1. Pendahuluan

1.1. Latar Belakang

Perkembangan yang sangat pesat di dunia teknologi informasi dan komunikasi menyebabkan bertaburannya dokumen elektronik yang dapat diakses secara *online*. Namun, kekayaan data tersebut justru seringkali membuat user kesulitan untuk memilah informasi yang berguna. Hal ini menyebabkan setiap kali mencari informasi, user justru akan banyak membaca dokumen yang menarik namun tidak relevan dengan pencariannya. Oleh sebab itu, saat ini user membutuhkan sistem peringkasan teks yang handal, yaitu sistem yang secara efektif mampu meringkas informasi yang ditemukan pada beberapa dokumen menjadi lebih pendek, namun tetap tidak kehilangan makna dari dokumen tersebut, atau yang biasa disebut ringkasan. Sehingga user tidak perlu menghabiskan waktu banyak untuk mencari informasi yang mereka butuhkan dalam sebuah artikel yang panjang, cukup dengan membaca ringkasan dari artikel yang mereka temukan.

Text summarization adalah bidang keilmuan yang khusus mengkaji tentang permasalahan di atas, dimana text summarization ini mampu memangkas kumpulan informasi menjadi informasi yang dibutuhkan dengan menseleksi informasi yang paling penting [13]. Berdasarkan jumlah dokumen yang diringkas, teknik meringkas teks dibagi menjadi dua macam, yaitu *single-document summarization* dan *multi-document summarization* [13]. *Single-document summarization* hanya akan meringkas satu buah dokumen menjadi sebuah paragraf yang lebih singkat, sedangkan *multi-document summarization* akan meringkas sekumpulan dokumen, ringkasan informasi didapatkan dari klaster beberapa dokumen sumber.

Centroid-Based Summarization, adalah sebuah metode yang menghasilkan kalimat ringkasan dengan menggunakan centroid di setiap klaster untuk memilih kalimat yang paling mendekati dan relevan dengan topik klaster. Menurut penelitian yang dilakukan oleh Dragomir R.Radev, dari berbagai evaluasi yang digunakan untuk pengujian hasil ringkasan yang dihasilkan dari metode ini, seperti CBSU (*Cluster-Based Relative Utility)* dan CSIS *(Cross Sentence Information Subsumption)*, ditemukan bahwa metode ini menghasilkan ringkasan yang secara kualitas sama dengan ringkasan yang dihasilkan oleh manusia [9]. Namun, sebuah ringkasan dari *multi-document* tidak dapat langsung dibuat tanpa terlebih dahulu mengelompokkannya dengan dokumen lain yang mempunyai kesamaan topik. Salah satu algoritma yang dapat digunakan untuk mengelompokkan dokumen adalah *Single Pass Clustering.* S. Rieber dalam bukunya, *The Single Pass Clustering Method*, menyebutkan bahwa metode ini sangat menarik karena mampu memangkas waktu yang digunakan ketika melakukan klastering. Jika sebuah metode klastering membutuhkan waktu 1 menit untuk menklasterkan 100 dokumen, maka untuk menklasterkan 1.000.000 dokumen, waktu yang dibutuhkan untuk algoritma dengan kompleksitas N^2 akan memakan waktu selama 170 tahun, algoritma N log N menghabiskan waktu 21 hari, sementara metode *Single-Pass* hanya akan

menghabiskan waktu 7 hari [14]. Selain itu, algoritma ini dikenal sebagai algoritma yang cukup sederhana karena hanya melalui data sebanyak satu kali [14]. Karena keunggulan tersebut, algoritma ini sering digunakan untuk kebutuhan pengelompokan dokumen, proses katalog, dll, utamanya yang menggunakan data yang besar.

Dari permasalahan di atas, penelitian ini menerapkan metode CBS untuk membuat ringkasan dari sekumpulan dokumen berita berbahasa Indonesia, yang data inputnya terlebih dahulu diklasterkan menggunakan metode *Single Pass clustering.* Dan kemudian dievaluasi dengan membandingkan hasil ringkasan dari CBS dengan hasil ringkasan oleh manusia, yaitu dengan menggunakan metode ROUGE-2 dan *Relative Utility*.

1.2. Perumusan masalah

Berdasarkan latar belakang yang telah dipaparkan sebelumnya, maka dapat dirumuskan beberapa masalah diantaranya:

1. Bagaimana menghasilkan ringkasan dari beberapa dokumen input yang belum didefinisikan topiknya menggunakan metode *Centroid-Based Summarization* (dengan bantuan algoritma *Single Pass Clustering* untuk proses pengelompokan dokumen)?
2. Bagaimana performansi hasil ringkasan yang dibandingkan dengan hasil ringkasan manusia (ahli) menggunakan metode ROUGE-2 dan *Relative Utility?*
3. Bagaimana performansi hasil peringkasan sistem yang dibandingkan dengan hasil ringkasan yang melalui proses pelabelan?
4. Bagaimana pengaruh tahap *preprocessing* terhadap hasil akurasi ringkasan?

1.3. Tujuan

Tujuan yang ingin dicapai pada tugas akhir ini adalah :

1. Menghasilkan ringkasan dari beberapa dokumen input yang belum didefinisikan topiknya menggunakan *Centroid-Based Summarization* (dengan bantuan algoritma *Single Pass Clustering* untuk proses pengelompokan dokumen).
2. Menganalisis performansi hasil ringkasan dengan membandingkan dengan hasil ringkasan manusia (ahli) menggunakan metode ROUGE-2 dan *Relative Utility.*
3. Menganalisis performansi hasil peringkasan sistem yang dibandingkan dengan hasil ringkasan yang melalui proses pelabelan.
4. Menganalisis pengaruh tahap *preprocessing* terhadap hasil akurasi ringkasan.

1.4. Batasan Masalah

Berdasarkan perumusan masalah di atas, pada penelitian ini ditentukan batasan masalah sebagai berikut:

1. Semua dokumen dalam dataset yang digunakan berbahasa Indonesia dan telah memiliki ejaan yang benar sesuai degan EYD.
2. Dataset yang digunakan adalah sekumpulan artikel berita yang diberitakan dari rentang tahun 2012-2013 dan dipublikasikan di situs berita online Indonesia, seperti www.kompas.com, www.detik.com, www.tempo.co, www.viva.com,

www.okezone.com, www.metrotvnews.com, dll, yang sebelumnya telah diolah menjadi dokumen berformat .txt.

3. Metode peringkasan yang digunakan adalah *extractive summarization* (peringkasan ekstraksi).
4. Fitur peringkasan yang digunakan adalah *centroid value, positional value,* dan *SimWithFirst.*
5. Evaluasi terhadap hasil ringkasan dilakukan dengan dua parameter, yaitu ROUGE-2 dan *Relative Utility.*
6. Nilai threshold yang digunakan untuk pengelompokan dokumen pada studi kasus ini adalah 0,7.

1.5. Metodologi Penyelesaian Masalah

Untuk mencapai tujuan yang ingin dicapai dan menyelesaikan rumusan masalah yang telah dipaparkan sebelumnya, pada penelitian ini penulis menerapkan pendekatan sistematis sebagai berikut:

1. Studi Literatur
   Untuk lebih memahami tentang peringkasan multi-document, maka dilakukan studi literatur yang sesuai dengan permasalahan, meliputi konsep peringkasan teks, peringkasan *multi-document*, *Single Pass Clustering*, *Centroid-Based Summarization*, dan konsep-konsep lain yang berkaitan dengan pengerjaan tugas akhir.
2. Pengumpulan Data
   Kumpulan data yang digunakan dalam tugas akhir ini adalah dokumen dalam format .txt yang diambil dari situs berita *online* Indonesia, seperti diantaranya www.kompas.com, www.detik.com, www.okezone.com, www.tempo.com, dan www.metrotvnews.com, data yang digunakan maksimal berumur 2 tahun semenjak dipublikasikan, yaitu berita pada tahun 2012 – 2013.
3. Perancangan Sistem
   Membuat desain sistem peringkasan *multi-document* dengan metode *extractive* yang akan dibangun menggunakan metode *Centroid-Based Summarization*.
4. Implementasi
   Mengimplementasikan sistem sesuai dengan perancangan di tahap sebelumnya menggunakan bahasa pemrograman PHP.
5. Testing
   Melakukan pengujian terhadap sistem yang dibangun menurut skenario yang telah dirancang sebelumnya.
6. Analisis Hasil Pengujian
   Melakukan pengujian terhadap sistem yang telah dibangun serta melakukan analisis terhadap hasil terhadap hasil ringkasan yang diperoleh.
7. Pembuatan laporan
   Pada tahap ini, dilakukan penyusunan laporan akhir serta dokumentasi dari penelitian yang dilakukan.