BAB I

PENDAHULUAN

1.1. Latar belakang

Membaca merupakan sebuah aktivitas setiap individu untuk mengambil intisari atau informasi yang ingin disampaikan oleh penulis kepada pembaca melalui tulisan [1]. Semakin panjang teks suatu sumber bacaan atau dokumen, proses perpindahan informasi dari penulis kepada pembaca akan semakin lama. Oleh karena itu, dengan memahami isi dokumen melalui ringkasan teks akan memerlukan waktu yang lebih singkat dibandingkan dengan membaca keseluruhan isi dokumen, sehingga ringkasan teks menjadi sangat penting.

Perkembangan teknologi komunikasi berdampak pada penggunaan internet untuk mempublikasikan informasi pada situs-situs di internet [2], salah satunya adalah artikel-artikel berita yang banyak diunggah di situs-situs surat kabar *online* seperti kompas.com, detik.com, republika.co.id dan okezone.com. Dengan semakin ringkasnya suatu artikel berita maka pembaca dapat dengan cepat mengetahui isi dari artikel tersebut. Namun, membuat ringkasan dokumen teks memerlukan waktu dan biaya peringkasan yang lebih besar bila dokumen yang diringkas berjumlah banyak serta isi dokumen yang panjang. Oleh karena itu, menurut Aristoteles ringkasan dokumen teks secara otomatis diperlukan untuk mengatasi masalah waktu baca dan biaya data [3].

Metode peringkasan teks terbagi menjadi dua macam yaitu peringkasan teks dengan metode abstraksi dan metode ekstraksi [4]. Metode abstraksi adalah mengambil intisari dari teks sumber dokumen kemudian dibuat ringkasan dengan menciptakan kalimat-kalimat baru yang merepresentasikan intisari teks asal dalam bentuk yang berbeda [5]. Sedangkan metode ekstraksi merupakan suatu teknik memilih kalimat atau frase dari teks asli yang memiliki skor tertinggi dan menggabungkannya kembali [6] menjadi ringkasan tanpa mengubah kalimat asal. Metode abstraksi membutuhkan pemrosesan lebih mendalam, khususnya pada *Natural Language Processing* (NLP), *Inference* dan *Natural Language Generation* yang sampai saat ini

masih belum matang [5]. Oleh karena itu, pada penelitian ini akan difokuskan pada peringkasan teks dengan metode ekstraksi.

Proses peringkasan teks dengan menggunakan metode ekstraksi terdiri dari tiga fase: analisis, transformasi dan sintesis [5]. Pada fase analisis dilakukan proses analisis terhadap inputan teks dan pemilihan fitur-fitur yang penting. Pada fase transformasi dilakukan proses pengubahan hasil pada fase analisis menjadi bentuk ringkasan yang representatif. Dan terakhir, pada fase sintesis dilakukan pemilahan terhadap ringkasan menjadi ringkasan yang sesuai dengan kebutuhan pengguna. Rasio peringkasan teks yang masih bisa diterima adalah 5-30% [5–7].

Genetic algorithm adalah algoritma pencarian yang dapat digunakan mencari kombinasi dari bobot-bobot fitur teks yang paling optimal [5], [8]. Seperti dalam penelitian Aristoteles yang meneliti tentang pembobotan fitur teks bahasa Indonesia dengan GA [3]. Dalam penelitiannya, mereka menyimpulkan bahwa GA dapat digunakan sebagai penentu bobot yang optimal pada fitur peringkasan teks bahasa Indonesia dan diperoleh akurasi 47,12% [3]. Pembobotan fitur teks ini bertujuan untuk mencari bobot-bobot yang optimal untuk masing-masing fitur teks. Hasil penelitian Fattah dan Ren juga menunjukkan bahwa pembobotan fitur teks yang dihasilkan dengan menggunakan teknik GA lebih akurat dibandingkan teknik *Mathematical Regression* (MR) [5]. Kiani-B dan Akbarzadeh memadukan logika fuzzy, GA dan *Genetic Programming* (GP) untuk peringkasan teks [9]. Suanmali memadukan logika fuzzy dengan GA untuk peringkasan teks dan keduanya diperoleh hasil yang signifikan [6]. Dalam implementasinya GA digunakan untuk pembobotan fitur teks sedangkan logika fuzzy digunakan untuk memilih skor tertinggi dari kalimat berdasarkan fitur-fitur teks [6], [9]. Oleh karena itu, dalam penelitian ini GA akan digunakan untuk melakukan pembobotan fitur teks.

Fitur-fitur teks untuk peringkasan teks dalam penelitian suanmali terdiri dari 8 fitur teks [6] yaitu: fitur judul, panjang kalimat, bobot kata, posisi kalimat, kemiripan antar kalimat, *proper noun*, kata tematik, dan data numerik. Menurut Suanmali metode ekstraksi untuk peringkasan teks dengan pembobotan GA hanya mampu mendapatkan konten utama melalui fitur-fitur teks, akan tetapi tidak dapat menangkap hubungan semantik antar kalimat [6]. Untuk mengatasi masalah tersebut,

Suanmali menambahkan *Semantic Role Labeling* (SRL) yang bertujuan untuk menangkap hubungan yang terjadi antar kalimat [6]. Dengan integrasi tersebut, hanya terjadi peningkatan akurasi sebesar 0,16% dari 49,80% menjadi 49,96%.

Salah satu cara lain dalam menangkap hubungan semantik antar kalimat adalah dengan *Vector Space Model* (VSM). Dimana VSM merupakan teknologi semantik baru [10] yang dapat digunakan untuk menentukan kemiripan antar term dan kalimat dalam suatu dokumen. VSM mampu mengekstrak *knowledge* secara otomatis dari korpus yang diberikan sehingga membutuhkan sumber daya yang lebih sedikit dibandingkan dengan metode semantik yang lain [10], [11]. Ide dasar dari VSM adalah tiap-tiap term dan kalimat dalam dokumen dinyatakan dalam titik-titik di dalam ruang vektor. Titik-titik yang berdekatan secara semantik dianggap dekat sebaliknya yang berjauhan dianggap memiliki hubungan semantik yang jauh [10]. Byung-Wong dan Ingyu Lee menambahkan dalam penelitiannya bahwa dengan mamadukan VSM dengan metode Monge-Elkan dapat menambah akurasi dalam pengenalan *abbreviation* dengan kombinasi tertentu [12].

Pada penelitian ini, akan dilakukan integrasi antara *genetic algorithm* (GA) dan *Vector Space Model* (VSM), dimana GA akan digunakan untuk menghasilkan konten utama dari dokumen teks. Sedangkan VSM akan digunakan untuk menangkap hubungan semantik yang terjadi antar kalimat. Sehingga diharapkan terjadi peningkatan akurasi peringkasan teks bahasa Indonesia.

1.2. Rumusan masalah

a. Umum

   Proses memahami intisari suatu sumber bacaan memerlukan waktu lebih lama dibandingkan dengan memahami melalui ringkasan dari sumber bacaan sehingga tidak efisien.

b. Khusus

   Peringkasan teks dengan integrasi antara *genetic algorithm* (GA) dan *Semantic Role Labeling* (SRL) dimana GA digunakan untuk mendapatkan konten utama dan SRL digunakan untuk menangkap hubungan semantik hanya memiliki akurasi sebesar 49,96%.

1.3. Tujuan penelitian

Berdasarkan latar belakang di atas, tujuan penelitian ini adalah sebagai berikut.

a. Umum

   Terdapat metode peringkasan teks yang lebih akurat sehingga proses memahami intisari suatu sumber bacaan menjadi lebih cepat dan efisien.

b. Khusus

   Terdapat metode peringkasan teks dalam bahasa Indonesia yang mampu mendapatkan konten utama melalui pembobotan *Genetic Algorithm* (GA) dan mampu menangkap hubungan semantik antar kalimat melalui *Vector Space Model* (VSM) sedemikian hingga akurasi hasil peringkasan teks lebih tinggi dari 49,96%.

1.4. Manfaat penelitian

a. Umum

   Mempermudah pembaca dalam memahami intisari dari berita media online berbahasa Indonesia melalui ringkasan teks.

b. Khusus

   Menambah metode baru yang lebih akurat dalam melakukan peringkasan teks otomatis bahasa Indonesia.