PERANCANGAN SISTEM PERINGKASAN ARTIKEL ILMIAH UNTUK MEMBANTU PROSES TINJUAN PUSTAKA MENGGUNAKAN GROBID DAN *LARGE LANGUAGE MODELS* (LLM) BERBASIS SciBERT

PROPOSAL TUGAS AKHIR

Laporan ini Disusun untuk Memenuhi Salah Satu Syarat Memperoleh Gelar Sarjana Strata 1 (SI) Program Studi Teknik Informatika Fakultas Teknologi Industri Universitas Islam Sultan Agung Semarang

Disusun Oleh :

Ellisa Mu'alifah

NIM 32602100041

PROGRAM STUDI TEKNIK INFORMATIKA

FAKULTAS TEKNOLOGI INDUSTRI

UNIVERSITAS ISLAM SULTAN AGUNG

SEMARANG

2025

FINAL PROJECT

DESIGN OF A SCIENTIFIC ARTICLES SUMMARIZING SYSTEM TO HELP THE LITERATURE REVIEW PROCESS USING GROBID AND LARGE LANGUAGE MODELS (LLM) BASED ON SciBERT

Proposed to complete the requirement to obtain a bachelor's degree (S1)
at Informatics Engineering Departement of Industrial Technology Faculty
Sultan Agung Islamic University

Arranged By :

Ellisa Mu'alifah
NIM 32602100041

MAJORING OF INFORMATICS ENGINEERING
INDUSTRIAL TECHNOLOGY FACULTY
SULTAN AGUNG ISLAMIC UNIVERSITY
SEMARANG
2025

LEMBAR PENGESAHAN TUGAS AKHIR

PERANCANGAN SISTEM PERINGKASAN ARTIKEL ILMIAH UNTUK MEMBANTU PROSES TINJUAN PUSTAKA MENGGUNAKAN GROBID DAN *LARGE LANGGUAGE MODELS* (LLM) BERBASIS SciBERT

ELLISA MU'ALIFAH
NIM 32602100041

Telah dipertahankan di depan tim penguji ujian sarjana tugas akhir
Program Studi Teknik Informatika
Universitas Islam Sultan Agung
Pada tanggal : 25 februari 2025

TIM PENGUJI UJIAN SARJANA :

Mustafa, ST,MM., M.Kom
NIK. 2106100040
(Ketua Penguji) 07/03/2025

Arief Marwanto, ST., M.Eng., Ph.D
NIK. 210600018
(Anggota Penguji) 06/3/2025

Sam Farisa C.H., ST, M.Kom
NIK. 210615046
(Pembimbing) 10/3/2025

Semarang, 11 Maret 2025
Mengetahui,

Kaprodi Teknik Informatika
Universitas Islam Sultan Agung

Moch. Taufik, ST, MIT
NIK. 210604034

SURAT PERNYATAAN KEASLIAN TUGAS AKHIR

Yang bertanda tangan dibawah ini :

Nama	:	Ellisa Mu'alifah
NIM	:	32602100041
Judul Tugas Akhir	:	PERANCANGAN SISTEM PERINGKASAN ARTIKEL ILMIAH UNTUK MEMBANTU PROSES TINJUAN PUSTAKA MENGGUNAKAN GROBID DAN *LARGE LANGUAGE MODELS* (LLM) BERBASIS SciBERT

Dengan bahwa ini saya menyatakan bahwa judul dan isi Tugas Akhir yang saya buat dalam rangka menyelesaikan Pendidikan Strata Satu (S1) Teknik Informatika tersebut adalah asli dan belum pernah diangkat, ditulis ataupun dipublikasikan oleh siapapun baik keseluruhan maupun sebagian, kecuali yang secara tertulis diacu dalam naskah ini dan disebutkan dalam daftar pustaka, dan apbila di kemudian hari ternyata terbukti bahwa judul Tugas Akhir tersebut pernah diangkat, ditulis ataupun dipublikasikan, maka saya bersedia dikenakan sanksi akademis. Demikian surat pernyataan ini saya buat dengan sadar dan penuh tanggung jawab.

Semarang, 11 Maret 2025

Yang Menyatakan,

Ellisa Mu'alifah

KATA PENGANTER

Dengan mengucap syukur Alhamdulillah atas kehadirat Allah swt atas limpahan rahmat dan karunia-Nya kepada penulis, sehingga penulis dapat menyelesaikan tugas akhir yang berjudul "Perancangan sistem peringkasan artikel ilmiah untuk membantu proses tinjuan Pustaka menggunakan GROBID dan *large language models* (LLM) berbasis SciBERT" Tugas akhir ini disusun sebagai salah satu syarat untuk memperoleh gelar sarjana strata 1 (S1) diprogram studi teknik informatika, fakultas teknologi industri, universitas islma sultan agung semarang.

Penulisan tugas akhir ini tidak epas dari dukungan dari berbagai pihak. Oleh karena itu, saya ingin menyampaikan ucapan terima kasih yang sebesar-besarnya kepada :

1. Rektor UNISSULA Bapak Prof. Dr. H. Gunarto, S.H., M.H yang mengizinkan penulis menimba ilmu di kampus ini.
2. Dekan Fakultas Teknologi Industri Ibu Dr. Novi Marlyana, S.T., M.T.
3. Dosen pembimbing I penulis Sam Farisa Chaerul Haviana, S.T., M.Kom yang telah meluangkan waktu dan memberi ilmu. Serta memberikan banyak nasehat dan Saran.
4. Orang tua penulis, Bapak Sutadi dan Ibu Aminatun, yang selalu memberikan restu, dukungan, serta Doa dalam menyelesaikan Tugas Akhir ini.
5. Para sahabat, teman-teman yang telah memberikan begitu banyak bantuan, semangat, inspirasi dan diskusi progres penyusunan Tugas Akhir.

Saya menyadari bahwa skripsi ini masih jauh dari sempurna. Oleh karena itu, saya mengharapkan kritik dan saran yang membangun bagi para pembaca. Semoga skripsi ini dapat memberikan manfaat bagi perkembangan ilmu pengetahuan dan memberikan inspirasi bagi para pembaca.

Semarang, 4 Maret 2025

Ellisa Mu'alifah

DAFTAR ISI

DAFTER GAMBAR

DAFTAR TABEL

ABSTRAK

Peringkasan teks otomatis atau *automated text summarization* merupakan metode yang efektif dalam mengekstrak inti dari dokumen teks guna mempercepat pemahaman informasi. Penelitian ini mengembangkan sistem peringkasan artikel ilmiah dengan integrasi GROBID dan SciBERT untuk mempercepat ekstraksi informasi relevan serta mengurangi waktu dan potensi hilangnya informasi penting dalam tinjuan pustaka. Penelitian ini bertujuan untuk meningkatkan efisiensi peneliti dalam melakukan tinjauan pustaka, menghemat waktu, serta mendukung pengambilan keputusan berbasis data. Metode yang digunakan meliputin ekstraksi data bibliografi menggunakan GROBID, Peringkasan teks ilmiah dengan SciBERT menggunakan metode ekstraktif dan Evaluasi menggunakan metrik ROUGE untuk menilai akurasi hasil peringkasan. Hasil penelitian menunjukkan bahwa integrasi kedua teknologi ini mampu menghasilkan ringkasan yang akurat dan relevan. Evaluasi dengan ROUGE menghasilkan skor *F1-score* sebesar 0.8084 untuk ROUGE-1, 0.6051 untuk ROUGE-2, dan 0.6184 untuk ROUGE-L, yang menunjukkan efektivitas model dalam merangkum teks ilmiah. Dengan demikian, sistem ini dapat mempercepat penelusuran literatur dan meningkatkan produktivitas penelitian ilmiah.

Kata Kunci: Peringkasan Teks Otomatis, GROBID, SciBERT, Evaluasi ROUGE

ABSTRACT

Automatic text summarization is an effective method for extracting the essence of text documents to speed up understanding of information. This research develops a scientific article summarization system with the integration of GROBID and SciBERT to speed up the extraction of relevant information and reduce the time and potential loss of important information in literature reviews. This research aims to increase researcher efficiency in conducting library observations, save time, and support data-based decision making. The methods used include bibliographic data extraction using GROBID, scientific text summarization with SciBERT using extractive methods and evaluation using the ROUGE metric to assess the accuracy of the summarization results. The research results show that the integration of these two technologies is able to produce accurate and relevant summaries. Evaluation with ROUGE produces an F1-score of 0.8084 for ROUGE-1, 0.6051 for ROUGE-2, and 0.6184 for ROUGE-L, which shows the effectiveness of the model in summarizing scientific texts. Thus, this system can speed up literature searches and increase scientific research productivity.

Keywords: Automatic Text Summarization, GROBID, SciBERT, Evaluation ROUGE

BAB I
PENDAHULUAN

1.1 Latar Belakang

Di era informasi saat ini, Peningkatan publikasi ilmiah Indonesia pada jurnal ilmiah mengalami peningkatan pesat. Pada tahun 2019-2020, Indonesia bahkan berada pada peringkat 1 di ASEAN (Asy'ari dkk., 2022). Hal ini menimbulkan tantangan bagi peneliti untuk melakukan tinjauan pustaka yang efektif dan efisien. Peneliti seringkali menghadapi kesulitan dalam menyaring informasi relevan dari tumpukan literatur yang ada, yang dapat mengakibatkan terbuangnya waktu dan potensi hilangnya informasi penting. Untuk mengatasi masalah ini, peringkasan informasi menjadi solusi yang bermanfaat. Peringkasan dokumen teks tersebut dapat dilakukan dengan 2 cara, yaitu peringkasan dokumen secara ekstraktif (*Extractive summarization*) dan peringkasan dokumen secara abstraktif (*Abstractive summarization*) (Yuliska & Syaliman, 2020). Akses terhadap informasi ilmiah yang berkualitas menjadi semakin penting untuk mendukung inovasi dan pengembangan ilmu pengetahuan Oleh karena itu, diperlukan solusi yang dapat membantu peneliti mengelola dan merangkum informasi dari artikel ilmiah dengan lebih baik.

Seiring dengan meningkatnya jumlah publikasi, tantangan ini semakan mendesak, sehingga ada kebutuhan mendesak untuk teknologi yang dapat memproses dan merangkum informasi dengan cepat. Dalam beberapa tahun terakhir, banyak penelitian telah dilakukan untuk mengembangkan sistem peringkasan otomatis. Meskipun beberapa metode tradisional menggunakan algoritma statistik dan teknik pemrosesan bahasa alami dasar, namun sering kali hasilnya tidak memadai dalam hal akurasi dan relevansi. Di sinilah teknologi seperti GROBID (Farisa & Haviana, 2019) dan SciBERT (Maheshwari dkk., 2021) menjadi relevan. GROBID (Farisa & Haviana, 2019) adalah alat yang sangat efisien dalam mengekstraksi data bibliografis dari artikel ilmiah dalam format PDF, memungkinkan identifikasi elemen-elemen penting seperti judul, abstrak, dan penulis. Studi sebelumnya menunjukkan bahwa GROBID dapat secara

signifikan mengurangi waktu yang dibutuhkan untuk mengekstraksi informasi dasar dari dokumen yang kompleks.

Sementara itu, SciBERT (Maheshwari dkk., 2021), yang dibangun di atas arsitektur BERT, dirancang khusus untuk menangani teks ilmiah. Berbeda dengan model bahasa umum, SciBERT dilatih pada korpus ilmiah, memungkinkan pemahaman yang lebih baik terhadap terminologi dan struktur bahasa di bidang ini. Penelitian sebelumnya menunjukkan bahwa SciBERT unggul dalam tugas-tugas seperti klasifikasi teks dan peringkasan dalam konteks ilmiah. Beberapa penelitian telah mengeksplorasi penggunaan teknologi *Natural Language Processing* (NLP) untuk meningkatkan efisiensi wawasan perpustakaan, dan model berbasis transformer seperti BERT dan SciBERT terbukti dapat menghasilkan ringkasan yang lebih akurat dan relevan dibandingkan metode tradisional (Pearce dkk., 2021). Selain itu, otomatisasi pada observasi perpustakaan dapat mengurangi beban kerja peneliti dan meningkatkan kualitas hasil penelitian.

Dengan mengintegrasikan GROBID dan SciBERT, penelitian ini bertujuan untuk mengembangkan sistem peringkasan artikel yang lebih efektif dan efisien. Sistem ini diharapkan dapat memberikan ringkasan yang akurat dan relevan, mengurangi beban kerja para peneliti, dan meningkatkan efisiensi dalam proses tinjauan pustaka. Pendekatan ini tidak hanya menawarkan solusi inovatif untuk tantangan saat ini tetapi juga membuka jalan bagi pengembangan alat yang lebih canggih di masa depan. Manfaat yang diharapkan dari penelitian ini meliputi penghematan waktu dan tenaga, peningkatan aksesibilitas informasi, dan dukungan yang lebih baik untuk pengambilan keputusan berbasis data dalam penelitian ilmiah. Dengan demikian, solusi ini dapat berkontribusi secara signifikan terhadap produktivitas dan kualitas penelitian akademik secara keseluruhan.

1.2 Perumusan Masalah

Bagaimana mengembangkan sistem peringkasan artikel ilmiah dengan menggunakan integrasi GROBID dan SciBERT untuk mempercepat ekstraksi informasi relevan dalam tinjauan pustaka, serta mengurangi terbuangnya waktu dan potensi hilangnya informasi penting?

1.3 Pembatasan Masalah

Batasan masalah ini bertujuan untuk memudahkan dan menghindari adanya kegiatan di luar sasaran, sehingga dalam pembuatan laporan ini perlu ditentukan suatu batasan masalah. Batasan masalah tersebut sebagai berikut :

1. Sistem hanya akan memproses artikel ilmiah dalam format PDF yang bisa di proses oleh GROBID, utamanya jurnal dan laporan penelitian jurnal. Dokumen dalam format lain (misalnya, Word atau gambar) tidak akan digunakan.
2. Sistem ini terbatas pada artikel ilmiah yang ditulis dalam bahasa indonesia. Penelitian tidak akan mencakup artikel dalam bahasa lain untuk menjaga konsistensi.

1.4 Tujuan

Tujuan penelitian ini adalah mengambangkan aplikasi peringkasan artikel ilmiah menggunakan integrasi GROBID dan SciBERT, serta melakukan validasi untuk memastikan aplikasi tersebut efektif dalam mengurangi waktu yang dibutuhkan dalam proses tinjuan Pustaka dan meningkatkan efisiensi ektraksi informasi relevan.

1.5 Manfaat

Manfaat penelitian ini mengembangkan sistem peringkasan artikel ilmiah dengan GROBID dan SciBERT untuk menghemat waktu, meningkatkan akses informasi relevan, dan mendukung pengembalian keputusan berbasis data, serta berkontribusi pada efisiensi proses tinjuan Pustaka dalam penelitian ilmiah.

1.6 Sistem Penulisan

Sistematika penulisan yang akan digunakan oleh penulis dalam sebuah pembuatan laporan tugas akhir adalah sebagai berikut:

BAB I PENDAHULUAN

Bab pertama tugas akhir ini akan berisikan latar belakang yang membahas dan menjelasakan mengenai urgensi dari peneliti. Pembahasan dan penjelasan

tersebut dibagi menjadi beberapa bagian yaitu, latar belakang, perumusan masalah, pembatasan masalah, tujuan, manfaat, dan sistematika penulisan.

BAB II TINJUAN PUSTAKA DAN DASAR TEORI

Bab kedua tugas akhir ini akan berisikan tinjuan pustaka dan dasar teori. Tinjuan bab ini adalah untuk menunjukan Pustaka penelitian sebelumnya dan dasar teori yang akan digunakan pada penelitian.

BAB III METODE PENELITIAN

Pada penelitian ini, bab ketiga akan membahas bagaimana metode yang akan digunakan pada penelitian. Termasuk perancangan kebutuhan yang berupa Flowchat dan user interface.

BAB IV HASIL DAN ANALISIS PENELITIAN

Bab ini menyajikan hasil implementasi sistem peringkasan, mencakup kinerja GROBID dalam mengekstraksi elemen dokumen, kualitas ringkasan yang dihasilkan oleh SciBERT.

BAB V KESIMPULAN DAN SARAN

Bab ini menuat kesimpulan dan saran yang dibuat setelah membahas dan melakukan penelitian.

BAB II
TINJAUAN PUSTAKA DAN DASAR TEORI

2.1 Tinjauan Pustaka

Dalam peneliti (Halimah dkk., 2022) melakukan peringkasan teks otomatis pada artikel berbahasa Indonesia menggunakan algoritma LexRank. Penelitian ini menggunakan dataset 300 artikel yang dibagi menjadi dua bagian (150 untuk pengembangan sistem dan 150 untuk pengujian) dan melibatkan beberapa tahap, termasuk pra pemrosesan teks, perhitungan bobot tf-idf, pembentukan graf, dan pemeringkasan kalimat. Hasil dari penelitian ini menunjukkan bahwa dengan tingkat kompresi 50%, nilai *f-measure* untuk metrik ROUGE-1, ROUGE-2, dan ROUGE-L masing-masing adalah 67,53%, 59,10%, dan 67,05%. Sedangkan untuk tingkat kompresi 30%, nilai rata-rata *f-measure* adalah 55,82%, 45,51%, dan 54,76%. Model berbasis LexRank menunjukkan performa yang baik dengan rata-rata ROUGE-L di atas 50%.

Pada penelitian sebelumnya yang berjudul "*An End-to-End Pipeline for Bibliography Extraction from Scientific Articles*" Penelitian ini menghasilkan sistem yang berfungsi untuk mengekstraksi informasi bibliografi lengkap dari artikel ilmiah dalam format PDF digital dan membagi informasi tersebut menjadi kutipan individu. Metode yang digunakan melibatkan model multimodel bernama *Language-independent Layout Transformer* (LiLT) untuk mendeteksi bibliografi dan versi SciBERT yang telah disesuaikan untuk membagi kutipan. Pipeline ini mencapai *F1-score* sebesar 94.6%, mengungguli alat lain seperti GROBID, dan dirancang untuk menangani tantangan format dan tata letak yang beragam, serta dapat digunakan untuk dokumen multibahasa dan berbasis gambar (Joshi dkk., 2023).

Pada penelitian yang dilakukan oleh (And & Expert, 2021) yang bertujuan untuk membangun sistem peringkasan teks otomatis untuk makalah ilmiah menggunakan metode *Term Frekuensi - Inverse Document Frekuensi* (TF-IDF) dan *Maximum Marginal Relevan* (MMR), Metode yang digunakan meliputi *preprocessing* data seperti segmentasi, case lipat, tokenisasi, dan penghapusan

stopwords, diikuti dengan penerapan TF-IDF dan MMR untuk memilih kalimat yang relevan. Dari hasil penelitian menunjukkan bahwa sistem peringkasan otomatis yang menggunakan MMR mencapai akurasi terbaik pada metrik ROUGE, dengan nilai tertinggi 64% untuk ROUGE-1.

Pada penelitian sebelumnya yang dilakukan oleh (Hendry dkk., 2023) berhasil membuat sistem peringkasan teks otomatis untuk artikel berita ekonomi berbahasa Indonesia menggunakan metode *Latent Semantic Analysis* (LSA). Sistem ini menerapkan pendekatan aljabar linear *Singular Value Decomposition* (SVD) untuk mengurangi noise dan mengekstrak kalimat-kalimat yang relevan dari dokumen. Hasil penelitian menunjukkan bahwa pada tingkat kompresi 10%, sistem ini mencapai nilai presisi 0,7916 dan akurasi 0,9015. Pada tingkat kompresi 30%, nilai rata-rata presisi, *Recall*, *f-measure*, dan akurasi adalah 0.475, 0.171828, 0.2264444, dan 0.787366.

Pada penelitian sebelumnya yang dilakukan (Callegari dkk., 2023) penelitian ini menggunakan metode BART, T5, dan Flan T5 digunakan untuk menghasilkan judul akademik dengan ringkasan teks. T5 dan Flan T5 mengubah tugas NLP menjadi masalah teks-ke-teks, sementara Flan T5 menunjukkan kinerja yang lebih baik dibandingkan T5. BART, dengan arsitektur *auto-regresif* dan *auto-encoding*, menghasilkan judul yang koheren dan relevan dengan konten abstrak. Penelitian ini menunjukkan bahwa hasil model T5 Large adalah yang terbaik untuk menghasilkan judul dari abstrak penelitian. Penilaian dilakukan menggunakan skor ROUGE dan evaluasi manusia menunjukkan bahwa judul yang dihasilkan oleh mesin kadang-kadang lebih baik daripada judul asli, menyoroti potensi generasi judul otomatis dalam publikasi akademik.

Berdasarkan berbagai penelitian, dapat disimpulkan bahwa model SciBERT sangat cocok diterapkan dalam perancangan sistem peringkasan artikel ilmiah untuk membantu proses tinjauan pustaka. SciBERT, yang dirancang khusus untuk korpus ilmiah, menunjukkan keunggulan dalam menangani relasi yang spesifik dalam domain ilmiah dengan peningkatan kinerja rata-rata sebesar 2,63% dibandingkan model BERT pada tugas ekstraksi relasi. Model ini dapat diintegrasikan dengan GROBID, yang sudah terbukti unggul dalam ekstraksi

bibliografi, untuk membangun pipeline berbasis *Large Language Model* (LLM) yang efisien dan relevan. Pendekatan ini dapat meningkatkan akurasi dan relevansi hasil peringkasan, mempermudah proses analisis literatur ilmiah, serta mempercepat pengambilan informasi dari dokumen berformat kompleks. Beberapa tinjauan Pustaka terdapat pada tabel 2.1 Tinjauan Pustaka.

Tabel 2. 1 Tinjuan Pustaka

No	Nama Peneliti dan Tahun	Judul	Metode Penelitian	Hasil
1.	(Li dkk., 2020)	*Teaching Natural Language Processing through Big Data Text Summarization with Problem-Based Learning*	Penelitian ini menggunakan algoritma *Latent Dirichlet Allocation* (LDA) untuk pemodelan topik, yang merupakan pendekatan popular di antara tim, *Latent Semantic Analysis* (LSA), sebagai alternatif yang digunakan oleh beberapa tim ketika LDA tidak memberikan hasil yang baik, Doc2Vec, diikuti dengan metode K-means untuk pengelompokan, *Vector Space Model* untuk perhitungan kesamaan dan pengelompokan menggunakan *scikit learn.*	Hasil penelitian ini menunjukan bahwa Algoritma yang digunakan dalam penelitian, seperti *Latent Dirichlet Allocation* (LDA) dan *Latent Semantic Analysis* (LSA), menunjukkan bahwa beberapa tim berhasil mencapai skor ROUGE yang tinggi dalam ringkasan yang mereka buat. Tim 9, misalnya, mencapai skor tertinggi untuk ROUGE-1, ROUGE-L, dan ROUGE-SU4, yang menunjukkan efektivitas metodologi mereka dalam merangkum data namun beberapa tim mengalami kesulitan dalam pemodelan dan pengelompokan, yang berdampak negatif pada kualitas ringkasan mereka.
2.	(Fatmalasari dan dkk., 2022)	Peringkasan Teks Artikel Ilmiah Berbahasa Indonesia dengan Metode Pembobotan Kalimat	Penelitian ini menggunakan metode pembobotan kalimat dengan algoritma TF-IDF (*Term Frequency-Inverse Document Frequency*) dan *Similarity* untuk meringkas teks.	Hasil dari penelitian ini menggunakan metode pembobotan kalimat dengan algoritma TF-IDF dan similarity untuk meringkas teks menunjukkan bahwa sistem peringkasan yang dikembangkan memiliki tingkat kepuasan pengguna yang tinggi, dengan nilai 82,6% untuk tampilan sistem, 80,2% untuk efisiensi kalimat yang dihasilkan, dan 83,7% untuk kepuasan penggunaan

3.	(Utomo dkk., 2022)	*TEXT SUMMARIZATION PADA ARTIKEL BERITA MENGGUNAKAN VECTOR SPACE MODEL DAN COSINE SIMILARITY*	Penelitian ini menggunakan metode *Vector Space Model* (VSM) dan *Cosine Similarity* (CS). VSM digunakan untuk memberikan bobot nilai pada semua kata yang ada di artikel. CS digunakan untuk menghitung kemiripan antara judul artikel dengan isi artikel.	Hasil dari penelitian ini menunjukan metode yang digunakan untuk membandingkan judul artikel system isi dokumen menggunakan *Similarity* dengan algoritma *Jaccard Similarity*. bahwa dari 104 kalimat yang ada pada artikel, diperoleh 5 kalimat yang mempunyai nilai kemiripan paling tinggi dengan judul. Lima kalimat ini dijadikan satu paragraf sebagai hasil dari proses peringkasan artikel
4.	(Poleksić & Martinčić-Ipšić, 2023)	*Effects of Pretraining Corpora on Scientific Relation Extraction Using BERT and SciBERT*	Penelitian ini menggunakan metode *F1-score* digunakan untuk membandingkan kinerja model dengan dengan mempertimbangkan kedua matrik yaitu *Recall* dan *Precision*. Akurasi (*Accuracy*) digunakan untuk menghitung seberapa banyak prediksi model yang benar (*True Positive* dan *True Negative*) dibandingkan dengan total prediksi. *Precision* sebagai rasio *True Positive* (TP) terhadap jumlah TP dan *False Positive* (FP). *Precision* dapat digunakan pada klasifikasi biner dan multi kelas dengan dua pendekatan *averaging*: micro dan macro OpenNRE Toolkit digunakan untuk ekstraksi relasi (*Relation Extraction*) dan menyediakan implementasi dasar yang dapat diperluas untuk tugas-tugas seperti tokenisasi, lapisan neural umum, modul encoder, pemrosesan data, pelatihan model, dan	Hasil penelitian ini menunjukkan bahwa model BERT sedikit lebih unggul dalam metrik mikro-rata dan akurasi, yang menunjukkan kemampuan yang lebih kuat dalam mengklasifikasikan relasi yang dominan (umum). Sebaliknya, SciBERT unggul dalam metrik makro-rata, menunjukkan bahwa model ini lebih efektif untuk relasi ilmiah yang kurang dominan (spesifik). Penelitian ini menyoroti pentingnya korpus pra-pelatihan dan menyimpulkan bahwa SciBERT lebih cocok untuk ekstraksi relasi ilmiah, sementara BERT lebih baik untuk relasi umum. Selain itu, SciBERT menunjukkan peningkatan kinerja rata-rata sebesar 2,63% dibandingkan BERT dalam domain ilmiah, menunjukkan bahwa penggunaan korpus pra-pelatihan yang relevan secara tematis meningkatkan kinerja model dalam tugas ekstraksi relasi.

			evaluasi. BERT dan SciBERT BERT digunakan sebagai model bahasa dasar untuk tugas klasifikasi relasi, dan SciBERT adalah varian yang dilatih pada korpus ilmiah untuk mendukung kasus penggunaan berbasis ilmiah.	
5.	(Gao dkk., 2023)	*Human-like Summarization Evaluation with ChatGPT*	Penelitian ini menggunakan ChatGPT untuk evaluasi ringkasan teks dengan pendekatan yang menyerupai evaluasi manusia. ChatGPT digunakan untuk melakukan penilaian menggunakan empat metode evaluasi manusia yang umum: *Likert scale scoring, pairwise comparison, Pyramid, dan binary factuality evaluation.*	Hasil penelitian ini menunjukkan bahwa ChatGPT memiliki kemampuan untuk melakukan evaluasi ringkasan teks menggunakan berbagai metode evaluasi manusia. Dalam beberapa kasus, ChatGPT mencapai korelasi yang lebih tinggi dengan penilaian manusia dibandingkan metrik evaluasi otomatis yang ada. Kinerja ChatGPT dalam evaluasi ringkasan sangat bergantung pada desain prompt.

2.2 Dasar Teori

2.2.1 Teknologi Peringkas Teks Otomatis

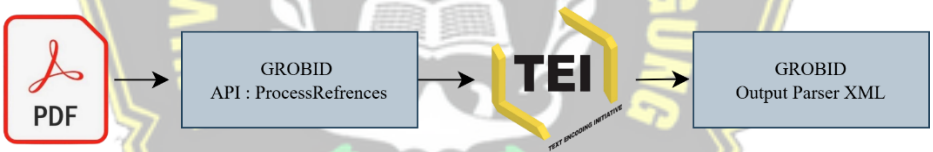
Teknologi peringkas teks otomatis adalah proses menggunakan algoritma atau model kecerdasan buatan yaitu *Artificial Intelligence* (AI) untuk merangkum teks panjang menjadi versi yang lebih singkat namun tetap mempertahankan informasi penting dan makna inti dari teks aslinya.

1. Ekstraktif dalam peringkasan teks merupakan metode yang memilih kalimat atau frasa penting secara langsung dari teks asli untuk dijadikan ringkasan. Pendekatan ini tidak mengubah struktur atau kata-kata asli, sehingga hasilnya sering kali berupa gabungan kalimat-kalimat yang ada dalam dokumen sumber (Yuliska & Syaliman, 2020).
2. Abstraktif menghasilkan ringkasan baru dengan menulis ulang informasi menggunakan pemahaman semantik teks. Proses ini melibatkan pemodelan ulang konten asli, sehingga ringkasannya dapat berisi kalimat-kalimat baru

yang tidak secara langsung ditemukan dalam dokumen sumber, namun tetap mewakili ide utama teks. (Hendry dkk., 2023) Natural Language Processing (NLP) telah menjadi teknologi penting dalam bidang analisis teks, yang mengubah data tekstual yang luas dan tidak terstruktur menjadi wawasan yang dapat ditindaklanjuti.

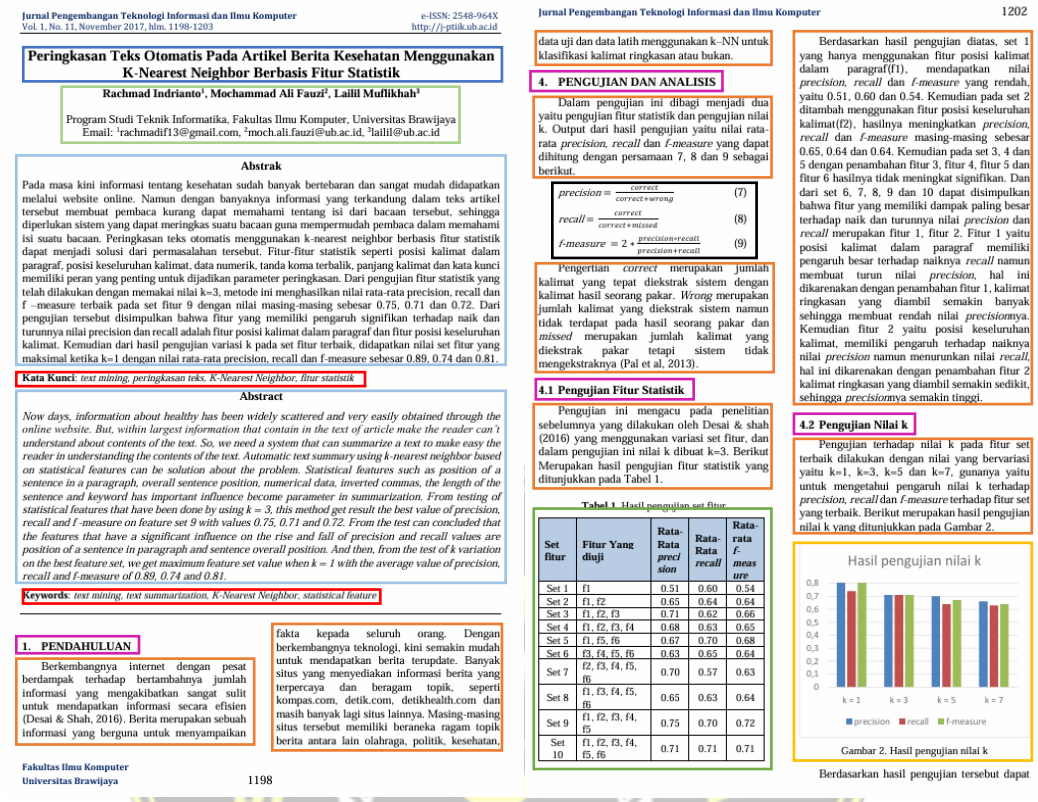
2.2.2 *GeneRation Of BIbliographic Data* (GROBID)

GeneRation Of Bibliographic Data GROBID adalah perpustakaan pembelajaran mesin untuk mengekstraksi, mengurai, dan menyusun ulang dokumen mentah seperti PDF ke dalam dokumen berkode XML/TEI terstruktur dengan fokus khusus pada publikasi teknis dan ilmiah. GROBID menyediakan API untuk ekstraksi entitas dari bagian Kepala dan Ekor (bibliografi) dari manuskrip PDF. GROBID populer digunakan untuk ekstraksi entitas dari artikel ilmiah dan berfungsi sebagai garis dasar yang kuat untuk kstraksi entitas dari liografi header dan bib. Alat ini telah ada selama lebih dari satu dekade dan dianggap sebagai alat standar di dunia akademis dan industri (Joshi dkk., 2023)



Gambar 2. 1 Ekstraksi data bibliografi GROBID (Joshi dkk., 2023)

Pada Gambar 2.1 menjelaskan alur kerja pemrosesan referensi dari file PDF menggunakan GROBID (*GeneRation Of BIbliographic Data*). Proses dimulai dengan memasukkan file PDF sebagai input. Selanjutnya, GROBID memanfaatkan API *process References* untuk mengekstraksi data referensi dari PDF tersebut. Data yang berhasil diekstraksi kemudian dikonversi ke dalam format TEI (*Text Encoding Initiative*), yaitu standar berbasis XML yang digunakan untuk merepresentasikan teks dengan struktur yang terorganisir, termasuk metadata bibliografis. Setelah itu, hasil dalam format TEI diproses lebih lanjut oleh GROBID Output Parser untuk menghasilkan data yang dapat digunakan, seperti JSON atau format lain yang sesuai untuk kebutuhan aplikasi tertentu. Proses ini sangat bermanfaat dalam otomatisasi pengolahan data bibliografi, terutama di bidang akademik dan penelitian.



Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer
Vol. 1, No. 11, November 2017, hlm. 1198-1203
e-ISSN: 2548-964X
http://j-ptiik.ub.ac.id

Peringkasan Teks Otomatis Pada Artikel Berita Kesehatan Menggunakan K-Nearest Neighbor Berbasis Fitur Statistik

Rachmad Indrianto[1], Mochammad Ali Fauzi[2], Lailil Muflikhah[3]

Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Brawijaya
Email: [1]rachmadif13@gmail.com, [2]moch.ali.fauzi@ub.ac.id, [3]lailil@ub.ac.id

Abstrak

Pada masa kini informasi tentang kesehatan sudah banyak bertebaran dan sangat mudah didapatkan melalui website online. Namun dengan banyaknya informasi yang terkandung dalam teks artikel tersebut membuat pembaca kurang dapat memahami tentang isi dari bacaan tersebut, sehingga diperlukan sistem yang dapat meringkas suatu bacaan guna mempermudah pembaca dalam memahami isi suatu bacaan. Peringkasan teks otomatis menggunakan k-nearest neighbor berbasis fitur statistik dapat menjadi solusi dari permasalahan tersebut. Fitur-fitur statistik seperti posisi kalimat dalam paragraf, posisi keseluruhan kalimat, data numerik, tanda koma terbalik, panjang kalimat dan kata kunci memiliki peran yang penting untuk dijadikan parameter peringkasan. Dari pengujian fitur statistik yang telah dilakukan dengan memakai nilai k=3, metode ini menghasilkan nilai rata-rata precision, recall dan f –measure terbaik pada set fitur 9 dengan nilai masing-masing sebesar 0.75, 0.71 dan 0.72. Dari pengujian tersebut disimpulkan bahwa fitur yang memiliki pengaruh signifikan terhadap naik dan turunnya nilai precision dan recall adalah fitur posisi kalimat dalam paragraf dan fitur posisi keseluruhan kalimat. Kemudian dari hasil pengujian variasi k pada set fitur terbaik, didapatkan nilai set fitur yang maksimal ketika k=1 dengan nilai rata-rata precision, recall dan f-measure sebesar 0.89, 0.74 dan 0.81.

Kata Kunci: *text mining, peringkasan teks, K-Nearest Neighbor, fitur statistik*

Abstract

Now days, information about healthy has been widely scattered and very easily obtained through the online website. But, within largest information that contain in the text of article make the reader can't understand about contents of the text. So, we need a system that can summarize a text to make easy the reader in understanding the contents of the text. Automatic text summary using k-nearest neighbor based on statistical features can be solution about the problem. Statistical features such as position of a sentence in a paragraph, overall sentence position, numerical data, inverted commas, the length of the sentence and keyword has important influence become parameter in summarization. From testing of statistical features that have been done by using k = 3, this method get result the best value of precision, recall and f -measure on feature set 9 with values 0.75, 0.71 and 0.72. From the test can concluded that the features that have a significant influence on the rise and fall of precision and recall values are position of a sentence in paragraph and sentence overall position. And then, from the test of k variation on the best feature set, we get maximum feature set value when k = 1 with the average value of precision, recall and f-measure of 0.89, 0.74 and 0.81.

Keywords: *text mining, text summarization, K-Nearest Neighbor, statistical feature*

1. PENDAHULUAN

Berkembangnya internet dengan pesat berdampak terhadap bertambahnya jumlah informasi yang mengakibatkan sangat sulit untuk mendapatkan informasi secara efisien (Desai & Shah, 2016). Berita merupakan sebuah informasi yang berguna untuk menyampaikan fakta kepada seluruh orang. Dengan berkembangnya teknologi, kini semakin mudah untuk mendapatkan berita terupdate. Banyak situs yang menyediakan informasi berita yang terpercaya dan beragam topik, seperti kompas.com, detik.com, detikhealth.com dan masih banyak lagi situs lainnya. Masing-masing situs tersebut memiliki beraneka ragam topik berita antara lain olahraga, politik, kesehatan,

data uji dan data latih menggunakan k-NN untuk klasifikasi kalimat ringkasan atau bukan.

4. PENGUJIAN DAN ANALISIS

Dalam pengujian ini dibagi menjadi dua yaitu pengujian fitur statistik dan pengujian nilai k. Output dari hasil pengujian yaitu nilai rata-rata *precision, recall* dan *f-measure* yang dapat dihitung dengan persamaan 7, 8 dan 9 sebagai berikut.

$$precision = \frac{correct}{correct+wrong} \quad (7)$$

$$recall = \frac{correct}{correct+missed} \quad (8)$$

$$f\text{-}measure = 2 * \frac{precision*recall}{precision+recall} \quad (9)$$

Pengertian *correct* merupakan jumlah kalimat yang tepat diekstrak sistem dengan kalimat hasil seorang pakar. *Wrong* merupakan jumlah kalimat yang diekstrak sistem namun tidak terdapat pada hasil seorang pakar dan *missed* merupakan jumlah kalimat yang diekstrak pakar tetapi sistem tidak mengekstraknya (Pal et al, 2013).

4.1 Pengujian Fitur Statistik

Pengujian ini mengacu pada penelitian sebelumnya yang dilakukan oleh Desai & shah (2016) yang menggunakan variasi set fitur, dan dalam pengujian ini nilai k dibuat k=3. Berikut Merupakan hasil pengujian fitur statistik yang ditunjukkan pada Tabel 1.

Tabel 1. Hasil pengujian set fitur

Set fitur	Fitur Yang diuji	Rata-Rata *precision*	Rata-Rata *recall*	Rata-rata *f-measure*
Set 1	f1	0.51	0.60	0.54
Set 2	f1, f2	0.65	0.64	0.64
Set 3	f1, f2, f3	0.71	0.62	0.66
Set 4	f1, f2, f3, f4	0.68	0.63	0.65
Set 5	f1, f5, f6	0.67	0.70	0.68
Set 6	f3, f4, f5, f6	0.63	0.65	0.64
Set 7	f2, f3, f4, f5, f6	0.70	0.57	0.63
Set 8	f1, f3, f4, f5, f6	0.65	0.63	0.64
Set 9	f1, f2, f3, f4, f5	0.75	0.70	0.72
Set 10	f1, f2, f3, f4, f5, f6	0.71	0.71	0.71

Berdasarkan hasil pengujian diatas, set 1 yang hanya menggunakan fitur posisi kalimat dalam paragraf(f1), mendapatkan nilai *precision, recall* dan *f-measure* yang rendah, yaitu 0.51, 0.60 dan 0.54. Kemudian pada set 2 ditambah menggunakan fitur posisi keseluruhan kalimat(f2), hasilnya meningkatkan *precision, recall* dan *f-measure* masing-masing sebesar 0.65, 0.64 dan 0.64. Kemudian pada set 3, 4 dan 5 dengan penambahan fitur 3, fitur 4, fitur 5 dan fitur 6 hasilnya tidak meningkat signifikan. Dan dari set 6, 7, 8, 9 dan 10 dapat disimpulkan bahwa fitur yang memiliki dampak paling besar terhadap naik dan turunnya nilai *precision* dan *recall* merupakan fitur 1, fitur 2. Fitur 1 yaitu posisi kalimat dalam paragraf memiliki pengaruh besar terhadap naiknya *recall* namun membuat turun nilai *precision*, hal ini dikarenakan dengan penambahan fitur 1, kalimat ringkasan yang diambil semakin banyak sehingga membuat rendah nilai *precision*nya. Kemudian fitur 2 yaitu posisi keseluruhan kalimat, memiliki pengaruh terhadap naiknya nilai *precision* namun menurunkan nilai *recall*, hal ini dikarenakan dengan penambahan fitur 2 kalimat ringkasan yang diambil semakin sedikit, sehingga *precision*nya semakin tinggi.

4.2 Pengujian Nilai k

Pengujian terhadap nilai k pada fitur set terbaik dilakukan dengan nilai yang bervariasi yaitu k=1, k=3, k=5 dan k=7, gunanya yaitu untuk mengetahui pengaruh nilai k terhadap *precision, recall* dan *f-measure* terhadap fitur set yang terbaik. Berikut merupakan hasil pengujian nilai k yang ditunjukkan pada Gambar 2.

Gambar 2. Hasil pengujian nilai k

Berdasarkan hasil pengujian tersebut dapat

Gambar 2. 2 Halaman beranotasi otomatis (Pisaneschi dkk., 2023)

Gambar 2. 2 adalah contoh halaman beranotasi otomatis. Dari GROBID dan PDFMiner kami mengekstrak sembilan kategori: judul (biru), penulis (hijau muda), abstrak (cyan), kata kunci (merah), subtitle (ungu), teks (Oren), gambar (kuning), rumus (hitam) dan tabel (hijau) (Pisaneschi dkk., 2023).

2.2.3 *Large Language Models* (LLM)

Model bahasa besar *Large Language Models* (LLM) adalah jenis model kecerdasan buatan yang dirancang untuk memahami dan menghasilkan teks dalam bahasa alami. LLM dilatih menggunakan data teks dalam jumlah besar dari berbagai sumber, memungkinkan mereka mengenali pola, konteks, dan makna dalam bahasa manusia.

Kemampuan LLM mencakup berbagai tugas seperti pemrosesan bahasa alami (NLP), seperti terjemahan, penulisan otomatis, peringkasan teks, dan menjawab pertanyaan. Dengan inovasi terkini dalam model arsitektur, pelatihan teknik, dan pemrosesan data, LLM telah menjadi alat penting di berbagai bidang, termasuk penelitian, bisnis, pendidikan, dan teknologi. (Naveed dkk., 2023)

Contoh LLM yang terkenal adalah GPT, BERT, dan SciBERT, yang masing-masing berkontribusi pada pengembangan solusi otomatis yang lebih cerdas dan efisien dalam memahami dan menyusun teks berbasis bahasa alami.

2.2.4 SciBERT

SciBERT adalah model bahasa yang telah dilatih sebelumnya pada BERT yang dirancang khusus untuk teks ilmiah. Model ini dilatih menggunakan korpus multi-domain yang terdiri dari 1,14 juta makalah ilmiah (18% ilmu komputer, 82% biomedis) dengan total 3,17 miliar token, menggunakan seluruh isi makalah, bukan hanya abstraknya. Dengan tokenisasi berbasis ScispaCy, SciBERT dioptimalkan untuk menangani teks ilmiah dan menunjukkan kinerja unggul dalam berbagai tugas NLP ilmiah, seperti penandaan urutan, klasifikasi kalimat, dan penguraian ketergantungan, dibandingkan dengan BERT (Kilimci & Yalcin, 2024).

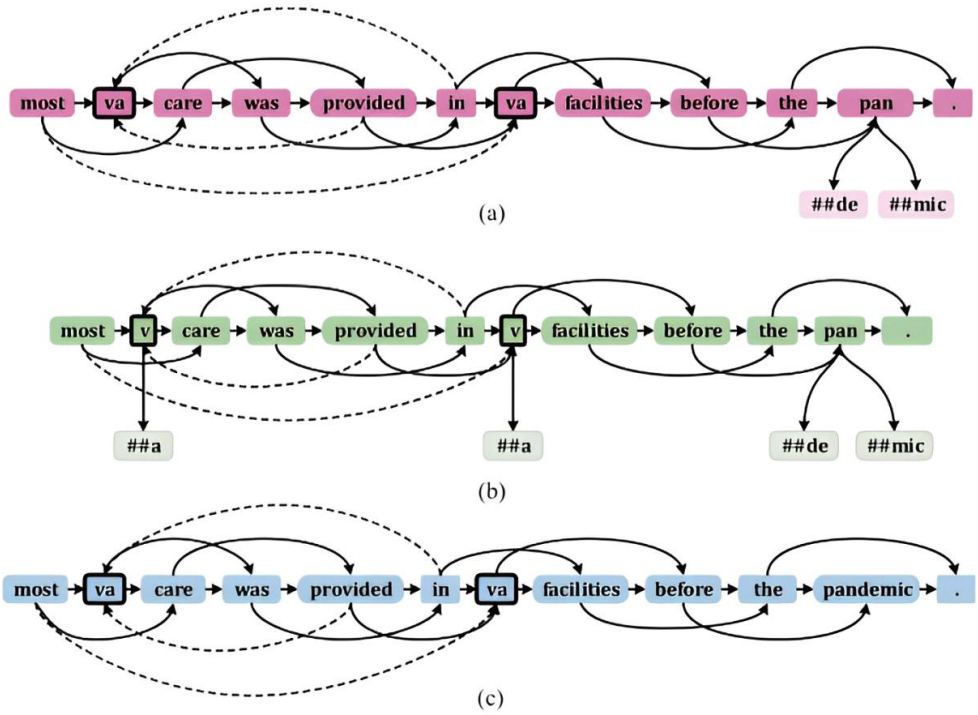
SciBERT menggunakan mekanisme *self-attention* berbasis *Transformer*. Mekanisme ini memungkinkan setiap kata dalam kalimat untuk saling berinteraksi dan berbagi informasi melalui tiga komponen utama:

(Key) $K = W^k x$

(Query) $Q = W^Q x$

(value) $V = W^v x$

$$\textit{Attention}\ (K, Q, V) = \textit{softmax}\left(\frac{QK^t}{\sqrt{d_k}}\right) V \quad (1)$$



Gambar 2. 3 Ilustrasi tokenisasi sub-kata dalam SciBERT (Cai dkk., 2022)

Gambar 2. 3 menggambarkan bagaimana proses tokenisasi bekerja dalam SciBERT untuk memecah kalimat menjadi token-token yang lebih kecil (Cai dkk., 2022). Pada gambar ini, kita dapat melihat tiga langkah tokenisasi yang berbeda:

1. Gambar (a) menunjukkan tahap awal tokenisasi, di mana kata-kata panjang seperti "*facilities*" dipecah menjadi beberapa bagian yang lebih kecil. Token-token ini ditandai dengan simbol "##", yang menandakan bahwa bagian ini merupakan sub-token dari kata yang lebih besar.
2. Gambar (b) melanjutkan proses pemecahan kata lebih lanjut, dengan menandai beberapa sub-token sebagai bagian dari kata yang lebih panjang, seperti "##a". Ini menunjukkan bagaimana kata yang lebih panjang dipecah menjadi bagian-bagian kecil agar model dapat memahaminya dengan lebih baik.
3. Gambar (c) menunjukkan tokenisasi yang lebih mendalam lagi, seperti kata "*pandemic*" yang dipisah menjadi dua bagian—"pan" dan "##demic". Proses ini memungkinkan SciBERT untuk menangani kata-kata yang tidak ada dalam kamus atau kata-kata teknis yang jarang ditemukan, yang sering ada dalam teks ilmiah.

Proses tokenisasi ini menggunakan tokenisasi berbasis *WordPiece*, yang memungkinkan SciBERT mengelola kata-kata baru atau jarang ditemukan dalam teks ilmiah dengan memecah kata-kata menjadi sub-token (Chen dkk., 2023). Simbol "##" menandakan bahwa token tersebut adalah bagian dari kata yang lebih besar, memudahkan model untuk memproses teks ilmiah dengan berbagai kata teknis yang spesifik.

2.2.5 Integrasi GROBID dan SciBERT

Integrasi antara GROBID dan SciBERT menawarkan pendekatan yang efektif untuk meningkatkan akurasi peringkasan teks ilmiah. GROBID sebagai alat *open source* yang dirancang untuk mengekstrak metadata dari artikel ilmiah mampu memberikan informasi penting seperti judul, penulis, abstrak dan referensi dengan tingkat presisi yang tinggi (Foppiano dkk., 2022). Dengan dukungan metadata yang akurat, pemahaman konteks artikel menjadi lebih komprehensif. SciBERT, di sisi lain, adalah model bahasa berbasis BERT yang dilatih khusus pada korpora ilmiah, sehingga memiliki kemampuan lebih baik dalam memahami terminologi dan struktur kalimat kompleks, yang umum dalam literatur akademis (Maheshwari dkk., 2021). Integrasi kedua alat ini memungkinkan peringkasan artikel menjadi lebih relevan dan tetap informatif, tidak hanya memadatkan isi tetapi juga menjaga esensi artikel aslinya, sehingga memberikan manfaat besar bagi peneliti dalam memahami literatur ilmiah dengan lebih efisien.

BAB III
METODOLOGI PENELITIAN

3.1 Metode Penelitian

Pada tahap pelatihan, penulis akan membuat sistem tinjauan pustaka dengan menggunakan GROBID untuk mengekstrak informasi bibliografi dan SciBERT untuk memahami konten artikel. Kemudian, pada tahap pengembangan aplikasi berbasis web, penulis akan menggunakan Streamlit untuk membangun antarmuka yang memungkinkan pengguna mengunggah artikel dan mendapatkan rekomendasi literatur yang relevan. Keluaran dari penelitian ini adalah sistem yang dapat membantu peneliti dalam proses tinjauan pustaka dengan lebih efisien dan akurat.

Gambar 3. 1 Tahapan Penelitian

Pada Gambar 3. 1 tampilan alur pada tahap 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.

3.1.1 Studi Literatur

Studi literatur dilakukan dengan menganalisis berbagai artikel ilmiah yang diperoleh dari platform Garuda untuk memahami peringkasan teks otomatis, ekstraksi dokumen dengan GROBID, dan penerapan SciBERT. Studi sebelumnya menunjukkan bahwa SciBERT lebih unggul dalam merangkum teks ilmiah dibandingkan metode tradisional. Integrasi GROBID dan SciBERT diharapkan dapat menghasilkan ringkasan yang akurat dan mendukung efisiensi tinjauan pustaka.

3.1.2 Pengumpulan Data

Pengumpulan data dilakukan dengan memilih artikel-artikel yang relevan dengan topik penelitian dari platform Garuda. Sebanyak 5 file PDF yang berbahasa

Indonesia digunakan dalam pengujian dan 1 file PDF untuk validasi penelitian ini. Artikel-artikel ini dipilih berdasarkan tentang Teknik informatika, kualitas jurnal yang terpublikasi, dan tahun publikasi terbaru, untuk memastikan data yang diambil adalah relevan dan berkualitas. Untuk keperluan validasi, saya mengambil 10 paragraf dari file PDF yang dipilih dan merangkumnya secara manual.

3.1.3 Modeling

1. Ekstraksi GROBID

GROBID *Application Program Interface* (API) untuk mengekstrak metadata dari artikel PDF. Dengan fokus utama pada pengambilan isi artikel per-paragraf. Proses dimulai dengan menggunakan GROBID *ProcessFullText* API, yang mengurai seluruh dokumen PDF. API ini menghasilkan file XML yang tersetruktur menggunakan format TEI (http:/www.tei-c org/ns/1.0), dimana setiap paragraf dalam teks artikel dipisahkan menjadi simpul (node) yang berbeda. Setelah file XML diterima, dokumen tersebut kemudian diurai untuk mengambil teks yang ada di setiap simpul paragraf secara terpisah. Teks-teks ini kemudian digunakan sebagai input dalam proses peringkasan tekst otomatis

```
                xmlns="http://www.tei-c.org/ns/1.0">
                <head n="1.">PENDAHULUAN</head>
                <p>Revolusi pada dunia ilmu pengetahuan dan teknologi telah memicu lahirnya pola baru dalam penyampaian maupun pe
nerimaan informasi, dimana pola penyampaian informasi yang lazim dilakukan sekarang ini adalah memanfaatkan komputer, monitor dan
televisi sebagai piranti medianya.</p>
                <p>Pada saat ini, mahasiswa selalu dibingungkan oleh masalah pelaksanaan perkuliahan. Baik mengenai waktu pelaksa
naan perkuliahan, ruangan yang akan digunakan, bahkan mengenai kepastian pelaksanaan perkuliahan yang akan berlangsung. Akibatnya
mahasiswa harus selalu ke bagian ke pengajaran hanya untuk menanyakan masalah ini. Hal ini sangat tidak efektif.</p>
                <p>Berdasarkan masalah di atas dan pentingnya pengaturan jadwal dengan baik, maka penulis membangun sistem inform
asi jadwal perkuliahan menggunakan media televisi. Sistem ini akan memberikan informasi mengenai pelaksanaan perkuliahan kepada m
ahasiswa melalui sebuah layar televisi. Informasi yang diberikan mencakup waktu pelaksanaan perkuliahan yang akan berlangsung, ba
hkan menginformasikan pengumumanpengumuman dan kegiatan-kegiatan yang akan berlangsung.</p>
                <p>Sebelumnya, penyampaian jadwal dan informasi kepada mahasiswa masih menggunakan media kertas dan papan pengumu
man. Hal ini jika dilihat dari sudut pandang waktu merupakan hal yang tidak efisien. Penyampaian informasi yang cepat, efisien da
n akurat juga dapat meningkatkan dan mendapatkan pengakuan dari masyarakat. Saat ini persaingan antar perguruan tinggi begitu ket
at dalam menghasilkan sumber daya manusia yang unggul dan berkualitas, baik teori maupun praktek, serta dituntut memiliki kemampu
an analisis dan logika berpikir dengan cermat dan tajam.
                    <ref type="bibr" target="#b1">[4]</ref> Sistem Informasi adalah kombinasi dari teknologi informasi dan aktivi
tas orang yang menggunakan teknologi itu untuk mendukung operasi dan manajemen. Dalam arti yang sangat luas, istilah sistem infor
masi yang sering digunakan merujuk kepada interaksi antara orang, proses algoritmik, data, dan teknologi. Dalam pengertian ini, i
stilah ini digunakan untuk merujuk tidak hanya pada penggunaan organisasi teknologi informasi dan komunikasi (TIK), tetapi juga u
ntuk cara di mana orang berinteraksi dengan teknologi ini dalam mendukung proses bisnis.
                </p>
            </div>
            <div
```

Gambar 3. 2 Hasil Konversi PDF dari file XML

Pada Gambar 3. 2 adalah tampilan hasil dari GROBID untuk mengonversi teks PDF menjadi format XML 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.

2. Peringkasan Text Otomatis

Peringkasan teks otomatis menggunakan SciBERT memanfaatkan kemampuan pemrosesan bahasa alami yang dirancang khusus untuk teks ilmiah. SciBERT menganalisis dokumen penelitian dengan memahami terminologi dan struktur kompleksnya, lalu menghasilkan ringkasan yang berisi poin-poin utama. Proses ini mencakup tokenisasi teks, ekstraksi embedding kalimat, serta perhitungan kemiripan dengan *cosine similarity* untuk memilih kalimat paling relevan, konteks menggunakan SciBERT, dan peringkasan secara ekstraktif, yaitu dengan memilih kalimat-kalimat penting dari dokumen. Dengan pelatihan berbasis literatur ilmiah, SciBERT memberikan hasil yang relevan dan akurat, mempermudah peneliti dalam menyaring informasi penting dari dokumen panjang secara efisien. Penelitian ini berfokus pada metode ekstraktif, yang menekankan pada pemilihan kalimat kunci untuk membentuk ringkasan yang informatif.

3.1.4 Evaluasi Ringkasan

Penelitian ini akan menilai kualitas ringkasan artikel yang dihasilkan oleh metode SciBERT dengan menggunakan skor ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*). Skor ini dipilih karena telah terbukti memiliki korelasi positif dengan penilaian manusia terhadap kualitas linguistik suatu teks. (Goodrich dkk., 2019). ROUGE mengevaluasi ringkasan yang dihasilkan oleh metode SciBERT dengan membandingkannya terhadap ringkasan manual yang dibuat oleh manusia. Skor yang lebih tinggi menunjukkan tingkat kesamaan yang lebih besar antara kedua ringkasan tersebut. (Moradi dkk., 2020).

Skor ROUGE mencakup *Precision, Recall*, dan *F1-score* untuk ROUGE-N, di mana N dapat bernilai 1, 2, atau L. Variabel N menunjukkan jumlah kata berurutan yang sama dalam kedua dokumen yang dibandingkan, yaitu ringkasan dari metode SciBERT dan ringkasan manual. Jika N bernilai 1, maka dihitung word-1-gram (*unigram*), sedangkan jika N bernilai 2, maka dihitung word-2-gram (*bigram*) yang muncul di kedua dokumen. Sementara itu, jika N bernilai L, maka digunakan Longest Common Subsequent (LCS), yaitu urutan kata terpanjang yang

sama dalam kedua ringkasan yang dibandingkan. Persamaan yang digunakan untuk menghitung skor ROUGE adalah sebagai berikut:

1. *Recall*

Recall mengukur jumlah kata yang sesuai, baik dalam bentuk *unigram*, *bigram*, atau *Longest Common Subsequent* (LCS), yang terdapat dalam ringkasan metode SciBERT dibandingkan dengan ringkasan manual yang dibuat oleh manusia.

$$ROUGE\text{-}1\ Recall = \frac{Jumplah\ unigram\ yang\ sama}{Keseluruhan\ kata\ di\ teks\ ringkasan\ validasi} \quad (2)$$

$$ROUGE\text{-}2\ Recall = \frac{Jumplah\ bigram\ kata\ sama}{Keseluruhan\ kata\ di\ teks\ ringkasan\ validasi} \quad (3)$$

$$ROUGE\text{-}L\ Recall = \frac{LCS\ (longest\ common\ subsequent)}{Keseluruhan\ kata\ di\ teks\ ringkasan\ validasi} \quad (4)$$

2. *Precision*

Precision digunakan untuk mengukur seberapa banyak kata yang relevan dengan membandingkan jumlah kata yang sama, baik dalam bentuk unigram, *bigram*, atau *Longest Common Subsequent* (LCS), terhadap total jumlah kata dalam ringkasan yang dihasilkan oleh sistem.

$$ROUGE\text{-}1\ precission = \frac{Jumlah\ unigram\ kata\ sama}{Keseluruhan\ kata\ di\ teks\ ringkasan\ validasi} \quad (5)$$

$$ROUGE\text{-}2\ precission = \frac{Jumlah\ biagram\ kata\ yang\ sama}{Keseluruhan\ kata\ di\ teks\ ringkasan\ validasi} \quad (6)$$

$$ROUGE\text{-}L\ precission = \frac{LCS\ (longest\ common\ subsequent)}{Keseluruhan\ kata\ di\ teks\ ringkasan\ validasi} \quad (7)$$

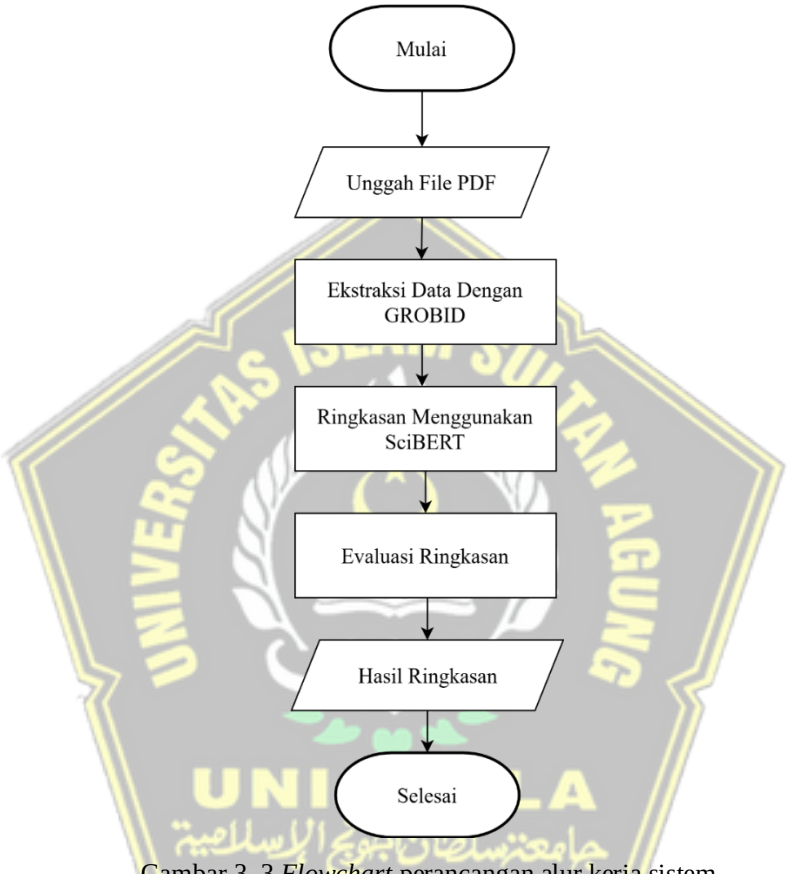
3. *F1-score*

F1-score, atau *F-measure*, merupakan metode yang digunakan untuk menghitung rata-rata *harmonik* antara *Recall* dan *Precision*.

$$F1\text{-}scores = 2\ X\ \frac{LCS\ (longest\ common\ subsequent)}{Keseluruhan\ kata\ di\ teks\ ringkasan\ validasi} \quad (8)$$

3.2 Perancangan Arsitektur sistem

Pada tahap ini, dilakukan analisis untuk merancang alur kerja sistem dalam memproses dan merangkum dokumen PDF secara otomatis. Proses ini divisualisasikan dalam bentuk *Flowchart*, seperti yang ditampilkan pada Gambar 3.3.



Gambar 3. 3 *Flowchart* perancangan alur kerja sistem

Gambar 3. 3 merupakan *flowchart* rancangan sistem yang akan dibanggun dimana tahapanya adalah :

1. Unggah File PDF

Pada tahap ini, pengguna mengunggah file PDF yang ingin diproses. File ini akan menjadi masukan utama untuk sistem.

2. Ekstraksi Data Dengan GROBID

File PDF yang telah diunggah akan diproses menggunakan GROBID. GROBID adalah alat berbasis *machine learning* yang digunakan untuk mengekstrak teks dan informasi terstruktur dari dokumen PDF, seperti metadata atau isi teks mentah.

3. Ringkasan Menggunakan SciBERT

Data teks hasil pemrosesan diringkas menggunakan model SciBERT. SciBERT adalah model NLP berbasis BERT yang dirancang khusus untuk menangani teks ilmiah dan menghasilkan ringkasan yang relevan.

4. Evaluasi Ringkasan

Hasil ringkasan kemudian dievaluasi untuk memastikan kualitas dan akurasi. Evaluasi ini bertujuan memverifikasi apakah ringkasan sudah sesuai dengan tujuan dan tidak ada kesalahan dalam proses sebelumnya.

5. Hasil Ringkasan

Setelah evaluasi selesai, sistem menghasilkan output berupa ringkasan akhir dari teks yang telah diproses.

3.3 Gambaran Sistem

Sistem peringkasan artikel ilmiah yang dirancang untuk membantu proses tinjauan pustaka ini akan berfungsi sebagai aplikasi berbasis web yang memungkinkan pengguna untuk mengunggah file PDF artikel ilmiah. Setelah diunggah, sistem akan memanfaatkan GROBID untuk mengekstrak data penting dari artikel, seperti judul, penulis, dan isi utama, yang kemudian disimpan dalam format terstruktur. Selanjutnya, data yang telah diekstrak akan diproses menggunakan model bahasa besar *Large Language Models* (LLM) berbasis SciBERT untuk menghasilkan ringkasan otomatis yang mencakup poin-poin kunci dari artikel tersebut. Proses ini meliputi tokenisasi dan pemahaman konteks untuk memastikan ringkasan yang dihasilkan relevan dan akurat. Dengan demikian, sistem ini tidak hanya mempermudah peneliti dalam mendapatkan informasi penting dari artikel ilmiah, tetapi juga meningkatkan efisiensi dalam proses tinjauan pustaka dengan menyediakan ringkasan yang padat dan informatif.

3.4 Analisis Kebutuhan Sistem

Kebutuhan sistem dilakukan dengan memastikan bahwa sistem memiliki spesifikasi perangkat keras dan perangkat lunak yang memadai untuk menjalankan proses peringkasan artikel. Berikut adalah rincian spesifikasi yang digunakan:

1. Perangkat keras

Komputer atau lep top : Spesifikasi prosesor Intel Core i7/AMD Ryzen 7, RAM minimal 16 GB, SSD 512 G B, dan GPU NVIDIA RTX 3060 atau lebih tinggi untuk mendukung pemrosesan data dan model AI secara optimal.

2. Perangkat Lunak

a. Python 3.7

Bahasa pemprograman yang digunakan pada penelitian ini adalah python 3.7 karena peggunaan Bahasa ini banyak digunakan dan didukung untuk penelitian *mechine lerning*

b. *Visual Studio Code*

Untuk Mengembangkan aplikasi dan melakukan penelitian diperlukan adanya *text editor*. Penelitian ini menggunakan *text editor visual studio code* karena mendukung banyak Bahasa pemprograman dan banyak *framework*.

c. GROBID

Untuk ekstraksi metadata artikel ilmiah diperlukan sebuah *tool* yang andal. Penelitian ini memngguakan GROBID karena mampu mengketraksi informasi dan akurat, mendukung format seperti PDF, sehingga sangat membantu dalam proses pengolahan data penelitian.

d. XML (Xml.etree.ElementTree)

Digunakan untuk parsing file XML, seperti file TEI XML dalam kode ini, agar dapat membaca struktur dan mengekstrak elemen-elemen yang relevan (judul, penulis, abstrak, dll.)

e. JSON

Digunakan untuk menyimpan hasil parsing ke dalam format JSON yang lebih terstruktur dan mudah dibaca atau digunakan oleh sistem lain.

f. NLTK (*Natural Language Toolkit*)

NLTK digunakan untuk memproses teks dalam analisis bahasa alami. Digunakan untuk mengelola teks, seperti menghapus kata-kata umum (*stopwords*) dan melakukan *stemming* (mengubah kata menjadi bentuk dasarnya.

g. *Torch*

Merupakan *library* dari PyTorch yang digunakan untuk komputasi numerik dan *machine learning*. Dalam konteks ini, torch digunakan untuk memanfaatkan model berbasis *deep learning*, seperti SciBERT, yang memerlukan tensor untuk melakukan perhitungan matriks dan model transformasi.

h. *Transformers*

Library dari *Hugging Face* untuk bekerja dengan model berbasis transformer. Dalam kode ini, digunakan untuk memanfaatkan *tokenizer* dan model SciBERT (*AutoTokenizer* dan *AutoModelForMaskedLM*) untuk memahami dan meringkas artikel ilmiah.

i. Sklearn.metrics.pairwise_cosine_similarity

Library dari *Scikit-learn* ini digunakan untuk menghitung kesamaan antar vektor berdasarkan *cosine similarity*. Dalam penelitian ini, ini digunakan untuk mengukur kemiripan antara teks, seperti antara ringkasan yang dihasilkan oleh model dan ringkasan manual.

j. ROUGE

Library yang digunakan untuk menghitung ROUGE (*Recall*-Oriented Understudy for Gisting Evaluation), yang merupakan metode untuk mengevaluasi kualitas ringkasan otomatis dengan membandingkannya dengan ringkasan manual. Dalam penelitian ini, ROUGE digunakan untuk mengukur kesamaan antara ringkasan yang dihasilkan oleh model dan ringkasan referensi.

k. Matplotlib

Digunakan untuk membantu visualisasi hasil evaluasi model, seperti skor ROUGE.

l. Seaborn

Digunakan untuk mempercantik tampilan grafik yang dibuat dengan matplotlib agar lebih informatif.

m. ReportLab

Digunakan untuk menyimpan hasil evaluasi dalam bentuk laporan PDF agar dapat didokumentasikan atau dibagikan.

n. *Streamlit*

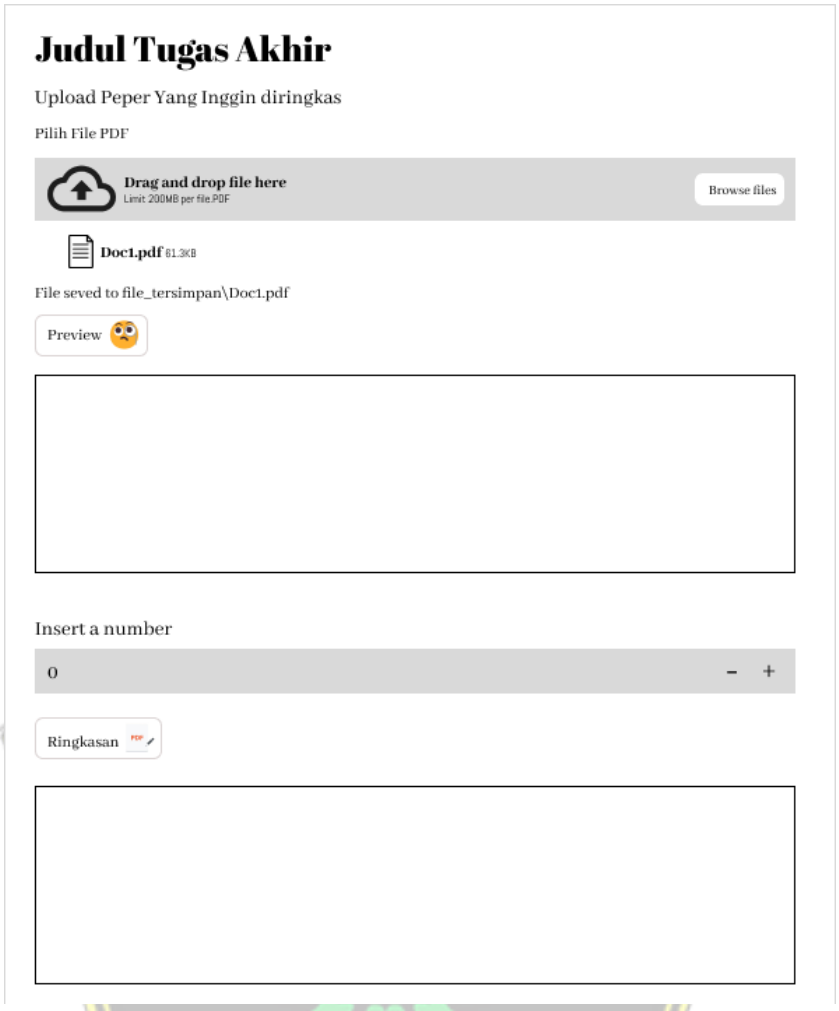
Digunakan untuk membangun antarmuka aplikasi berbasis web yang memungkinkan pengguna menjalankan sistem peringkasan ini melalui browser. Baris terakhir! streamlit `run app.py` digunakan untuk memulai aplikasi web.

3.5 Perancangan *User Interface*

Pada tahap ini, Menunjukan tampilan dari sistem, yang mencerminkan desain minimalis dan fokus pada kemudahan pengguna.

Gambar 3. 4 Tampilan awal sistem

Pada Gambar 3. 4 adalah tampilan awal sistem. Perancangan sistem peringkasan artikel ilmiah untuk membantu proses tinjauan pustaka dirancang dengan pendekatan minimalis yang fokus pada fungsionalitas dan kemudahan penggunaan. Pada halaman awal, terdapat judul "Peringkasan artikel ilmiah berbasis SciBERT" di bagian atas sebagai penanda utama. Di bawahnya, tersedia area unggah file berbentuk kotak dengan ikon dan teks "*Drag and drop file here*" untuk mempermudah pengguna mengunggah dokumen PDF yang akan diringkas. Selain itu, terdapat kolom input angka untuk menentukan jumlah ringkasan yang diinginkan, disertai tombol "+" dan "-" untuk menambah atau mengurangi nilai. Tombol "Ringkasan" dengan ikon PDF di bagian bawah mempermudah pengguna mengunduh hasil ringkasan yang telah dihasilkan.



Gambar 3. 5 Tampilan saat meringkas

Pada gambar 3. 5 adalah tampilan akhir setelah file PDF diunggah melalui area "*Drag and drop file here*", antarmuka akan menampilkan nama file yang berhasil disimpan. Ketika tombol "*Preview*" ditekan, isi file PDF akan ditampilkan di area besar berbentuk persegi panjang dengan latar abu-abu di tengah halaman, memudahkan pengguna untuk membaca isi dokumen. Selanjutnya, pengguna dapat memasukkan angka pada kolom "*Insert a number*" untuk menentukan tingkat peringkasan. Setelah tombol "Ringkasan" ditekan, hasil ringkasan akan muncul di area yang sama, menggantikan tampilan isi file PDF. Elemen-elemen ini dirancang untuk memberikan pengalaman interaktif dan intuitif bagi pengguna.

BAB IV
HASIL DAN ANALISIS PENELITIAN

4.1 Hasil pengumpulan Data

Penelitian ini menggunakan 5 data PDF artikel ilmiah berbahasa Indonesia dari platform garuda yang relevan dengan topik Teknik informatika.

Tabel 4. 1 contoh data PDF artikel ilmiah

No	PDF Sumber	Hasil PDF
1	Garuda (1)	

Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer
Vol. 1, No. 11, November 2017, hlm. 1198-1203
e-ISSN: 2548-964X
http://j-ptiik.ub.ac.id

Peringkasan Teks Otomatis Pada Artikel Berita Kesehatan Menggunakan K-Nearest Neighbor Berbasis Fitur Statistik

Rachmad Indrianto[1], Mochammad Ali Fauzi[2], Lailil Muflikhah[3]

Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Brawijaya
Email: [1]rachmadif13@gmail.com, [2]moch.ali.fauzi@ub.ac.id, [3]lailil@ub.ac.id

Abstrak

Pada masa kini informasi tentang kesehatan sudah banyak bertebaran dan sangat mudah didapatkan melalui website online. Namun dengan banyaknya informasi yang terkandung dalam teks artikel tersebut membuat pembaca kurang dapat memahami tentang isi dari bacaan tersebut, sehingga diperlukan sistem yang dapat meringkas suatu bacaan guna mempermudah pembaca dalam memahami isi suatu bacaan. Peringkasan teks otomatis menggunakan k-nearest neighbor berbasis fitur statistik dapat menjadi solusi dari permasalahan tersebut. Fitur-fitur statistik seperti posisi kalimat dalam paragraf, posisi keseluruhan kalimat, data numerik, tanda koma terbalik, panjang kalimat dan kata kunci memiliki peran yang penting untuk dijadikan parameter peringkasan. Dari pengujian fitur statistik yang telah dilakukan dengan memakai nilai k=3, metode ini menghasilkan nilai rata-rata precision, recall dan f –measure terbaik pada set fitur 9 dengan nilai masing-masing sebesar 0.75, 0.71 dan 0.72. Dari pengujian tersebut disimpulkan bahwa fitur yang memiliki pengaruh signifikan terhadap naik dan turunnya nilai precision dan recall adalah fitur posisi kalimat dalam paragraf dan fitur posisi keseluruhan kalimat. Kemudian dari hasil pengujian variasi k pada set fitur terbaik, didapatkan nilai set fitur yang maksimal ketika k=1 dengan nilai rata-rata precision, recall dan f-measure sebesar 0.89, 0.74 dan 0.81.

Kata Kunci: *text mining, peringkasan teks, K-Nearest Neighbor, fitur statistik*

Abstract

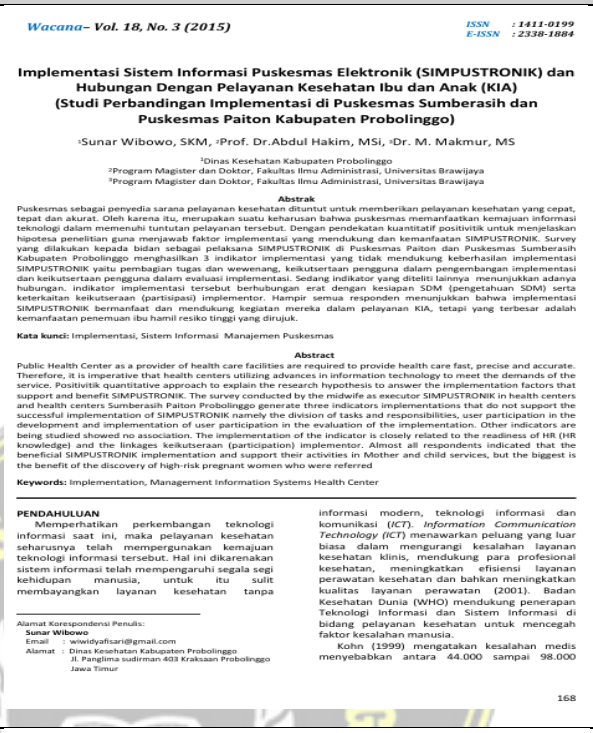
Now days, information about healthy has been widely scattered and very easily obtained through the online website. But, within largest information that contain in the text of article make the reader can't understand about contents of the text. So, we need a system that can summarize a text to make easy the reader in understanding the contents of the text. Automatic text summary using k-nearest neighbor based on statistical features can be solution about the problem. Statistical features such as position of a sentence in a paragraph, overall sentence position, numerical data, inverted commas, the length of the sentence and keyword has important influence become parameter in summarization. From testing of statistical features that have been done by using k = 3, this method get result the best value of precision, recall and f -measure on feature set 9 with values 0.75, 0.71 and 0.72. From the test can concluded that the features that have a significant influence on the rise and fall of precision and recall values are position of a sentence in paragraph and sentence overall position. And then, from the test of k variation on the best feature set, we get maximum feature set value when k = 1 with the average value of precision, recall and f-measure of 0.89, 0.74 and 0.81.

Keywords: *text mining, text summarization, K-Nearest Neighbor, statistical feature*

1. PENDAHULUAN

Berkembangnya internet dengan pesat berdampak terhadap bertambahnya jumlah informasi yang mengakibatkan sangat sulit untuk mendapatkan informasi secara efisien (Desai & Shah, 2016). Berita merupakan sebuah informasi yang berguna untuk menyampaikan fakta kepada seluruh orang. Dengan berkembangnya teknologi, kini semakin mudah untuk mendapatkan berita terupdate. Banyak situs yang menyediakan informasi berita yang terpercaya dan beragam topik, seperti kompas.com, detik.com, detikhealth.com dan masih banyak lagi situs lainnya. Masing-masing situs tersebut memiliki beraneka ragam topik berita antara lain olahraga, politik, kesehatan,

No	PDF Sumber	Hasil PDF
2	Garuda (2)	Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN: 2548-964X Vol. 2, No. 11, November 2018, hlm. 4414-4420 http://j-ptiik.ub.ac.id **Implementasi *Fuzzy K-Nearest Neighbour (FK-NN)* Untuk Pemilihan Keminatan Mahasiswa Teknik Informatika (Studi Kasus : Program Studi Teknik Informatika Fakultas Ilmu Komputer Universitas Brawijaya)** **Dhony Lastiko Widyastomo[1], Indriati[2], Rizal Setya Perdana[3]** Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Brawijaya Email:dhonylastiko@gmail.co.id, [2]indriati.tif@ub.ac.id, [3]rizalespe@ub.ac.id **Abstrak** Pengambilan keminatan merupakan salah satu tahapan yang harus dilalui oleh seorang mahasiswa dalam menempuh masa studinya. Program studi Informatika Universitas Brawijaya memiliki 4 keminatan yang terdiri dari keminatan yang berbeda. Sayangnya karena kurangnya pengetahuan dan berbagai hambatan, menyebabkan mahasiswa mendapat masalah dalam pemilihan keminatan yang berakibat pada kesulitan proses belajar yang dilalui oleh mahasiswa. Untuk memberikan suatu solusi, dibutuhkan sistem klasifikasi keminatan yang dapat memberikan rekomendasi keminatan berdasaran kemampuan mahasiswa. Proses klasifikasi keminatan menggunakan metode *fuzzy k-nearest neighbor* menghitung nilai jarak tiap kelas target yang diinginkan dengan memanfaatkan nilai K untuk menghasilkan keluaran berupa keminatan yang berdasar pada nilai prasyarat 4 keminatan yang ada pada program studi Informatika Fakultas Ilmu Komputer Brawijaya. Berdasarkan hasil penelitian menggunakan 200 data mahasiswa pada lulusan Teknik Informatika fakultas Ilmu Komputer, akurasi terbesar yang didapatkan oleh sistem klasifikasi sebesar 87,5% pada nilai K=3. Dengan nilai akurasi terkecil sebesar 62,5% pada nilai K=10. **Kata kunci**: *keminatan, klasifikasi, Fuzzy K-Nearest Neighbor* **Abstract** *Concentration selection is one of few steps for a students to finish their studies. Informatics programee have 4 concentration consist of Artificial Inteligent(AI), Software Engineering (SE), Network and Game. Unfortunately because the limited and many internal problems from the students causing some problem for the concetration selection. To solve the problem of selection, a system who can give a classification is needed to give the solution. A classification for concentracion selection uses fuzzy k-nearest neighbor for its method. The method works with calculate the number of K value to Process the classification of 4 study concentration and resulting the recomendation class of concentration class based on the student data. Based on the research of study using 200 data of the students of Informatics engineering, from 2011 to 2013, the biggest accuracy was produced by K value=3 and have 87,5% accuracy. While the lowest precentages of accuracy was produced by K value=10 with the averages of 67,5% accuracy.* **Keywords:** *concentration, studies, classification, Fuzzy K-Nearest Neighbor* **PENDAHULUAN** Dunia perkuliahan adalah masa terakhir bagi seseorang untuk mengemban ilmu dan semua pengetahuan yang akan berguna bagi kehidupannya nanti. Fakultas Ilmu Komputer merupakan salah satu fakultas yang berdiri di Universitas Brawijaya yang memiliki misi menghasilkan lulusan yang memiliki kompetensi di bidang TIK, berjiwa enterpreneur dan dapat dipercaya sehingga mampu bekerjasama dan memberikan kontribusi di Mahasiswa Teknik Informatika diarahkan agar memiliki karakter yang khas sebagai kekuatan untuk bersaing pada dunia nyata. Karakter lulusan Teknik Informatika ini disusun berdasarkan *Computing Curricula 2013* yang dikombinasikan dengan karakter khas mahasiswa maupun lulusan Universitas Brawijaya dan serta Fakultas Ilmu Komputer
3	Garuda (3)	Jurnal Sains Komputer & Informatika (J-SAKTI) Volume (1) No. 1 Maret 2017 ISSN:2548-9771/EISSN: 2549-7200 http://ejurnal.tunasbangsa.ac.id/index.php/jsakti **SISTEM INFORMASI JADWAL PERKULIAHAN MENGGUNAKAN MEDIA TELEVISI (STUDI KASUS PADA JURUSAN TEKNIK INFORMATIKA UPN "VETERAN" YOGYAKARTA)** **Sundari Retno Andani, Subastian Wibowo, Poningsih** Program Studi Manajemen Informatika, AMIK Tunas Bangsa Pematangsiantar Jln. Jenderal Sudirman Blok A No. 1,2,3 Pematangsiantar Magister STMIK AMIKOM Yogyakarta Jln. Ring Road Utara, Condong Catur, Sleman, Yogyakarta Sundari.ra@amiktunasbangsa.ac.id, subastian.w@students.amikom.ac.id, poningsih@amiktunasbangsa.ac.id ***Abstract*** *Students are always confused with a schedule of lectures,* about *time of lectures, room that will be used, even about come or not the lecturer on class. As a result, students must always go to the education department to inquire this issue. It is very ineffective. But there is no system that helps students in overcoming this problem. In this paper, the authors build a system with the title is the information system of the schedules of lectures using television. The system is built using Delphi 6.0 programming language and uses Microsoft Access 2003 as the database. To operate this system requires a CPU and a television screen as an output device. This system provides information on the schedule of lectures to the students through a television screen. Information provided includes schedule of lectures, room that will be used, the certainty of the lecture will take place, and informing announcements and activities that will take place. These systems also support the effectiveness of the performance of educational staff in the conduct of daily operations.* ***Keywords****: information system, schedule of lectures* ***Abstrak*** Mahasiswa selalu dibingungkan dengan jadwal perkuliahan, baik mengenai waktu perkuliahan, ruangan yang akan digunakan, bahkan mengenai datang atau tidaknya dosen yang mengajar. Akibatnya mahasiswa harus selalu ke bagian pendidikan untuk menanyakan masalah ini. Hal ini sangat tidak efektif. Namun belum ada sistem yang membantu mahasiswa dalam mengatasi masalah ini. Pada jurnal ini, penulis membangun sebuah sistem dengan judul sistem informasi jadwal perkuliahan menggunakan media televisi. Sistem ini dibuat dengan menggunakan bahasa pemrograman *Delphi 6.0* dan menggunakan *Microsoft Access 2003* sebagai databasenya. Untuk

No	PDF Sumber	Hasil PDF
4	Garuda (4)	
5	Garuda (5)	

Wacana – Vol. 18, No. 3 (2015)

ISSN : 1411-0199
E-ISSN : 2338-1884

Implementasi Sistem Informasi Puskesmas Elektronik (SIMPUSTRONIK) dan Hubungan Dengan Pelayanan Kesehatan Ibu dan Anak (KIA) (Studi Perbandingan Implementasi di Puskesmas Sumberasih dan Puskesmas Paiton Kabupaten Probolinggo)

[1]Sunar Wibowo, SKM, [2]Prof. Dr.Abdul Hakim, MSi, [3]Dr. M. Makmur, MS

[1]Dinas Kesehatan Kabupaten Probolinggo
[2]Program Magister dan Doktor, Fakultas Ilmu Administrasi, Universitas Brawijaya
[3]Program Magister dan Doktor, Fakultas Ilmu Administrasi, Universitas Brawijaya

Abstrak

Puskesmas sebagai penyedia sarana pelayanan kesehatan dituntut untuk memberikan pelayanan kesehatan yang cepat, tepat dan akurat. Oleh karena itu, merupakan suatu keharusan bahwa puskesmas memanfaatkan kemajuan informasi teknologi dalam memenuhi tuntutan pelayanan tersebut. Dengan pendekatan kuantitatif positivitik untuk menjelaskan hipotesa penelitian guna menjawab faktor implementasi yang mendukung dan kemanfaatan SIMPUSTRONIK. Survey yang dilakukan kepada bidan sebagai pelaksana SIMPUSTRONIK di Puskesmas Paiton dan Puskesmas Sumberasih Kabupaten Probolinggo menghasilkan 3 indikator implementasi yang tidak mendukung keberhasilan implementasi SIMPUSTRONIK yaitu pembagian tugas dan wewenang, keikutsertaan pengguna dalam pengembangan implementasi dan keikutsertaan pengguna dalam evaluasi implementasi. Sedang indikator yang diteliti lainnya menunjukkan adanya hubungan. indikator implementasi tersebut berhubungan erat dengan kesiapan SDM (pengetahuan SDM) serta keterkaitan keikutseraan (partisipasi) implementor. Hampir semua responden menunjukkan bahwa implementasi SIMPUSTRONIK bermanfaat dan mendukung kegiatan mereka dalam pelayanan KIA, tetapi yang terbesar adalah kemanfaatan penemuan ibu hamil resiko tinggi yang dirujuk.

Kata kunci: Implementasi, Sistem Informasi Manajemen Puskesmas

Abstract

Public Health Center as a provider of health care facilities are required to provide health care fast, precise and accurate. Therefore, it is imperative that health centers utilizing advances in information technology to meet the demands of the service. Positivitik quantitative approach to explain the research hypothesis to answer the implementation factors that support and benefit SIMPUSTRONIK. The survey conducted by the midwife as executor SIMPUSTRONIK in health centers and health centers Sumberasih Paiton Probolinggo generate three indicators implementations that do not support the successful implementation of SIMPUSTRONIK namely the division of tasks and responsibilities, user participation in the development and implementation of user participation in the evaluation of the implementation. Other indicators are being studied showed no association. The implementation of the indicator is closely related to the readiness of HR (HR knowledge) and the linkages keikutseraan (participation) implementor. Almost all respondents indicated that the beneficial SIMPUSTRONIK implementation and support their activities in Mother and child services, but the biggest is the benefit of the discovery of high-risk pregnant women who were referred

Keywords: Implementation, Management Information Systems Health Center

PENDAHULUAN

Memperhatikan perkembangan teknologi informasi saat ini, maka pelayanan kesehatan seharusnya telah mempergunakan kemajuan teknologi informasi tersebut. Hal ini dikarenakan sistem informasi telah mempengaruhi segala segi kehidupan manusia, untuk itu sulit membayangkan layanan kesehatan tanpa informasi modern, teknologi informasi dan komunikasi (*ICT*). *Information Communication Technology (ICT)* menawarkan peluang yang luar biasa dalam mengurangi kesalahan layanan kesehatan klinis, mendukung para profesional kesehatan, meningkatkan efisiensi layanan perawatan kesehatan dan bahkan meningkatkan kualitas layanan perawatan (2001). Badan Kesehatan Dunia (WHO) mendukung penerapan Teknologi Informasi dan Sistem Informasi di bidang pelayanan kesehatan untuk mencegah faktor kesalahan manusia.

Kohn (1999) mengatakan kesalahan medis menyebabkan antara 44.000 sampai 98.000

Alamat Korespondensi Penulis:
Sunar Wibowo
Email : wiwidyafisari@gmail.com
Alamat : Dinas Kesehatan Kabupaten Probolinggo
Jl. Panglima sudirman 403 Kraksaan Probolinggo
Jawa Timur

Jurnal Pendidikan dan Konseling
Volume 4 Nomor 4 Tahun 2022
E-ISSN: 2685-936X dan P-ISSN: 2685-9351
Universitas Pahlawan Tuanku Tambusai

Penggunaan Media *Online* dalam Meningkatkan Kemampuan Vocabulary pada Mahasiswa Teknik Informatika Semester 2 Fakultas Teknik Universitas Wiraraja

Hanifatur Rizqi[1], Ach. Andiriyanto[2]
[1,2]Universitas Wiraraja
Email: Hanifarizqi7@gmail.com, aryauri@wiraraja.ac.id

Abstrak

Vocabulary atau kosakata adalah perbendaharaan kata yang memiliki makna beragam kumpulan kata yang dimiliki seseorang, entitas, ataupun negara dalam bahasa tertentu. Kosakata merupakan bagian penting dalam pembelajaran karena pengusaan vocabulary atau kosakata dalam bahasa inggris pasti selalu berhubungan dengan empat keterampilan berbahasa seperti mendengarkan (listening), berbicara (speaking), membaca (reading), dan menulis (writing). Dengan menggunakan media online, peneliti berharap dapat memberu pengaruh positif terhadap motivasi belajar mahasiswa sehingga tercipta pembelajaran yang baik. Dalam penelitian ini, peneliti menggunakan *True Experimental Design*, dimana sampel yang digunakan adalah 15 mahasiswa kelompok eksperimen (Kelompok A) dan 15 mahasiswa kelompok kontrol (Kelompok B) yang diambil secara random dari seluruh populasi. Berdasarkan pengumpulan data hasil t-test menunjukkan bahwa hasil uji-T yang dilaksanakan sebelum *treatment (pretest)* adalah 0,08 sedangkan hasil tes setelah *treatment (posttest)* adalah 0,096. Sedangkan tabel T menunjukkan 1,584. Sehingga dapat disimpulkan bahwa penggunaan Media Online kurang berpengaruh dalam meningkatkan vocabulary atau kosakata bahasa Inggris pada mahasiswa Teknik Informatika Universitas Wiraraja.

Kata Kunci : *Media Online, Vocabulary*

Abstract

Vocabulary has the meaning of various collections of words that are owned by a person, entity, or country in a particular language. Vocabulary is an important part of teaching learning because mastery of vocabulary in English must always be related to the four language skills such as listening, speaking, reading, and writing. By using online media, researchers hope to have a positive influence on student learning motivation, so that good learning is created. In this study, the researcher used a True Experimental Design, where the samples used were 15 experimental group students (Group A) and 15 control group students (Group B) which were taken randomly from the entire population. Based on data collection, the results of the T-test showed that the result of the t-test conducted before treatment (pretest) was 0.08, while the result of the test after treatment (posttest) was 0.096. While the T table shows 1.584. So it can be concluded that the use of online media is less influential in improving English vocabulary for Informatics Engineering students at Wiraraja University.

Pada Tabel 4. 1 dapat di lihat bahwa 5 artikel PDF yang digunakan dalam penelitian ini diambil dari platform garuda dan masing-masing artikel memberikan kontribusi yang relevan terkait topik teknik informatika.

4.2 Hasil Modeling

Pada tahap ini, dilakukan serangkaian proses untuk mengonversi dokumen TEI XML ke JSON, meringkas isi teks menggunakan SciBERT, serta menyimpan hasilnya untuk dokumentasi lebih lanjut. Berikut adalah langkah-langkah implementasi preprocessing model:

1. Extraksi GROBID
   a. Membaca Dokumen XML

```
def parse_tei_to_json(tei_file_path):
    """
    Parse TEI XML file to JSON format.
    """
    # Parse file XML dan ambil elemen root
    tree = ET.parse(tei_file_path)
    root = tree.getroot()
    # DNamespace untuk TEI (perlu karena elemen-elemen
      menggunakan namespace ini)
    namespace = {'tei': 'http://www.tei-c.org/ns/1.0'}
```

Pada *source code* diatas digunakan untuk membaca dokumen XML menggunakan pustaka `ElementTree`. Fungsi ini bertanggung jawab untuk membuka dan mengolah file XML ager dapat diproses lebih lanjut.

   b. Parsing XML ke JSON

```
# Inisialisasi struktur data JSON
data = {}

# Ekstrak judul (title)
title_elem = root.find('.//tei:titleStmt/tei:title',
namespace)
data['title'] = title_elem.text.strip() if title_elem is
not None else None
```

```
# Ekstrak informasi penulis (authors)
authors = []
for author in
root.findall('.//tei:titleStmt/tei:author', namespace):
surname = author.find('tei:persName/tei:surname',
namespace)
forename = author.find('tei:persName/tei:forename',
namespace)
affiliation_elems =
author.findall('tei:affiliation/tei:orgName', namespace)
affiliations = [aff.text.strip() for aff in
affiliation_elems if aff is not None]
authors.append({
'forename': forename.text.strip() if forename is not
None else None,
'surname': surname.text.strip() if surname is not None
else None,
'affiliations': affiliations
})
data['authors'] = authors

# Ekstrak abstrak
abstract_elem =
root.find('.//tei:profileDesc/tei:abstract', namespace)
data['abstract'] =
''.join(abstract_elem.itertext()).strip() if
abstract_elem is not None else None

# Ekstrak konten body
body = []
for div in root.findall('.//tei:body/tei:div',
namespace):
head = div.find('tei:head', namespace)
paragraphs = [p.text.strip() for p in
div.findall('tei:p', namespace) if p.text]
body.append({
section_title': head.text.strip() if head is not None
else None,
```

```
        'content': paragraphs
        })
        data['body'] = body

        return data
```

Pada *source code* diatas merupakan tampilan untuk parsing file XML berformat TEI menjadi format JSON. Pertama, dilakukan *import library* `xml.etree.ElementTree` untuk membaca file XML. Selanjutnya, fungsi `ET.parse` digunakan untuk memuat file XML dan menginisiasi elemen *root*. *Namespace* khusus TEI didefinisikan agar elemen-elemen dalam XML dapat diakses.

Fungsi ini dimulai dengan mengekstrak elemen `title` untuk mendapatkan judul dokumen. Setelah itu, elemen `author` diekstrak untuk mendapatkan informasi penulis, termasuk nama depan, nama belakang, dan afiliasi. Kemudian, elemen abstract diambil untuk mendapatkan `abstrak` dokumen dalam bentuk teks. Selanjutnya, isi utama dokumen (`body`) diekstrak dengan mengambil setiap elemen `div`, yang mencakup judul seksi dan daftar paragrafnya. Data yang berhasil diekstrak disusun ke dalam format JSON dan dikembalikan oleh fungsi ini.

c. Merangkum Isi Paragraf

```
def extractive_summary_with_SciBERT(paragraph, tokenizer,
model, max_sentences=3):
    """
    Merangkum paragraf menggunakan pendekatan ekstraktif
    dengan SciBERT.
    """
    # Memecah paragraf menjadi kalimat-kalimat
     sentences = nltk.tokenize.sent_tokenize(paragraph)

    # Tokenisasi setiap kalimat
    inputs = tokenizer(sentences, return_tensors="pt",
    padding=True, truncation=True, max_length=512)
    # Mendapatkan embedding dari SciBERT untuk setiap kalimat
```

```
 with torch.no_grad():
      outputs = model(**inputs)
      sentence_embeddings =
      outputs.last_hidden_state.mean(dim=1)
# Menghitung kemiripan antara kalimat-kalimat dengan teks
asli (paragraf)
original_embedding = model(**tokenizer(paragraph,
return_tensors="pt", truncation=True,
max_length=512)).last_hidden_state.mean(dim=1)

# Memisahkan tensor dari graf komputasi dan
mengonversinya ke numpy
original_embedding_np =
original_embedding.detach().cpu().numpy()
sentence_embeddings_np =
sentence_embeddings.detach().cpu().numpy()

# Menghitung kemiripan antara kalimat dan teks asli
menggunakan cosine similarity
similarity_scores =
cosine_similarity(original_embedding_np,
sentence_embeddings_np

# Menyortir kalimat berdasarkan skor kemiripan
ranked_sentences = sorted([(sentences[i],
similarity_scores[0][i]) for i in range(len(sentences))],
key=lambda x: x[1], reverse=True)

# Ambil kalimat dengan skor kemiripan tertinggi
summary = " ".join([ranked_sentences[i][0] for i in
range(min(max_sentences, len(ranked_sentences)))])

return summary
```

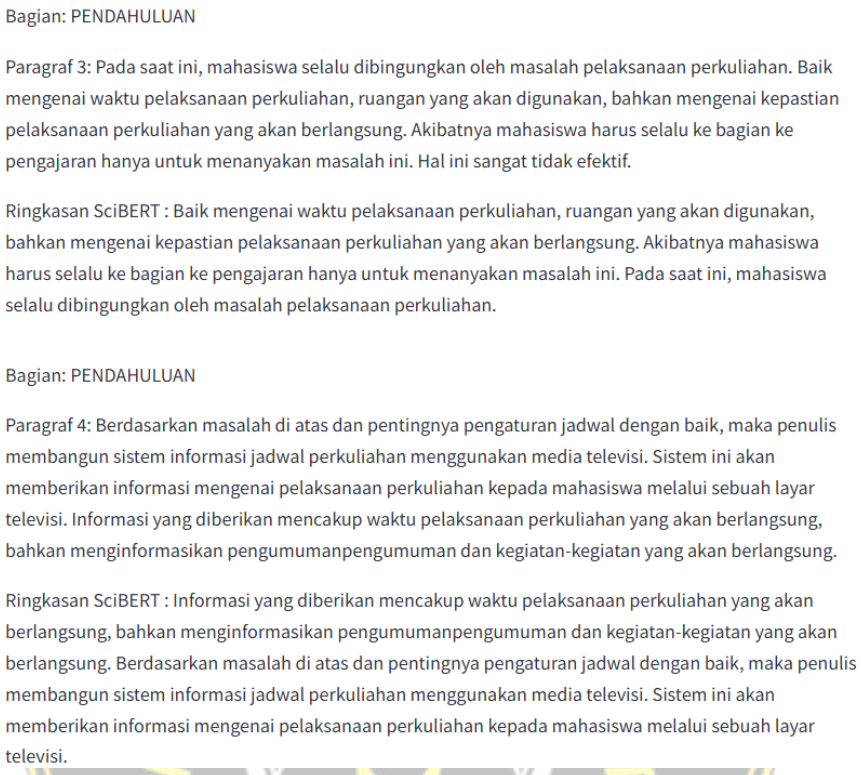
Pada source code di atas, terdapat sebuah fungsi Python bernama `extractive_summary_with_SciBERT`, yang digunakan untuk merangkum paragraf teks secara otomatis menggunakan model SciBERT. Proses dimulai

dengan tokenisasi paragraf menggunakan tokenizer SciBERT, yang mengubah paragraf menjadi bentuk token agar bisa diproses oleh model. Selanjutnya, model melakukan *embedding* untuk setiap kalimat dengan menggunakan SciBERT dan menghasilkan representasi vektor dari setiap kalimat. Kemudian, fungsi memilih kalimat yang paling relevan dengan teks asli berdasarkan kemiripan (*similarity*) antara kalimat tersebut dengan paragraf utuh menggunakan *cosine similarity*.

Dengan cara ini, kalimat yang memiliki kemiripan tertinggi dengan teks asli akan diprioritaskan dalam ringkasan. Setelah memilih kalimat-kalimat yang relevan, kalimat-kalimat tersebut disusun ulang dalam urutan kemiripan tertinggi dan digabungkan menjadi sebuah ringkasan. Untuk memastikan ringkasan tidak terlalu panjang, angka `max_sentences` digunakan untuk membatasi jumlah kalimat yang diambil (default: 3 kalimat). Fungsi ini mengembalikan ringkasan teks yang lebih pendek namun tetap mempertahankan informasi utama dari paragraf asli. Dengan menggunakan pendekatan ekstraktif ini, fungsi berfokus pada memilih kalimat yang relevan dengan teks asli, bukan memodifikasi atau merangkum bagian informasi secara bebas.

2. Hasil Peringkasan Teks Otomatis

Hasil peringkasan teks otomatis yang ditampilkan pada gambar 4.1 menunjukan bahawa sistem telah berhasil melakukan peringkasan menggunakan model SciBERT.



Gambar 4. 1 hasil Ringkasan menggunakan model SciBERT

Gambar 4. 1 adalah Hasil peringkasan teks otomatis menggunakan SciBERT cukup efektif karena mempersingkat teks tanpa menghilangkan inti dari informasi aslinya. Model ini dapat menangkap poin-poin utama dengan baik, sehingga maknanya tetep jelas. Namun ada beberapa kekurangan, seperti perubahan susunan informasi atau penghapusan bagian tertentu yang sebenarnya bisa memberi konteks tambahan. Hal ini bisa membuat pembaca kehilangan pemahaman yang lebih lengkap. Jadi, meskipun SciBERT sangat berguna untuk meringkas teks, penggunaanya tetep harus di sesuaikan, terutama jika struktur dan alur cerita dalam teks asli sangat penting untuk di pertahankan. Selain itu hasil ringkasan otomatis tetep berbentuk narasi yang lengkap, sehingga tetep mudah di fahami dan bisa digunakan sesuai kebutuhan, terutama jika inggin merangkum teks panjang dengan cepat.

4.3 Hasil Evaluasi

Pada tabel di bawah menunjukan perbandinga ringkasan manual dan sistem artikel ilmiah per paragraf.

1. Hasil Evaluasi Sistem

Gambar 4. 2 Hasil ringkasan manual dan sistem

Ringkasan Manual	Ringkasan Sistem
Revolusi di dunia ilmu pengetahuan dan teknologi telah memicu lahirnya pola baru dalam penyampaian maupun penerimaan informasi, dimana pola penyampaian informasi yang lazim dilakukan sekarang adalah memanfaatkan komputer, monitor dan televisi sebagai piranti medianya.	Revolusi di dunia ilmu pengetahuan dan teknolog telah memicu lahirnya pola baru dalam penyampaian maupun penerimaan informasi, dimana pola penyampaian informasi yang lazim dilakukan sekarang ini adalah memanfaatkan komputer, monitor dan televisi sebagai piranti medianya.
mahasiswa dibingungkan oleh masalah pelaksanaan perkuliahan, seperti waktu, ruangan, dan kepastian jadwal. Akibatnya, mahasiswa harus sering ke bagian pengajaran untuk menanyakan hal ini, yang tidak efektif	baik mengenai waktu pelaksanaan perkuliahan, ruangan yang akan digunakan, bahkan mengenai kepastian pelaksanaan perkuliahan yang akan berlangsung. akibatnya siswa haru selalu ke bagian pengajaran hanya untuk menanyakan masalah ini. pada saat ini, siswa selalu dibingungkan oleh masalah pelaksanaan perkuliahan.
berdasarkan masalah pengaturan jadwal, penulis membangun sistem informasi jadwal perkuliahan menggunakan media televisi. Sistem ini memberikan informasi kepada mahasiswa melalui layar televisi, mencakup waktu pelaksanaan perkuliahan, serta pengumuman dan kegiatan yang akan berlangsung.	informasi yang diberikan mencakup waktu pelaksanaan perkuliahan yang akan berlangsung, bahkan menginformasikan pengumumanpengumuman dan kegiatan-kegiatan yang akan berlangsung. berdasarkan masalah di ata dan pentingnya pengaturan jadwal dengan baik, maka penuli membangun sistem informasi jadwal perkuliahan menggunakan media televisi. sistem ini akan memberikan informasi mengenai pelaksanaan perkuliahan kepada mahasiswa melalui sebuah layar televisi.
penyampaian jadwal dan informasi kepada mahasiswa sebelumnya menggunakan media kertas dan	saat ini persaingan antar perguruan tinggi begitu ketat dalam menghasilkan sumber daya manusia yang unggul dan

papan pengumuman, yang tidak efisien dari segi waktu. Penyampaian informasi yang cepat, efisien, dan akurat dapat meningkatkan pengakuan dari masyarakat. Persaingan antar perguruan tinggi semakin ketat dalam menghasilkan sumber daya manusia unggul dan berkualitas, dengan kemampuan analisis dan logika berpikir yang tajam.	berkualitas, baik teori maupun praktek, serta dituntut memiliki kemampuan analisi dan logika berpikir dengan cermat dan tajam. penyampaian informasi yang cepat, efisien dan akurat juga dapat meningkatkan dan mendapatkan pengakuan dari masyarakat. sebelumnya, penyampaian jadwal dan informasi kepada mahasiswa masih menggunakan media kerta dan papan pengumuman.
Sistem informasi terdiri dari komponen-komponen yang disebut blok bangunan, yaitu: blok input, blok model, blok keluaran, blok teknologi, blok basis data, dan blok kendali.	sistem informasi terdiri dari komponen-kompenen yang disssebut dengannn issstilah blok bangunan (build block), yaitu blok input (input block), blok model (model block), blok keluaran (output block), blok teknolog (technolog block), blok basi data (databas block) dan blok kendali (control block).

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 4. 2 Hasil Evaluasi

Metode ROUGE	***Precision***	***Recall***	***F1-score***
ROUGE-1	0.7109	0.9371	0.8084
ROUGE-2	0.5319	0.7018	0.6051
ROUGE-L	0.5438	0.7168	0.6184

Hasil pengukuran menggunakan metrik ROUGE menunjukkan kualitas ringkasan otomatis yang cukup baik. Pada ROUGE-1, yang mengukur kemiripan kata individu, sistem mencatat *Precision* 0.7109, *Recall* 0.9371, dan *F1-score* 0.8084, menunjukkan kemampuan yang baik dalam menangkap kata-kata penting

dari dokumen asli. Untuk ROUGE-2, yang mengukur kemiripan pasangan kata (*bigrams*), hasilnya adalah *Precision* 0.5319, *Recall* 0.7018, dan *F1-score* 0.6051, menunjukkan bahwa sistem cukup baik dalam mempertahankan konteks lokal antar kata, meskipun ada penurunan dibandingkan ROUGE-1. Pada ROUGE-L, yang mengukur kesamaan urutan kata terpanjang (longest common subsequence), sistem memperoleh *Precision* 0.5438, *Recall* 0.7168, dan *F1-score* 0.6184, menandakan bahwa sistem cukup baik dalam mempertahankan struktur kalimat atau urutan kata yang relevan. Secara keseluruhan, meskipun terdapat sedikit penurunan di beberapa metrik, nilai *F1-score* yang tercatat menunjukkan kualitas ringkasan otomatis yang sangat baik.

a. Analisis Perbandingan Ringkasan Manual dan Ringkasan Sistem

Pada tabel di atas menunjukan perbandingan anatra ringkasan manual dan ringkasan sistem seberapa detail dan kelengkapan informasi. Ringkasan manual lebih ringkas dan langsung pada inti, menyampaikan informasi secara efisien tanpa pengulangan, sementara ringkasan sistem lebih terperinci dan kadang mengulang informasi atau menambah penjelasan tambahan, seperti penggunaan istilah bahasa Inggris. Meskipun sistem memberikan informasi yang lebih lengkap, hal ini membuatnya terasa lebih panjang dan pengulangan informasi yang tidak perlu. Perbedaan ini terjadi karena ringkasan manual berfokus pada kesederhanaan, sedangkan sistem berusaha memberikan detail yang lebih jelas dan mendalam.

Berdasarkan perbandingan antara ringkasan manual dan sistem, sistem peringkasan berbasis aplikasi menawarkan solusi dengan menyajikan ringkasan yang lebih terperinci dan lengkap. Meskipun terkadang terdapat pengulangan informasi yang tidak perlu, sistem ini memungkinkan peneliti untuk dengan cepat mengakses informasi yang relevan tanpa harus membaca seluruh artikel ilmiah. Ini sangat membantu ketika peneliti harus memproses banyak artikel secara efisien, menghemat waktu, dan fokus pada bagian yang paling penting dan relevan.

2. Hasil evaluasi Artikel PDF Baru

Pada tabel di bawah menunjukan perbandinga ringkasan manual pada PDF garuda (1) dan ringkasan sistem pada garuda (2) artikel ilmiah per paragraf.

Ringkasan Manual Garuda (1)	Ringkasan Sistem Garuda (2)
Revolusi di dunia ilmu pengetahuan dan teknologi telah memicu lahirnya pola baru dalam penyampaian maupun penerimaan informasi, dimana pola penyampaian informasi yang lazim dilakukan sekarang adalah memanfaatkan komputer, monitor dan televisi sebagai piranti medianya.	Perkembangan internet yang semakin cepat dan terus mengalami inovasi setiap saat, mendorong berkembangnya informasi dan memunculkan berbagai situs-situs berita online baik nasional maupun skala lokal di Indonesia. Tercatat pada sebuah riset 2016 oleh Indonesia Digital Association (IDA) terkait sumber yang digunakan konsumen dalam pencarian berita, sumber online memperoleh peringkat yang tertinggi
Mahasiswa dibingungkan oleh masalah pelaksanaan perkuliahan, seperti waktu, ruangan, dan kepastian jadwal. Akibatnya, siswa harus sering ke bagian pengajaran untuk menanyakan hal ini, yang tidak efektif	Kemudahan ini menyebabkan informasi semakin banyak dan beragam yang tersedia secara online, berita biasanya terdiri dari teks yang panjang, hingga membutuhkan waktu untuk memahami inti (informasi penting atau ide pokok) dari berita tersebut, untuk itu diperlukan sebuah peringkasan teks otomatis yang dapat membantu mengekstrak informasi penting dari isi berita. Salah satu bidang yang mampu mengatasi masalah ini ialah Text Summarization (Peringkasan Teks).
berdasarkan masalah pengaturan jadwal, penulis membangun sistem informasi jadwal perkuliahan menggunakan media televisi. Sistem ini memberikan informasi kepada siswa melalui layar televisi, mencakup waktu pelaksanaan perkuliahan, serta pengumuman dan kegiatan yang akan berlangsung.	Penelitian peringkasan dokumen pertama kali dimulai pada tahun 1958 yakni The Automatic Creation of Literature Abstrak oleh Luhn, penelitian tersebut menghasilkan sebuah ringkasan dengan menghitung nilai frekuensi kata dan kalimat untuk menentukan hasil ringkasan yang terbaik. Di tahun yang sama Baxendle menemukan fakta bahwa 85% kalimat yang mengadung topik dari isi berita berada pada awal kalimat dan 7% pada akhir kalimat
penyampaian jadwal dan informasi kepada mahasiswa sebelumnya menggunakan media kertas dan papan pengumuman, yang tidak efisien dari segi waktu. Penyampaian informasi yang cepat, efisien, dan akurat dapat meningkatkan pengakuan masyarakat. Persaingan antar perguruan tinggi semakin ketat dalam menghasilkan sumber daya manusia unggul dan berkualitas, dengan kemampuan analisis dan logika berpikir yang tajam.	Umumnya peringkasan dokumen diklasifikasi menjadi dua yaitu peringkasan ekstraktif dan peringkasan abstraktif
Sistem informasi terdiri dari komponen-komponen yang disebut blok bangunan, yaitu: blok input, blok model, blok keluaran, blok teknologi, blok basis data, dan blok kendali.	Peringkasan dokumen dengan LSA menggunakan Singular Value Decomposition (SVD) untuk menemukan kesamaan semantik kata dan kalimat pada sebuah dokumen. Salah satu algoritma yang digunakan untuk menyelesaikan peringkasan ekstraktif adalah Latent Semantic Analysis (LSA). SVD adalah model hubungan antara kata dan kalimat

Gambar 4. 3 Contoh hasil manual dan sistem PDF yang berbeda

Setelah mengimplementasikan sistem peringkasan artikel ilmiah, evaluasi dilakukan menggunakan metrik ROUGE. Pengujian dilakukan dengan membandingkan 10 paragraf hasil ringkasan manual dari PDF Garuda (1) dan 10 paragraf hasil ringkasan sistem dari PDF Garuda (2). Berikut hasil evaluasi menggunakan metrik ROUGE

Tabel 4. 3 Contoh evaluasi artikel PDF baru

Metode ROUGE	***Precision***	***Recall***	***F1-score***
ROUGE-1	0.1535	0.3322	0.2099
ROUGE-2	0.0113	0.0246	0.0155
ROUGE-L	0.0711	0.1538	0.0972

Hasil evaluasi menggunakan metrik ROUGE menunjukkan bahwa kualitas ringkasan otomatis masih rendah. Pada ROUGE-1, yang mengukur kesamaan kata individu antara ringkasan dan teks referensi, sistem mencatat Precision sebesar 0.1535, Recall 0.3322, dan F1-score 0.2099, menunjukkan bahwa hanya sebagian kecil kata-kata penting yang berhasil ditangkap. Untuk ROUGE-2, yang menilai kemiripan pasangan kata (bigrams), hasilnya lebih rendah dengan Precision 0.0113, Recall 0.0246, dan F1-score 0.0155, mengindikasikan bahwa sistem kurang mampu mempertahankan hubungan antar kata dalam konteks lokal. Sementara itu, pada ROUGE-L, yang mengukur kesamaan berdasarkan urutan kata terpanjang yang cocok, diperoleh Precision 0.0711, Recall 0.1538, dan F1-score 0.0972, yang menunjukkan kelemahan sistem dalam mempertahankan struktur kalimat yang relevan. Secara keseluruhan, nilai F1-score yang rendah di semua metrik menunjukkan bahwa kualitas ringkasan masih perlu ditingkatkan agar lebih akurat dan sesuai dengan teks referensi.

a. Analisis Perbandingan Ringkasan Manual PDF garuda (1) dan Ringkasan Sistem PDF Garuda (2).

Perbandingan antara ringkasan manual dari PDF Garuda (1) dan ringkasan sistem dari PDF Garuda (2) menunjukkan bahwa validasi sistem sangat bergantung pada proses filtering. Karena dokumen yang diringkas

berbeda, hasil validasi menjadi tidak sesuai. Ringkasan manual lebih akurat karena dibuat berdasarkan pemahaman manusia, sementara sistem peringkasan otomatis hanya mengandalkan algoritma yang memilih informasi berdasarkan pola tertentu. Evaluasi dengan metrik ROUGE menghasilkan skor rendah, yang menandakan bahwa perbedaan sumber dokumen memengaruhi kualitas ringkasan. Oleh karena itu, untuk validasi yang lebih akurat, penting memastikan dokumen yang dibandingkan berasal dari sumber yang sama agar proses filtering dalam sistem bekerja secara optimal.

4.4 Hasil Implementasi Aplikasi Streamlit

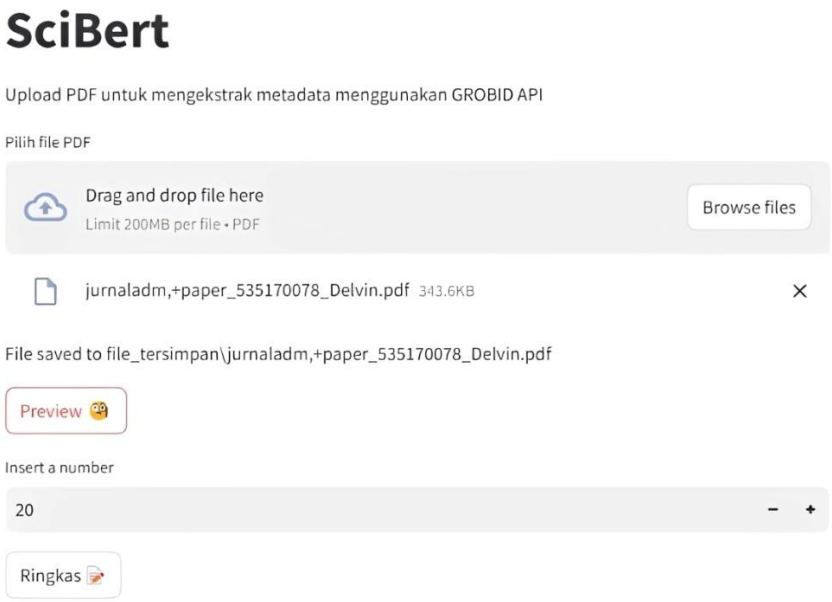
Setelah mengumpulkan data artikel ilmiah yang dibutuhkan sebagai bahan uji sistem, selanjutnya dilakukan implementasi pada sistem peringkasan artikel ilmiah. Berikut ini merupakan hasil tangkapan layar hasil implementasi sistem yang menggunakan teknologi GROBID untuk ekstraksi metadata dan teks, serta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

1. Halaman Utama

Gambar 4. 4 Halaman Utama

Gambar 4. 2 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*" atau *drag and drop*. Terdapat *input* numerik untuk mengatur panjang ringkasan sebelum menekan tombol "Ringkas" untuk memulai proses.

2. Halaman Saat Format PDF Dimasukan



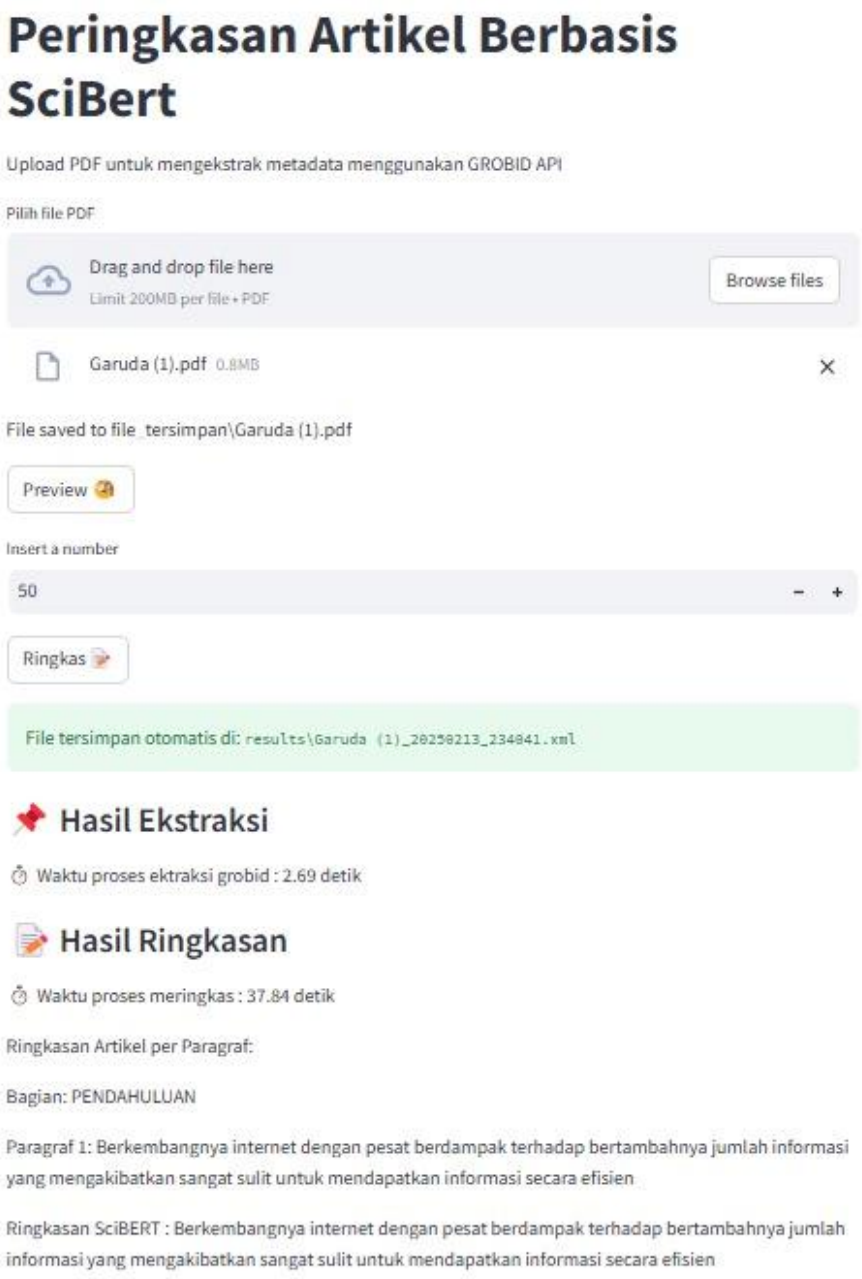
Gambar 4. 5 Tampilan Menggunakan File PDF

Pada gambar 4. 3 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 4. 6 Tampilan ketika klik preview

Pada gambar 4. 4 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.

3. Halaman Ringkasan artikel



Gambar 4. 7 Halaman Ringkasan Artikel

Pada Gambar 4. 5 kita bisa melihat tampilan aplikasi berbasis Streamlit yang digunakan untuk merangkum artikel PDF menggunakan model SciBERT. Setelah pengguna mengklik tombol "Ringkas", pertama-tama sistem akan mengekstrak data dari PDF dan mengubahnya menjadi format JSON. Proses ini akan menunjukkan berapa lama waktu yang dibutuhkan untuk mengonversi PDF ke

format JSON. Setelah itu, aplikasi akan melanjutkan dengan meringkas artikel per paragraf, seperti pendahuluan, metode penelitian, hasil dan pembahasan, serta kesimpulan. Sistem juga menampilkan berapa lama waktu yang diperlukan untuk menghasilkan ringkasan tersebut. Jadi, selain mendapatkan ringkasan yang lebih singkat, pengguna juga bisa melihat durasi waktu yang dibutuhkan untuk setiap langkah dalam proses tersebut.

4.5 Hasil Pengujian Sistem

1. Pengujian kinerja dan Kecepatan Sistem

Pada pengujian ini, kami mengevaluasi waktu yang dibutuhkan oleh sistem dalam mengonversi file PDF menjadi format JSON dan menghasilkan ringkasan dari teks yang ada. Data yang dikumpulkan melibatkan pengolahan berbagai jumlah file PDF, mulai dari 1 file hingga 5 file, dengan ukuran file yang bervariasi.

Tabel 4. 4 Hasil Pengujian Kinerja Sistem

Data Artikel Garuda	Batas Kata	Ukuran File	Waktu Pemprosesan PDF To JSON (Detik)	Waktu hasil Ringkasan (Detik)	Jumlah
PDF Garuda (1)	25	0.8MB	2.01	21.47	23.48
PDF Garuda (1)	50	0.8MB	2.10	22.57	24.67
PDF Garuda (2)	25	1.0MB	2.20	19.41	21.61
PDF Garuda (3)	50	1.0MB	2.13	8.06	10.19
PDF Garuda (3)	25	1.9MB	1.95	8.81	10.76
PDF Garuda (3)	50	1.9MB	1.88	9.54	11.42
PDF Garuda (4)	25	1.0MB	2.30	14.92	17.22
PDF Garuda (4)	50	1.0MB	2.14	15.75	17.89
PDF Garuda (5)	25	272.6KB	2.17	28.75	30.92

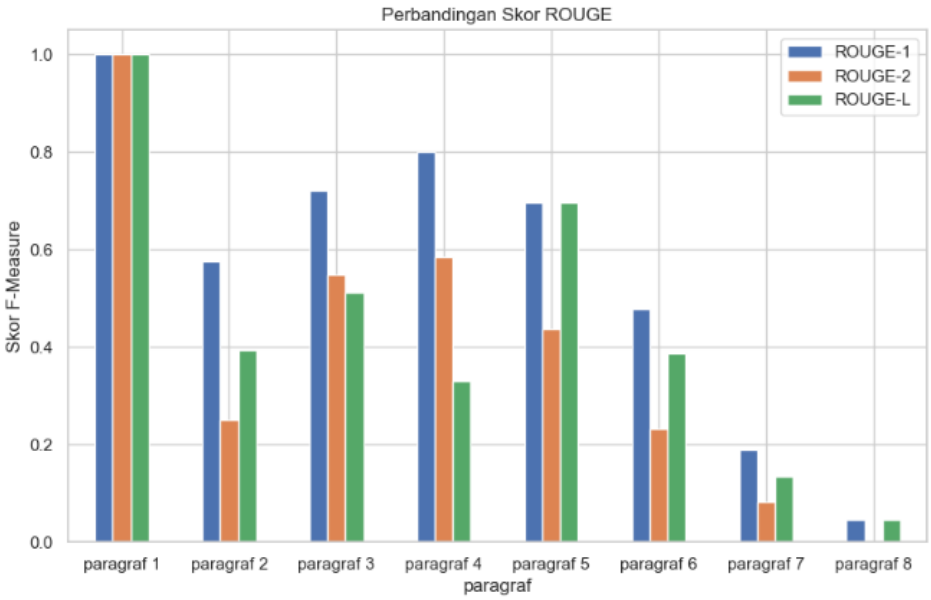
PDF Garuda (5)	50	272.6KB	2.22	29.24	31.46

Sistem ini menunjukkan kinerja yang baik dan efisien dalam memproses file PDF menjadi JSON dan menghasilkan ringkasan, meskipun ada sedikit peningkatan waktu pemrosesan seiring dengan bertambahnya jumlah kata per file. Kecepatan sistem tetap berada dalam batas waktu yang wajar, dengan sistem mampu menangani hingga 50 kata per file dalam waktu yang masih dapat diterima untuk tujuan praktis. Hal ini menunjukkan bahwa sistem dirancang untuk menangani volume data yang moderat secara efisien tanpa menurunkan kualitas hasil ringkasan secara signifikan.

4.6 Analisis Pembahasan

Evaluasi kinerja sistem peringkasan artikel ilmiah yang dikembangkan menggunakan konbinasi teknologi GROBID dan SciBERT. Evaluasi dilakukan dengan mengukur kualitas ringkasan yang dihasilkan oleh sistem dibandingkan dengan ringkasan manual menggunakan matrik ROUGE (*Recall Oriented Understudy for Gisting Evaluation*)

Sistem ini telah diuji menunjukan bahawa sistem mampu menangkap informasi penting dalam dokumen dengan cukup baik. Meskipun ada beberapa perbedaan ringkasan otomatis dan manual, sistem ini memberikan keunggulan dalam efesiensi waktu dan kepraktisan dan memperoleh informasi utama dari artikel ilmiah.



Gambar 4. 8 Perbandingan Skor ROUGE

Gambar 4. 6 adalah hasil evaluasi dengan matrik ROUGE menunjukan variasi akurasi sistem peringkasan dalam menangkap informasi dari dokumen ilmiah. Paragraf 1 memiliki skor tinggi di semua matrik, menandakan ringkasan yang akurat. Namun, pada paragraf 2 hingga 6, skor ROUGE -2 menurun, menunjukan kualitas sistem dalam mempertahankan kesamaan konteks pada bigram. Paragraf 7 dan 8 mencatat skor rendah, mengindikasikan ketidak efektifan sistem dalam merangkum informasi kompleks. Secara keseluruhan, sistem berkerja cukup baik pada bagian awal dokumen tetapi memerlukan perbaikan dengan menangani teks yang lebih rumit.

BAB V
KESIMPULAN DAN SARAN

5.1 Kesimpulan

Kesimpulan dari penelitian ini menunjukkan bahwa sistem peringkasan artikel ilmiah berbasis GROBID dan SciBERT telah berhasil dikembangkan dan terbukti efektif dalam membantu proses tinjauan pustaka. Dengan mengintegrasikan GROBID untuk ekstraksi metadata dan SciBERT untuk peringkasan teks secara ekstraktif, sistem ini mampu mengurangi waktu yang dibutuhkan dalam memahami dokumen ilmiah serta meningkatkan efisiensi pencarian informasi. Evaluasi menggunakan metrik ROUGE menunjukkan bahwa hasil ringkasan memiliki tingkat akurasi yang tinggi dengan *F1-score* mencapai 0.8084 untuk ROUGE-1, 0.6051 untuk ROUGE-2, dan 0.6184 untuk ROUGE-L, yang mengindikasikan kesesuaian ringkasan otomatis dengan ringkasan manual. Sistem ini juga terbukti mampu menangani pemrosesan dokumen dalam format PDF dengan kecepatan yang cukup baik.

5.2 Saran

Saran yang dapat diterapkan untuk mengembangkan sistem ini lebih lanjut di masa mendatang yaitu meningkatkan kemampuan ekstraksi teks dari PDF hasil pemindaian (*scanned* PDF) yang tidak dapat disalin dengan mengintegrasikan teknologi *Optical Character Recognition* (OCR). Hal ini memungkinkan sistem mengenali dan mengonversi teks dari gambar menjadi format yang dapat diproses lebih lanjut. Selain itu, akurasi ringkasan masih perlu ditingkatkan karena model cenderung menghasilkan ringkasan yang terlalu rinci. Oleh karena itu, diperlukan penyesuaian metode peringkasan agar lebih fokus pada poin utama, dengan memilih kalimat yang benar-benar mewakili isi dokumen secara ringkas dan padat. Selain itu, sistem evaluasi perlu dikembangkan lebih lanjut agar hasil evaluasi dapat langsung ditampilkan setelah proses peringkasan selesai, sehingga validasi lebih efisien dan tidak perlu dilakukan secara manual.

DAFTAR PUSTAKA

And, I., & Expert, D. (2021). *Peringkasan Otomatis Makalah Menggunakan Maximum Marginal Relevance INFORMASI ARTIKEL A B S T R A K* (Vol. 3, Nomor 01). http://index.unper.ac.id

Asy'ari, M., Bilad, M. R., & Muhali, M. (2022). Standar Isi Artikel Penelitian: Komponen Detail untuk Dipublikasikan di Jurnal Ilmiah. *Empiricism Journal*, *3*(1), 1–8. https://doi.org/10.36312/ej.v3i1.668

Cai, X., Liu, S., Yang, L., Lu, Y., Zhao, J., Shen, D., & Liu, T. (2022). COVIDSum: A linguistically enriched SciBERT-based summarization model for COVID-19 scientific papers. *Journal of Biomedical Informatics*, *127*. https://doi.org/10.1016/j.jbi.2022.103999

Callegari, E., Vajdecka, P., Xhura, D., & Ingason, A. K. (2023). *Enhancing Academic Title Generation Using SciBERT and Linguistic Rules*. https://huggingface.co/datasets/

Chen, Q., Yao, H., Zhou, D., Li, S., & Dong, L. (2023). Extracting fact-condition relation from geological papers via deep structured semantic model with multi-grained representation. *Computers and Geosciences*, *178*. https://doi.org/10.1016/j.cageo.2023.105416

Farisa, S., & Haviana, C. (2019). *Obtaining Reference\'s Topic Congruity in Indonesian Publications using Machine Learning Approach*. http://garuda.ristekdikti.go.id/

Fatmalasari dan, D., Rosyking Lumbanraja, F., Ilmu Komputer, J., Matematika dan Ilmu Pengetahuan Alam, F., Lampung Jalan Soemantri Brojonegoro No, U., Meneng, G., & Lampung, B. (2022). *Peringkasan Teks Artikel Ilmiah Berbahasa Indonesia dengan Metode Pembobotan Kalimat* (Vol. 3, Nomor 3).

Foppiano, L., de Castro, P. B., Suarez, P. O., Terashima, K., Takano, Y., & Ishii, M. (2022). *Automatic extraction of materials and properties from superconductors scientific literature*. https://doi.org/10.1080/27660400.2022.2153633

Gao, M., Ruan, J., Sun, R., Yin, X., Yang, S., & Wan, X. (2023). *Human-like Summarization Evaluation with ChatGPT*. http://arxiv.org/abs/2304.02554

Goodrich, B., Rao, V., Liu, P. J., & Saleh, M. (2019). Assessing the factual accuracy of generated text. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 166–175. https://doi.org/10.1145/3292500.3330955

Halimah, Surya Agustian, & Siti Ramadhani. (2022). Peringkasan teks otomatis (automated text summarization) pada artikel berbahasa indonesia menggunakan algoritma lexrank. *Jurnal CoSciTech (Computer Science and Information Technology)*, *3*(3), 371–381. https://doi.org/10.37859/coscitech.v3i3.4300

Hendry, M., Sianturi, F., Ridok, A., & Santoso, E. (2023). *Peringkasan Teks Otomatis menggunakan Metode Latent Semantic Analysis pada Artikel Berita Ekonomi berbahasa Indonesia* (Vol. 7, Nomor 5). http://j-ptiik.ub.ac.id

Joshi, B., Symeonidou, A., Mazin Danish, S., & Hermsen Elsevier, F. (2023). *An End-to-End Pipeline for Bibliography Extraction from Scientific Articles*. https://www.pdflib.com/products/tet/

Kilimci, Z. H., & Yalcin, M. (2024). ACP-ESM: A novel framework for classification of anticancer peptides using protein-oriented transformer approach. *Artificial Intelligence in Medicine*, *156*. https://doi.org/10.1016/j.artmed.2024.102951

Li, L., Geissinger, J., Ingram, W. A., & Fox, E. A. (2020). Teaching Natural Language Processing through Big Data Text Summarization with Problem-Based Learning. *Data and Information Management*, *4*(1), 18–43. https://doi.org/10.2478/dim-2020-0003

Maheshwari, H., Singh, B., & Varma, V. (2021). *SciBERT Sentence Representation for Citation Context Classification*. https://scikit-learn.org/stable/

Moradi, M., Dashti, M., & Samwald, M. (2020). Summarization of biomedical articles using domain-specific word embeddings and graph ranking. *Journal of Biomedical Informatics*, *107*. https://doi.org/10.1016/j.jbi.2020.103452

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., & Mian, A. (2023). *A Comprehensive Overview of Large Language Models*. http://arxiv.org/abs/2307.06435

Pearce, K., Zhan, T., Komanduri, A., & Zhan, J. (2021). *A Comparative Study of Transformer-Based Language Models on Extractive Question Answering*. http://arxiv.org/abs/2110.03142

Pisaneschi, L., Gemelli, A., & Marinai, S. (2023). Automatic generation of scientific papers for data augmentation in document layout analysis. *Pattern Recognition Letters*, *167*, 38–44. https://doi.org/10.1016/j.patrec.2023.01.018

Poleksić, A., & Martinčić-Ipšić, S. (2023). *Effects of Pretraining Corpora on Scientific Relation Extraction Using BERT and SciBERT*. http://ceur-ws.org

Utomo, M. S., Wibowo, J. S., & Wahyudi, E. N. (2022). TEXT SUMMARIZATION PADA ARTIKEL BERITA MENGGUNAKAN VECTOR SPACE MODEL DAN COSINE SIMILARITY. *Dinamika Informatika*, *14*(1).

Yuliska, Y., & Syaliman, K. U. (2020a). Literatur Review Terhadap Metode, Aplikasi dan Dataset Peringkasan Dokumen Teks Otomatis untuk Teks Berbahasa Indonesia. *IT Journal Research and Development*, *5*(1), 19–31. https://doi.org/10.25299/itjrd.2020.vol5(1).4688

Yuliska, Y., & Syaliman, K. U. (2020b). Literatur Review Terhadap Metode, Aplikasi dan Dataset Peringkasan Dokumen Teks Otomatis untuk Teks Berbahasa Indonesia. *IT Journal Research and Development*, *5*(1), 19–31. https://doi.org/10.25299/itjrd.2020.vol5(1).4688