan untuk memberikan gambaran struktur laporan.

BAB II LANDASAN TEORI

Bab ini memuat teori-teori yang menjadi dasar penelitian. Dimulai dengan penjelasan tentang Machine Learning sebagai pendekatan umum, kemudian algoritma *Naive Bayes* dan *K-Nearest Neighbor* (KNN) yang digunakan dalam klasifikasi. Selanjutnya, dibahas metode TF-IDF untuk representasi teks, konsep klasifikasi, serta pengertian *email* dan spam. Tinjauan pustaka menyajikan ringkasan penelitian sebelumnya yang relevan, menjadi dasar ilmiah dalam penelitian ini.

BAB III METODOLOGI PENELITIAN

Bab ini menjelaskan pendekatan dan langkah penelitian. Kerangka penelitian menggambarkan alur kerja penelitian, mulai dari identifikasi masalah, kajian literatur, pengumpulan data, proses klasifikasi, hingga analisis hasil. Alat bantu penelitian meliputi perangkat keras seperti laptop dan perangkat lunak seperti Google Colab. Alur eksperimen menjelaskan tahapan teknis dari praproses data, implementasi algoritma, hingga evaluasi kinerja model.

BAB IV HASIL DAN PEMBAHASAN

Bab ini menyajikan hasil penelitian secara rinci. Deskripsi *dataset* memberikan gambaran data yang digunakan, seperti ukuran dan sumber data. Implementasi algoritma menjelaskan proses penerapan *Naive Bayes* dan KNN pada *dataset*. Evaluasi performa membahas metrik seperti akurasi, *precision*, *recall*, dan F1-score, sedangkan

analisis hasil menginterpretasikan temuan berdasarkan kinerja masing-masing algoritma.

BAB V PENUTUP

Bab ini berisi kesimpulan dari penelitian, mencakup jawaban atas rumusan masalah dan pencapaian tujuan penelitian. Selain itu, saran disampaikan untuk pengembangan metode atau penelitian lanjutan yang dapat memperluas aplikasi algoritma yang telah digunakan.