

1. Pendahuluan

1.1 Latar Belakang Masalah

Melihat era sekarang dimana orang cenderung tidak lagi membaca berita dari majalah atau koran konvensional, sebagian besar orang mulai beralih untuk membaca berita dari *gadget* secara *online*, sehingga membuat penyedia jasa berita *online* semakin banyak dijumpai, ada yang berbasis *website* atau aplikasi yang semakin mudah untuk diakses dimanapun dan kapanpun. Berita yang disajikan beragam dan tentu saja berita tersebut berasal dari berbagai sumber dan sudut pandang penulis yang berbeda juga. Hal tersebut akan membuat pembaca mencari berita dari berbagai sumber untuk melengkapi informasi yang bisa didapatkan. Terkadang ada berita yang dijelaskan panjang lebar dan terlihat berbeda dari sumber berita sebelumnya namun sebenarnya memiliki informasi yang sama. Sebagai contoh, dengan memanfaatkan mesin pencari atau *search engine* 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 pengguna untuk mendapatkan informasi yang benar-benar diinginkan dan penting dari setiap dokumen, sebab dengan banyaknya dokumen yang ada maka semakin banyak pula waktu dan usaha yang dibutuhkan untuk membacanya dan menyaring informasi yang diinginkan. Oleh sebab itu, dibutuhkan sebuah solusi berupa peringkasan multi-dokumen sehingga pengguna bisa dengan cepat mendapatkan informasi-informasi penting yang ada dari semua dokumen tersebut tanpa harus membaca satu per satu sehingga pencarian informasi menjadi efektif.

Peringkasan dokumen dapat dilakukan dengan melihat nilai informasi suatu kalimat dari dokumen tersebut berdasarkan fitur yang diterapkan pada kalimat yang ada. Fitur tersebut diantaranya adalah fitur linguistik dan statistikal seperti posisi kalimat. Karena pada saat melakukan peringkasan dokumen sering kali ditentukan dari posisi kalimat itu, yang selanjutnya disebut dengan *basic feature*. Selain itu ada fitur yang melihat kalimat berdasarkan keterkaitan kalimat itu dengan kalimat lain dan pengaruh kalimat tersebut dalam suatu dokumen, yang selanjutnya disebut *complex feature*. Dengan mengolah informasi fitur maka akan didapatkan nilai kalimat dari setiap kalimat yang ada dan dapat dibuat model untuk dapat diterapkan di dokumen lain.

Kunci dari sistem peringkasan ini adalah ekstraksi kalimat, yaitu mengambil kalimat yang merepresentasikan isi dokumen dengan mempertahankan bentuk dan tidak mengubah konten dari dokumen. 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 [3].

Conditional Random Fields (CRF) merupakan salah satu model probabilistik untuk mengatasi segmentasi dan pemberian label pada data sekuens [6]. CRF menerapkan peringkasan dengan kategori *extract-summary* dimana hasil peringkasannya menggunakan kalimat-kalimat yang telah diekstraksi dari sebuah dokumen, sehingga mengurangi kemungkinan hilangnya informasi. Hal tersebut membuat CRF dapat diterapkan pada peringkasan berita.[9] Dengan menggunakan CRF maka dapat diterapkan fitur-fitur yang dapat mendukung peringkasan multi-dokumen, contohnya adalah *Similarity to Neighboring*

Sentences, position, dan thematic word. Setelah mendapatkan *probability* dari setiap kalimat kemudian ditentukan *threshold* untuk mengklasifikasikan kalimat dari seluruh dokumen kedalam dua kelas yaitu *summary* dan *non-summary*.

Dengan adanya permasalahan diatas dan dengan solusi yang ada maka diperlukan adanya penelitian dengan judul “Peringkasan Multi-Dokumen Berita Berbahasa Indonesia Menggunakan *Conditional Random Fields*”.

1.2 Perumusan Masalah

Berdasarkan landasan teori yang ada maka dapat dirumuskan masalah yang diselesaikan untuk penelitian ini antara lain :

- a. Bagaimana menerapkan metode *Conditional Random Fields* untuk membangun sistem peringkasan multi-dokumen?
- b. Bagaimana performansi metode *Conditional Random Fields* terhadap nilai *threshold*, jumlah data latih dan fitur yang digunakan?

1.3 Tujuan

Tujuan yang akan dicapai pada penelitian ini adalah :

- a. Membangun sistem peringkasan multi-dokumen dengan menerapkan metode *Conditional Random Fields* agar pencarian informasi menjadi lebih efektif dan efisien.
- b. Menganalisis pengaruh *threshold*, jumlah data latih dan fitur yang digunakan agar sistem menghasilkan ringkasan dengan akurasi yang baik sehingga pencarian informasi menjadi lebih efektif dan efisien.

1.4 Batasan Masalah

Adapun batasan masalah pada penelitian ini adalah :

- a. Peringkasan dilakukan pada dokumen yang menggunakan kata baku dalam Bahasa Indonesia dan tidak menggunakan kata singkatan.
- b. Data set diambil secara manual dari web penyedia berita antara lain www.bola.kompas.com, www.bola.okezone.com, www.bola.viva.co.id dan www.tempo.com/bola.
- c. Panjang ringkasan yang dihasilkan tidak ditentukan.

1.5 Metodologi Penelitian

Untuk mendapatkan hasil akhir yang sesuai dengan tujuan, disusunlah metodologi penelitian sebagai berikut :

- a. Studi Literatur
Dilakukan untuk memahami peringkasan multi-dokumen, metode *Conditional Random Fields* (CRF) dan lain-lain yang diperlukan dalam penyusunan penelitian ini dapat berupa buku, jurnal, paper ataupun artikel.
- b. Pengumpulan data
Dilakukan pencarian dan pengumpulan data yang dibutuhkan. Data yang digunakan adalah dokumen berita.
- c. Analisa perancangan

Merancang sistem peringkasan multi-dokumen dimulai dari *input*, proses peringkasan, hingga menghasilkan keluaran berupa dokumen yang sudah teringkaskan.

d. Implementasi

Implementasi dari rancangan yang telah dibuat.

e. Pengujian dan analisis

Melakukan pengujian dengan menghitung nilai *Precision*, *Recall* dan *F-Measure* untuk membandingkan ringkasan dari sistem dengan ringkasan ahli. Dan melakukan analisa terhadap akurasi *output* sistem.

f. Kesimpulan

Membuat kesimpulan hasil penelitian berdasarkan tujuan yang telah ditentukan sebelumnya.

g. Penyusunan laporan

Dokumentasi selama pengerjaan penelitian berupa laporan.