Pembobotan Fitur Ekstraksi Pada Peringkasan Teks Bahasa Indonesia Menggunakan Algoritma Genetika

Zulkifli [1], Agung Toto Wibowo [2], Gia Septiana [3]
[123]Fakultas Informatika, Universitas Telkom, Bandung, 40257, Indonesia.
[1] zulkifli.lea@gmail.com, [2] agungtoto@telkomuniversity.ac.id, [3] giaseptiana@ telkomuniversity.ac.id

Abstrak— Hal yang penting dari peringkasan teks otomatis adalah bagaimana cara menetukkan informasi penting dari sebuah dokumen. Informasi penting dapat diperoleh dengan menggunakan teknik ekstraksi. Teknik ekstraksi merupakan teknik peringkasan secara lengkap yang terdiri dari urutan-urutan kalimat yang disalin dan memilih bagian-bagian kalimat penting dari dokumen asli. Peringkasan teks otomatis dengan teknik ekstraksi dapat dilakukan dengan menggunakan beberapa fitur ekstraksi teks. Tiap-tiap fitur ekstraksi teks memiliki tingkat pengaruh yang berbeda-beda terhadap hasil ringkasan sistem. Oleh karena itu, dibutuhkanlah sebuah algoritma optimasi untuk menentukkan tingkat kepentingan atau nilai bobot dari tiap-tiap fitur ekstraksi. Salah satu algoritma optimasi yang dapat digunakan adalah algoritma genetika. Dengan dataset yang sama untuk CR 30%, penelitian yang telah dilakukan oleh Aristoteles, Marlina, Rivaldi dan Wibowo menghasilkan maksimal akurasi berturut-turut 47%, 42,8%, 44% dan 52,47%. Pada TA ini, algoritma genetika digunakan untuk melakukan optimasi bobot fitur ekstraksi pada peringkasan teks bahasa Indonesia. Pada tahap pelatihan algoritma genetika mampu melakukan optimasi bobot fitur ekstraksi teks yang menghasilkan akurasi sekitar 46%. Sedangkan pada tahap pengujian, sistem menghasilkan ringkasan dengan akurasi 46% untuk sepuluh fitur teks. Setelah dilakukan observasi model kromosom terbaik, sistem menghasilkan ringkasan dengan akurasi 53% untuk delapan fitur ekstraksi teks.

***Kata kunci*—** peringkasan, teks, otomatis, fitur, ekstraksi teks, algoritma genetika.

I. PENDAHULUAN

Membaca berita adalah salah satu aktifitas yang dilakukan oleh seseorang untuk mendapatkan intisari dari berita. Untuk mendapatkan intisari dari berita biasanya seseorang harus membaca seluruh teks yang ada pada dokumen tersebut. Namun, pada kenyataanya hanya dengan membaca ringkasan dari sebuah dokumen seseorang memperoleh intisari dari berita. Membaca dan memahami keseluruhan teks membutuhkan waktu yang cukup lama. Oleh karena itu, ringkasan teks pada sebuah dokumen sangat penting untuk mengatasi masalah waktu baca tersebut. Akan tetapi, untuk membuat sebuah ringkasan dokumen membutuhkan biaya dan waktu pula. Sehingga diperlukan sebuah sistem yang dapat melakukan peringkasan teks dokumen secara otomatis agar proses lebih efisien.

Peringkasan teks otomatis adalah proses mengurangi teks pada dokumen menggunakan *program* komputer untuk membuat ringkasan yang berisikan poin-poin penting dimana hasil ringkasan tidak lebih dari setengah dokumen asli [14]. Terdapat dua bagian dari kriteria peringkasan teks yaitu ekstraksi dan abstraksi. Teknik ekstraksi yaitu teknik peringkasan secara lengkap yang terdiri dari urutan-urutan kalimat yang disalin dan memilih bagian-bagian kalimat penting dari dokumen asli [10]. Sedangkan teknik abstraksi adalah teknik peringkasan dengan mengambil informasi penting dari dokumen kemudian menghasilkan ringkasan yang menggunakan kalimat baru yang tidak terdapat pada dokumen asli [2].

Pada peringkasan teks menggunakan teknik ekstraksi memiliki beberapa bagian penting dalam proses peringkasan yaitu bagaimana cara menentukan kalimat-kalimat yang penting dalam sebuah dokumen [19]. Salah satu caranya dapat menggunakan beberapa fitur ekstraksi teks [1] seperti posisi kalimat, koneksi antar kalimat, kalimat positif dan sebagainya. Akan tetapi, pada kenyataanya tiap-tiap fitur ekstraksi teks memiliki tingkat pengaruh yang berbeda-beda terhadap hasil ringkasan sistem. Oleh karena itu, dibutuhkanlah sebuah algoritma optimasi untuk menentukkan tingkat kepentingan atau nilai bobot dari tiap-tiap fitur ekstraksi. Salah satu algoritma optimasi yang dapat digunakan adalah algoritma genetika.

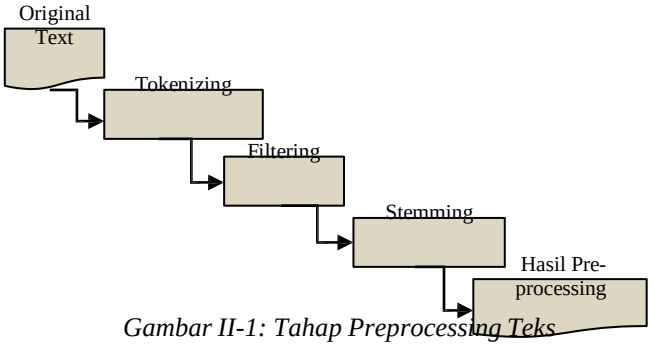
Dengan dataset yang sama dan CR 30%, penelitian yang telah dilakukan oleh [3], [12], [16] dan [21] menghasilkan maksimal akurasi berturut-turut 47%, 42,8%, 44% dan 52,47%. Pada tugas akhir ini menggunakan algoritma genetika untuk melakukan optimasi bobot fitur ekstraksi pada peringkasan teks bahasa Indonesia dikarenakan permasalahan bobot fitur ekstraksi tersebut dapat direpresentasikan dalam bentuk kromosom dan telah berhasil digunakan dalam peringkasan teks untuk dokumen bahasa Inggris [2].

II. LANDASAN TEORI

A. *Preprocessing*

Tahap *preprocessing* teks terdiri dari beberapa tahap yaitu *tokenizing*, *filtering*, *tagging*, dan *stemming*. Pada tugas akhir

ini hanya menggunakan proses *tokenizing*, *filtering*, dan *stemming*. Proses *tagging* tidak digunakan karena ketiga proses yang telah disebutkan sebelumnya sudah cukup untuk mendapatkan data yang terstruktur. Berikut ini penjelasan tahap-tahap *preprocessing*:



Gambar II-1: Tahap Preprocessing Teks

Tokenizing

Pada proses *tokenizing*, kata-kata yang ada di dalam dokumen harus dipecah-pecah terlebih dahulu menjadi bagian-bagian yang lebih kecil berupa kata tunggal yang memiliki arti atau biasa disebut token. Proses *tokenizing* pada tugas akhir ini dilakukan per kalimat. Selain itu dilakukan juga pengubahan huruf-huruf yang ada di dalam dokumen menjadi huruf kecil (*case folding*) serta dilakukan penghilangan tanda baca. Hal ini perlu dilakukan terlebih dahulu untuk mempermudah proses pengolahan lebih lanjut.

Filtering

Text filtering bertujuan untuk mengambil kata-kata yang dapat merepresentasikan isi dokumen dengan cara membuang kata-kata yang dianggap tidak penting yang biasa disebut stopwords. Stopwords dapat berupa kata sambung, kata depan, dan kata seru seperti “di”, “yang”, “dan”, “ke”, “wah”, “serta”, “wow”, dan lain lain.

Stemming

Stemming adalah proses yang dilakukan untuk mengambil bentuk dasar dari suatu kata yang telah melalui proses *filtering* [11]. Banyak metode yang dapat digunakan untuk melakukan *stemming* pada dokumen berbahasa Indonesia salah satunya adalah algoritma Nazief dan Adriani. Algoritma ini berdasarkan aturan-aturan yang mengelompokkan imbuhan yang diperbolehkan dan dilarang untuk digunakan. Pada tugas akhir ini, menggunakan librari untuk melakukan *stemming* yang dapat di download di: https://github.com/SeptiyanAndika/ENHANCED-CS.

B. Fitur Ekstraksi Teks

Pada tugas akhir ini menggunakan fitur ekstraksi untuk menghitung skor tiap-tiap kalimat dalam dokumen. Untuk setiap kalimat dalam dokumen, skor kalimat dihitung berdasarkan fitur ekstraksi dimana nilai dari tiap-tiap fitur dinormalisasikan sehingga nilainya berada dalam range [0,1]. Normalisasi ini dilakukan agar nilai dari tiap-tiap fitur ekstraksi tidak memiliki gap atau selisih yang besar.

Adapun fitur-fitur ekstraksi yang digunakan pada tugas akhir ini yaitu posisi kalimat (f1), *positive keyword* pada kalimat (f2), *negative keyword* pada kalimat (f3), kemiripan antar kalimat (f4), kalimat yang menyerupai judul (f5), kalimat yang mengandung nama entiti (f6), kalimat yang mengandung angka (f7), panjang kalimat (f8), koneksi antar kalimat (f9), penjumlahan bobot koneksi antar kalimat (f10). Penjelasan dari tiap-tiap fitur adalah sebagai berikut ini:

Posisi Kalimat (F1)

Posisi kalimat adalah letak kalimat dalam sebuah paragraf. Pada tugas akhir ini diasumsikan bahwa kalimat pertama dari sebuah paragraf adalah yang paling penting [6]. Fitur ini dapat dihitung menggnakan rumus berikut ini:

$$( ) \quad —— \quad (\quad)$$

Diasumsikan bahwa s adalah kalimat, N adalah jumlah kalimat, i adalah posisi kalimat ke-i dan ( ) adalah skor sebuah kalimat s berdasarkan fitur f1.

Positive Keyword (F2)

Positive keyword adalah kata yang sering muncul pada sebuah paragraf [12]. Fitur ini dapat dihitung menggunakan rumus (2.2) :

$$( ) \quad \frac{(\qquad\qquad)}{\sum \quad (\qquad\qquad)} (\quad)$$

Dengan ( ) adalah jumlah kata dalam suatu kalimat yang mengandung *keyword* dibagi dengan jumlah kata dalam seluruh kalimat yang mengandung *keyword*, dengan *keyword* merupakan banyaknya kata yang muncul dalam suatu paragraf.

Negatif Keyword (F3)

Kebalikan dari *positive keyword*, *negative keyword* adalah kata yang sedikit muncul dalam kalimat paragraf [12]. *Negative keyword* dapat dihitung dengan menggunakan rumus (2.3) :

$$( ) \quad \frac{(\qquad\qquad)}{\sum \quad (\qquad\qquad)} (\quad)$$

dengan ( ) adalah jumlah kata dalam suatu kalimat yang mengandung *keyword* dibagi dengan jumlah kata dalam seluruh kalimat yang mengandung *keyword*, dengan *keyword* merupakan banyaknya kata yang muncul dalam suatu paragraf.

Kemiripan Antar Kalimat (F4)

Kemiripan antar kalimat adalah daftar kata-kata yang dapat dicocokkan antara kalimat yang satu dengan kalimat yang lainnya dalam dokumen atau dengan kata lain merupakan kata yang muncul dalam kalimat sama dengan kata yang muncul dalam kalimat lain [3]. Fitur ini dihitung menggunakan rumus (2.4)

$$( )$$
$$\left| \frac{\cap}{\cup} \right| (\quad)$$

Kalimat yang Menyerupai Judul Dokumen (F5)

Kalimat yang menyerupai judul dokumen adalah kumpulan kata yang dapat dicocokkan antara kalimat satu dengan judul

kalimat atau dengan kata lain merupakan kata yang muncul dalam kalimat sama dengan kata yang ada dalam judul dokumen [17]. Fitur ini dapat dihitung dengan menggunakan rumus (2.5)

( )

| ∩ / ∪ | ( )

Kalimat yang Mengandung Entiti (F6)

Nama entiti adalah sebuah kumpulan kata yang memiliki makna atau membentuk nama sebuah intitusi. Misalnya adalah Universitas Telkom yang merupakan kumpulan kata yang memiliki makna sebuah institusi perguruan tinggi [6] Fitur ini dihitung menggunakan rumus (2.6)

(( ))

( ) ( ) ( )

Kalimat yang Mengandung Data Numerik (F7)

Pada peringkasan teks mempertimbangkan data numerik karena kalimat yang memiliki angka numerik biasanya penting dan sangat mungkin berada pada ringkasan dokumen [6]. Fitur ini dapat dihitung menggunakan rumus (2.7) :

( )

( ) ( ) ( )

Panjang Kalimat (F8)

Fitur ini bertujuan untuk menghilangkan kalimat-kalimat yang terlalu pendek. Fitur ini dihitung berdasarkan jumlah kata dalam kalimat dibagi jumlah kata unik dalam paragraf [5,6]. Fitur ini dihitung menggunakan rumus (2.8):

( )

( ) ( )

Koneksi Antar Kalimat (F9)

Koneksi antar kalimat adalah banyaknya *link* dari suatu kalimat yang terhubung dengan kalimat yang lain. Atau dengan kata lain merupakan jumlah kalimat yang memiliki kata yang sama dengan kalimat lain dalam satu dokumen [3]. Fitur ini dihitung menggunakan rumus (2.9):

( )

Σ ( )

Penjumlahan Bobot Koneksi Antar Kalimat (F10)

Fitur ini berfungsi menjumlahkan kata dalam suatu kalimat dengan kata yang sama dari kalimat lain dalam suatu dokumen [12]. Fitur ini dihitung menggunakan rumus (2.10):

( )

Σ ( )

C. Pembobotan Fitur Ekstraksi Teks

Pembobotan fitur ekstraksi teks adalah sebuah pendekatan yang dilakukan untuk menentukkan kepentingan suatu fitur dari terhadap akurasi hasil ringkasan sistem nantinya. Pembobotan fitur pada tugas akhir ini menggunakan algoritma genetika untuk mendapatkan bobot yang optimal untuk tiap-tiap fiturnya. Skor untuk tiap kalimat dapat dihitung menggunakan rumus (2.11) :

( ) Σ

( )

Diasumsikan w_i adalah bobot fitur ke-i dan f_i adalah fitur ekstraksi ke-i.

D. Evaluasi Hasil Ringkasan Sistem

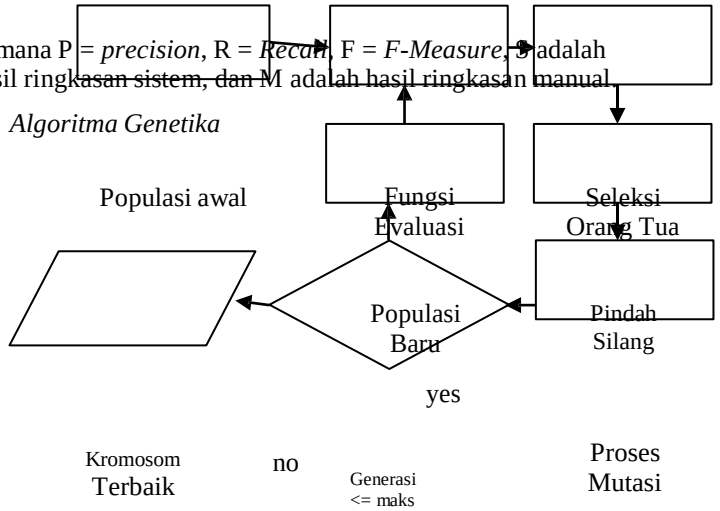
Terdapat dua teknik untuk mengevaluasi hasil ringkasan teks yaitu *extrinsic evaluation* dan *instrinsic evaluation* [8]. *Extrinsic* evaluation adalah proses penilaian efisiensi hasil ringkasan berdasarkan pada fungsi tertentu dan ditujukan untuk penilaian relevansi atau pemahaman membaca. Sedangkan *instrinsic evaluation* merupakan metode yang berdasarkan perhitungan antara sistem (peringkasan teks oleh sistem) dengan hasil ringkasan manual yang dibuat oleh manusia.

Penelitian ini menggunakan metode *instrinsic evaluation* dengan menggabungkan metode *recall* (R), *precision* (P), dan *F-Measure* (F) pada hasil ringkasan oleh manusia (ringkasan manual) dengan hasil ringkasan yang dibuat oleh mesin. Berdasarkan [3] untuk menghitung besarnya *precision*, *recall* dan *F-Measure* diperlihatkan pada formula berikut ini :

∩ ∩ ( ) ( )

Dimana P = *precision*, R = *Recall*, F = *F-Measure*, S adalah hasil ringkasan sistem, dan M adalah hasil ringkasan manual.

E. Algoritma Genetika



fitur yang akan diteliti dengan cara mengalikan bobot dengan fitur ekstraksi. Pembobotan ini sangat berpengaruh

Gambar II-2: Proses Algoritma Genetika

Dalam menyelesaikan suatu permasalahan, algoritma genetika bekerja mengikuti langkah-langkah berikut [13] :

1. Mulai dengan menciptakan populasi awal secara acak sebanyak n kromosom

2. Hitung nilai *fitness* dari masing-masing kromosom yang ada di dalam populasi.
3. Ulangi langkah berikut untuk membentuk populasi baru:
   a. Pilih sepasang *parent* dari populasi saat ini, lalu lakukan *crossover* atau pindah silang sesuai dengan *crossover rate* sehingga menghasilkan sepasang *offspring* atau anak .
   b. Lakukan mutasi terhadap populasi saat ini sesuai dengan *mutation rate*.
   c. Lakukan seleksi terhadap individu-individu yang ada di dalam populasi awal ditambah dengan *offspring* untuk membentuk sebuah populasi baru.

4. Gunakan populasi baru untuk melanjutkan penggunaan algoritma genetika.
5. Kembali ke langkah 2 dan ulang hingga kondisi akhir terpenuhi.

III. PERANCANGAN SISTEM

Penelitian ini dilakukan dengan tiga tahap yaitu : tahap pengumpulan dokumen teks, tahap penelitian, dan tahap pengujian.

A. Pengumpulan Dokumen

Penelitian ini membutuhkan inputan berupa dokumen bahasa Indonesia dengan dokumen bertipe file teks xml. Pada penelitian ini digunakan dokumen sebanyak 150 dokumen berita nasional. 100 dokumen untuk penelitian dan 50 dokumen untuk pengujian.

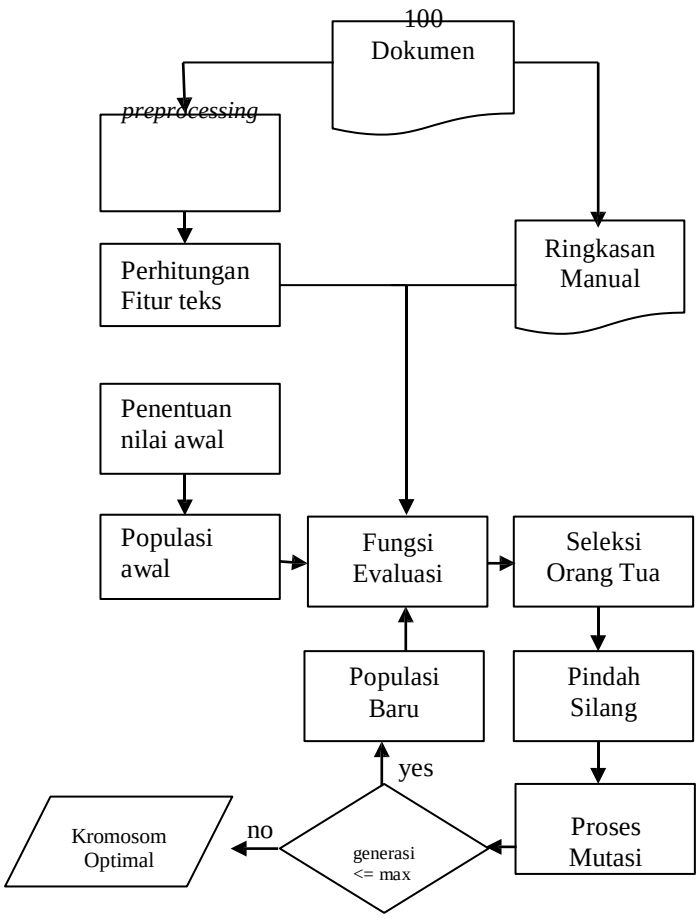
Peringkasan dokumen secara manual adalah proses peringkasan yang dilakukan oleh seseorang yang ahli dalam bidangnya dengan menggunakan prinsip-prinsip peringkasan sesuai dengan kaidah Bahasa Indonesia. Pada TA ini, menggunakan dokumen yang berasal dari berita *online* harian yang didapat dari korpus penelitian [15]. Ringkasan dokumen tersebut nantinya akan digunakan untuk menghitung akurasi dengan cara membandingkan hasil ringkasan sistem dengan hasil ringkasan manual. Berikut ini contoh dokumen :

```
<?xml version="1.0"?>
<ROOT>
<DOCUMENT>
<TITLE>Pemerintah Belum Miliki Draf RUU Antiterorisme</TITLE>
<TEXT>Jakarta, Kompas - Menteri Kehakiman dan Hak Asasi Manusia (Menkeh dan HAM) Yusril Ihza Mahendra menegaskan, pemerintah belum memiliki rancangan atau draf RUU Antiterorisme seperti yang disebut berbagai pihak.

"Draf itu belum ada sama sekali. Oleh karena itu, masyarakat tidak perlu buang waktu membahas bahan yang disebut sebagai draf RUU Antiterorisme," katanya kepada pers menjelang Sidang Kabinet di Gedung Utama Sekretariat Negara Jakarta, Senin (8/4).</TEXT>
</DOCUMENT>
</ROOT>
```

B. Pelatihan

Tahap ini tediri dari empat bagian utama yaitu: peringkasan dokumen secara manual, *preprocessing* data, ekstraksi fitur dan pemodelan algoritma genetika. Berikut ini *flowchart* tahap pelatihan :



Gambar III-1: Flowchart Proses Pelatihan

Pada tahap pelatihan, digunakan 100 dokumen sebagai inputan. Sebelum melakukan ekstraksi fitur dilakukan, dokumen akan di *preprocessing* dengan cara menerapkan konsep *text mining.* Ekstraksi fitur teks merupakan suatu proses untuk mendapat ekstraksi teks dari dokumen [3]. Hasil dari fitur teks adalah ektraksi teks seperti posisi kalimat (f1), *positive keyword* (f2), *negative keyword* (f3), kemiripan antar-kalimat (f4), kalimat yang menyerupai judul dokumen (f5), kalimat yang mengandung entiti (f6), kalimat yang mengandung data numerik (f7), panjang relatif kalimat (f8), koneksi antar-kalimat (f9) dan penjumlahan bobot koneksi antar-kalimat (f10). *Output* dari proses ekstraksi fitur teks adalah nilai skor dari masing-masing fitur. Nilai skor tersebut nanti akan digunakan untuk perhitungan fungsi evaluasi pada algoritma genetika.

Tahap selanjutnya, algoritma genetika digunakan untuk melakukan pencarian bobot yang optimal pada tiap ekstraksi fitur teks. Ringkasan manual dan ekstraksi fitur teks digunakan untuk menghitung nilai fitness yang berfungsi untuk mengevaluasi kromosom.

Berikut ini langkah-langkah proses algoritma genetika pada tugas akhir ini :

a. Proses algoritma genetika dimulai dengan memberi nilai awal populasi atau jumlah individu, peluang mutasi, peluang rekombinasi, *tournament size*, range nilai gen, dan variabel *elitism*.

b. Populasi awal dibangkitkan secara acak sebanyak n kromosom sesuai dengan nilai populasi yang telah ditentukkan sebelumnya, dimana tiap kromosom merepresentasikan bobot-bobot untuk sebuah ringkasan secara keseluruhan. Sedangkan gen pada kromosom merepresentasikan bobot untuk tiap ekstraksi fitur teks. Pada tugas akhir ini, kromosom direpresentasikan sebagai kombinasi seluruh fitur bobot dalam bentuk sebagai berikut:

W1	W2	W3	W4	W5	W6	W7	W8	W9	W10

Gambar III-2: merupakan representasi kromosom pada pembobotan ekstraksi fitur teks dengan w1 bobot pada ekstraksi fitur teks (f1), w2 bobot pada ekstraksi fitur teks (f2), dan seterusnya

Nilai-nilai gen yang ada pada kromosom digunakan untuk menghitung nilai skor sebuah kalimat dengan mengalikan nilai bobot ke-i dengan nilai fitur ke-i. Menggunakan rumus d bawah ini. Perhitungan nilai skor kalimat telah dijelaskan pada bab 3.1.4

( ) ( ) ( )
( ) ( )
( ) ( )
( ) ( )
( ) ( ) ( )

c. Tiap kromosom dievaluasi oleh rata-rata *F-measure*, dimana nilai *precision* dan *recall* diperoleh dari irisan hasil ringkasan yang dibuat oleh sistem dan hasil ringkasan manual. Berdasarkan [3] untuk menghitung besarnya *precision*, *recall* dan *F- Measure* untuk satu buah dokumen diperlihatkan pada formula berikut ini :

$$\frac{\cap}{} \qquad \frac{\cap}{} \qquad \frac{}{(\quad)} (\quad)$$

Dimana P = *precision*, R = *Recall*, F = *FMeasure*, S adalah hasil ringkasan sistem, dan M adalah hasil ringkasan manual. Untuk contoh perhitungan *F-measure* telah dijelaskan pada bab 3.1.5. Untuk setiap kromosom, proses *F-measure* dilakukan sebanyak 100 kali sesuai dengan jumlah dokumen pelatihan. Sehingga fungsi *fitness* untuk suatau kromosom adalah sebagai berikut ini:

$$\frac{\sum}{} (\quad)$$

Dimana adalah fungsi *fitness* untuk suatu kromosom, adalah indeks dokumen, adalah jumlah dokumen yang digunakan, dan adalah hasil perhitungan *precision* dan *recall* pada formula (2.12)

d. Setelah kromosom dievaluasi, dilakukan seleksi orang tua menggunakan metode *tournament selection*. Metode ini berfungsi untuk memilih kromosom-kromosom mana saja yang akan dipilih menjadi orang tua untuk proses pindah silang yang akan menghasilkan kromosom anak yang baru.

e. Peluang pindah silang yang digunakan pada penelitian ini dalam selang [0,1]. Teknik yang digunakan pada pindah silang adalah teknik pindah silang *uniform*/seragam.

f. Setelah proses pindah silang, proses mutasi dilakukan pada seluruh kromosom baru. Untuk setiap gen pada kromosom dilakukan pembangkitan bilangan acak dalam interval [0,1]. Jika jumlah gen adalah L, maka diperlukan pembangkitan bilangan acak sebanyak L kali.

g. Jumlah generasi diterapkan pada proses algoritma genetika untuk mendapatkan bobot ekstraksi fitur teks yang optimal. Diasumsikan ketika generasi telah mencapai maksimum generasi, maka akan dihasilkan kromosom terbaik atau optimal, jika generasi kurang dari maksimum generasi, maka proses algoritma terus dilakukan.

IV. PENGUJIAN DAN ANALISIS

A. Pengujian

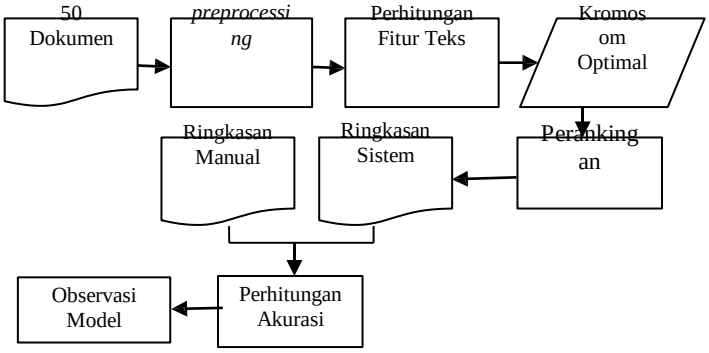
Pengujian yang dilakukan pada sistem terdiri dari dua tahap. Tahap pertama adalah pengujian pada tahap pelatihan. Pada tahap ini bertujuan untuk menemukan kombinasi parameter GA yang menghasilkan kromosom dengan nilai *fitness* yang paling optimal. Tahap kedua adalah menentukkan model kromosom terbaik pada tahap pengujian dengan menggunakan dokumen yang berbeda dengan dokumen yang digunakan pada tahap pelatihan sebelumnya.

Observasi Parameter GA

Langkah pengujian ini adalah melakukan observasi untuk mendapatkan parameter GA yang menghasilkan model kromosom yang paling optimal dengan cara melakukan kombinasi parameter iterasi, jumlah individu, dan elitisme. Untuk nilai iterasi yang digunakan yaitu 10, 100, 500 1000, dan 1500. Untuk jumlah individu yaitu menggunakan nilai 5, 10, 50 dan 100. Sedangkan untuk elitisme bernilai *true* or *false*. Ketiga parameter tersebut akan digunakan untuk nilai CR yaitu 10%, 20%, dan 30%

Observasi Model Kromosom Terbaik

Langkah pengujian ini bertujuan untuk menentukan bobot fitur teks yang penting dalam peringkasan teks sehingga bobot-bobot fitur teks yang tidak penting dapat diabaikan dalam peringkasan teks. Pada tahap ini menggunakan 50 dokumen yang berbeda dengan tahap pelatihan sebelumnya. Pada tahap ini diambil 10 model kromosom optimal hasil dari tahap pelatihan. 10 kromosom tersebut nantinya akan digunakan pada perankingan bobot untuk menentukan bobot-bobot fitur teks yang penting dan tidak penting. Berikut ini *flowchart* untuk tahap ini :



Gambar IV-1: Flowchart Pengujian Model Kromosom

Tahap pengujian ini mengunakan 50 dokumen berita berbahasa Indonesia (dokumen yang digunakan pada tahap ini berbeda dengan dokumen yang digunakan pada tahap pelatihan). Dokumen-dokumen tersebut selanjutnya di *preprocessing*. Proses selanjutnya yaitu ekstraksi fitur teks. Proses *preprocessing* dan ekstraksi fitur teks ini sama dengan yang dilakukan pada tahap pelatihan yang telah dijelaskan sebelumnya. Proses peringkasan teks secara otomatis didasari model (kromosom terbaik) yang yang diperoleh dari hasil perankingan bobot. Model kromosom direpresentasikan sebagai bobot *(w1,w2,…,w10)* yang diterapkan pada fungsi skor untuk setiap kalimat.

Scoring kalimat merupakan salah satu tahap untuk menghasilkan sebuah ringkasan. Setelah seluruh kalimat dihitung skornya, kalimat diurutkan secara *descending* berdasarkan hasil *scoring* dan jumlah kalimat yang ditetapkan sebagai hasil ringkasan sistem menggunakan top-skor sesuai dengan *compression rate* (CR) yang digunakan.

B. Analisis

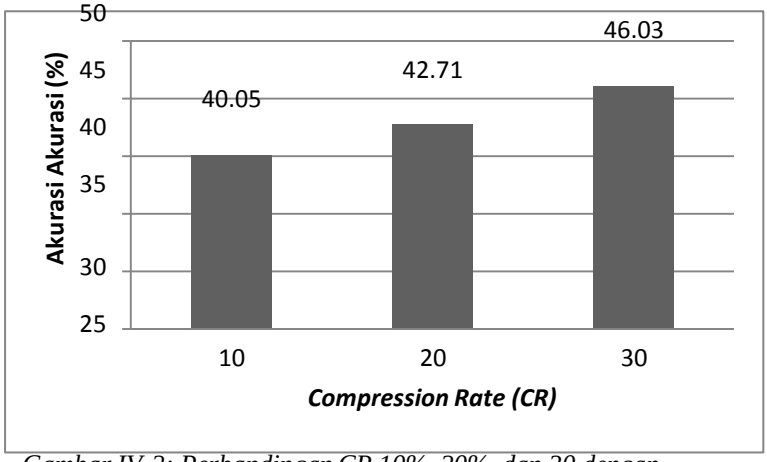
Analisis Hasil Observasi Parameter GA

Berdasrakan hasil observasi parameter GA yang telah dilakukan, kombinasi parameter GA yang optimal yaitu jumlah individu=50, iterasi=500, dan elitisme=*true*. Kombinasi parameter tersebut digunakan pada tahap pelatihan. Berikut ini hasil pelatihan dengan model kromosom yang optimal untuk tiap-tiap CR 10%, 20% dan 30% dengan menggunakan parameter jumlah individu=50, iterasi=500, dan elitisme=*true*:

Tabel IV-1:Hasil Pelatihan Untuk Tiap-Tiap CR

CR	Rata2 Best *Fitness*	*Fitness* Global	Fmeasure	Akurasi
10%	37.6326	40.0536	0.4005	40.05
20%	40.371	42.7144	0.4271	42.71
30%	44.719	46.0308	0.4603	46.03

Dari tabel IV-1, dapat dilihat bahwa percobaan yang menghasilkan *fitness* terbaik pada CR 10%, 20% dan 30% yaitu dengan menggunakan parameter jumlah individu=50, iterasi=500 dan elitisme=*true*.



Gambar IV-2: Perbandingan CR 10%, 20%, dan 30 dengan jumlah individu=50, iterasi=500, dan elitism=true

Berdasarkan gambar IV-2, CR 30% memiliki akurasi paling tinggi dibandingkan dengan hasil akurasi dari CR 10%, dan CR 20%. Oleh karena itu, pada tahap pelatihan ini akan menggunakan model kromosom optimal CR 30%. Pelatihan ini dilakukan sebanyak 10 kali untuk menghasilkan 10 model kromosom optimal dengan menggunakan CR 30%. Alasan dilakukannya 10 kali percobaan pelatihan adalah karena nilai output kromosom tiap percobaan berbeda-beda. Sebagai contoh kemungkinan pada percobaan pertama fitur yang memiliki nilai tertinggi adalah fitur f1, sedangkan untuk percobaan kedua kemungkinan fitur yang memiliki nilai tertinggi adalah fitur f2.

Analisis Hasil Observasi Model Kromosom Terbaik

Berdasarkan hasil observasi parameter GA, model kromosom yang memiliki akurasi paling besar adalah model kromosom dengan CR 30%. Oleh karena itu, pada tahap ini akan menggunakan CR 30%. Selanjutnya, diambil sepuluh model kromosom optimal dari tahap pelatihan. Sepuluh model kromosom tersebut digunakan pada perankingan bobot untuk menentukan bobot-bobot fitur teks yang penting dan tidak penting. Mekanisme perankingan bobot yang digunakan yaitu dengan memberikan nilai antara 1-10. 10 untuk bobot yang memiliki nilai tertinggi dan 1 untuk bobot yang memiliki nilai terendah. Sehingga maksimal total bobot dari sepuluh model kromosom adalah 100.

Setelah hasil penjumlahan bobot diperoleh, langkah selanjutnya melakukan perankingan bobot berdasasrkan total bobot secara *descending*. Bobot yang memiliki total bobot terbesar menduduki ranking teratas. Berikut ini tabel hasil perankingan bobot.

Tabel IV-2: Tabel Hasil Perankingan Bobot

Ranking	Bobot Fitur	Total Bobot
1	W5	93
2	W6	82
3	W4	74
4	W7	55
5	W8	46
6	W9	45
7	W10	45

8	W1	38
9	W2	37
10	W3	35

Tabel IV-2 merupakan hasil perankingan bobot. Bobot fitur W5 memiliki total bobot terbesar yaitu 93 sehingga bobot fitur W5 berada pada ranking pertama. Hal ini menandakkan bahwa bobot fitur W5 memiliki pengaruh yang sangat besar terhadap hasil ringkasan sistem. Sebaliknya bobot fitur W3 memiliki total bobot terkecil yaitu 35 sehingga bobot fitur W5 berada pada ranking terakhir. Hal ini menandakkan bahwa

ringkasan sistem.

Tujuan dari perangkingan bobot fitur teks adalah

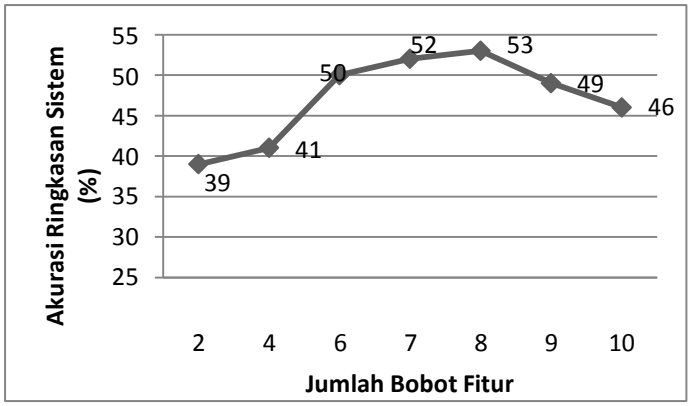
peringkasan teks. Berdasarkan hasil ranking tersebut, dilakukan pengujian kombinasi bobot untuk menentukkan pengaruh bobot terhadap hasil ringkasan sistem. Kombinasi bobot dilakukan berdasarkan hasil ranking. Berikut ini kombinasi bobot yang digunakan besarta akurasinya yang

menggunakan 50 dokumen pengujian:

Tabel IV-3: Tabel Kombinasi Bobot

Jumlah Bobot	Kombinasi Bobot	Akurasi
2	(F5,F6)	39.0
4	(F4,F5,F6,F7)	41.0
6	(F4,F5,F6,F7,F8,F9)	50.0
7	(F4,F5,F6,F7,F8,F9,F10)	52.0
8	(F1,F4,F5,F6,F7,F8,F9,F10)	53.0
9	(F1,F2,F4,F5,F6,F7,F8,F9,F10)	49.0
10	(F1,F2,F3,F4,F5,F6,F7,F8,F9,F10)	46.0

Tabel IV-3 merupakan kombinasi bobot beserta model kromosom dari kombinasinya. Dengan menggunakan 2 bobot fitur, ringkasan sistem menghasilkan akurasi sebesar 39%, 4 bobot fitur menghasilkan akurasi 41%, 6 bobot fitur menghasilkan akurasi 50%, 7 bobot fitur menghasilkan akurasi 52%, 8 bobot fitur menghasilkan akurasi 53%, 9 bobot fitur menghasilkan akurasi 49%,dan 10 bobot fitur menghasilkan akurasi 46%. Jika ditransformasikan kedalam bentuk grafik menjadi berikut ini:



Gambar IV-3: Grafik Pengaruh Jumlah Kombinasi Bobot Terhadap Akurasi Ringkasan Sistem

Pada gambar IV-3, dapat dilihat bahwa akurasi ringkasan sistem pada saat menggunakan 2 bobot fitur adalah 39%. Terjadi peningkatan 2% ketika menggunakan 4 bobot fitur, peningkatan 9% ketika menggunakan 6 bobot fitur, peningkatan 2% ketika menggunakan 7 bobot fitur, peningkatan 1% ketika menggunakan 8 bobot fitur, mengalami penurunan 4% ketika menggunakan 9 bobot fitur, dan mengalami penurunan 3% ketika menggunakan 10 bobot fitur.

Dari gambar 4-7, akurasi ringkasan sistem terbesar yaitu pada saat menggunakan 8 bobot fitur. Oleh karena itu, dapat disimpulkan bahwa peringkasan sistem dapat dilakukan hanya dengan menggunakan 8 fitur teks saja yaitu (F1,F4,F5,F6,F7,F8,F9,F10) tanpa harus menggunakan seluruh fitur teks. Fitur F2 dan F3 dapat diabaikan karena ketika menggunakan fitur F2 dan F3, akurasi ringkasan sistem mengalami penurunan.

V. KESIMPULAN DAN SARAN

A. Kesimpulan

Kesimpulan dari tugas akhir ini adalah sebagai berikut:

1. Algoritma genetika dapat digunakan untuk mengoptimasi bobot fitur ekstraksi pada peringkasan teks berbahasa Indonesia.
2. Parameter *crossover rate*, *mutation rate*, iterasi, jumlah individu dan elitisme berpengaruh terhadap akurasi ringkasan sistem dengan nilai parameter optimal berturut-turut 0,5 ; 0,1 ; 500 ; 50 ; *true*.
3. Peringkasan sistem dapat dilakukan hanya dengan menggunakan 8 fitur teks saja yaitu F1, F4, F5, F6, F7, F8, F9, F10 tanpa harus menggunakan seluruh

fitur teks. Fitur Teks F2 dan F3 dapat diabaikan karena ketika menggunakan fitur teks positif *keyword* (F2) dan negative *keyword* (F3), akurasi ringkasan sistem pada tahap pengujian mengalami penurunan.

B. Saran

Berikut ini saran dari penulis untuk pengembangan Tugas Akhir ini :

1. Mencari fitur ekstraksi baru kemudian menambahkannya dengan fitur ekstraksi yang ada pada Tugas Akhir ini.
2. Menggunakan Algoritma lainnya untuk pembobotan fitur ekstraksi teks pada peringkasan bahasa Indonesia.

REFERENCES

[1] Alguliev, R. (2009). *Evolutionary Algorithm for Extractive Text Summarization*. Baku, Azerbaijan: Scientific Research.

[2] Al-Hashemi, R. (June 2010). *Text Summarization Extraction System (TSES)*. International Arab Journal of e-Technology, (pp. Vol. 1, No. 4,).

[3] Aristoteles, Yeni, H., Ridha, A., & Adisantoso, J. (May 2012). *Text Feature Weighting for Summarization of Documents*. IJCSI International Journal of Computer Science Issues, (pp. Vol. 9, Issue 3, No 1).

[4] Berker, M., & Güngör, T. (2012). *Using Genetic Algorithms With Lexical Chains For Automatic Text Summarization*. Istanbul: Bogaziçi University.

[5] Dehkordi, P. K., Kumarci, F., & Khosravi, H. (2009). *Text Summarization Based on Genetic Programming*. International Journal of Computing and ICT Research.

[6] Fattah, M. A., & Ren, F. (2008). *Automatic Text Summarization*. World Academy of Science, Engineering and Technology 13 .

[7] Feldman, R., & Sanger, J. (2007). *Advanced Approaches in Analyzing Unstructured Data*. New York: Cambridge University Press.

[8] Hassel, M. (2004). *Evaluation of Automatic Text Summarization*. Stockholm, Sweden.

[9] HS, Widjono. (2007). *Bahasa Indonesia Mata Kuliah Pengembangan Kepribadian di PT*. Jakarta: Grasindo.

[10] Jezek, K., & Steinberger, J. (2008). *Automatic text summarization (The state of the art 2007 and new challenges)*. Vaclav Snasel : Znalosti. 1-12.

[11] Manning, C. D., Raghavan, P., & Schütze, H. (2009). *An Introduction to Information Retrieval*. Cambridge: Cambridge University Press.

[12] Marlina, M. (2012). *Sistem Peringkasan Dokumen Berita Bahasa Indonesia*. Bogor: Institut Pertanian Bogor.

[13] Purnanto, A. D. (2012). *Peringkasan Dokumen Berita Bahasa Indonesia Menggunakan Algoritma Genetika*. Malang: Universitas Brawijaya.

[14] Radev, D., Hovy, E., & McKeown, K. *Introduction to the special issue on text*. Computer linguist. 28(4).

[15] Ridha, A. (2002). *Pengindeksan Otomatis Densan Istilah Tunggal Untuk Dokumen Berbahasa Indonesia*. Bogor: Intsitut Pertanian Bogor.

[16] Rivaldi, Pinandhita. 2013. *Peringkas Dokumen Berbahasa Indonesia Berbasis Kata Benda Dengan Bm25*. Fakultas Matematika Dan Ilmu Pengetahuan Alam Institut Pertanian Bogor. Bogor.

[17] Silla, C. N., Pappa, G. L., Freitas, A. A., & Kaestner, C. A. (2004). *Automatic Text Summarization with Genetic*. Canterbury: University of Kent.

[18] Susanti, R. Y. (2009). *Analisis Perbandingan Peringkasan Teks Menggunakan Metode Lexical Chain Dan Metode TF-IDF dengan Ekstraksi Frase Utama*. Bandung: Insitut Teknologi Telkom.

[19] Suyanto. (2008). *Soft Computing Membangun Mesin ber-IQ Tinggi*. Bandung: Informatika Bandung.

[20] Suyanto. (2011). *Artificial Intelligence Searching Reasoning Planning Learning*. Bandung: Informatika Bandung.

[21] Wibowo, Septiandi. 2013. Peringkasan Teks Bahasa Indonesia Dengan Pemilihan Fitur C4.5 Dan Klasifikasi Naive Bayes. Fakultas Matematika Dan Ilmu Pengetahuan Alam Institut Pertanian Bogor. Bogor.

[22] Zaefarian, R. (2006). *A New Algorithm for Term Weighting in Text Summarization Process*. Tehran: Department of IE, Sharif University of Technology.

1. Zulkifli. Lahir di Ampana, Sulawesi Tengah pada tanggal 04 Pebruari 1992. Sebelumnya penulis telah mendapatkan gelar diploma teknik informatika di Institut Teknologi Telkom Bandung pada tahun 2012.