PENERAPAN ALGORITMA LEXRANK DALAM PERINGKASAN OTOMATIS TEKS DOKUMEN AKADEMIK

Silvana Maretha Br. Simbolon, Mansur AS
Ilmu Komputer, Universitas Negeri Medan
Jl. William Iskandar Ps. V, Indonesia
silvanasimbolon9@gmail.com

ABSTRAK

Penelitian ini bertujuan mengevaluasi kinerja algoritma *Continuous LexRank* dalam menghasilkan ringkasan ekstraktif dari dokumen regulatif akademik berbahasa Indonesia. Dokumen yang digunakan adalah Pedoman Akademik Universitas Negeri Medan Tahun 2019--2020 yang terdiri atas 1.414 kalimat dan memiliki karakteristik ketentuan yang berdiri sendiri, berbeda secara fundamental dari teks naratif yang umumnya dijadikan korpus uji peringkasan otomatis. *Continuous LexRank* diterapkan pada pembentukan graf kemiripan antarkalimat dengan pembobotan menggunakan IDF-*Modified Cosine Similarity* dan pemeringkatan berbasis PageRank dengan *damping factor* 0,85. Evaluasi dilakukan pada dua ukuran ringkasan (Top-40 dan Top-70) menggunakan metrik Precision, Recall, dan F1-score terhadap ringkasan acuan yang divalidasi oleh validator ahli. Hasil evaluasi menunjukkan Top-70 memberikan kinerja terbaik dengan F1-*score* 0,3768, lebih tinggi dibandingkan Top-40 yang memperoleh 0,2821. Validasi ahli menunjukkan rata-rata skor 3,00 (kategori "Baik") dengan keterwakilan informasi 57,5% pada Top-40 dan 64,28% pada Top-70. *Continuous LexRank* menunjukkan kemampuan mengekstraksi kerangka informasi sentral dokumen regulatif, namun terbatas dalam mengendalikan redundansi dan konteks rujukan, sehingga lebih optimal diposisikan sebagai pratinjau informasi dokumen regulatif yang panjang.

Kata kunci : *peringkasan teks otomatis, LexRank, dokumen regulatif, pedoman akademik, F1-score, validasi ahli*

1. PENDAHULUAN

Pedoman akademik merupakan dokumen regulatif yang memuat aturan dan ketentuan sebagai dasar dalam pelaksanaan kegiatan akademik di perguruan tinggi [1] [2]. Dokumen ini umumnya memiliki struktur yang panjang dengan ketentuan yang cenderung berdiri sendiri, menjadikannya objek yang menantang secara komputasional bagi pendekatan peringkasan otomatis, karena berbeda secara fundamental dari teks naratif yang umumnya dijadikan korpus uji [3].

Peringkasan otomatis ekstraktif (*automatic extractive summarization*) merupakan pendekatan yang mempertahankan kalimat asli dari dokumen sumber tanpa melakukan penulisan ulang, sehingga lebih sesuai untuk dokumen regulatif yang menuntut ketepatan makna [4]. Salah satu metode ekstraktif yang banyak digunakan adalah LexRank, yang merepresentasikan kalimat sebagai simpul dalam graf dan menentukan kepentingan kalimat berdasarkan keterhubungan antarkalimat menggunakan IDF-*Modified Cosine Similarity* serta pemeringkatan berbasis PageRank [5]. Asumsi dasarnya adalah bahwa kalimat dengan keterhubungan tinggi terhadap kalimat lain cenderung bersifat sentral dan representatif terhadap isi dokumen.

Dalam implementasinya, *LexRank* memiliki dua varian utama yang dibedakan berdasarkan penerapan ambang batas (*threshold*) pada graf kemiripan. Varian pertama menerapkan *threshold* tetap, di mana hanya pasangan kalimat dengan nilai kemiripan di atas batas tertentu yang dihubungkan dalam graf. Kondisi ini berisiko menghilangkan keterhubungan antarkalimat yang sebenarnya relevan apabila nilai kemiripannya berada tepat di bawah batas yang ditetapkan [6]. Varian kedua, yaitu *Continuous LexRank*, tidak menerapkan *threshold* tetap sehingga seluruh pasangan kalimat tetap terhubung dalam graf dengan bobot sesuai nilai kemiripannya masing-masing. *Continuous LexRank* dipilih dalam penelitian ini agar keterhubungan antarkalimat dapat diperhitungkan secara menyeluruh tanpa adanya informasi yang terpotong akibat batasan nilai kemiripan [7].

Dokumen Pedoman Akademik Universitas Negeri Medan Tahun 2019–2020 merupakan dokumen regulatif berbahasa Indonesia yang terdiri atas 281 halaman dan mencakup enam bab utama. Setelah melalui proses ekstraksi dan kurasi, dokumen ini menghasilkan 1.414 kalimat yang menjadi unit analisis dalam penelitian ini. Karakteristik dokumen ini panjang, formal, dan kalimatnya berdiri sendiri menjadikannya objek yang tepat untuk menguji apakah asumsi kerja LexRank berlaku pada jenis dokumen ini.

Namun, evaluasi kinerja LexRank pada dokumen regulatif akademik berbahasa Indonesia masih belum banyak dilakukan. Dokumen regulative akademik khususnya dokumen Pedoman Akademik Universitas Negeri Medan memiliki karakter campuran yang memadukan bagian deskriptif dengan ketentuan normatif yang relatif berdiri sendiri, kondisi yang secara struktural dapat memengaruhi keterhubungan antarkalimat dalam graf kemiripan.

Asumsi kerja LexRank yang mengandalkan kemiripan antarkalimat belum tentu berlaku pada jenis dokumen ini, sehingga perlu diuji secara langsung

pada dokumen tersebut. Maka dari itu, penelitian ini difokuskan pada evaluasi kinerja algoritma *Continuous LexRank* dalam menghasilkan ringkasan ekstraktif dari dokumen Pedoman Akademik Universitas Negeri Medan Tahun 2019–2020 yang terdiri atas 1.414 kalimat. Evaluasi dilakukan menggunakan metrik Precision, Recall, dan F1-score terhadap ringkasan acuan (*gold summary*) yang divalidasi oleh validator ahli, serta penilaian ahli untuk menilai keterwakilan isi ringkasan terhadap dokumen sumber.

2. TINJAUAN PUSTAKA

Beberapa penelitian terdahulu telah menerapkan *LexRank* pada dokumen yang bersifat formal dan regulatif. Setiawan et al. [7] melakukan studi komparatif metode peringkasan berbasis graf pada korpus berbahasa Indonesia dan menemukan bahwa LexRank dapat diterapkan pada teks berbahasa Indonesia, meskipun kinerjanya masih di bawah metode berbasis relevansi judul pada konteks teks naratif berita.

Quadir Md et al. [8] menguji *LexRank* pada dokumen kebijakan privasi yang memiliki karakteristik bahasa formal dan normatif, dan menemukan bahwa *LexRank* menghasilkan konsentrasi kata kunci penting yang lebih tinggi dibandingkan KL *Summarizer* pada dokumen kebijakan privasi yang bersifat formal dan normatif. Gregório et al. [9] menerapkan varian *Guided LexRank* pada dokumen hukum dan menunjukkan bahwa pendekatan berbasis *LexRank* dapat digunakan untuk mengekstraksi bagian-bagian penting dari dokumen formal yang panjang.

Halimah et al. [10] menerapkan algoritma LexRank pada peringkasan teks otomatis terhadap 300 artikel berbahasa Indonesia dari berbagai topik. Hasil pengujian dengan *compression rate* 50% menghasilkan rata-rata f-*measure* sebesar 67,53% pada ROUGE-1, 59,10% pada ROUGE-2, dan 67,05% pada ROUGE-L. Temuan ini mengonfirmasi kompatibilitas metode LexRank dengan karakteristik bahasa Indonesia yang juga digunakan dalam penelitian ini.

2.1. Peringkasan Teks Otomatis

Peringkasan teks otomatis (*automatic text summarization*) adalah proses menghasilkan ringkasan dari dokumen yang panjang secara otomatis untuk menampilkan informasi penting dalam bentuk yang lebih singkat dan terfokus [11] [12]. Secara umum, metode peringkasan dibedakan menjadi dua pendekatan utama. Pertama, peringkasan ekstraktif yang memilih kalimat-kalimat paling penting dari teks asli tanpa melakukan modifikasi bentuk atau parafrasa [13] [14]. Kedua, peringkasan abstraktif yang menyusun ringkasan dengan menghasilkan kalimat baru melalui parafrase atau perumusan ulang informasi menggunakan model *deep learning* atau *transformer* [15] [16]. Untuk dokumen regulatif yang menuntut ketepatan makna, pendekatan ekstraktif lebih sesuai karena mempertahankan kalimat asli sehingga isi aturan tidak mengalami perubahan makna.

2.2. LexRank

LexRank merupakan metode peringkasan teks otomatis berbasis ekstraktif yang pertama kali diperkenalkan oleh Erkan dan Radev dengan mengadaptasi konsep PageRank untuk peringkasan teks. Metode ini memodelkan hubungan antarkalimat menggunakan pendekatan graf, di mana setiap kalimat direpresentasikan sebagai simpul (*node*) dan hubungan antarkalimat dibentuk sebagai sisi (*edge*) yang bobotnya mencerminkan tingkat kemiripan [6]. Gagasan dasarnya adalah bahwa kalimat yang memiliki konektivitas tinggi dengan kalimat lain cenderung bersifat sentral dan representatif terhadap isi dokumen.

Kemiripan antarkalimat dihitung menggunakan IDF-*Modified Cosine Similarity* yang memberi bobot lebih besar pada kata-kata yang jarang muncul dalam korpus, sehingga ukuran kemiripan lebih mencerminkan kesamaan makna daripada sekadar kesamaan kata permukaan [17] [6].

$$idf - cos(x,y) = \frac{\sum_{w \in x \cap y} (tf_{w,x} \cdot tf_{w,y} \cdot (idf(w)^2)}{\sqrt{\sum_{i \in x} \left(tf_{i,x} \cdot idf(i)\right)^2} \cdot \sqrt{\sum_{j \in y} (tf_{j,y} \cdot idf(j))^2}} \quad (1)$$

Keterangan:

- x,y adalah dua kalimat yang dibandingkan
- w adalah kata yang muncul pada kedua kalimat (irisan kosakata)
- $tf_{(w,x)}$ adalah frekuensi kata *w* dalam kalimat *x*
- $tf_{(w,y)}$ adalah frekuensi kata *w* dalam kalimat *y*
- $idf_{(w)}$ adalah nilai *inverse document frequency* dari kata *w*

Penelitian ini mengadopsi pendekatan *Continuous LexRank* di mana seluruh pasangan kalimat tetap terhubung dan bobot sisi ditentukan langsung dari nilai kemiripannya tanpa ambang batas (*threshold*) tetap [6]. Pendekatan ini dipilih karena selaras dengan karakteristik dokumen regulatif akademik yang kalimatnya bersifat mandiri, sehingga perbedaan nilai kemiripan antarkalimat relatif kecil dan penerapan *threshold* tetap berisiko menghilangkan informasi penting [18].

2.3. Perangkingan Kalimat

Setelah graf kemiripan terbentuk, skor kepentingan setiap kalimat ditentukan menggunakan algoritma PageRank yang dihitung secara iteratif hingga mencapai konvergensi [19]. Konsep utamanya adalah bahwa sebuah kalimat akan dianggap penting apabila terhubung dengan banyak kalimat penting lainnya dalam graf. Semakin banyak dan kuat koneksi yang dimiliki sebuah kalimat, semakin tinggi skor PageRank-nya [6].

$$PR(u) = \frac{1-d}{N} + d \sum_{v \in adj(u)} \frac{PR(v) w_{v \to u}}{\Sigma_k w_{v \to k}} \quad (2)$$

Keterangan:

- $PR(u)$ adalah skor kepentingan kalimat u

- d adalah *damping factor* (biasanya 0.85)
- N adalah jumlah total simpul (kalimat) dalam graf
- $adj(u)$ adalah himpunan kalimat yang memiliki kemiripan dengan u
- $w_{v \to u}$ adalah bobot kemiripan antara kalimat v dan kalimat u
- $\Sigma_k w_{v \to k}$ adalah jumlah bobot keluar dari simpul v
- $\frac{1-d}{N}$ adalah distribusi acak merata ke semua simpul

Kalimat dengan skor PageRank tertinggi kemudian dipilih sebagai kandidat ringkasan akhir sesuai ukuran ringkasan yang ditetapkan.

2.4. Evaluasi Sistem

Evaluasi kinerja sistem dilakukan menggunakan tiga metrik utama yaitu Precision, Recall, dan F1-score yang mengukur sejauh mana ringkasan sistem mendekati ringkasan acuan (*gold summary*).

a. Precision mengukur proporsi kalimat dalam ringkasan sistem yang juga muncul dalam ringkasan acuan.

$$Presisi = \frac{Ringkasan\ Hasil\ Sistem \cap Ringkasan\ Acuan}{Ringkasan\ Hasil\ Sistem} \quad (3)$$

b. Recall mengukur proporsi kalimat penting dalam ringkasan acuan yang berhasil diambil oleh sistem.

$$Recall = \frac{Ringkasan\ Hasil\ Sistem \cap Ringkasan\ Acuan}{Ringkasan\ Acuan} \quad (4)$$

c. F1-score adalah rata-rata harmonik dari Precision dan Recall yang memberikan gambaran lebih seimbang tentang kinerja sistem.

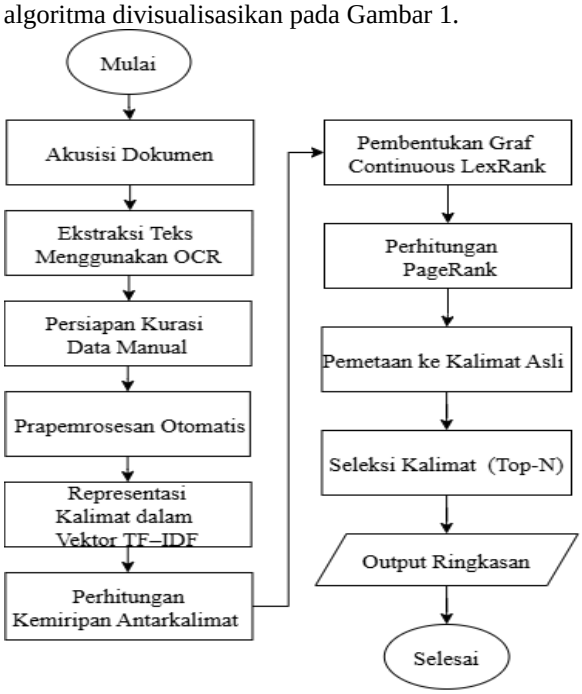
$$F1 = 2 \times \frac{Presisi \times Recall}{Presisi + Recall} \quad (5)$$

Dalam penelitian ini, kesesuaian kalimat ditentukan berdasarkan kesamaan teks setelah normalisasi, bukan berdasarkan posisi kalimat dalam dokumen, sehingga kalimat yang muncul lebih dari sekali tetap dapat dicocokkan secara adil.

3. METODE PENELITIAN

Penelitian ini merupakan penelitian evaluatif yang bertujuan menilai kinerja algoritma *Continuous LexRank* dalam menghasilkan ringkasan ekstraktif dari dokumen Pedoman Akademik Universitas Negeri Medan Tahun Akademik 2019–2020. Dokumen tersebut terdiri atas 281 halaman yang memuat enam bab utama yaitu Pendahuluan, Organisasi dan Personalia, Sarana Pelayanan, Tata Tertib Kehidupan Kampus, Penyelenggaraan Pendidikan, Fakultas Pascasarjana, Jurusan Dan Program Studi yang mencakup ketentuan mengenai sistem pembelajaran, kurikulum, peraturan akademik, serta prosedur administratif yang berlaku bagi mahasiswa dan sivitas akademika dalam format PDF berbasis citra (*scanned document*), sehingga ekstraksi teks dilakukan menggunakan mesin OCR *Tesseract* melalui Google Colab dengan dukungan bahasa Indonesia.

Hasil OCR kemudian melalui tahap kurasi manual yang meliputi pembersihan *noise*, seleksi bagian utama dokumen, koreksi kesalahan karakter fatal, dan segmentasi kalimat secara manual dalam format *one-sentence-per-line*. Proses ini menghasilkan dua versi korpus paralel versi teks asli bersih untuk penyusunan ringkasan akhir dan versi teks mesin untuk proses komputasional yang divalidasi memiliki jumlah baris identik guna menjamin konsistensi indeks pada tahap pemetaan. Alur keseluruhan tahapan algoritma divisualisasikan pada Gambar 1.



Gambar 1. Alur Tahapan Algoritma *Continuous LexRank*

Dari alur pada Gambar 1, dapat dilihat bahwa proses peringkasan otomatis dimulai dari akuisisi dokumen dan ekstraksi teks menggunakan OCR, dilanjutkan dengan persiapan kurasi data secara manual dan prapemrosesan otomatis. Teks yang telah diproses kemudian direpresentasikan dalam vektor TF-IDF sebagai dasar perhitungan kemiripan antarkalimat menggunakan IDF-*Modified Cosine Similarity*. Hasil kemiripan tersebut digunakan untuk membentuk graf *Continuous LexRank*, yang selanjutnya menjadi input perhitungan skor PageRank secara iteratif. Kalimat dengan skor tertinggi dipilih melalui seleksi Top-N, dipetakan kembali ke teks asli, dan disusun menjadi *output* ringkasan akhir.

Prapemrosesan otomatis dilakukan terhadap versi teks mesin, meliputi *case folding*, tokenisasi, *stopword removal* menggunakan daftar *stopword* Bahasa Indonesia dari *library NLTK*, dan *stemming* menggunakan *library* Sastrawi. Setiap kalimat kemudian direpresentasikan dalam vektor TF-IDF, di mana nilai IDF dihitung berdasarkan distribusi *term* antarkalimat dalam dokumen yang sama (*single document*). Kemiripan antarkalimat dihitung menggunakan IDF-*Modified Cosine Similarity*, dan hasilnya disusun dalam matriks kemiripan berukuran 1.414 × 1.414. Matriks tersebut dinormalisasi secara baris untuk membentuk *Transition Matrix* sebagai

dasar pembentukan graf *Continuous LexRank* tanpa *threshold* tetap, sehingga seluruh pasangan kalimat tetap terhubung dalam graf berbobot.

Skor kepentingan kalimat dihitung menggunakan PageRank dengan *damping factor* 0,85, *threshold* konvergensi 10^{-6}, dan maksimum iterasi 100 menggunakan metode iteratif (*power method*). Skor hasil komputasi dipetakan kembali ke versi teks asli berdasarkan kesesuaian indeks baris. Kalimat dengan skor tertinggi dipilih menggunakan pendekatan Top-N dengan dua variasi ukuran, yaitu Top-40 dan Top-70, untuk menganalisis pengaruh volume ringkasan terhadap keterwakilan informasi.

Ringkasan acuan (*gold summary*) disusun secara manual dengan pendekatan ekstraktif sebanyak 70 kalimat yang diurutkan berdasarkan indeks kemunculan dalam dokumen asli. Jumlah 70 kalimat ditetapkan untuk menyamakan ukuran dengan skenario Top-70, sehingga evaluasi kinerja sistem dapat dilakukan secara adil dengan jumlah kalimat yang setara antara ringkasan sistem dan ringkasan acuan.

Ringkasan acuan divalidasi melalui dua tahap peninjauan oleh seorang validator ahli bidang Bahasa Indonesia. Pada tahap pertama, validator menilai ringkasan acuan dan memberikan catatan perbaikan, sehingga dilakukan revisi substansial terhadap pemilihan kalimat agar lebih merepresentasikan isi dokumen secara menyeluruh. Pada tahap kedua, validator menilai kembali kedua ringkasan secara final dan menyatakan ringkasan acuan layak digunakan sebagai standar pembanding.

Evaluasi kinerja sistem dilakukan menggunakan metrik *Precision*, *Recall*, dan F1-*score* pada tingkat kesesuaian pemilihan kalimat, serta penilaian ahli menggunakan instrumen skala Likert empat tingkat terhadap lima indikator: relevansi topik, representativitas informasi, kecukupan informasi, efektivitas ringkasan, dan kelayakan keseluruhan.

4. HASIL DAN PEMBAHASAN

4.1. Deskripsi Data

Dokumen Pedoman Akademik Universitas Negeri Medan Tahun 2019–2020 tersedia dalam format PDF berbasis citra (*scanned document*) sehingga ekstraksi teks dilakukan melalui OCR. Hasil OCR kemudian melalui tahap kurasi manual sebelum digunakan sebagai korpus analisis. Perubahan jumlah data pada setiap tahap ditunjukkan pada Tabel 1.

Tabel 1. Perubahan Jumlah Data

Tahap	Jumlah
Hasil OCR	16.186 baris teks mentah
Setelah seleksi konten	1.596 baris isi dokumen
Setelah segmentasi kalimat	1.414 kalimat

Dari Tabel 1 dapat dilihat bahwa proses kurasi mereduksi 16.186 baris teks mentah menjadi 1.414 kalimat final. Penurunan ini terjadi karena bagian non-analitis seperti sampul, daftar isi, lampiran, kalender akademik, dan elemen administratif lainnya tidak disertakan dalam korpus. Dengan 1.414 kalimat, sistem membentuk matriks kemiripan berukuran 1.414 × 1.414 atau hampir dua juta pasangan perhitungan, dan seluruhnya tetap terhubung dalam graf karena pendekatan *Continuous LexRank* tidak memotong pasangan kalimat berdasarkan nilai kemiripan minimum.

Selain dari sisi kuantitas, karakteristik kalimat dalam dokumen pedoman akademik juga memiliki sifat khusus. Sebagian besar kalimat berbentuk butir enumerasi yang relatif pendek namun padat secara normatif. Struktur seperti ini berbeda dengan teks naratif atau berita, karena banyak kalimat memiliki pola leksikal yang serupa. Kondisi tersebut dapat memengaruhi perhitungan kemiripan dalam LexRank, karena keterkaitan antarkalimat sangat bergantung pada kesamaan term yang muncul. Pada kalimat yang relatif pendek, jumlah *term* pembanding menjadi lebih terbatas sehingga variasi nilai kemiripan antarkalimat dapat berbeda dibandingkan pada teks yang lebih panjang dan deskriptif. Oleh karena itu, karakteristik dokumen regulatif menjadi faktor penting dalam interpretasi hasil peringkasan otomatis.

4.2. Hasil Perhitungan Skor PageRank

Perhitungan PageRank dilakukan secara iteratif menggunakan *power method* dengan *damping factor* 0,85. Konvergensi tercapai pada iterasi ke-17, di mana perubahan nilai PageRank antar iterasi tidak lagi berarti hingga iterasi ke-21. Kondisi ini menunjukkan bahwa graf kemiripan yang terbentuk memiliki struktur keterhubungan yang stabil untuk mendistribusikan skor sentralitas ke seluruh kalimat. Kalimat-kalimat dengan skor PageRank tertinggi disajikan pada Tabel 2.

Tabel 2. Kalimat dengan Skor PageRank Tertinggi

Rank	Indeks	Skor Page Rank	Kalimat
1	222	0.001934	Bagian Akademik menyelenggarakan fungsi Pelaksanaan layanan pendidikan, penelitian, dan pengabdian kepada masyarakat.
2	140	0.001805	Dalam melaksanakan tugas pokok tersebut, Unimed menyelenggarakan fungsi pelaksanaan dan pengembangan pendidikan tinggi, pelaksanaan penelitian dalam rangka pengembangan ilmu pengetahuan, teknologi dan/atau kesenian, serta pelaksanaan pengabdian kepada masyarakat.

Rank	Indeks	Skor Page Rank	Kalimat
3	304	0.001746	Lembaga Penelitian dan Pengabdian kepada Masyarakat adalah unsur pelaksana akademik yang melaksanakan sebagian tugas dan fungsi Unimed di bidang penelitian dan pengabdian kepada masyarakat yang berada di bawah Rektor.
4	223	0.001746	Bagian Akademik menyelenggarakan fungsi Pelaksanaan evaluasi kegiatan pendidikan, penelitian dan pengabdian kepada masyarakat.
5	177	0.001728	Dekan Dalam mempunyai tugas memimpin penyelenggaraan pendidikan, penelitian, pengabdian kepada masyarakat, membina tenaga kependidikan, mahasiswa tenaga kependidikan dan administrasi fakultas.
6	146	0.001697	Rektor mempunyai tugas memimpin penyelenggaraan pendidikan, penelitian dan pengabdian kepada masyarakat, membina tenaga pendidik, mahasiswa, tenaga kependidikan dan hubungannya dengan lingkungan.

Dari Tabel 2 dapat dilihat bahwa kalimat-kalimat dengan skor PageRank tertinggi didominasi oleh frasa "pendidikan", "penelitian", dan "pengabdian kepada masyarakat". Kata-kata tersebut tersebar di seluruh bab dokumen sehingga kalimat yang memuatnya memiliki kemiripan leksikal tinggi dengan banyak kalimat lain dan memperoleh skor sentralitas tinggi dalam graf. Rentang skor yang sempit (0,001697–0,001934) menunjukkan tidak ada kalimat yang sangat dominan, hal ini merupakan karakteristik khas *Continuous LexRank* di mana seluruh kalimat tetap terhubung dan berkontribusi dalam distribusi sentralitas.

4.3. Analisis Pola Ringkasan

Perbandingan antara ringkasan Top-40 dan Top-70 memperlihatkan perbedaan pola yang signifikan. Pada Top-40, ringkasan didominasi oleh kalimat struktural yang berkaitan dengan tugas dan fungsi unit kerja seperti Rektor, Dekan, dan Bagian Akademik. Kalimat-kalimat ini memiliki pola redaksi serupa seperti "mempunyai tugas" dan "menyelenggarakan fungsi", sehingga kemiripan leksikalnya antara satu kalimat dengan kalimat lain tinggi dan menghasilkan skor sentralitas tinggi dalam graf. Kondisi ini juga memunculkan redundansi langsung, di mana kalimat "Pelaksanaan dan pengembangan pendidikan tinggi" muncul dua kali dalam ringkasan karena kalimat tersebut hadir lebih dari sekali dalam dokumen sumber dan keduanya memperoleh skor sentralitas yang hampir identik.

Pada Top-70, cakupan informasi meluas ke kalimat dengan skor sentralitas menengah yang memuat informasi lebih operasional, seperti peran dosen pembimbing akademik, mekanisme Sistem Kredit Semester (SKS), dan fungsi dosen pembimbing skripsi. Informasi ini menggunakan istilah yang lebih spesifik dan muncul pada bagian dokumen yang lebih terbatas, sehingga tingkat keterhubungannya dalam graf lebih rendah dibandingkan kalimat struktural. Namun, perluasan jumlah kalimat juga meningkatkan potensi redundansi karena mekanisme PageRank tidak memiliki mekanisme pengurangan kalimat yang berulang.

4.4. Evaluasi Kinerja Sistem

Evaluasi kinerja dilakukan pada tingkat kesesuaian pemilihan kalimat antara ringkasan sistem dan ringkasan acuan. Sebuah kalimat dinyatakan cocok (*match*) apabila teksnya identik dengan salah satu kalimat dalam *gold summary* setelah melalui proses normalisasi, bukan berdasarkan kesamaan indeks posisi. Hasil evaluasi pada dua ukuran ringkasan disajikan pada Tabel 3.

Tabel 3. Hasil Evaluasi Kinerja Ringkasan Sistem

Ukuran Ringkasan	Precision	Recall	F1-score	Ukuran Ringkasan
Top-40	0,2821	0,2821	0,2821	0,2821
Top-70	0,3768	0,3768	0,3768	0,3768

Nilai Precision, Recall, dan F1-score yang identik pada setiap skenario terjadi secara matematis karena jumlah kalimat ringkasan sistem ditetapkan sama persis dengan jumlah kalimat ringkasan acuan (n=40 dan n=70). Dalam kondisi himpunan yang sama besar, setiap kesalahan seleksi (*False Positive*) secara otomatis menyebabkan hilangnya satu kalimat relevan (*False Negative*), sehingga rasio ketepatan dan kelengkapan selalu bernilai sama.

Ringkasan Top-70 menunjukkan kinerja lebih unggul dengan F1-*score* 0,3768 dibandingkan Top-40 yang memperoleh 0,2821, dengan peningkatan sebesar 33,6%. Hal ini menunjukkan bahwa cakupan ringkasan yang lebih luas meningkatkan ketercakupan informasi terhadap ringkasan acuan. Pada Top-40, ringkasan didominasi oleh kalimat struktural yang bersifat umum dan berulang seperti mandat universitas dan fungsi unit kerja, karena kalimat-kalimat tersebut memiliki kemiripan leksikal tinggi dengan banyak kalimat lain sehingga mendapat skor sentralitas tinggi.

Pada Top-70, sistem mulai menjangkau kalimat dengan skor sentralitas menengah yang memuat informasi lebih spesifik dan operasional. Meskipun demikian, skor F1-*score* 0,3768 mengindikasikan adanya perbedaan mendasar antara kriteria seleksi berbasis sentralitas graf dan pertimbangan semantik manusia. Untuk memahami nilai F1-score ini secara proporsional, perlu dipertimbangkan konteks ukuran korpusnya.

Ringkasan acuan disusun oleh peneliti secara ekstraktif dari pool 1.414 kalimat yang sama,

kemudian divalidasi kelayakannya oleh validator ahli melalui dua tahap peninjauan hingga dinyatakan representatif. Jika seleksi dilakukan secara acak, probabilitas kesesuaian antara dua himpunan 70 kalimat dari keseluruhan kalimat dokumen hanya sekitar 5%. Fakta bahwa sistem berhasil mencapai irisan sebesar 37% menunjukkan bahwa LexRank tidak bekerja secara acak, melainkan menangkap pola yang cukup konsisten dengan penilaian peneliti yang telah tervalidasi.

Perbedaan yang tersisa mencerminkan perbedaan definisi 'kalimat penting' antara mesin dan manusia, LexRank memilih berdasarkan sentralitas leksikal, sementara manusia memilih berdasarkan signifikansi informasi. Skor F1-score lebih tepat dipahami sebagai ukuran konvergensi dua pendekatan berbeda, bukan indikator kegagalan sistem.

Analisis terhadap ukuran ringkasan menunjukkan bahwa peningkatan jumlah kalimat dari Top-40 ke Top-70 berpengaruh terhadap jenis informasi yang tertangkap. Pada Top-40, ringkasan lebih banyak berisi informasi yang bersifat struktural dan administratif, seperti tugas rektor atau fungsi unit kerja, yang memiliki pola redaksi serupa dan berulang. Hal ini terjadi karena LexRank menghitung sentralitas berdasarkan kemiripan leksikal, kalimat yang sering muncul dan memiliki kemiripan dengan banyak kalimat lain akan memperoleh skor lebih tinggi.

Sebaliknya, informasi teknis yang lebih spesifik dan bersifat operasional, seperti aturan SKS atau skripsi, cenderung muncul lebih sedikit dan bersifat unik sehingga skor sentralitasnya lebih rendah. Informasi tersebut baru lebih banyak muncul ketika jumlah kalimat diperluas menjadi Top-70. Temuan ini menunjukkan bahwa dalam dokumen regulatif, kalimat yang memiliki tingkat keterhubungan tinggi dalam graf belum tentu merupakan kalimat yang paling informatif secara substantif.

Berdasarkan temuan tersebut, LexRank menunjukkan potensi yang cukup baik sebagai alat bantu pratinjau informasi untuk memahami struktur dan konteks umum dokumen akademik yang panjang secara cepat. Namun, karena masih memiliki keterbatasan dalam menjaga konteks dan cenderung memilih kalimat yang repetitif, ringkasan ini belum dapat sepenuhnya menggantikan pembacaan dokumen asli, terutama untuk kebutuhan yang memerlukan ketelitian tinggi.

LexRank memilih kalimat berdasarkan tingkat keterhubungan dalam graf kemiripan, sehingga cenderung memilih kalimat yang mengandung istilah berulang. Sebaliknya, validator memilih kalimat berdasarkan bobot informasi dan signifikansi isi, termasuk kalimat yang unik dan tidak memiliki kemiripan leksikal tinggi dengan kalimat lain. Perbedaan pendekatan inilah yang menyebabkan tidak seluruh kalimat pilihan manusia dapat terdeteksi oleh sistem.

4.5. Keterwakilan Informasi

Selain evaluasi berbasis metrik Precision, Recall, dan F1-score, dilakukan juga analisis keterwakilan informasi berdasarkan penandaan yang diberikan validator secara langsung pada kalimat-kalimat dalam ringkasan sistem. Penandaan ini dimanfaatkan sebagai bahan analisis tambahan untuk melihat tingkat keterwakilan informasi pada kedua ukuran ringkasan. Hasil perhitungan keterwakilan pada masing-masing ukuran ringkasan adalah sebagai berikut:

$$Top-40 = \frac{23}{40} \times 100\% = 57{,}5\%$$

$$Top-70 = \frac{45}{70} \times 100\% = 64{,}28\%$$

Berdasarkan hasil tersebut, keterwakilan informasi pada Top-40 mencapai 57,5%, di mana 23 dari 40 kalimat dinilai representatif oleh validator. Sementara itu, pada Top-70 sebanyak 45 dari 70 kalimat dinilai representatif dengan persentase 64,28%. Peningkatan ini menunjukkan bahwa perluasan jumlah kalimat memungkinkan sistem menangkap informasi yang lebih beragam. Meskipun demikian, perluasan tersebut juga disertai masuknya kalimat yang kurang relevan, sehingga tidak semua kalimat tambahan bersifat representatif. Adanya kalimat yang tidak mewakili ini tidak bertentangan dengan skor validasi 3 (Baik) yang diberikan validator, karena skor tersebut menilai kelayakan ringkasan secara keseluruhan, sedangkan analisis keterwakilan ini melihat pada tingkat kalimat secara individual.

Analisis lebih lanjut terhadap pola kalimat menunjukkan adanya perbedaan yang cukup jelas antara kalimat yang dinilai representatif dan yang tidak. Kalimat yang dinilai representatif umumnya berupa definisi formal seperti definisi program studi, pernyataan tugas pokok pejabat, serta ketentuan spesifik yang merujuk pada entitas konkret dalam struktur universitas. Sebaliknya, kalimat yang dinilai tidak representatif menunjukkan beberapa pola. Sebagian bersifat terlalu generik dan pendek tanpa informasi konkret, sebagian lain terlalu spesifik pada satu unit atau fakultas tertentu sehingga tidak mencerminkan isi dokumen secara menyeluruh, dan sebagian lainnya merupakan pernyataan visi atau misi yang bersifat normatif tanpa ketentuan operasional yang jelas.

Kondisi ini berkaitan dengan cara kerja algoritma LexRank yang menghitung sentralitas kalimat berdasarkan distribusi kata (TF-IDF) dan *cosine similarity*. Dalam dokumen Pedoman Akademik, istilah-istilah umum seperti “pendidikan” dan “penelitian” muncul sangat sering dan tersebar di seluruh bab dokumen. Akibatnya, kalimat yang memuat kata-kata tersebut memiliki kemiripan leksikal dengan banyak kalimat lain sehingga memperoleh skor sentralitas yang tinggi dalam graf. Keterbatasan utama sistem terletak pada ketidakmampuannya membedakan antara kalimat yang benar-benar informatif dengan kalimat yang hanya bersifat repetitif.

4.6. Validasi Ahli

Validasi ahli dilakukan dalam dua tahap peninjauan oleh seorang dosen bidang Bahasa Indonesia yang memiliki keahlian dalam analisis teks akademik. Pada tahap pertama, ringkasan acuan direvisi secara substansial berdasarkan masukan validator karena belum sepenuhnya merepresentasikan poin-poin penting setiap bab. Pada tahap kedua, validator menilai kembali kedua ringkasan secara final dan menyatakan ringkasan acuan layak digunakan sebagai standar pembanding. Hasil penilaian final disajikan pada Tabel 4.

Tabel 4. Hasil Penilaian Validator Ahli

Objek Evaluasi	Total Skor	N	Rata-rata	Kategori
Ringkasan Acuan	27	9	3,00	Baik
Ringkasan Sistem LexRank	15	5	3,00	Baik

Perbedaan jumlah indikator antara ringkasan acuan (9 indikator) dan ringkasan sistem (5 indikator) disebabkan oleh perbedaan fokus penilaian. Ringkasan acuan dinilai lebih ketat karena berfungsi sebagai standar pembanding, mencakup aspek tambahan seperti selektivitas informasi, struktur, dan kelayakan sebagai acuan evaluasi. Ringkasan sistem dinilai berdasarkan aspek yang relevan dengan kualitas *output* algoritma secara substantif. Kedua ringkasan memperoleh rata-rata skor 3,00 yang berada pada kategori "Baik". Hasil ini menunjukkan bahwa ringkasan acuan layak digunakan sebagai *gold summary*, dan ringkasan sistem *Continuous LexRank* cukup mampu merepresentasikan informasi utama dokumen.

4.7. Pembahasan

Berdasarkan pengujian yang telah dilakukan, ditemukan adanya perbedaan antara hasil pengukuran kinerja sistem dan penilaian validator. Secara komputasi, nilai F1-score berada pada 0,3768. Namun, secara substantif, validator ahli memberikan penilaian kategori 'Baik'. Perbedaan ini menunjukkan bahwa evaluasi berbasis kesamaan teks kalimat belum sepenuhnya mampu menggambarkan kualitas ringkasan secara menyeluruh.

Karakteristik dokumen regulatif secara langsung memengaruhi pola seleksi LexRank. Frasa baku yang berulang seperti "menyelenggarakan pendidikan dan penelitian" membuat algoritma lebih mudah menangkap pola umum dokumen dibandingkan ketentuan spesifik yang jarang disebut. Hal ini mengkonfirmasi bahwa asumsi kerja LexRank kalimat dengan keterhubungan tinggi bersifat sentral dan representatif tidak sepenuhnya berlaku pada dokumen regulatif akademik yang strukturnya berdiri sendiri dan tidak naratif.

Selain itu, sifat ekstraktif LexRank menghasilkan *dangling reference* di mana kalimat yang diawali "hal tersebut" atau "ketentuan ini" kehilangan konteksnya ketika diambil tanpa kalimat sebelumnya. Secara keseluruhan, Continuous LexRank menunjukkan kemampuan dalam mengekstraksi kerangka informasi sentral dokumen regulatif, dan berpotensi dimanfaatkan sebagai pratinjau informasi sebelum pembacaan dokumen asli secara menyeluruh. Temuan ini selaras dengan Setiawan et al. yang menemukan LexRank stabil pada korpus berbahasa Indonesia, namun memperluas pemahaman bahwa kestabilan tersebut tidak serta merta menghasilkan ketercakupan informasi yang tinggi pada dokumen regulatif akademik [7].

5. KESIMPULAN DAN SARAN

Penelitian ini menunjukkan bahwa algoritma *Continuous LexRank* dapat diterapkan sebagai metode peringkasan otomatis pada dokumen regulatif akademik berbahasa Indonesia, mengisi celah yang belum banyak dieksplorasi sebelumnya, dengan memanfaatkan pembobotan IDF-*Modified Cosine Similarity* dan pemeringkatan berbasis PageRank, serta mencapai konvergensi pada iterasi ke-17 hingga ke-21. Evaluasi kinerja menunjukkan bahwa ringkasan Top-70 memberikan hasil terbaik dengan F1-score 0,3768, lebih tinggi dibandingkan Top-40 yang memperoleh 0,2821 dengan selisih peningkatan 33,6%, mengindikasikan bahwa cakupan ringkasan yang lebih luas meningkatkan ketercakupan informasi terhadap ringkasan acuan. Validasi ahli menempatkan ringkasan sistem pada kategori "Baik" dengan keterwakilan informasi 57,5% pada Top-40 dan 64,28% pada Top-70. Terdapat perbedaan mendasar antara kriteria seleksi berbasis sentralitas graf dan pertimbangan semantik manusia, LexRank memilih berdasarkan keterhubungan leksikal, sementara manusia mempertimbangkan kepentingan dan kebermanfaatan informasi. Perbedaan ini merupakan karakteristik pendekatan ekstraktif berbasis sentralitas, bukan kegagalan algoritma. Keterbatasan utama metode ini terletak pada kecenderungan memilih kalimat repetitif dan hilangnya konteks rujukan *(dangling reference)*, sehingga ringkasan lebih optimal diposisikan sebagai pratinjau informasi dokumen regulatif yang panjang. Untuk penelitian selanjutnya, disarankan penambahan tahap pascapemrosesan untuk mengurangi redundansi dan masalah rujukan terputus, integrasi pendekatan berbasis semantik atau metode hibrida untuk meningkatkan kualitas seleksi kalimat, penggunaan metrik evaluasi yang lebih toleran terhadap variasi redaksi seperti ROUGE-2 atau BERTScore, serta pengujian pada jenis dokumen regulatif lain untuk menguji konsistensi kinerja metode dalam domain yang berbeda.

DAFTAR PUSTAKA

[1] J. Singh, S. Fazili, R. Jain, and M. S. Akhtar, "EROS: Entity-Driven Controlled Policy Document Summarization," *arXiv:2403.00141v1*, Feb. 2024, doi: https://doi.org/10.48550/arXiv.2403.00141.

[2] M. Sie, R. Beek, M. Bots, S. Brinkkemper, and A. Gatt, "Summarizing Long Regulatory Documents with a Multi-Step Pipeline," in *Proceedings of the 6th Workshop on Natural Legal Language Processing (NLLP 2024)*, Association for Computational Linguistics, Nov. 2024, pp. 18–32. doi: https://doi.org/10.48550/arXiv.2408.09777.

[3] F. Ariai, J. Mackenzie, and G. Demartini, "Natural Language Processing for the Legal Domain: A Survey of Tasks, Datasets, Models, and Challenges," *ACM Comput. Surv.*, vol. 58, no. 6, Dec. 2025, doi: 10.1145/3777009.

[4] A. P. Widyassari *et al.*, "Review of automatic text summarization techniques & methods," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 4, pp. 1029–1046, May 2022, doi: 10.1016/j.jksuci.2020.05.006.

[5] A. M. Qaroush, L. Naser, M. Mali, and A. Naji, "Semantic-aware hybrid graph-based extractive summarization for arabic texts," *Journal of King Saud University - Computer and Information Sciences*, vol. 37, no. 10, Dec. 2025, doi: 10.1007/s44443-025-00359-x.

[6] G. Erkan and D. R. Radev, "LexRank: Graph-based Lexical Centrality as Salience in Text Summarization," *Journal of Artificial Intelligence Research*, vol. 22, pp. 457–479, Apr. 2004.

[7] A. Setiawan, Z. Abidin, and M. Imamudin, "Impact of Preprocessing on Indonesian Extractive Summarization Using LexRank, TextRank, DivRank, and Cosine Similarity," *G-Tech: Jurnal Teknologi Terapan*, vol. 9, no. 4, pp. 2311–2321, Oct. 2025, doi: 10.70609/g-tech.v9i4.8306.

[8] A. Quadir Md *et al.*, "Data-Driven Analysis of Privacy Policies Using LexRank and KL Summarizer for Environmental Sustainability," *Sustainability (Switzerland)*, vol. 15, no. 7, p. 5941, Apr. 2023, doi: 10.3390/su15075941.

[9] F. Gregório, R. Castro, K. Belloze, R. P. Lopes, and E. Bezerra, "GLARE: Guided LexRank for Advanced Retrieval in Legal Analysis," *ArXiv*, Sep. 2024, Accessed: Mar. 15, 2026. [Online]. Available: https://arxiv.org/abs/2409.15348

[10] Halimah, Surya Agustian, and Siti Ramadhani, "Peringkasan teks otomatis (automated text summarization) pada artikel berbahasa indonesia menggunakan algoritma lexrank," *Jurnal CoSciTech (Computer Science and Information Technology)*, vol. 3, no. 3, pp. 371–381, Dec. 2022, doi: 10.37859/coscitech.v3i3.4300.

[11] N. Hendrastuty and Azhari, "Text Summarization in Multi Document Using Genetic Algorithm," *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 15, no. 4, p. 327, Oct. 2021, doi: 10.22146/ijccs.66026.

[12] Y. M. Sari and N. S. Fatonah, "Peringkasan Teks Otomatis pada Modul Pembelajaran Berbahasa Indonesia Menggunakan Metode Cross Latent Semantic Analysis (CLSA)," *Jurnal Edukasi dan Penelitian Informatika)*, vol. 7, no. 2, Aug. 2021, doi: https://doi.org/10.26418/jp.v7i2.47768.

[13] M. R. Hadwirianto, F. Hamami, and O. N. Pratiwi, "Extractive Text Summarization Terhadap Artikel Berita Indonesia Berbasis Machine Learning," in *e-Proceeding of Engineering*, Telkom University., Oct. 2024, p. 3941. Accessed: Mar. 15, 2026. [Online]. Available: https://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/23804

[14] J. J. Sihombing, A. Arnita, S. I. Al Idrus, and D. Y. Niska, "Implementation of text summarization on indonesian scientific articles using textrank algorithm with TF-IDF web-based," *Journal of Soft Computing Exploration*, vol. 5, no. 3, pp. 310–319, Dec. 2024, doi: 10.52465/joscex.v5i3.475.

[15] I. M. Karo Karo, A. Perdana, and S. Dewi, "Implementasi Text Summarization Pada Review Aplikasi Digital Library System Menggunakan Metode Maximum Marginal Relevance," *JEKIN (Jurnal Teknik Informatika)*, vol. 4, no. 1, Apr. 2024, doi: https://doi.org/10.58794/jekin.v4i1.671.

[16] M. Yani, N. Siti Khodijah, and M. Mustamiin, "Aplikasi Peringkas Teks Bahasa Indonesia Menggunakan Model Text-to-Text Transfer Transformer (T5)," *IKRAITH-INFORMATIKA*, vol. 9, no. 2, Jul. 2025, doi: 10.37817/ikraith-informatika.v9i2.

[17] A. Munshi, A. Mehra, and A. Choudhury, "LexRank Algorithm: Application in Emails and Comparative Analysis," *International Journal of New Technology and Research*, vol. 7, no. 5, May 2021, doi: 10.31871/ijntr.7.5.9.

[18] Y. More, R. Gulage, P. Majage, Y. Patil, C. Shinde, and H. Solanke, "Study Of Different Algorithms Used For Text Summarization," *ALOCHANA JOURNAL*, vol. 13, no. 11, p. 295, 2024.

[19] S. C. Suvarna, M. Srivastava, B. Jaganathan, and P. Shukla, "PageRank Algorithm using Eigenvector Centrality- New Approach," *ArXiv*, Jan. 2022, doi: https://doi.org/10.48550/arXiv.2201.05469