

# BAB 1

## PENDAHULUAN

### 1.1. Latar Belakang Masalah

Berita merupakan sebutan yang mengacu kepada informasi yang disebarkan oleh surat kabar, radio, televisi, internet, dan media lainnya. Ratusan berita dituliskan setiap harinya di berbagai portal berita Indonesia berbasis online. Perkembangan teknologi menyebabkan banyaknya portal berita yang beralih dari media cetak menjadi media elektronik yang dapat diakses secara online menggunakan internet. **Error! Reference source not found.** Hal ini menimbulkan terbitnya berita yang semakin banyak dan beragam. **Error! Reference source not found.** Dengan melakukan klasifikasi berita yang cepat dan akurat, editor dapat dimudahkan dalam proses penyuntingan berita, **Error! Reference source not found.** sedangkan pembaca dapat dengan mudah memperoleh berita dengan kategori tertentu yang diinginkan. **Error! Reference source not found.**

Penelitian dalam klasifikasi berita berbahasa Indonesia telah banyak dilakukan dengan berbagai metode seperti *KNN*, **Error! Reference source not found.****Error! Reference source not found.** *SVM*, **Error! Reference source not found.****Error! Reference source not found.** *Cosine Similarity*, **Error! Reference source not found.** *ANN*, **Error! Reference source not found.** *MLP*, **Error! Reference source not found.** dan Naive Bayes. **Error! Reference source not found.****Error! Reference source not found.** Algoritma Naive Bayes merupakan metode klasifikasi populer dan masuk dalam sepuluh algoritma terbaik dalam *data mining*. **Error! Reference source not found.** Naive Bayes sangat cocok diterapkan pada kasus pengklasifikasian berita dengan masukan berupa teks atau dokumen. **Error! Reference source not found.****Error! Reference source not found.** Algoritma ini sangat populer digunakan karena kesederhanaan dan kecepatannya dalam proses perhitungan. **Error! Reference source not found.****Error! Reference source not found.** Meskipun memiliki perhitungan yang sederhana, algoritma Naive Bayes memiliki akurasi dan kecepatan yang tinggi. **Error! Reference source not found.****Error! Reference source not found.** Untuk mengimbangi

pesatnya perkembangan berita dengan kategori yang semakin beragam, performa model klasifikasi berita juga harus dapat dikembangkan. Naive Bayes memiliki kelemahan dalam pemilihan fitur. **Error! Reference source not found.** Penggunaan fitur yang tidak relevan akan berpengaruh pada menurunnya tingkat akurasi dikarenakan sifat independensi Naive Bayes. Untuk menangani kelemahan Naive Bayes tersebut, optimasi dengan teknik seleksi fitur dapat menjadi salah satu solusi. **Error! Reference source not found.**

Pada kasus klasifikasi teks, setiap kata dan frasa dari seluruh dokumen adalah fitur, sehingga jumlah fitur dalam kasus ini sangatlah banyak. **Error! Reference source not found.** Memilih fitur mana yang sebaiknya digunakan pada proses pelatihan dapat mempengaruhi hasil keputusan secara signifikan. **Error! Reference source not found.** Seleksi fitur (*features selection*) adalah tahap yang sangat penting dalam klasifikasi karena akan sangat berpengaruh pada hasil klasifikasi. **Error! Reference source not found.** Dengan memilih fitur-fitur yang relevan, model klasifikasi dapat mengarahkan hasil klasifikasi pada keputusan yang benar begitu pula sebaliknya. Seleksi fitur memiliki banyak keuntungan seperti pengurangan waktu pelatihan, penurunan kompleksitas, dan peningkatan akurasi. **Error! Reference source not found.** Pada kasus klasifikasi teks, beberapa penelitian meninggalkan saran untuk mengimplementasikan seleksi fitur, **Error! Reference source not found.** sementara penelitian lainnya memberikan saran untuk menggunakan metode seleksi fitur lain supaya dapat mengurangi fitur yang tidak relevan dengan harapan dapat meningkatkan akurasi. **Error! Reference source not found.**

Algoritma genetika merupakan salah satu algoritma yang dapat digunakan untuk melakukan seleksi fitur. Algoritma genetika sebagai cabang dari algoritma evolusi merupakan metode *adaptive* yang dapat digunakan untuk memecahkan suatu pencarian nilai dalam sebuah masalah optimasi. **Error! Reference source not found.** Algoritma genetika memiliki performa yang baik dalam menyelesaikan masalah optimasi. **Error! Reference source not found.** Terdapat beberapa kelebihan dari algoritma genetika dibandingkan algoritma optimasi tradisional lainnya. Kelebihan algoritma genetika dua diantaranya yaitu kemampuan untuk menangani permasalahan kompleks dan paralel. Algoritma genetika dapat menangani berbagai macam optimasi tergantung pada fungsi

objektifnya (*fitness*) seimbang atau tidak seimbang, linier atau tidak linier, berkesinambungan atau tak berkesinambungan, atau dengan *random noise*. Algoritma genetika telah terbukti menjadi teknik seleksi fitur yang mengungguli metode *forward selection*, *backward elimination*, *greedy forward selection* dan *greedy backward selection*. **Error! Reference source not found.** Algoritma genetika sebagai teknik seleksi fitur juga telah diterapkan pada beberapa penelitian. **Error! Reference source not found.****Error! Reference source not found.** Selain itu, teknik seleksi fitur menggunakan algoritma genetika juga cocok diterapkan pada kasus klasifikasi menggunakan algoritma Naive Bayes seperti pada penelitian **Error! Reference source not found.** yang dapat menghasilkan akurasi terbaik pada setiap percobaan.

Berdasarkan proses evaluasinya, algoritma seleksi fitur dibagi menjadi 3 metode, yaitu: *filter*, *wrapper*, dan *embedded*. **Error! Reference source not found.** Pada penelitian **Error! Reference source not found.**, Nelly dkk menduga bahwa seleksi fitur dengan peringkasan justru telah menghilangkan fitur kata yang relevan dalam menentukan kelas dari data yang diuji sehingga berdampak pada menurunnya akurasi. Teknik seleksi fitur tersebut termasuk ke dalam metode *filter*, sementara algoritma genetika termasuk ke dalam metode *wrapper*. Metode *filter* memiliki keunggulan dalam kecepatan komputasi namun keputusan yang dihasilkan kurang akurat karena metode *filter* menggunakan beberapa fungsi evaluasi matematis tanpa mengujikan fitur-fitur terpilih pada algoritma klasifikasi, sementara metode *wrapper* memiliki kinerja yang lebih baik dari metode *filter* karena menggunakan algoritma klasifikasi saat proses evaluasi fitur tetapi memiliki beban komputasi yang jauh lebih tinggi. **Error! Reference source not found.** Dengan berbagai pertimbangan yang telah dipaparkan sebelumnya, maka penelitian ini akan menerapkan Algoritma Naive Bayes sebagai metode klasifikasi dan algoritma genetika sebagai teknik seleksi fitur. Penelitian ini diharapkan mampu mengatasi kelemahan sifat independensi dari Naive Bayes dengan menggunakan kemampuan metode *wrapper* dalam mereduksi fitur, sehingga seleksi fitur menggunakan algoritma genetika mampu menghasilkan fitur-fitur optimal untuk mengurangi beban komputasi dan meningkatkan akurasi dari klasifikasi berita berbahasa Indonesia menggunakan algoritma Naive Bayes.

## 1.2. Rumusan Masalah

Berdasarkan latar belakang masalah yang telah diuraikan, maka dapat dirumuskan masalah dalam penelitian ini yaitu bagaimana pengaruh seleksi fitur menggunakan algoritma genetika terhadap klasifikasi berita berbahasa Indonesia menggunakan algoritma Naive Bayes.

## 1.3. Maksud dan Tujuan

Maksud dari penelitian ini adalah untuk menganalisis pengaruh teknik seleksi fitur algoritma genetika pada klasifikasi berita berbahasa Indonesia menggunakan algoritma Naive Bayes.

Tujuan yang akan dicapai dalam penelitian ini adalah untuk mengetahui kompleksitas, akurasi, dan waktu eksekusi pengujian dari model klasifikasi berita berbahasa Indonesia menggunakan algoritma Naive Bayes sebelum dan setelah dioptimasi dengan teknik seleksi fitur algoritma genetika.

## 1.4. Batasan Masalah

Pembahasan permasalahan diharapkan tidak menyimpang dari pokok permasalahan, sehingga diperlukan batasan masalah. Adapun batasan dari penulisan penulisan laporan skripsi ini adalah sebagai berikut:

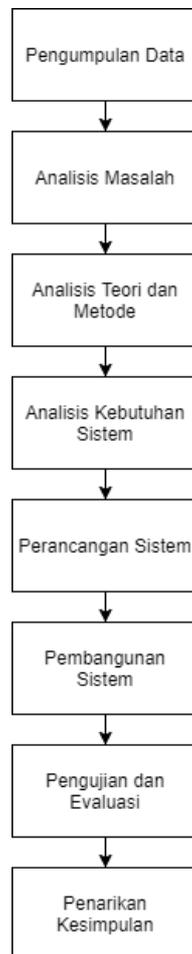
- a. Dokumen yang digunakan adalah berita berbahasa Indonesia sejumlah 300 buah.
- b. Kategori awal yang disediakan berjumlah 8 buah yaitu otomotif, bisnis, pendidikan, kesehatan, gaya hidup, pengetahuan, olahraga, dan teknologi.
- c. Algoritma *stemming* ditangani oleh *library* sastrawijs.
- d. *Stopword* bahasa indonesia yang akan dihilangkan dalam proses *preprocessing* didapat dari penelitian sebelumnya sebanyak 758 buah *stopword*. **Error!**

**Reference source not found.**

## 1.5. Metode Penelitian

Metode penelitian merupakan suatu proses yang digunakan untuk memecahkan suatu masalah yang logis, dimana memerlukan data-data untuk mendukung terlaksananya suatu penelitian. Metode penelitian yang digunakan dalam penelitian ini adalah analisis kuantitatif. Metode analisis kuantitatif adalah metode penelitian yang digunakan untuk meneliti populasi atau sampel tertentu menggunakan instrumen penelitian dengan analisis data bersifat kuantitatif / statistik, tujuannya untuk menguji hipotesis yang telah

ditetapkan. **Error! Reference source not found.** Diagram alur penelitian dapat dilihat pada Gambar 1.1.



Gambar 1.1 Diagram Alur Penelitian

### 1.5.1. Metode Pengumpulan Data

Metode pengumpulan data yang digunakan dalam penelitian ini adalah berikut ini:

#### a. Studi Literatur

Studi literatur merupakan metode pengumpulan data yang dilakukan untuk mendapatkan pengetahuan tentang pengolahan teks, klasifikasi Naive Bayes, teknik seleksi fitur, cara kerja algoritma genetika, *functional programming*, *object oriented programming*, bahasa pemrograman JavaScript, *runtime environment* Node.js, dan *tools* perancangan draw.io. Studi literatur juga diperlukan untuk mendapatkan informasi penelitian-penelitian sebelumnya tentang klasifikasi teks.

#### b. Web Scraping

Teknik pengumpulan data dilakukan dengan cara melakukan scraping terhadap portal berita [tribunnews.com](http://tribunnews.com) terhadap berita berbahasa Indonesia sebanyak 300 buah berita dari 8 kategori populer, yaitu: otomotif, bisnis, pendidikan, kesehatan, gaya hidup, pengetahuan, olahraga, dan teknologi.

### **1.5.2. Metode Analisis Kebutuhan Sistem**

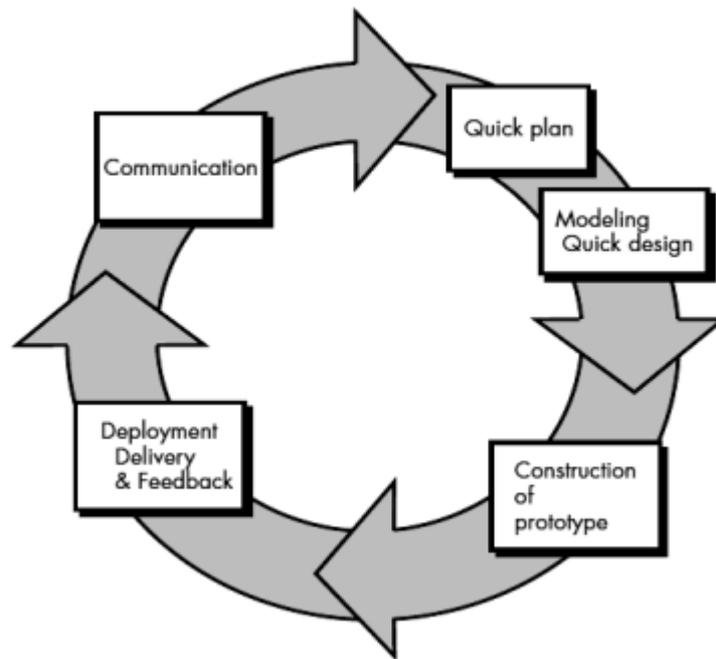
Analisis kebutuhan sistem yaitu kegiatan melakukan analisis pada metode yang digunakan yaitu Algoritma Naive Bayes dan genetika untuk melakukan klasifikasi berita berbahasa Indonesia serta melakukan penyusunan tahapan proses yang akan dilakukan.

### **1.5.3. Metode Perancangan Sistem**

Metode perancangan sistem yang digunakan dalam penelitian dan sistem yang akan dibangun adalah perancangan *Object Oriented Analysis Design*. Perancangan *Object Oriented Analysis Design* merupakan perancangan yang berorientasi obyek, sehingga perancangan sistem akan menggunakan *Use Case Diagram*, *Use Case Scenario*, *Activity Diagram*, *Class Diagram* dan *Sequence Diagram*.

### **1.5.4. Metode Pembangunan Perangkat Lunak**

Dalam penelitian ini metode pembangunan perangkat lunak yang digunakan adalah metode *prototyping*, karena metode ini menyajikan gambaran lengkap dari suatu sistem perangkat lunak yang akan dibangun, penentuan kebutuhan lebih mudah diwujudkan saat perancangan. Tahapan metode *prototyping* dapat dilihat pada Gambar 1.2.



Gambar 1.2 Tahapan Metode *Prototyping*

Berikut ini adalah langkah-langkah yang dilakukan dalam metode *prototyping*.

### 1. Pengumpulan Kebutuhan dan Analisis

Tahapan ini merupakan kegiatan untuk menganalisis dan mengumpulkan kebutuhan data dengan cara membaca buku dan jurnal tentang metode yang digunakan dalam klasifikasi berita berbahasa Indonesia. Selain itu mengumpulkan sampel dokumen berita dengan teknik *web scraping*.

### 2. Perancangan Cepat

Tahapan ini merupakan perancangan awal dari sistem yang dibangun, seperti perancangan antarmuka aplikasi. Perancangan dibuat berdasarkan dengan data-data yang telah dikumpulkan pada tahap pengumpulan kebutuhan dan analisis.

### 3. Membangun Prototipe

Tahapan ini merupakan pengimplementasian dari tahap desain aplikasi yang sudah dibuat pada tahapan perancangan cepat. Tahapan ini merupakan proses pembuatan sistem dalam bahasa pemrograman JavaScript.

#### 4. Evaluasi Prototipe

Tahapan ini akan dilakukan pengujian terhadap aplikasi yang telah dibuat oleh peneliti. Hal ini dilakukan untuk melihat metode yang telah diterapkan akan berjalan dengan baik atau tidak.

#### 5. Perubahan Desain dan Prototipe

Tahapan ini merupakan perbaikan terhadap aplikasi dari segi fungsionalitas dan juga metode-metode yang digunakan belum berjalan dengan baik pada tahapan evaluasi prototipe.

#### 6. Pembangunan Aplikasi

Tahapan ini merupakan tahapan akhir dari model prototipe, aplikasi pada tahap ini sudah berjalan sesuai dengan tujuan yang diharapkan. Pada tahapan ini juga aplikasi sudah bisa melakukan klasifikasi kategori berita berbahasa Indonesia.

### 1.5.5. Metode Pengujian

Metode pengujian perangkat lunak yang digunakan yaitu metode *black box*. *Black box* adalah pengujian perangkat lunak yang berfokus pada fungsionalitas dari perangkat lunak yang dibangun. Pengujian *black box* memungkinkan perekayasa perangkat lunak mendapatkan serangkaian *input* yang sepenuhnya menggunakan semua persyaratan fungsional untuk suatu program. Pengujian yang dilakukan pada metode *black box* yaitu:

- a. Fungsi-fungsi yang tidak benar
- b. Kesalahan *interface*
- c. Kesalahan dalam struktur data atau akses *database*
- d. Kesalahan inisialisasi dan tujuan akhir

Metode pengujian tujuan penelitian yang akan digunakan yaitu pengujian performa model klasifikasi. Parameter yang akan dibandingkan sebagai acuan ialah kompleksitas, akurasi, dan total waktu eksekusi pengujian dari model klasifikasi Naive Bayes sebelum dan setelah dioptimasi menggunakan teknik seleksi fitur algoritma genetika.

### **1.5.6. Metode Penarikan Kesimpulan**

Kesimpulan adalah intisari dari hasil penelitian dan pernyataan mengenai hubungan hasil penelitian dengan hipotesis, termasuk alasan-alasan yang menyebabkan hasil penelitian berbeda dengan hipotesis. Metode pengambilan kesimpulan pada penelitian ini yaitu dilihat dari hasil penelitian dan pengujian yang telah dilakukan sebelumnya. Pengambilan keputusan ini bertujuan memberikan ringkasan hasil akhir penelitian yang dilakukan sesuai dengan maksud dan tujuan yang telah ditentukan.

### **1.6. Sistem Penulisan**

Sebagai acuan bagi penulis agar penulisan skripsi ini dapat terarah dan tersusun sesuai dengan yang penulis harapkan, maka sistematika penulisannya adalah sebagai berikut :

## **BAB 1 PENDAHULUAN**

Pada bab ini berisi uraian latar belakang masalah, identifikasi masalah, maksud dan tujuan, batasan masalah, metode penelitian, tahap pengumpulan data, dan sistematika penulisan.

## **BAB 2 LANDASAN TEORI**

Pada bab ini akan membahas berbagai konsep-konsep dasar dan teori-teori pendukung yang berhubungan dengan pembangunan model klasifikasi, antara lain: klasifikasi teks, seleksi fitur, *preprocessing*, *features extraction*, algoritma *naive bayes*, algoritma genetika, *confusion matrix*, *performance measure*, dan *k-fold cross validation*.

## **BAB 3 ANALISIS DAN PERANCANGAN SISTEM**

Pada bab ini akan membahas tentang analisis kebutuhan dalam pembangunan sistem serta perancangan sistem. Analisis kebutuhan meliputi analisis masalah, analisis sistem yang akan dibangun, analisis data masukan, analisis metode, analisis kebutuhan non-fungsional, analisis kebutuhan fungsional, serta analisis basis data. Sementara, perancangan sistem meliputi perancangan antarmuka dan jaringan semantik.

## **BAB 4 IMPLEMENTASI DAN PENGUJIAN SISTEM**

Pada bab ini berisi hasil implementasi algoritma Naive Bayes dan genetika kedalam perangkat lunak berdasarkan perancangan sistem yang sudah dibuat, serta hasil pengujian

sistem untuk mengetahui apakah sistem yang dibangun sudah sesuai tujuan. Implementasi sistem meliputi implementasi perangkat keras, implementasi perangkat lunak, implementasi basis data, dan implementasi antarmuka. Sedangkan pengujian sistem meliputi pengujian *black box* dan pengujian performa model klasifikasi.

## **BAB 5 KESIMPULAN DAN SARAN**

Pada bab ini berisi kesimpulan yang diperoleh dari hasil pengujian sistem dan model klasifikasi, serta saran untuk pengembangan penelitian berikutnya.