Jurnal Rekayasa Sistem Informasi dan Teknologi
Volume 2, No 3 – Februari 2025
e-ISSN : 3025-888X

PERANCANGAN SISTEM PERINGKASAN ARTIKEL ILMIAH MENGGUNAKAN GROBID DAN *LARGE LANGGUAGE MODELS* (LLM) BERBASIS SCIBERT

Ellisa Mu'alifah[1*], Sam Farisa Chaerul Haviana[2]
Universitas Islam Sultan Agung, Semarang[1,2]
ellisamualifah@gmail.com

Received: 02-02-2025	Revised: 07-02-2025	Approved: 18-02-2025

ABSTRAK

Penelitian ini bertujuan untuk mengembangkan sistem tinjauan pustaka berbasis kecerdasan buatan yang memanfaatkan GROBID untuk ekstraksi informasi bibliografi dan SciBERT untuk menghasilkan ringkasan artikel ilmiah. Sistem ini dirancang sebagai aplikasi berbasis web menggunakan Streamlit, yang memungkinkan pengguna mengunggah artikel dalam format PDF dan mendapatkan rekomendasi literatur yang relevan. Metode penelitian yang digunakan meliputi studi literatur, pengumpulan data dari repositori jurnal ilmiah Garba Rujukan Digital (Garuda), serta pengolahan data menggunakan teknik pemrosesan bahasa alami (NLP). Data dalam format PDF dikonversi menjadi format JSON melalui GROBID untuk mengekstrak elemen penting, seperti judul, penulis, abstrak, dan isi artikel. Selanjutnya, SciBERT digunakan untuk melakukan peringkasan teks otomatis dengan metode ekstraktif yang menyoroti kalimat-kalimat utama dalam dokumen. Evaluasi sistem dilakukan menggunakan metrik ROUGE untuk mengukur kesamaan antara ringkasan sistem dan ringkasan manual, dengan hasil F1-score yang menunjukkan tingkat akurasi tinggi dalam mempertahankan esensi artikel. Hasil penelitian ini diharapkan dapat meningkatkan efisiensi dan akurasi dalam proses tinjauan pustaka, sehingga memberikan manfaat bagi para peneliti dalam menyaring informasi secara lebih efektif.

Kata Kunci: *GROBID, SciBERT, tinjauan pustaka, peringkasan teks otomatis, Streamlit, ROUGE*

PENDAHULUAN

Di era informasi saat ini, publikasi ilmiah Indonesia mengalami peningkatan dengan pesat, Indonesia menduduki peringkat pertama di ASEAN pada tahun 2019-2020 (Riyana, Mala, and Sutantri 2024). Hal ini menimbulkan tantangan bagi peneliti untuk melakukan tinjauan pustaka yang efektif dan efisien. Peneliti sering kesulitan dalam menyaring informasi relevan dari banyak literatur, yang bisa membuang waktu dan menghilangkan informasi penting. Peringkasan dokumen teks tersebut dapat dilakukan dengan dua cara yaitu ekstraktif dan abstraktif (Mhd. Ansor Lubis 1, Muhammad Yasin Ali Gea 2 2018)V. Teknologi seperti GROBID dan SciBERT dapat menjadi solusi inovatif untuk mengatasi permasalahan ini. GROBID mampu mengekstraksi data bibliografis dari artikel ilmiah dalam format PDF secara efisien. Sedangkan SciBERT, yang dirancang khusus untuk teks ilmiah, unggul dalam memahami terminologi dan struktur bahasa ilmiah. Studi menunjukkan bahwa integrasi kedua teknologi ini menghasilkan ringkasan yang lebih akurat dibandingkan metode tradisional (Hidayatullah et al. 2024).

Dalam era digital yang berkembang pesat, jumlah artikel ilmiah yang dipublikasikan meningkat secara signifikan setiap tahunnya (Fitria and Subakti 2022). Hal ini menyebabkan tantangan bagi akademisi, peneliti, dan praktisi dalam mengakses, memahami, serta menyaring informasi yang relevan dari literatur ilmiah yang sangat banyak (et al. 2023). Oleh karena itu, diperlukan solusi yang mampu mengekstrak informasi inti dari artikel ilmiah secara otomatis dan efisien.

Salah satu metode yang digunakan untuk merangkum teks secara otomatis adalah peringkasan berbasis kecerdasan buatan (Artificial Intelligence/AI) (Dhia Yusrana et al. 2024). Large Language Models (LLM), khususnya yang telah dilatih pada literatur ilmiah

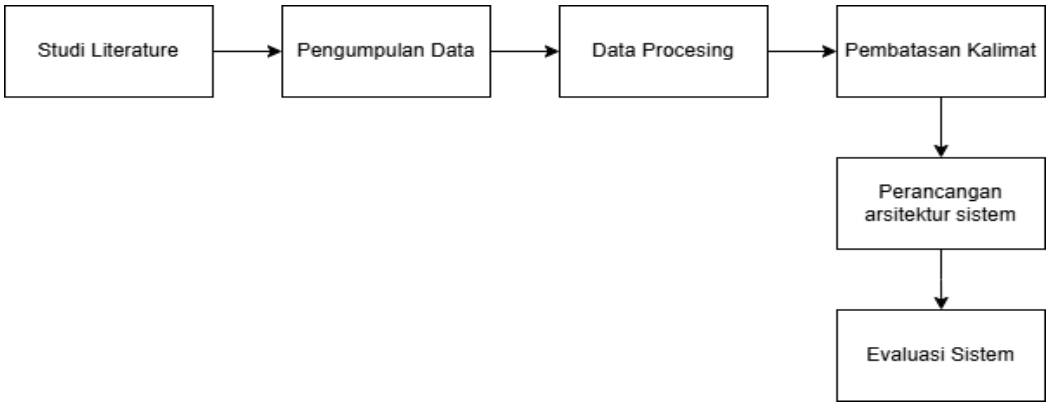
seperti **SciBERT**, memiliki kemampuan dalam memahami konteks dan struktur artikel ilmiah dengan lebih baik dibandingkan model bahasa umum. SciBERT adalah varian dari BERT (*Bidirectional Encoder Representations from Transformers*) yang dilatih secara khusus pada teks akademik, sehingga lebih akurat dalam memahami istilah dan struktur dalam artikel ilmiah. Selain itu, GROBID (GeneRation Of BIbliographic Data) merupakan alat open-source yang dapat digunakan untuk mengekstrak metadata dan struktur dokumen dari artikel ilmiah dalam format PDF (Novryzal and Kamilia 2024).

Berdasarkan penelitian sebelumnya, model transformator berbasis seperti SciBERT telah menunjukkan kinerja yang lebih baik dibandingkan dengan metode lain dalam menangani teks ilmiah. Model berbasis NLP seperti BERT dan SciBERT mampu meningkatkan akurasi peringkasan dan relevansi informasi dalam teks akademik dibandingkan dengan metode berbasis statistik tradisional (Kurniawan et al. 2023). Selain itu, algoritma LexRank mencapai nilai f-measure ROUGE-L sebesar 67,05% pada peringkasan teks otomatis artikel ilmiah berbahasa Indonesia. Model Language-independent Layout Transformer (LiLT) dan SciBERT dapat mencapai F1-score sebesar 94,6% dalam ekstraksi informasi biblio (Joshi et al. 2024).

Beberapa penelitian lain juga telah mengembangkan berbagai teknik peringkasan menggunakan algoritma seperti TF-IDF dan Maximum Marginal Relevance (MMR), yang menunjukkan hasil yang cukup baik dalam peringkasan teks berbahasa Indonesia. Ada penelitian yang menggunakan metode Latent Semantic Analysis (LSA) pada artikel berita ekonomi berbahasa Indonesia juga mencatat presisi sebesar 0,7916 dan akurasi 0,9015 pada tingkat kompresi 10%. Sementara itu, penelitian oleh Callegari dkk. (2023) menunjukkan bahwa model T5 Large adalah yang paling efektif dalam menghasilkan judul dari penelitian abstrak dengan skor ROUGE tertinggi dibandingkan model BART dan Flan T5. Dengan memanfaatkan kemampuan GROBID dan SciBERT, penelitian ini bertujuan untuk mengembangkan sistem peringkasan artikel ilmiah yang akurat dan relevan, sehingga mampu mengurangi beban kerja peneliti dan meningkatkan produktivitas akademik. Penelitian ini dibatasi pada artikel berformat PDF berbahasa Indonesia yang diambil dari Garuda, menggunakan metode peringkasan ekstraktif berbasis SciBERT dan bertujuan merancang aplikasi peringkasan otomatis yang efektif. Manfaat utama penelitian ini meliputi penghematan waktu, peningkatan aksesibilitas informasi, dan dukungan terhadap pengambilan keputusan berbasis data dalam penelitian ilmiah.

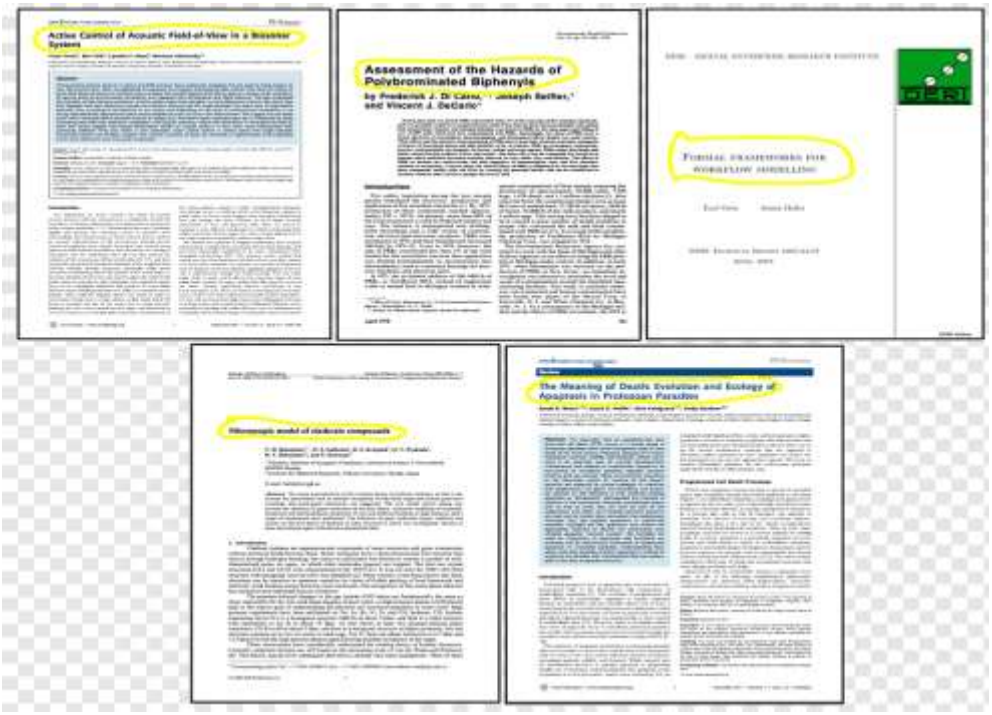
METODE PENELITIAN

Penelitian ini mengembangkan sistem tinjauan Pustaka menggunakan GROBID untuk mengekstrak informasi bibliografi dan SciBERT untuk menghasilkan konten artikel. Kemudian, pada tahap pengembangan aplikasi berbasis web, penulis akan memanfaatkan Streamlit untuk membangun antarmuka yang memungkinkan pengguna mengunggah artikel dan mendapatkan rekomendasi literatur yang relevan. Hasil dari penelitian ini diharapkan dapat menghasilkan sistem yang dapat membantu peneliti dalam proses tinjuan Pustaka dengan lebih efisien dan akurat.



Gambar 1. Tahapan Penelitian

Dengan menggunakan pendekatan ini, sistem akan memanfaatkan kemampuan GROBID dalam mengekstrak data bibliografi dan kemampuan SciBERT dalam memahami teks ilmiah, sehingga dapat memberikan rekomendasi literatur yang lebih tepat dan relevan bagi pengguna. Studi literatur dilakukan untuk memahami konsep dasar peringkasan teks, berbagai algoritma yang tersedia, serta aplikasi GROBID dan SciBERT dalam konteks peringkasan artikel ilmiah. Penelitian terdahulu menunjukkan bahwa kombinasi metode berbasis machine learning dan NLP dapat menghasilkan ringkasan yang efektif. Pengumpulan data dalam penelitian ini dilakukan dengan mengambil artikel dalam format PDF dari situs web Garba Rujukan Digital (Garuda), yang merupakan repositori jurnal ilmiah di Indonesia. Artikel yang dipilih berbahasa Indonesia dan memiliki topik yang relevan dengan bidang penelitian di teknik informatika. Pemilihan artikel didasarkan pada kesesuaian topik, kualitas jurnal, dan keterbaruan publikasi untuk memastikan data yang digunakan memiliki validitas akademik yang tinggi [17].



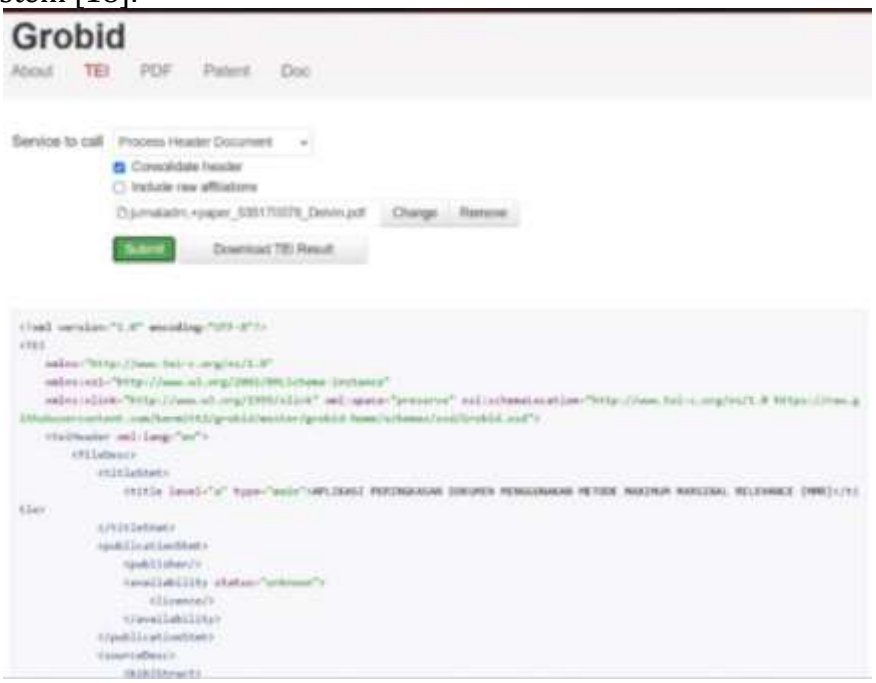
Gambar 2. Tata Letak artikel ilmiah

Gambar 2 menunjukkan contoh tata letak artikel ilmiah yang terorganisasi, dengan judul yang jelas, abstrak ringkas, dan struktur isi yang sistematis. Penyesuaian tata letak ini memudahkan pembaca dalam memahami inti penelitian dan menemukan informasi penting secara efisien.

Data Preprocessing

- Fitur Upload PDF

  Sistem memanfaatkan GROBID untuk mengubah file PDF artikel ilmiah menjadi data terstruktur dalam format JSON. File PDF yang diunggah menggunakan dikirim ke GROBID melalui API, di mana elemen penting seperti judul, penulis, abstrak, dan bagian teks utama diekstraksi. Hasil ekstraksi ini kemudian disimpan pada format JSON yang memuat informasi terstruktur dan mudah diproses lebih lanjut, seperti untuk pembuatan ringkasan otomatis atau analisis data. Proses ini memastikan dokumen PDF dapat diolah secara efisien oleh sistem [18].



Gambar 3. Konversi JSON dari File PDF

GROBID mengonversi teks PDF menjadi format JSON melalui API dengan akurasi identifikasi paragraf mencapai 98%, berkat model machine learning yang dilatih khusus untuk dokumen akademik. Hal ini memungkinkan sistem menghasilkan data yang akurat dan relevan, sehingga mendukung efisiensi dalam pemrosesan dokumen ilmiah.

- Peringkasan Text Otomatis (*Automated text summarization*)

  Peringkasan teks otomatis menggunakan SciBERT, yang dirancang khusus untuk pemrosesan bahasa alami dalam konteks teks ilmiah. SciBERT menganalisis dokumen penelitian dengan memahami terminologi dan struktur kompleksnya, kemudian menghasilkan ringkasan yang berisi poin-poin utama. Proses ini meliputi tokenisasi teks dan encoding konteks, dengan fokus pada metode ekstraktif yang memilih kalimat-kalimat penting dari dokumen. Dengan pelatihan berbasis literatur ilmiah, SciBERT memberikan hasil yang relevan dan akurat, mempermudah peneliti dalam menyaring informasi penting secara efisien [19].

Pembatasan Kalimat

Peringkasan artikel ilmiah bertujuan untuk menjaga ringkasan tetap padat, relevan, dan mudah dipahami tanpa mengurangi informasi penting. Umumnya, panjang ringkasan berkisar antara 10% hingga 30% dari teks asli, mencakup poin utama seperti pendahuluan, metode, hasil, dan kesimpulan. Untuk pengaturan kata dalam ringkasan, disarankan agar ringkasan memiliki maksimal 50 kata per-paragraf, tergantung pada panjang artikel asli. Dengan demikian, peneliti dapat menyesuaikan panjang ringkasan sesuai kebutuhan sambil memastikan bahwa informasi kunci tetap tersampaikan jelas.

Gambar 4. Perancanngan alur kerja system

Proses pelatihan melibatkan beberapa tahap utama:

- Preprocessing Teks dan Ekstraksi Data: File PDF diunggah dan diproses menggunakan GROBID untuk mengekstrak teks mentah. Teks ini kemudian melalui tahap preprocessing, seperti normalisasi dan pembersihan data.
- Pelatihan Model: Teks hasil preprocessing diringkas menggunakan model SciBERT. Model ini dilatih untuk menghasilkan ringkasan yang relevan dan akurat dari teks ilmiah.
- Validasi Model: Hasil ringkasan dievaluasi untuk memastikan kualitas dan akurasi. Proses validasi dilakukan dengan menyesuaikan parameter hingga mencapai performa optimal.

Evaluasi Sistem

Pengujian dilakukan menggunakan metode evaluasi *Recall-Oriented Understudy for Gisting Evaluation* (ROUGE). ROUGE menghitung jumlah n-gram kata yang tumpeng tindih (*overlap)* antara ringkasan sistem dan ringkasan referensi. Adapun teknik penghitungan ROUGE-N antara ringkasan sistem dan sekumpulan ringkasan manual dapat dilihat pada persamaan beriut :

$$ROUGE - N = \frac{\sum S \in \{ReferenceSummaries\}. \in S \sum Countmatch(gram)n}{\sum S \in \{ReferenceSummaries\}.gram \in S \sum Count(gram)n} \quad (1)$$

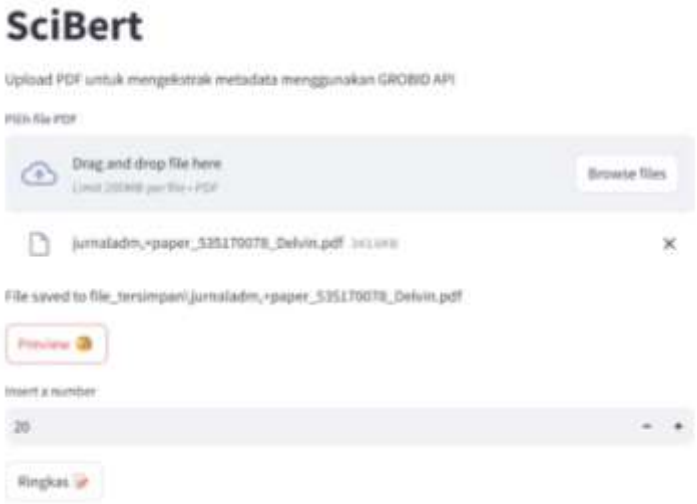
Di mana N adalah panjang dari N-gram, *Countmatch*(gramn) adalah jumlah N-gram yang sama antara ringkasan sistem dan ringkasan referensi, *Count*(gramn) adalah jumlah N-gram dalam ringkasan referensi [20].

HASIL PENELITIAN DAN PEMBAHASAN

Setelah mengumpulkan data artikel ilmiah yang dibutuhkan sebagai bahan uji sistem, selanjutnya dilakukan implementasi pada sistem peringkasan artikel ilmiah. Berikut ini merupakan hasil tangkapan layar hasil implementasi sistem yang menggunakan teknologi Grobid untuk ekstraksi metadata dan teks, serta *Large Language Models* (LLM) berbasis SciBERT untuk peringkasan konten artikel ilmiah.

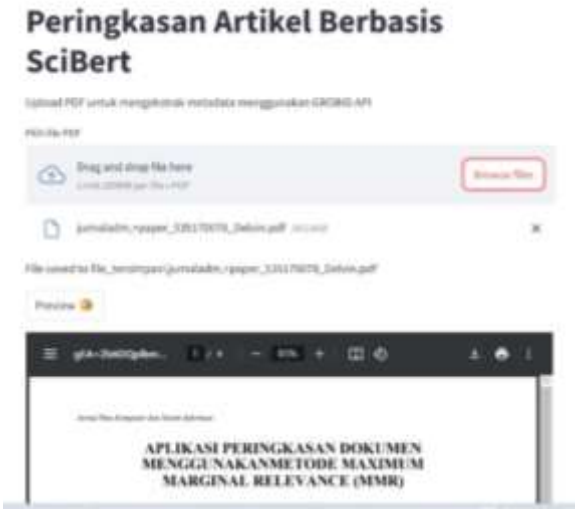
Gambar 5. Halaman utama aplikasi

Gambar 5 menunjukkan tampilan utama aplikasi "Peringkasan Artikel Berbasis SciBERT." Aplikasi ini memungkinkan pengguna mengunggah file PDF untuk diekstrak metadatanya menggunakan GROBID dan diringkas dengan SciBERT. Pengguna dapat memilih file melalui tombol "*Browse files*". Terdapat *input* numerik untuk mengatur panjang ringkasan sebelum menekan tombol "Ringkas" untuk memulai proses.



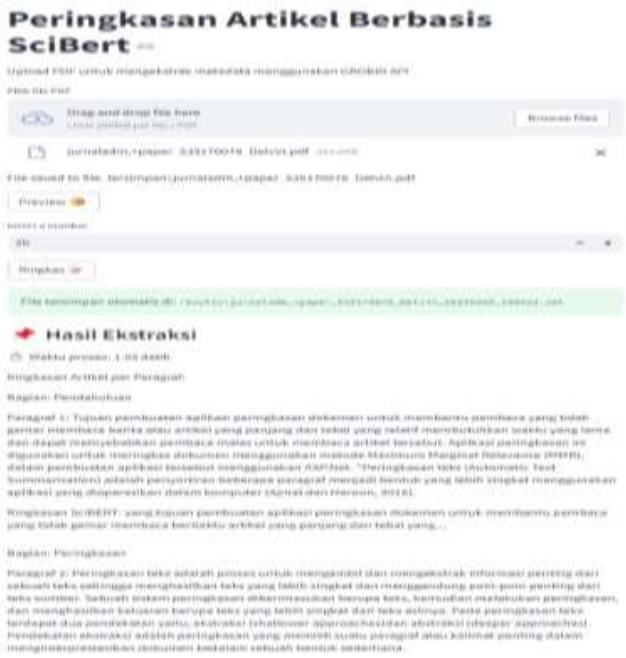
Gambar 6. Tampilan mengunggah file PDF

Pada gambar 6 terlihat tampilan halaman untuk mengunggah file PDF menggunakan sistem "PDF Metadata Extractor using GROBID." Pengguna dapat memilih file PDF dengan tombol "*Browse files*". Setelah file dipilih, sistem secara otomatis menyimpan file tersebut secara otomatis dalam format XML di folder hasil ekstraksi. Pengguna juga dapat meninjau dokumen dengan tombol "*Preview*" sebelum melanjutkan ke proses peringkasan.



Gambar 7. Tampilan ketika "Klik preview"

Pada gambar 7 Setelah pengguna mengklik tombol "*Preview*", tampilan dokumen PDF yang telah diunggah akan muncul di dalam aplikasi. Pengguna dapat melihat isi dokumen secara langsung sebelum melanjutkan ke proses peringkasan. Fitur ini memungkinkan pengguna untuk memverifikasi bahwa file yang diunggah sudah benar sebelum diproses lebih lanjut oleh sistem.



Gambar 8. Halaman Ringkasan artikel

Pada Gambar 8 terlihat tampilan hasil ekstraksi artikel yang dihasilkan oleh aplikasi berbasis Streamlit setelah pengguna mengklik tombol "Ringkas". Sistem memproses dokumen PDF dan menampilkan hasil ekstraksi dalam bentuk ringkasan per paragraf, mencakup informasi seperti waktu pemrosesan, jumlah kata dalam ringkasan, serta isi ringkasan yang dihasilkan menggunakan model SciBERT. Ringkasan ini mencakup berbagai bagian artikel, seperti pendahuluan, metode penelitian, hasil dan pembahasan, serta kesimpulan. Dengan menggunakan model SciBERT, sistem menghasilkan ringkasan yang lebih singkat dan padat tanpa menghilangkan makna penting dari paragraf asli, membantu pengguna memahami inti artikel dengan lebih efisien.

Setelah mengimplementasikan sistem peringkasan artikel ilmiah, evaluasi dilakukan menggunakan metrik ROUGE. Pengujian dilakukan dengan membandingkan 10 paragraf hasil ringkasan manual dan 10 paragraf hasil ringkasan sistem. Berikut hasil evaluasi menggunakan metrik ROUGE:

Tabel 1.
Hasil evaluasi representasi kualitas ringkasan

Metode ROUGE	Precision	Recall	F1-Score
ROUGE-1	0.9225	0.9261	0.9243
ROUGE-2	0.8560	0.8594	0.8577
ROUGE-L	0.9225	0.9261	0.9243

ROUGE-1 Mengukur kemiripan kata individu antara ringkasan sistem dan ringkasan manual, dengan F1-Score sebesar 0.9243. Ini menamdakan bahawa sistem memiliki tingkat akurasi yang sangat tinggi dalam menangkap kata-kata penting dari dokumen asli. ROUGE -2 mengukur kemiripan pasangan kata (*bigrams*), dengan nilai F1- Score sebesar 0.8577. Meskipun nilai ini sedikit lebih rendah dibandingkan ROUGE-1, hal ini menunjukan bahawa sistem cukup baik dalam mempertahanlan konteks lokal anatar kata. ROUGE -L mengukur kesamaan urutan kata terpanjang (longest common subsequence), juga menghasilkan nilai F1-Score sebesar 0.9243, sama seperti ROUGE-1. Ini menunjukan bahwa sistem sangat baik dalam mempertahanlan struktur kalimat atau urutan kata yang relevan. Secara keseluruhan nilai F1-Score, kualitas ringkasan otomatis dinilai sangat baik.

KESIMPULAN

Penelitian ini menunjukkan bahwa sistem peringkasan artikel ilmiah berbasis teknologi GROBID dan SciBERT dapat memberikan solusi efektif dalam mempermudah proses tinjauan pustaka bagi peneliti. Dengan menggunakan GROBID untuk mengekstraksi data bibliografi dan SciBERT untuk menganalisis serta merangkum konten artikel ilmiah, sistem ini mampu menghasilkan ringkasan yang lebih akurat dan relevan dibandingkan dengan metode tradisional. Pengujian sistem menggunakan metrik ROUGE menunjukkan bahwa sistem ini memiliki kualitas ringkasan yang sangat baik, dengan nilai F1-Score yang tinggi untuk setiap kategori. Implementasi sistem ini diharapkan dapat menghemat waktu peneliti dan meningkatkan efisiensi dalam mengakses informasi ilmiah, serta mendukung pengambilan keputusan berbasis data dalam penelitian.

DAFTAR PUSTAKA

Dhia Yusrana, Rafif et al. 2024. "Implementasi Kecerdasan Buatan Dalam Pengembangan Aplikasi Mobile Binarytalkhub." *JATI (Jurnal Mahasiswa Teknik Informatika)* 8(4): 6075–81.

Fitria, Wida, and Ganjar Eka Subakti. 2022. "Carenzino, Ikhsan, Edo Galasro Limbong, and Duane Masaji Raharja. "Motion Comic Pengenalan Ilmuwan Muslim Abbas Ibnu Firnas." *Jurnal Penelitian Keislaman* 18(2): 143–57.

Hidayatullah, Naufal Mufadhdhal et al. 2024. "Analisis Bibliometrik : Penelitian Technology Acceptance Model Tahun 2014-2023 Menggunakan Bibliometrik Dan Vosviewer." *Comdent: Communication Student Journal* 2(1): 138–58.

Joshi, Bikash, Anthi Symeonidou, Syed Mazin Danish, and Floris Hermsen. 2024. "An End-to-End Pipeline for Bibliography Extraction from Scientific Articles." : 101–6.

Kurniawan, Moh Heri et al. 2023. "Artificial Intelligence (AI) Dalam Pelayanan Keperawatan: Studi Literatur Artificial Intelligence (AI) in Nursing Services: A Literature Review." *Faletehan Health Journal* 10(1): 77–84. www.journal.lppm-stikesfa.ac.id/ojs/index.php/FHJ.

Mhd. Ansor Lubis 1, Muhammad Yasin Ali Gea 2, & Nur Muniifah. 2018. "Akibat Hukum Perjanjian Perdagangan Bebas Di Asean Indonesia-Malaysia." *Jurnal Ilmiah Penegakan Hukum* 5(2): 94–100.

Novryzal, Andress, and Nada Kamilia. 2024. "Pembangunan Aplikasi Whatsapp Gateway Official Dumas Menggunakan Artificial Intelligence (Studi Kasus Unit Kerja Asdep Dumas , Kemensetneg)."

Restiana Restiana, and Retno Sayekti. 2023. "Memahami Tren Penelitian Artificial Intelligence Di Perpustakaan Melalui Analisis Bibliometrik Pada Publikasi Ilmiah Internasional Tahun 2019-2023." *UNILIB : Jurnal Perpustakaan* 14(2): 83–93.

Riyana, Dhea Isti, Iva Khoiril Mala, and Sutantri Sutantri. 2024. "Peran Ekonomi Digital Terhadap Kinerja Pasar Modern Di Indonesia." *E-Bisnis: Jurnal Ilmiah Ekonomi dan Bisnis* 17(1): 23–31.