1. Pendahuluan

1.1 Latar Belakang

Media berita memiliki peran penting dalam penyebaran informasi dimasyarakat. Pada era dengan kemajuan teknologi ini, suatu informasi baru dapat dengan cepat disebar dengan memanfaatkan media elektronik dan internet. Sebagai contoh, dengan memanfaatkan mesin pencari di internet, hanya dengan mengetikkan suatu topik sebagai query maka puluhan hingga ratusan dokumen yang berhubungan dengan topik tersebut dapat langsung diakses. Namun, kemudahan pengaksesan dokumen berita ini justru memunculkan kesulitan bagi user untuk mendapatkan informasi yang benar-benar diinginkan dan penting dari setiap dokumen, sebab dengan banyaknya dokumen yang ada maka secara otomatis semakin banyak pula waktu dan usaha yang dibutuhkan untuk membacanya dan menyaring informasi yang diinginkan. Oleh sebab itu, dibutuhkan sebuah solusi berupa peringkasan beberapa dokumen (multi-dokumen) sehingga user bisa dengan cepat mendapatkan informasi-informasi utama yang ada dari semua dokumen tersebut tanpa harus membaca satu per satu, sehingga pencarian informasi menjadi lebif efektif dan efisien.

Kunci dari sistem peringkasan ini adalah ekstraksi kalimat, yaitu mengambil kalimat-kalimat yang dapat merepresentasikan isi dokumen dan mempertahankan bentuk serta konten kalimat tersebut. Metode peringkasan ekstraktif bisa diklasifikasikan kedalam dua kelompok, yaitu *supervised methods* yang mengandalkan pasangan dokumen dan ringkasan yang tersedia, dan untuk peringkasannya adalah dengan mengklasifikasikan kalimat kandidat ke dalam kalimat penting dan tidak penting berdasarkan fitur yang digunakan, sedangkan *unsupervised methods* bertujuan untuk mengambil kalimat berdasarkan pengelompokan semantik yang diekstrak dari dokumen [1].

Ada beberapa pendekatan *machine learning* yang diterapkan untuk membangun sistrem peringkasan ektraktif. Salah satu metode yang terkemuka adalah *Support Vector Regression (SVR). SVR* adalah model regresi dari *Support Vector Machine (SVM)* yang berdasarkan *Structural Risk Minimization(SRM)* yaitu prinsip induksi untuk pemilihan model yang digunakan untuk belajar dari set data training yang terbatas [4]. Optimasi SVR tidak bergantung pada dimensi dan vektor ruang input, sehingga untuk kasus *nonlinier*, SVR dapat memetakan input ke suatu ruang dimensi tinggi.

Teknik yang biasanya digunakan untuk meringkas dokumen adalah dengan menjadikan peringkat dokumen sebagai sebuah masalah klasifikasi biner yang model klasifikasinya belajar dari set kalimat “penting” dan “tidak penting” untuk membangun model klasifikasi yang akan digunakan untuk mengidentifikasi kalimat kunci [12]. Teknik lainnya yang juga bisa digunakan

adalah model perankingan. Namun, kebanyakan model klasifikasi dan model prankingan mengubah biner atau pemesanan informasi ke nilai kontinu yang digunakan dalam klasifikasi dan perankingan [12]. Pendekatan alternatif yang bisa digunakan adalah dengan model regresi yang mempelajari fungsi kontinu yang langsung memperkirakan kepentingan suatu kalimat, yang lebih baik dikarakteristik sebagai kontinu daripada diskrit. Pada Tugas Akhir ini, model regresi yang diimplementasikan adalah *Support Vector Regression* (SVR). SVR digunakan untuk mengkombinasikan fitur secara otomatis dan melakukan pembobotan kalimat.

Pengerjaan Tugas Akhir ini dibagi menjadi tiga modul, yaitu *Preprocessing* yang terdiri dari tahap segmentasi seluruh dokumen kedalam kalimat-kalimat dan tahap *text preprocessing* untuk setiap kalimat. Modul kedua adalah ekstraksi kalimat yang terdiri dari tiga langkah yaitu kalkulasi fitur, penilaian *(scoring*) kalimat, dan pemilihan kalimat. Modul terakhir adalah pembentukan ringkasan yang dilakukan melalui tahap penghilangan duplikasi informasi dan pengurutan kalimat. Setelah proses pembangunan ringkasan berhasil, kemudian dilakukan evaluasi dengan membandingkan hasil ringkasan dari sistem dengan ringkasan yang dibuat oleh manusia, yang dianggap sebagai ringkasan ideal. Pengujian ini dilakukan dengan metode ROUGE-2.

1.2 Perumusan Masalah

Berdasarkan latar belakang di atas, masalah-masalah yang diselesaikan antara lain:

1. Bagaimana menerapkan metode *Support Vector Regression* untuk melakukan peringkasan multi-dokumen?
2. Bagaimana pengaruh parameter toleransi dan parameter γ yang ada pada metode *Support Vector Regression* dengan kernel RBF terhadap hasil akurasi peringkasan kalimat?
3. Fitur apa saja yang paling berpengaruh untuk mendapatkan hasil peringkasan yang paling akurat?

1.3 Tujuan

Tujuan dari tugas akhir ini adalah:

1. Menerapkan metode Support Vector Regression untuk melakukan peringkasan dokumen.
2. Menganalisis parameter yang berpengaruh pada metode Support Vector Regression, yaitu Tolerance dan γ, serta kombinasi kedua parameter tersebut pada proses peringkasan dokumen.
3. Menganalisis fitur yang berpengaruh untuk mendapatkan hasil peringkasan yang paling akurat.

1.4 Batasan Masalah

Adapun batasan masalah di Tugas Akhir ini sebagai berikut:

1. Peringkasan dilakukan pada dokumen berbahasa Indonesia yang menggunakan bahasa baku.
2. Dataset yang digunakan adalah berita dengan topik utama "Pertandingan Sepak Bola" dan berasal dari web bola.kompas.com, bola.viva.co.id, sportsatu.com, bola.okezone.com, dan beritasatu.com.
3. Ringkasan acuan dan label kalimat pada data training dibuat dan ditentukan secara manual oleh satu orang expert dari bahasa Indonesia.
4. Daftar kata digunakan pada kamus WordNet hanya kata tunggal.
5. Panjang ringkasan yang dihasilkan tidak ditentukan.
6. Sistem tidak menangani duplikasi informasi yang terjadi pada hasil ringkasan.
7. Sistem tidak menangani dokumen yang mengandung kata yang disingkat dan tidak lengkap.

1.5 Metodologi Penyelesaian

Penyelesaian Tugas Akhir ini dilakukan dengan menggunakan metodologi sebagai berikut:

1. Studi Literatur
   a. Pencarian referensi yang berhubungan dengan peringkasan multi-dokumen, metode Support Vector Regression (SVR), dan ROUGE.
   b. Pendalaman materi yang berhubungan dengan Tugas Akhir.
2. Pengumpulan Requirement
   Pada tahap ini, dilakukan pencarian dan pengumpulan data-data yang dibutuhkan untuk mendukung penyelesaian masalah. Data yang dikumpulkan adalah dokumen berita dengan topik utama Pertandingan Sepak Bola. Data diambil dari website dan secara manual di-*copy-paste* kedalam file teks.
3. Perancangan Sistem
   Merancang sistem peringkasan dokumen, dimulai dari input, proses peringkasan, hingga sistem menghasilkan output berupa ringkasan dokumen.
4. Implementasi
   Pada tahap ini dilakukan implementasi berdasarkan perancangan yang telah dibuat.
5. Pengujian

Melakukan pengujian sistem dengan ROUGE evaluation untuk membandingkan ringkasan yang dihasilkan sistem dengan dengan ringkasan manusia.

6. Analisis Hasil Pengujian
   Melakukan analisa terhadap akurasi dari output sistem, serta pengaruh parameter dan fitur dalam menghasilkan rinngkasan.
7. Penyusunan Laporan
   a. Pengambilan kesimpulan terhadap keseluruhan proses dan hasil yang didapat.
   b. Pengumpulan dokumentasi.
   c. Penyusunan laporan Tugas Akhir.