BAB II

TINJAUAN PUSTAKA

Penelitian awal dalam bidang *automatic text summarization* dimulai dengan pembuatan metode *term frequency* oleh Luhn pada tahun 1958. Metode ini berasumsi bahwa frekuensi kata di dalam dokumen menunjukan indikasi pada tingkat kepentingannya. Dari penelitian tersebut, terdapat beberapa poin penting yang masih digunakan hingga sekarang yaitu proses *stemming* kata menjadi bentuk kata dasarnya sebagai langkah pertama, dan kemudian dilakukan penghapusan *stop words* (Luhn, 1958). Pada tahun yang sama, Baxendale melakukan penelitian dan menambahkan fitur posisi kalimat dalam dokumen sebagai salah satu faktor penentu. Ia meneliti 200 paragraf kemudian menemukan sebanyak 85% kalimat yang merupakan inti paragraf terdapat pada kalimat pertama, sedangkan 7% pada kalimat terakhir. Disimpulkan bahwa topik inti dari suatu paragraf terdapat pada salah satu dari kalimat pertama atau kalimat terakhir paragraf (Jezek & Steinberger, 2008).

Penelitian berikutnya yaitu pada tahun 1969 oleh Edmunson, dengan menjumlahkan bobot term frequency, posisi kalimat, kalimat judul dan *key phrases*. Contoh *key phrases* tersebut yaitu "*important*", "*results are*", "paper introduces", dan lain-lain. Secara umum, kalimat dimulai dengan *key phrases* dapat menjadi indikator yang baik pada konten yang signifikan dari dokumen teks (Silvia, et al., 2014).

Penelitian pada algoritma *automatic text summarization* terus berkembang dan dapat diklasifikasikan menjadi beberapa metode. Salah satu metode peringkasan secara unsupervised yang pernah dikembangkan adalah metode peringkasan menggunakan LSA (Latent Semantik Analysis) (Gupta & Gambhir, 2016). Metode ini menggunakan metode SVD (Singular Value Decomposition) untuk proses dekomposisi matriks. Salah satu matriks yang dihasilkan adalah matriks yang merepresentasikan topik dalam suatu kalimat. Matriks tersebut cenderung bersifat non sparse dan berisi bilangan negatif dan nonnegatif pada elemen-elemennya. Karena sifat matriks yang non sparse (padat) itulah, kecenderungan suatu kalimat mengandung suatu topik tertentu lebih sulit dikenali.

Gupta dan Lehal (2010) menjelaskan metode Term Frequency-Inverse Document Frequency (TF-IDF) menggunakan teori bahwa term di dalam dokumen proporsional secara berbanding terbalik dengan jumlah dokumen dalam korpus yang mengandung term tersebut. Salah satu sistem ekstraksi yang menggunakan metode ini adalah ANES yang dibuat pada tahun 1995. Untuk metode selanjutnya yaitu cluster-based, (Kumar & Salim, 2012) mengungkapkan bahwa clustering adalah mengelompokkan objek yang serupa ke dalam kelas-kelas tertentu. Metode ini banyak digunakan dalam multiple document summarization.

Metode berikutnya adalah machine learning. Gupta dan Lehal (2010) memberi pernyataan bahwa kalimat-kalimat diklasifikasikan sebagai kalimat summary dan kalimat non-summary berdasarkan pada keberadaan kriteria-

kriteria tertentu. Probabilitas klasifikasi "dipelajari" secara statistik dari training data dengan menggunakan aturan Bayes, SVM, dan lain-lain. Sedangkan untuk metode graph, (Kumar & Salim, 2012) menyatakan bahwa graph digunakan untuk merepresentasikan koneksi antara objek-objek yang ada. Kalimat merupakan objek dalam graph dan koneksi merupakan kesamaan antar kalimat. Contoh implementasinya adalah TextRank.

Penelitian lain menggunakan metode TF-IDF (Term Frequency-Inverse Document Frequency) dengan berasumsi bahwa term di dalam suatu dokumen proporsional berbanding terbalik dengan jumlah dokumen dalam korpus yang mengandung term tersebut (Gupta & Gambhir, 2016). Pada tahun 2008, Fattah & Ren melakukan penelitian dan berhasil membuat ringkasan teks menggunakan 10 fitur pembobotan teks dengan metode algoritme genetika. Hasil penelitian tersebut menunjukkan bahwa peringkasan teks yang dihasilkan dengan menggunakan teknik algoritme genetika lebih baik dibandingkan teknik regresi matematika.

Metode lain milik Berker dan Gungor (2012) yaitu Using Genetic Algorithms with Lexical Chain for Automatic Text Summarization menggunakan lexical chains dan pembobotan features dengan Genetic Algorithm. Sentence features yang digunakan adalah sentence location (F1), sentence relative length (F2), average TF (F3), average TF-IDF (F4), similarity to title (F5), cue words (F6), named entities (F7), numerical data (F8), sentence centrality (F9), synonim links (F10), dan co-occurence links (F11). Untuk penentuan bobot features digunakan Genetic Algorithm. Bagi setiap dokumen, nilai features dihitung per kalimat. Pada iterasi Genetic Algorithm, bobot feature awal dibuat secara acak.

Nilai kalimat dihitung dan ringkasan diekstraksi serta dievaluasi untuk setiap dokumen. Proses ini diulangi dan rata-rata presisi menunjukkan performa dari iterasi tersebut. Hasil iterasi terbaik akan dipilih oleh GA.

Beberapa metode yang dipaparkan diatas merupakan metode yang diterapkan pada peringkasan teks bahasa Inggris. Tetapi dalam penelitian pada teks bahasa Indonesia belum banyak dilakukan. Diantaranya yaitu metode graph dan algoritma exhaustive oleh Budhi, Intan, Silvia, dan Stevanus (2007) yang menerapkan konsep *virtual graph*. Dalam penelitian tersebut, diterapkan peringkasan teks dengan menggunakan jenis ringkasan informatif yang hanya menyangkut informasi berdasarkan aspek penting dari artikel tersebut dengan mencari dan memasukkan sebanyak mungkin semua topik yang relevan. Ringkasan akan mengabaikan topik yang tidak relevan dan rincian yang tidak penting atau informasi pendukung. Hasil ringkasan hanya menunjukan bagian yang paling penting dari artikel sumber. Umumnya, jenis ringkasan digunakan untuk ikhtisar artikel. Salah satu tahapan proses yang dilakukan yaitu menggunakan *Term Frequency-Inverse Document Frequency* (TF-IDF) dan penggunaan algoritma exhaustive untuk membuat graph (Budhi, Intan, Silvia, & Stevanus, 2007).

Penelitian lain oleh Aristoteles, Herdiyeni, Ridha dan Adisantoso (2012) membuat *automatic text summarizer* berbahasa Indonesia dengan menggunakan algoritma genetika. Terdapat sebelas fitur pembobotan yang digunakan yaitu posisi kalimat, kata kunci positif, kata kunci negatif, kemiripan dengan kalimat lain, kemiripan kalimat dengan judul, entitas nama, data angka, panjang kalimat,

path dari node, penjumlahan kemiripan pada setiap node, dan fitur latent semantic. Penelitian tersebut dilakukan dengan tiga tahap yaitu : tahap pengumpulan dokumen, tahap pelatihan, dan tahap pengujian. Penentuan bobot dilakukan dengan menggunakan algoritma genetika. Bobot yang optimal dapat dijadikan model untuk peringkasan teks. Algoritme genetika atau *genetic algorithm* adalah algoritme pencarian yang didasari pada mekanisme genetik alamiah dan seleksi alamiah. Siklus dari algoritme genetika terdiri beberapa bagian yaitu: populasi awal, evaluasi fitness, seleksi individu, pindah silang (crossover), mutasi (mutation), dan populasi baru. Analisis pembobotan fitur teks menunjukkan bahwa melalui penggunaan fitur kata kunci positif, kemiripan dengan kalimat lain, kemiripan dengan judul dan komponen *latent semantic* dapat memberikan hasil yang hampir sama dengan penggunaan kesebelas fitur. Semua fitur digunakan untuk melakukan training model *genetic algorithm* untuk mendapat kombinasi bobot yang tepat bagi setiap komponen.

Pada tahun 2013, Fereirra dkk meneliti dan melakukan evaluasi terhadapa metode *sentence scoring* dengan menggunakan lima belas fitur pembobotan. Dalam metode ini, ada tiga pendekatan yang diterapkan yaitu: *word scoring*, menetapkan bobot untuk kata-kata yang paling penting. Pendekatan awal dalam pembobotan kalimat didasarkan pada kata-kata. Setiap kata akan diberikan bobot, dan bobot untuk setiap kalimat adalah jumlah dari semua bobot kata penyusunnya. Fitur pembobotan dalam pendekatan ini terdiri dari word frequency, upper case, proper noun, word co-occurrence, dan lexical similarity. Pendekatan berikut adalah *sentence scoring*, penetapan bobot kalimat dengan memverifikasi

fitur kalimat seperti posisi kalimat dalam dokumen, kesamaan dengan judul, dll. Pendekatan yang terakhir adalah *graph scoring*, penetapan bobot kalimat dengan menganalisis hubungan antara kalimat. Metode ini telah diuji oleh Fereirra dkk menggunakan tiga model dataset, dan hasil uji menunjukan lima fitur pembobotan terbaik yaitu *Word Frequency*, *TF-IDF*, *Lexical Similarity*, *Sentence Length*, dan *Text-Rank Score*.