BAB III

METODE PENELITIAN

3.1. Tahap pengumpulan data

Data awal dalam penelitian ini adalah dokumen berupa artikel teks berita online dalam bahasa Indonesia yang dikumpulkan secara acak dari portal berita online www.kompas.com dan www.republika.co.id mulai dari bulan Nopember 2012 sampai Maret 2013.

Korea Utara berencana meluncurkan sebuah roket jarak jauh antara tanggal 10 dan 22 Desember 2012. Demikian dikabarkan kantor berita Korea Utara, *KCNA*, Sabtu (1/12/2012). Dalam pernyataan yang disiarkan *KCNA*, Komite Teknologi Angkasa Luar Korea mengatakan akan meluncurkan satelit baru setelah para ilmuwan mempelajari kesalahan saat pelulncuran roket yang gagal April lalu.

"Para ilmuwan dan teknisi Korea Utara telah menganalisa kesalahan yang kami buat dalam peluncuran roket April lalu. Kami kini bekerja keras untuk meningkatkan kemampuan dan ketepatan satelit dan roket pembawanya," demikian pernyataan komite.

Komite menambahkan peluncuran roket ini adalah untuk menempatkan sebuah satelit observasi di orbit Bumi. Pada April lalu, Korea Utara sempat menggegerkan dunia terkait rencana peluncuran roket jarak jauh Unha-3 yang juga diklaim Pyongyang untuk menempatkan sebuah satelit di orbit Bumi. PBB dan AS bersikukuh bahwa roket yang diluncurkan itu adalah misil balistik jarak jauh yang merupakan varian misil antar benua Taepoding-2.

Uji coba pada April lalu itu mempengaruhi semua upaya internasional terkait masalah nuklir Korea Utara. AS bahkan menghentikan rencana bantuan pangan yang sangat dibutuhkan negeri itu. Dewan Keamanan PBB telah memperingatkan Korea Utara agar tidak melaksanakan uji coba misil balistiknya.

"Kami semua sepakat bahwa tak ada perlunya Korea Utara melanjutkan uji coba misilnya," kata Ketua Komite Sanksi untuk Korea Utara, Jose Felipe Moraes Cabral.

Gambar 3.1 Contoh data awal

Korea Utara berencana meluncurkan sebuah roket jarak jauh antara tanggal 10 dan 22 Desember 2012. Komite menambahkan peluncuran roket ini adalah untuk menempatkan sebuah satelit observasi di orbit Bumi. Uji coba pada April lalu itu mempengaruhi semua upaya internasional terkait masalah nuklir Korea Utara.

Gambar 3.2 Contoh data ringkasan manual

Sebanyak 40 buah dokumen tersebut di atas akan digunakan dalam penelitian ini dengan rincian sebagai berikut.

1. Data Pelatihan

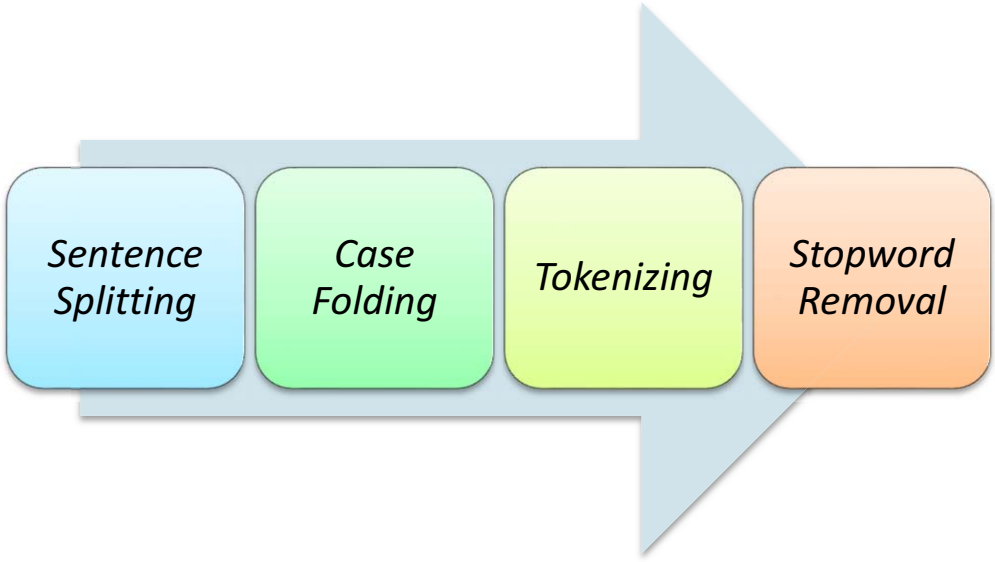
   Terdiri dari 20 dokumen teks berikut ringkasan manualnya.
2. Data Pengujian

   Terdiri dari 20 dokumen teks berikut ringkasan manualnya.

Dokumen disimpan dalam bentuk rtf dan diringkas dengan pemampatan teks (*compression rate*) sebesar 30% dari data asal sesuai dengan kriteria peringkasan teks yaitu di bawah 50%. Peringkasan manual akan dilakukan oleh seorang yang kompeten dalam bahasa Indonesia dalam hal ini guru bahasa Indonesia. Contoh Data awal seperti tampak pada gambar 3.1 dan gambar 3.2.

3.2. Pengolahan data awal

Data yang telah dikumpulkan akan diolah terlebih dahulu seperti proses gambar 3.3.



Gambar 3.3 *Text Preprocessing*

3.2.1. Sentence splitting

Pada tahap ini, dokumen teks akan dipecah ke dalam bentuk kalimat-kalimat.

3.2.2. Case folding

Pada tahap ini kalimat akan diubah menjadi huruf-huruf kecil, serta menghilangkan karakter selain huruf a-z dan data numerik.

3.2.3. Tokenizing

Pada tahap ini kalimat akan dipecah menjadi bentuk kata perkata.

3.2.4. Stopword removal

Kata-kata yang dianggap tidak penting akan dihapus pada tahap ini. Kata-kata tersebut terdaftar pada *stoplist*.

3.2.5. Stemming

Pada tahap ini kata akan diubah ke bentuk dasarnya dengan algoritma porter. Langkah-langkah algoritma porter untuk Bahasa Indonesia adalah sebagai berikut.

1. Hapus *Particle* ("-lah", "-kah", "-tah" atau "-pun"),
2. Hapus *Possesive Pronoun*.
3. Hapus awalan pertama. Jika tidak ada lanjutkan ke langkah 4a, jika ada cari maka lanjutkan ke langkah 4b.
4. a. Hapus awalan kedua, lanjutkan ke langkah 5a.
   b. Hapus akhiran, jika tidak ditemukan maka kata tersebut diasumsikan sebagai *root word*. Jika ditemukan maka lanjutkan ke langkah 5b.
5. a. Hapus akhiran. Kemudian kata akhir diasumsikan sebagai *root word*
   b. Hapus awalan kedua. Kemudian kata akhir diasumsikan sebagai *root word*

3.3. Tahap pelatihan

Tahap pelatihan terbagi menjadi dua bagian utama, yaitu: penskoran fitur teks dan pemodelan dengan GA.

3.3.1. Pemilihan fitur teks

Dalam tahap ini, akan dihitung skor kalimat berdasarkan fitur-fitur teks. Tujuh fitur teks yang dipakai adalah panjang kalimat (f_1), kalimat yang mengandung data numerik (f_2), kemiripan antar kalimat (f_3), bobot kata (f_4), kata tematik (f_5), posisi kalimat (f_6) dan kalimat yang menyerupai judul (f_7) sebagaimana dipaparkan dalam Bab 2. Dari proses ini akan didapatkan tujuh buah nilai fitur untuk tiap-tiap kalimat.

3.3.2. Pembobotan fitur teks

Fitur-fitur teks yang telah diperoleh pada tahap sebelumnya akan direpresentasikan sebagai gen-gen dalam suatu kromosom seperti gambar 3.2.

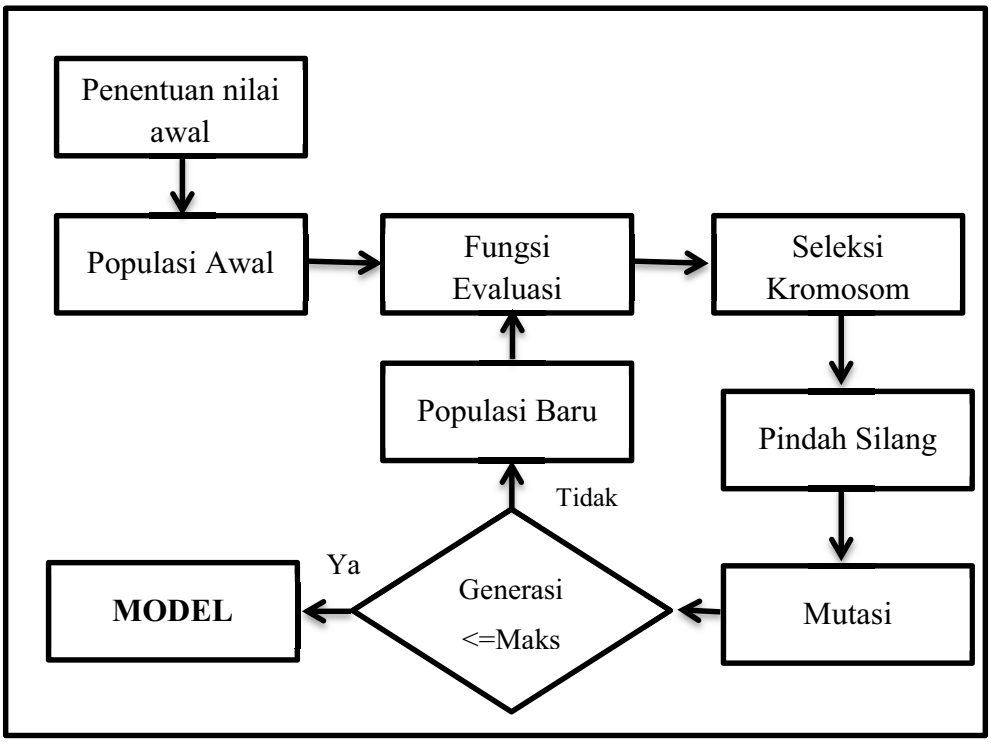
W_1	W_2	W_3	W_4	W_5	W_6	W_7

Gambar 3.4 Representasi Kromosom pada Pembobotan Fitur Teks

Pencarian bobot yang optimal dari masing-masing fitur teks dilakukan dengan GA. Bobot yang diperoleh akan diterapkan pada persamaan (3.1) yang berfungsi sebagai skor akhir dari masing-masing kalimat.

$$Score\ (S_i) = W_1 * Score_{f1}(S_i) + W_2 * Score_{f2}(S_i) + W_3 * Score_{f3}(S_i) + W_4 * Score_{f4}(S_i) + W_5 * Score_{f5}(S_i) + W_6 * Score_{f6}(S_i) + W_7 * Score_{f7}(S_i) \quad (3.1)$$

Dengan melakukan pengujian ringkasan teks yang diperoleh dalam tahap ini dan ringkasan teks dengan cara manual, dapat dihitung *fitness function* yang berfungsi untuk mengevaluasi kromosom.



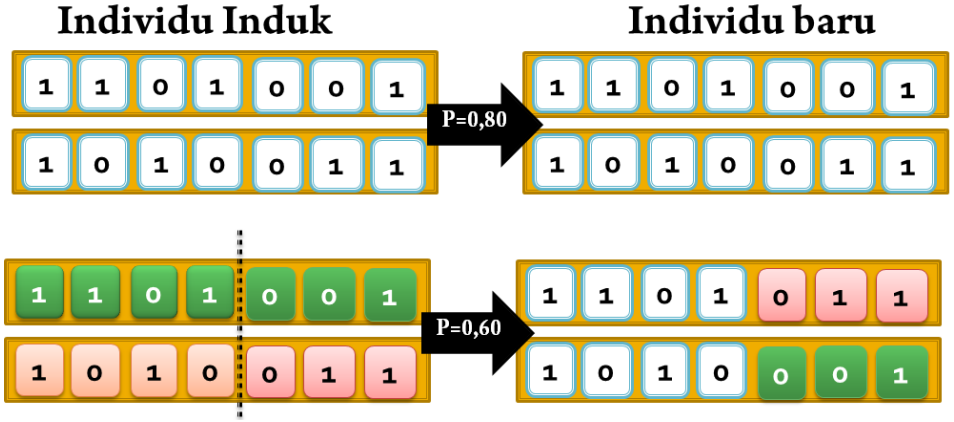
Gambar 3.5 Proses dalam GA

Proses GA dimulai dengan memberi nilai awal populasi. Tiap populasi berisi 20 kromosom. Sebuah kromosom direpresentasikan sebagai kombinasi dari fitur bobot

dalam bentuk (W_1, W_2, W_3, W_4, W_5, W_6, W_7) seperti tampak pada gambar 3.5. Sedangkan proses yang terjadi pada GA selengkapnya dapat dilihat pada gambar 3.6.

Nilai awal yang diberikan pada tahap pemodelan GA adalah 20 kromosom awal dengan peluang pindah silang sebesar 0,70 dan peluang mutasi 0,10 dengan 100 generasi akan diterapkan. Berikut ini langkah-langkah pencarian bobot optimal dengan GA:

a. Populasi awal dibangkitkan secara acak sebanyak 20 kromosom, dimana tiap kromosom merepresentasikan kombinasi dari skor fitur teks. Kombinasi bobot yang ada pada kromosom diterapkan pada (3.1) yang digunakan untuk mendapatkan nilai skor tiap kalimat.
b. Tiap kromosom dievaluasi dengan *precision rate*, dimana untuk mencari nilai *precision* seperti pada (3.2). Untuk setiap kromosom, perhitungan *precision rate* dilakukan pada 20 data *training*.
c. Setelah dievaluasi, maka dilakukan proses seleksi kromosom menggunakan minimum *fitness function*. Seleksi kromosom ini berfungsi untuk memilih kromosom-kromosom mana saja yang akan dipilih untuk proses pindah silang, mutasi dan mendapatkan calon induk yang baik.
d. Teknik yang digunakan pada pindah silang adalah teknik pindah silang satu titik (*one cut point*). Peluang pindah silang yang digunakan pada penelitian ini adalah 0,70. Pindah silang terjadi jika peluang yang dihasilkan kromosom yang dijadikan induk lebih kecil dari peluang pindah silang yang telah ditentukan.



Gambar 3.6 *Crossover* dengan *one cut point*

e. Peluang mutasi yang digunakan adalah 0,10. Sehingga diharapkan tiap generasi terdapat 14 gen yang mengalami mutasi.

Gambar 3.7 Ilustrasi Mutasi gen

f. 100 generasi diterapkan dalam proses GA untuk mendapatkan bobot ekstraksi fitur teks yang optimal.

3.4. Tahap pengujian

Tahap pengujian menggunakan program Java Netbeans. Dalam tahap ini akan dihitung nilai presisi dari metode peringkasan dokumen teks bahasa Indonesia dengan integrasi antara GA dan VSM. Penelitian ini dikatakan berhasil jika nilai presisi dari integrasi GA dan VSM lebih tinggi dari 49,96%.

3.5. Evaluasi

Setelah melakukan eksperimen dan pengujian metode, akan dilakukan tahap evaluasi. Dengan melihat tingkat akurasi pada *precision rate*. Untuk perhitungan *precision rate, recall rate* dan *f-measure* menurut Baeza dan Ribeiro [22] dapat digunakan persamaan seperti di bawah ini.

$$P = \frac{|S \cap T|}{|S|};\ R = \frac{|S \cap T|}{|T|};\ F_{untuk\ \beta=1} = \frac{(\beta^2+1)PR}{(\beta^2 P+R)} = \frac{2PR}{(P+R)} \qquad (3.2)$$

Diasumsikan bahwa S adalah hasil ringkasan teks dari sistem dan T hasil ringkasan teks secara manual. β adalah bobot dari *precision* (P) dan *recall* (R), β<1 penekanan pada precision dan β>1 penekanan pada recall.