# BAB III

# METODE PENELITIAN DAN PERANCANGAN SISTEM

## 3.1 Desain Sistem

Desain Sistem pada penelitian ini dapat dijelaskan sebagai berikut:

1. Pengumpulan Data

   Data yang digunakan dalam penelitian ini diambil dari portal berita *online* CNN Indonesia. Data yang diperoleh merupakan kumpulan berita yang didapatkan dengan menggunakan *web scraping*.

2. *Preprocessing*

   Proses yang dilakukan dalam tahapan ini adalah sebagai berikut:

   a. *Case Folding*, yaitu untuk menyeragamkan bentuk huruf.
   b. *Tokenizing*, yaitu pemenggalan suku kata
   c. *Stopword*, yaitu menghilangkan kata yang tidak deskriptif.
   d. *Stemming*, yaitu mengubah suku kata menjadi kata dasar.

3. Pelabelan

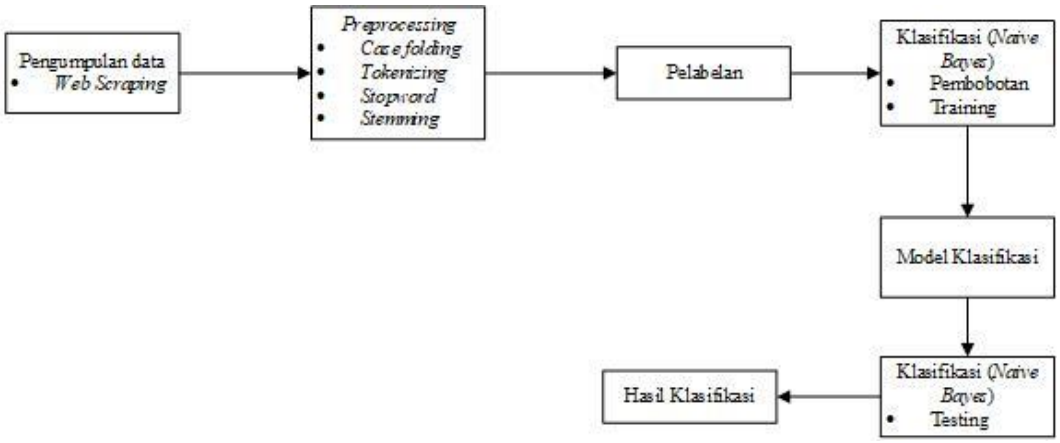
   Pelabelan adalah tahapan di mana berita diberi label yang nantinya akan digunakan pada proses *training* di tahap klasifikasi. Terdapat empat label yang disediakan yaitu politik untuk berita yang berisi informasi mengenai politik, olahraga untuk berita yang berisi informasi mengenai olahraga, teknologi untuk berita yang berisi informasi mengenai teknologi, dan ekonomi untuk berita yang hanya berisi informasi mengenai ekonomi. Pelabelan dilakukan dengan menggunakan program aplikasi Microsoft Excel. Hasil dari tahapan ini adalah kumpulan berita yang memiliki label.

4. Klasifikasi

Proses klasifikasi dibedakan menjadi dua proses yaitu:

a. *Training*, proses ini digunakan untuk melatih algoritma klasifikasi yang digunakan yaitu algoritma *Naïve Bayes* agar mampu melakukan prosesnya sesuai dengan yang diharapkan. Pada tahap ini pertama-tama akan dilakukan proses pembobotan terhadap kumpulan berita hasil pelabelan menggunakan perhitungan TF-IDF dengan hanya menghitung TF (*term frequency*)-nya saja. Selanjutnya akan dihasilkan model klasifikasi yang nantinya digunakan pada tahap *testing.*

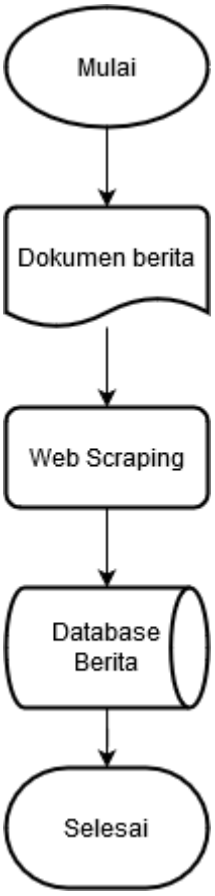
b. *Testing*, proses ini dilakukan untuk melakukan pengklasifikasian terhadap *dataset* dengan memanfaatkan model klasifikasi yang dihasilkan pada proses *training.* Hasil pada tahap ini adalah kumpulan berita yang telah diklasifikasikan ke dalam kategori berita. Untuk jelasnya dapat dilihat pada Gambar 3.1.




Gambar 3.1 Desain Sistem

### 3.1.1 Pengumpulan Data

Data yang digunakan dalam penelitian ini diambil dari portal berita CNN Indonesia dengan memanfaatkan *web scraping*. Berita yang diambil adalah kategori dokumen berita berupa olahraga, teknologi, ekonomi, dan politik. Untuk lebih jelasnya dapat dilihat pada Gambar 3.2 *flowchart* pengumpulan data berikut:



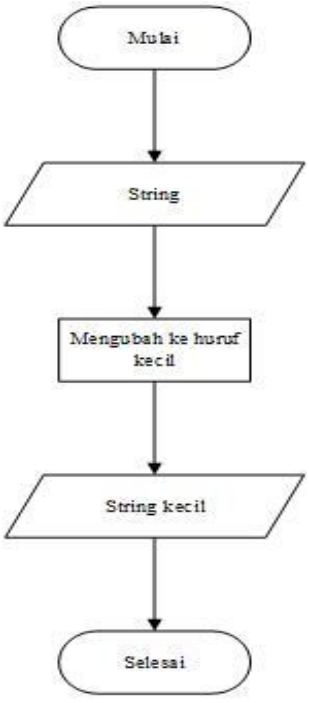

Gambar 3.2 *Flowchart* Pengumpulan Data

Tahap *preprocessing* adalah tahapan di mana aplikasi melakukan seleksi data yang akan diproses pada setiap dokumen. Proses *preprocessing* meliputi (1) *Case Folding* (menyeragamkan bentuk huruf) (2) *Tokenizing* (pemenggalan suku kata) (3) *Stopword* (menghilangkan kata yang tidak deskriptif) (4) *Stemming*

(mengubah suku kata menjadi kata dasar). Yang nanti akan diberikan label, pembobotan dan proses klasifikasi menggunakan *Naïve Bayes Classifier*.

### 3.1.2 Case Folding

Tidak semua dokumen teks konsisten dalam menggunakan huruf kapital. Oleh karena itu peran *case folding* dibutuhkan untuk mengonversikan keseluruhan teks dalam dokumen menjadi suatu bentuk standar (huruf kecil atau *lowercase*). Sebagai contoh, *user* yang ingin mendapatkan informasi tentang "BERITA" dan mengetik "BeRiTa", "BERITA", atau "berita" tetap diberikan hasil yang sama yakni "berita". *Case folding* adalah mengubah semua huruf dokumen menjadi huruf kecil. Hanya huruf 'a' sampai dengan 'z' yang diterima. Karakter selain huruf akan dihilangkan. Lebih jelasnya dapat dilihat pada Gambar 3.3 *Flowchart case folding*.




Gambar 3.3 *Flowchart Case Folding*

### 3.1.3 Tokenizing

Tahap *tokenizing* digunakan untuk memisahkan kalimat yang ada dalam *string* menjadi potongan kata tunggal. Contoh dari tahap *tokenizing* dapat dilihat pada Tabel 3.1. Contoh *Tokenizing* sebagai berikut:

Tabel 3.1 Contoh *Tokenizing*

| Teks Input | Teks Output |
|---|---|
| bela anies tak ke bogor, gerindra menyindir ‘orang mau jadi menteri jokowi’ | bela \| anies \| tak \| ke \| bogor \| gerindra \| menyindir \| orang \| mau \| jadi \| menteri \| jokowi |

*Tokenizing* secara garis besar memecah sekumpulan karakter dalam suatu teks ke dalam satuan kata. Sebagai contoh, karakter *whitespace*, seperti *enter*, tabulasi, spasi dianggap sebagai pemisah kata. Namun untuk karakter tunggal (‘), titik (.), titik koma (;), titik dua (:), atau lainnya, dapat memiliki peran yang cukup banyak sebagai pemisah kata.

### 3.1.4 Stopword

Pada tahap ini dilakukan pembuangan kata-kata yang kurang penting atau kata-kata yang sering muncul (*Stopword*), seperti kata penghubung dan kata keterangan yang bukan merupakan kata unik misalnya “sebuah”, “oleh”, “pada”, dan sebagainya. Contoh dari tahap *stopword* dapat dilihat pada Tabel 3.2. Contoh *Stopword* sebagai berikut:

Tabel 3.2 Contoh *Stopword*

| Hasil *Tokenizing* | Hasil *Filtering* |
|---|---|
| bela \| anies \| tak \| ke \| bogor \| gerindra \| menyindir \| orang \| mau \| jadi \| menteri \| jokowi | bela \| anies \| bogor \| gerindra \| menyindir \| orang \| menteri \| jokowi |

### 3.1.5 Stemming

Tahap *Stemming* adalah proses menghapus imbuhan, awalan, akhiran yang bertujuan untuk mengubah kata-kata sesuai dengan kata dasarnya. Contoh dari tahap *stemming* dapat dilihat pada Tabel 3.3. Contoh *Stemming* sebagai berikut:

Tabel 3.3 Contoh *Stemming*

| Hasil *Filtering* | Hasil *Stemming* |
|---|---|
| bela \| anies \| bogor \| gerindra \| menyindir \| orang \| menteri \| jokowi | bela \| anies \| bogor \| gerindra \| sindir \| orang \| menteri \| jokowi |

### 3.1.6 Pembobotan Kata

Dalam klasifikasi berita, pembobotan kata digunakan untuk mendapatkan suatu kategori. Salah satu metode pembobotan adalah TF-IDF (*Term Frequency – Inverse Document Frequency*).

Nilai bobot suatu kata (term) menyatakan kepentingan bobot tersebut dalam merepresentasikan judul. Pada pembobotan TF-IDF, bobot akan semakin besar jika

frekuensi kemunculan kata semakin tinggi, tetapi bobot akan berkurang jika kata tersebut semakin sering muncul pada berita lainnya.

Rumus TF-IDF diketahui sebagai berikut:

$$idf = \log(\frac{N}{df}) \qquad ...(3.1)$$

N = Berita

Df = Banyaknya berita di mana suatu kata (*term*) muncul

### 3.1.7 Naïve Bayes Classifier

Tahap ini merupakan tahap penentuan keterhubungan antara kata-kata pada data. Tahap ini menggunakan sebuah algoritma *Naïve Bayes Classifier*. *Naïve Bayes Classifier* terdiri dari dua proses dalam proses klasifikasi datanya. Kedua proses itu adalah proses pembelajaran *Naïve Bayes Classifier* dan Proses klasifikasi *Naïve Bayes Classifier*.

a. Proses Prior Probabilitas

   Tahap ini melakukan perhitungan pada kata yang terdapat data tes menggunakan Rumus 2.4 dan Rumus 2.5.

b. Proses Klasifikasi *Naïve Bayes Classifier*

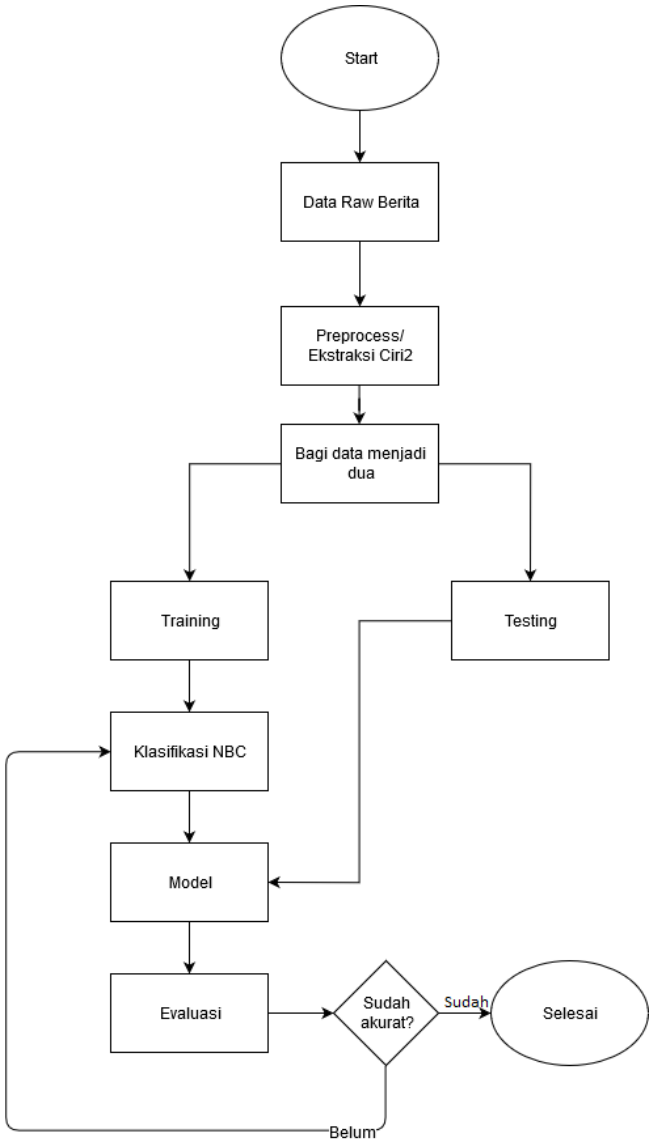
   Secara umum proses ini menentukan kelas dari data test dengan menggunakan Rumus 2.3.

### 3.1.8 Perancangan Alur Sistem

Sistem yang dibangun adalah implementasi algoritma Naïve Bayes Classifier untuk klasifikasi kategori berita pada website cnnindonesia.com. Klasifikasi ini dibangun bertujuan untuk mengetahui kategori pada sebuah artikel berita secara otomatis. Pada alur perancangan sistem akan dimulai dari

pengumpulan data berita, lalu sistem akan mengekstraksi data berita menjadi dua yaitu data untuk latihan (*training*) dan pengujian (*testing*). Data *training* akan di klasifikasikan dengan *Naïve Bayes Classifier* sampai data yang dilatih sudah akurat.

Berikut adalah perancangan alur sistem, pada Gambar 3.4:



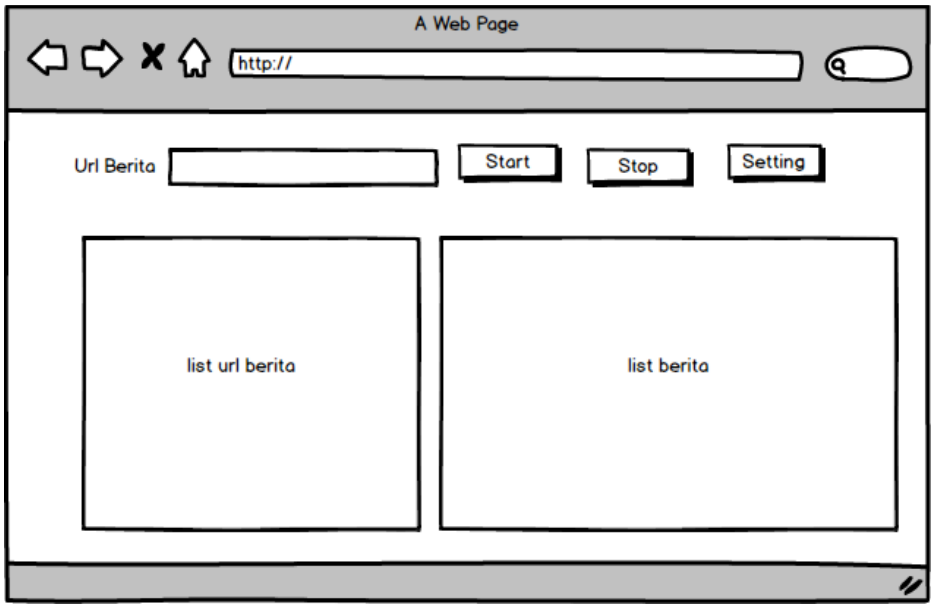

Gambar 3.4 Perancangan Alur Sistem

### 3.1.9 Desain Interface

Desain *interface* dari program yang berbasis web pada penelitian ini adalah sebagai berikut:

1. Pengumpulan Data dan Training

   Pengumpulan data dan data *training* adalah formulir untuk melakukan pengumpulan data berita dan melakukan pembobotan untuk disimpan ke dalam *database*. Halaman pengumpulan data dan *training* akan menampilkan semua hasil pengumpulan *web scraping*. *Input* berupa alamat URL dari portal berita yang ingin diambil. Halaman pada menu ini akan terlihat seperti pada Gambar 3.5.

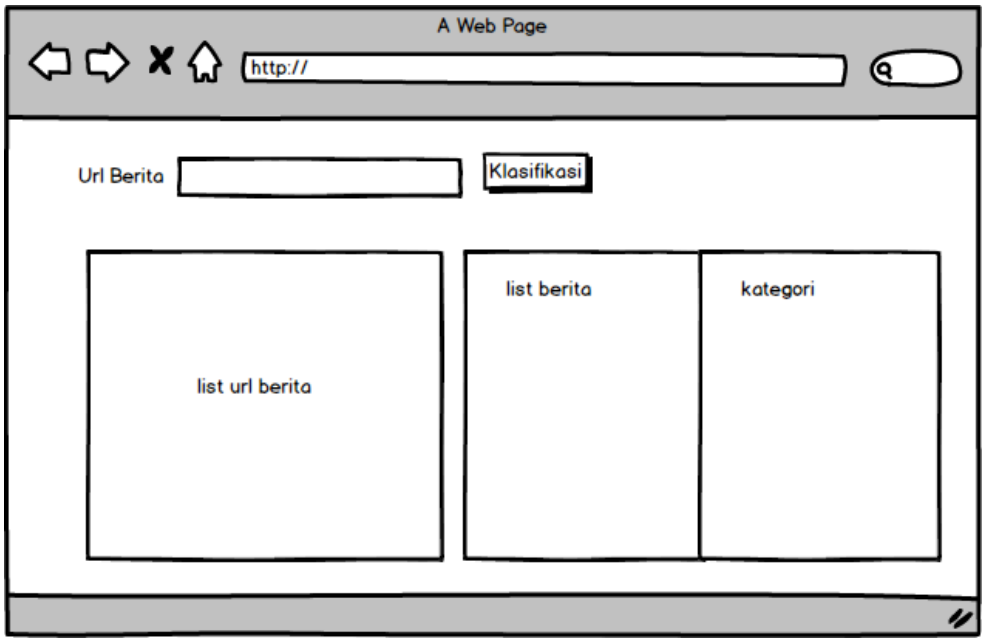



Gambar 3.5 Halaman Pengumpulan Data dan Training

2. Form Testing

   Form testing berfungsi untuk melakukan pencarian sesuai dengan *query* berita. Bentuk input *query* mungkin berisi kata, frasa atau kalimat.

Setelah proses klasifikasi dokumen akan ditampilkan menurut urutan yang paling relevan sesuai dengan *input*. Halaman pada menu ini akan terlihat seperti pada Gambar 3.6.




Gambar 3.6 Halaman Data Testing