BAB I

PENDAHULUAN

1.1. Latar Belakang Masalah

Informasi merupakan salah satu bagian terpenting dalam sejarah kehidupan manusia. Informasi tentang obat-obatan, makanan yang aman untuk dikonsumsi, kisah yang mengandung makna, hingga informasi tentang langkah-langkah melakukan sesuatu menjadi bagian yang tidak bisa dilepaskan dari pengetahuan manusia dan perkembangan peradabannya. Dalam ranah ilmu pengetahuan, informasi yang berharga dan dapat dipertanggungjawabkan kebenarannya ditulis dalam bentuk jurnal ilmiah, artikel atau dokumen. Informasi-informasi ini pun perlu disampaikan secara masif untuk tujuan penelitian lebih lanjut.

Semakin terbiasanya masyarakat Indonesia menggunakan teknologi komputer maupun *smartphone* dan banyaknya informasi yang beredar di *Internet*, berhasil menyediakan kemudahan untuk mendapatkan pengetahuan yang mereka inginkan dengan sentuhan jari. Hal ini juga berlaku dalam penyaluran dokumen ilmiah untuk penelitian. Tercatat pada situs *sinta.kemdikbud.go.id* ada sebanyak 392.293 publikasi dokumen oleh peneliti Indonesia yang ter-indeks di *scholar.google.com* pada tahun 2020 (Anonim. 2017. SINTA – Science and Technology Index). Namun, dari begitu banyak ragam dokumen ilmiah, tidak semuanya berisi informasi yang dibutuhkan oleh pengguna *Internet.*

Untuk menyelesaikan permasalahan pencarian informasi dari dokumen dalam jumlah besar, penelitian di bidang *Information Retrieval* (IR) atau "Sistem Temu Kembali Informasi" telah dikembangkan sejak tahun 1950-an. Menurut Manning (2009), diacu dalam Setiyadi (2018) "Sistem temu kembali informasi adalah sistem yang dapat mencari dokumen berdasarkan kata kunci, namun sistem temu kembali informasi akan menemui kendala jika harus mencari informasi spesifik dari dokumen yang jumlahnya banyak dengan jumlah teks penyusun dokumen yang jumlahnya tidak sedikit." Maka dari itu, perlu dilakukan peringkasan teks terlebih dahulu untuk dokumen-dokumen yang tercatat pada

indeks mesin pencarian untuk kemudian dihitung menggunakan algoritma tertentu sehingga mendapatkan urutan indeks berdasarkan nilai kesesuaian informasi.

Peringkasan teks itu sendiri adalah proses menyederhanakan sejumlah dokumen dengan menghilangkan informasi yang berulang dan/atau tidak diperlukan (Christopher dan Yusliani, 2016). Pada ringkasan teks terdapat dua jenis *input* atau sumber untuk dilakukan peringkasan, yakni dokumen tunggal dan multi-dokumen (Allahyari, dkk., 2017). Peringkasan dokumen tunggal meringkas dokumen asli hingga kurang dari 50% atau lebih sedikit dari teks asalnya, sedangkan peringkasan multi-dokumen perlu dilakukan proses untuk suatu kumpulan dokumen yang nantinya akan disajikan dalam bentuk ringkasan.

Umumnya, pendekatan untuk otomatisasi peringkasan teks terbagi menjadi dua: pendekatan ekstraktif dan pendekatan abstraktif. Peringkasan ekstraktif berfokus kepada mengidentifikasi kalimat-kalimat penting pada dokumen aslinya. Sedangkan peringkasan abstraktif mencoba untuk menyusun kalimat-kalimat penting ke dalam bentuk atau susunan kata baru yang berbeda daridokumen aslinya namun tetap bermakna sama. Meskipun peringkasan dokumen secara tradisional oleh manusia bukanlah merupakan peringkasan secara ekstraktif, penelitian tentang peringkasan teks lebih berfokus kepada pendekatan ekstraktif dikarenakan hasil yang diperoleh dari pendekatan ekstraktif lebih baik daripada pendekatan abstraktif (Erkan dan Radev, 2004:457-459, diacu dalam Allahyari, dkk., 2017).

Sebelum memperurutkan hasil ringkasan, diperlukan sebuah proses pembobotan kata. Hal ini dilakukan karena dalam kumpulan dokumen bisa saja sebagian informasi yang tercantum didalamnya dapat masuk ke dalam banyak kategori. Sedangkan kita harus menentukan dengan tepat kategori dokumen yang diringkas sesuai dengan informasi yang ada didalamnya. Pendekatan *Term Frequency - Inverse Document Frequency* (TF-IDF) sudah umum digunakan untuk menentukan bobot kata pada dokumen berdasarkan keunikannya (Zhang, dkk., 2005 6A(1):49-55). Unik disini adalah hasil dari proses TF-IDF dilihat dari relevansi terhadap kata-kata yang terdapat di dokumen, dokumen itu sendiri serta kategori tertentu.

Dikarenakan teknologi komputer saat ini belum mengerti *semantic* kalimat seperti manusia. Dibutuhkan sebuah model algoritma yang berbasis nilai *vector*. Model *Markov Random Walk* telah berhasil digunakan untuk peringkasan multi-dokumen dengan bantuan *clustering* K-Means. Model tersebut mulanya dibangun dengan sebuah grafik terarah maupun tak terarah kemudian diterapkan pada algoritma *ranking* berbasis grafik untuk menghitung nilai kalimat-kalimat yang telah diekstraksi (Wan dan Yang, 2008:299).

Berdasarkan penjabaran diatas, penelitian ini hendak melakukan sebuah analisis terhadap model *Markov Random Walk* dengan memanfaatkan pendekatan *cluster-based* dan peringkasan ekstraktif terhadap multi-dokumen berbahasa Indonesia serta variasi proses TF-IDF. Data set yang akan digunakan adalah 50 jurnal ilmiah berbahasa Indonesia dengan topik yang sama.

1.2. Identifikasi Masalah

Berdasarkan penjelasan latar belakang, maka pokok permasalahan yang akan diteliti antara lain:

1. Peringkasan teks diperlukan untuk mendukung pencarian informasi secara efektif dan efisien
2. Kebutuhan peringkasan dokumen otomatis pada dokumen berbahasa Indonesia
3. Analisa terhadap variasi proses peringkasan diperlukan untuk mengembangkan studi di bidang peringkasan teks
4. Analisa Model *Cluster Conditional Markov Random Walk* dengan variasi TF-IDF diperlukan untuk mengetahui langkah yang optimal untuk ringkasan dokumen.

1.3. Pembatasan Masalah

Untuk melakukan pembatasan terhadap penelitian yang akan dilakukan, dibuat batasan masalah sebagai berikut:

1. Peringkasan teks dilakukan pada multi-dokumen berjumlah 50 jurnal ilmiah berbahasa Indonesia
2. Kumpulan dokumen memiliki topik yang sama yaitu *Natural Language Processing*
3. Peringkasan teks menggunakan pendekatan ekstraktif

4. Peringkasan teks memakai model *Markov Random Walk* berbasis *cluster* dengan variasi proses TF-IDF

1.4. Rumusan Masalah

Rumusan masalah berdasarkan identifikasi sebelumnya pada penelitian ini, bagaimana hasil peringkasan teks multi-dokumen berbasis *cluster* denganmodel *Markov Random Walk* dan metode ekstraktif jika proses TF-IDF divariasikan?

1.5. Tujuan Penelitian

Penelitian ini ditujukan untuk mengetahui performa model Markov Random Walk berbasis *cluster* dan mencari variasi TF-IDF yang menghasilkan peringkasan teks dengan persentase hilang informasi terkecil juga nilai topik yang tinggi dari kumpulan jurnal ilmiah berbahasa Indonesia.

1.6. Manfaat Penelitian

Manfaat penelitian ini dapat dijadikan referensi pada kajian penelitian ilmu pengetahuan khususnya dibidang pengembangan peringkasan dokumen dalam menentukan variasi TF-IDF yang maksimal dengan model *cluster based Markov Random Walk* serta mempermudah pencarian dokumen dengan banyak informasi ke dalam topik yang diringkas dengan baik.