Jurnal Informatika: Jurnal pengembangan IT

Vol. 11, No. 1, 2026
ISSN : 2477-5126, DOI:10.30591/jpit.v11i1.10030

Pipeline NLP End-to-End untuk Peringkasan Abstraktif dan Ekstraksi Entitas Berita Berbahasa Indonesia Berbasis Model Transformer

Cuncun Setia[1], Novi Rukhviyanti[2]
[1,2]Department of Informatics, STMIK Indonesia Mandiri, Indonesia
[1]setiacuncun@gmail.com, [2]novi.rukhviyanti@stmik-im.ac.id

Info Artikel

Riwayat Artikel:

Received 2025-12-25
Revised 2026-01-21
Accepted 2026-02-06

Abstract – *The rapid growth of online news content poses challenges for readers to capture the core information quickly and accurately. This research proposes and implements an automated end-to-end pipeline that integrates three main stages: data acquisition, abstractive text summarization, and Named Entity Recognition (NER). The mT5 model is employed to generate coherent and concise summaries, while the BERT model is applied to extract key entities, including persons, organizations, and locations. The pipeline was evaluated using 100 news articles from the Egindo portal. Experimental results show that the system achieves an average text reduction of 62.47%, with a ROUGE-1 F1 score of 0.473. For NER tasks, the pipeline reached a Micro-F1 score close to 0.70, outperforming traditional approaches such as TextRank and CRF. These results demonstrate that the integration of Transformer-based models within a structured pipeline significantly improves summarization quality and entity extraction accuracy. The study contributes a practical NLP solution for the Indonesian language, providing a functional prototype that can be applied to online media analysis and media intelligence applications.*

Keywords: *Abstractive Summarization; Automated Pipeline; Entity Extraction; Indonesian NLP; Transformer Models.*

Corresponding Author:

Cuncun Setia

Email: setiacuncun@gmail.com

Abstrak – *Pertumbuhan informasi pada portal berita daring menimbulkan tantangan bagi pembaca untuk memahami inti konten secara cepat dan akurat. Penelitian ini mengusulkan dan mengimplementasikan alur kerja otomatis end-to-end yang menggabungkan tiga tahap utama: akuisisi data, peringkasan teks abstraktif, dan ekstraksi entitas bernama (Named Entity Recognition /NER). Model mT5 digunakan untuk menghasilkan ringkasan yang koheren dan ringkas, sedangkan model BERT diterapkan untuk mengidentifikasi entitas penting seperti tokoh, organisasi, dan lokasi. Alur kerja diuji menggunakan 100 artikel berita dari portal Egindo. Hasil eksperimen menunjukkan bahwa sistem mampu mereduksi panjang teks rata-rata sebesar 62,47% tanpa kehilangan esensi informasi, dengan skor ROUGE-1 F1 mencapai 0,473. Pada tugas NER, sistem menghasilkan Micro-F1 mendekati 0,70, lebih tinggi dibandingkan metode tradisional seperti TextRank dan CRF. Temuan ini membuktikan bahwa integrasi model Transformer dalam alur kerja terstruktur mampu meningkatkan kualitas peringkasan dan akurasi NER secara signifikan. Penelitian ini memberikan kontribusi praktis berupa rancangan sistem NLP terintegrasi untuk bahasa Indonesia yang dapat digunakan pada analisis media daring dan intelijen informasi.*

Kata Kunci: *Ekstraksi Entitas, NLP Indonesia, Pipeline Otomatis, Peringkasan Abstraktif, Transformer*

I. PENDAHULUAN

Perkembangan pesat teknologi informasi telah mendorong peningkatan signifikan jumlah konten pada portal berita daring[1], yang memicu fenomena kelebihan informasi (*information overload*). Kondisi ini menyulitkan pembaca untuk menangkap gagasan utama berita secara cepat dan akurat, terutama ketika harus menyaring informasi dari berbagai topik. Dalam konteks ini, *Natural Language Processing* (NLP)[2], [3], [4] menawarkan solusi melalui peringkasan teks abstraktif yang menghasilkan ringkasan koheren serta *Named Entity Recognition* (NER)[5], [6] yang menyoroti entitas penting guna memperjelas konteks suatu peristiwa.

Penelitian sebelumnya umumnya masih berfokus pada satu tugas secara terpisah, seperti peringkasan ekstraktif tradisional (misalnya TextRank)[7] atau NER berbasis Conditional Random Fields (CRF)[8], yang memiliki keterbatasan dalam hal koherensi ringkasan dan akurasi ekstraksi entitas. Meskipun kemajuan arsitektur Transformer, seperti mT5 untuk peringkasan dan BERT untuk NER, telah terbukti meningkatkan kinerja masing-masing tugas[9], sebagian besar studi belum mengintegrasikan kedua proses tersebut dalam satu sistem terpadu. Akibatnya, masih terdapat kesenjangan penelitian terkait terbatasnya pipeline NLP end-to-end yang menggabungkan akuisisi berita, peringkasan abstraktif, dan ekstraksi entitas bernama untuk bahasa Indonesia, padahal aplikasi nyata seperti pemantauan media membutuhkan solusi yang otomatis dan konsisten.

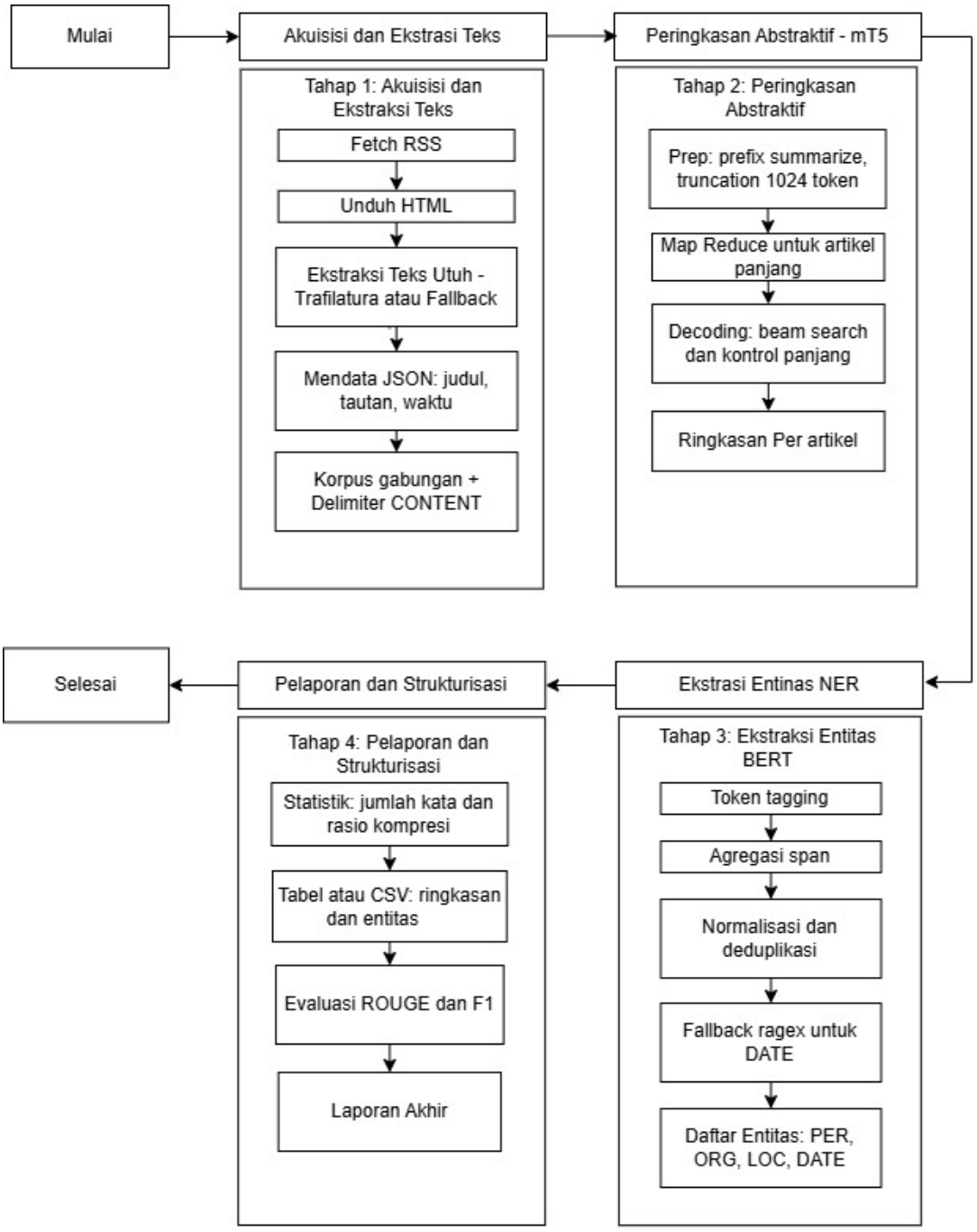
Berdasarkan kesenjangan tersebut, penelitian ini dimotivasi untuk merancang pipeline NLP end-to-end yang mengintegrasikan pengumpulan berita daring, peringkasan abstraktif, dan ekstraksi entitas dalam satu alur kerja otomatis. Penelitian ini bertujuan mengevaluasi fungsionalitas pipeline melalui studi kasus pada portal berita

Egindo.com dengan pendekatan evaluasi kuantitatif. Kontribusi utama penelitian ini meliputi integrasi model Transformer secara *end-to-end* untuk peringkasan dan NER berbahasa Indonesia, penyediaan dataset beranotasi yang memperkaya sumber daya NLP nasional, serta perancangan pipeline modular yang dapat digunakan kembali untuk penelitian lanjutan maupun aplikasi industri.

II. METODE

Bab ini menguraikan rancangan sistem, prosedur akuisisi dan pemrosesan data, konfigurasi model, format keluaran, orkestrasi eksekusi, lingkungan perangkat lunak, serta prosedur evaluasi. Seluruh langkah dirancang replikabel dan terlacak dari data mentah hingga laporan, dengan artefak pendukung disediakan sebagai bahan pelengkap.

A. Rancangan Sistem dan Alur Kerja



Gambar 1. Arsitektur alur kerja peringkasan abstraktif dan ekstraksi entitas

Alur kerja (Gambar 1) dirancang secara modular dan dapat direproduksi, yang terdiri atas empat tahapan utama yang dieksekusi secara *end-to-end*, yaitu: (1) akuisisi data dan ekstraksi teks penuh, (2) peringkasan abstraktif, (3) pengenalan entitas bernama (*Named Entity Recognition* /NER), dan (4) pelaporan serta penstrukturan hasil[10], [11]. Setiap tahapan menyediakan antarmuka input–output yang terstandarisasi dalam format JSON atau CSV[12], sehingga setiap komponen dapat diganti atau ditingkatkan tanpa memerlukan perubahan pada keseluruhan alur kerja. Orkestrasi proses diimplementasikan menggunakan skrip shell yang bersifat idempoten untuk memastikan urutan eksekusi yang konsisten, termasuk pembersihan artefak sebelumnya, pengambilan data RSS, proses peringkasan, pemrosesan NER, hingga pembangkitan keluaran akhir.

Alur data diawali dengan pengambilan entri RSS beserta konten HTML yang terkait, yang selanjutnya diekstraksi menjadi artikel teks penuh dan metadata, meliputi judul, URL, serta waktu publikasi. Hasil pemrosesan tersebut disimpan dalam satu berkas korpus terintegrasi dengan menggunakan delimiter khusus serta blok metadata JSON tersemat guna memfasilitasi pemrosesan secara batch. Untuk artikel dengan panjang teks yang besar, modul peringkasan menerapkan strategi map–reduce dengan mengintegrasikan beam search decoding dan pengendalian panjang keluaran untuk menjaga konsistensi ringkasan. Modul NER berbasis BERT diperkuat dengan strategi agregasi span serta mekanisme *fallback* berbasis regular expression (regex) untuk ekstraksi entitas temporal. Seluruh keluaran alur kerja disimpan dalam format terstruktur sehingga mudah diakses untuk keperluan evaluasi dan pelaporan.

Pipeline ini juga mendukung deteksi perangkat secara otomatis. Apabila GPU tersedia, sistem mengaktifkan eksekusi presisi rendah (low-precision execution) untuk meningkatkan efisiensi memori dan mengurangi waktu pemrosesan; sedangkan apabila hanya CPU yang tersedia, eksekusi dilakukan menggunakan presisi standar. Desain ini memastikan bahwa pipeline dapat dijalankan secara konsisten pada berbagai lingkungan komputasi serta tetap mudah dikembangkan untuk mendukung sumber berita tambahan.

B. Akuisisi Data dan Ekstraksi Teks Penuh

Sumber data diperoleh dari sebuah portal berita daring melalui umpan RSS (studi kasus: Egindo). Untuk setiap entri RSS, sistem mengunduh halaman berita yang bersesuaian dan mengekstraksi teks penuh dari konten HTML menggunakan adaptor ekstraksi konten yang memprioritaskan pustaka trafilatura, dengan mekanisme cadangan berbasis HTML parsing yang diterapkan apabila diperlukan. Hasil ekstraksi kemudian diserialisasikan ke dalam satu berkas korpus terintegrasi menggunakan delimiter unik antarartikel serta blok metadata berformat JSON yang tersemat, mencakup judul, URL, waktu publikasi, waktu pengambilan data, dan panjang karakter, yang diikuti oleh penanda -*CONTENT*- sebelum isi artikel, sebagaimana ditunjukkan pada Gambar 1. Proses deduplikasi antariterasi dilakukan menggunakan pengenal heksadesimal yang diturunkan dari URL artikel atau GUID yang disediakan pada umpan RSS. Keluaran dari tahap ini berupa korpus gabungan yang memuat metadata lengkap beserta teks artikel asli. Struktur korpus dan contoh entri disajikan pada bagian lampiran.

Pemilihan Egindo.com sebagai sumber data didasarkan pada tiga pertimbangan utama. Pertama, penulis memperoleh izin resmi dari dewan redaksi Egindo untuk memanfaatkan konten berita dalam konteks penelitian, sehingga penggunaan data memenuhi standar etika akademik dan ketentuan hukum yang berlaku. Kedua, Egindo memiliki cakupan kategori berita yang relatif lebih terbatas dan terfokus dibandingkan portal berita berskala besar, sehingga memudahkan pengelompokan artikel serta menjaga konsistensi pada tahap peringkasan dan ekstraksi entitas. Ketiga, Egindo menyediakan umpan RSS yang stabil dan dapat diakses secara publik, sehingga mendukung pemrosesan data secara otomatis dan menjadikannya sesuai untuk penelitian berbasis pipeline NLP end-to-end.

C. Peringkasan Abstraktif

Proses peringkasan dalam penelitian ini dilakukan menggunakan model multibahasa mT5 XLSum[13] dalam kerangka *sequence-to-sequence*. Setiap masukan dipersiapkan dengan menambahkan awalan perintah (*instruction prompt*) *"summarize:"* sebelum dilakukan proses tokenisasi. Panjang masukan dibatasi hingga 1.024 token melalui mekanisme pemotongan (*truncation*) untuk menjaga konsistensi serta mencegah kelebihan konteks (*context overflow*). Pada tahap dekoding, digunakan teknik *beam search*[14], [15], [16] dengan empat *beam* guna menghasilkan keluaran ringkasan yang lebih koheren.

Panjang ringkasan dibatasi dengan jumlah minimum sekitar 40 token dan maksimum 150 token, sehingga ringkasan yang dihasilkan tetap informatif namun ringkas. Untuk artikel yang melebihi batas jendela konteks model, diterapkan strategi *map–reduce* dengan cara membagi teks sumber menjadi beberapa segmen yang lebih kecil, merangkum setiap segmen secara terpisah, mengagregasikan ringkasan antara, dan kemudian merangkum kembali hasil agregasi tersebut untuk memperoleh ringkasan akhir yang lebih padat dan komprehensif.

Selain itu, sistem ini dilengkapi dengan mekanisme *retry* yang diaktifkan ketika ringkasan awal yang dihasilkan terlalu pendek. Mekanisme ini secara dinamis menyesuaikan panjang target ringkasan berdasarkan proporsi panjang teks sumber, sehingga ringkasan yang dihasilkan dapat merepresentasikan informasi inti artikel secara memadai. Evaluasi kuantitatif terhadap hasil peringkasan, termasuk rasio kompresi dan skor ROUGE, disajikan pada Bab 3 (Tabel 2 dan Tabel 3).

D. Ekstraksi Enti tas Bernama (NER)

Proses pengenalan entitas bernama (*Named Entity Recognition* /NER) diterapkan pada teks sumber asli, bukan pada ringkasan yang dihasilkan, dengan menggunakan model multibahasa berbasis BERT[17] yang dikonfigurasikan dengan strategi agregasi *span*. Tahapan *post-processing* meliputi perluasan batas kata untuk mengatasi fragmentasi yang disebabkan oleh tokenisasi *subword*, normalisasi kapitalisasi pada kata benda khusus, serta deduplikasi untuk mempertahankan bentuk entitas yang paling representatif.

Entitas yang diekstraksi selanjutnya dipetakan ke dalam empat kategori utama, yaitu *PERSON* (PER), *ORGANIZATION* (ORG), *LOCATION* (LOC), dan *DATE* (TIME), dengan penerapan ambang kepercayaan (*confidence threshold*) sebesar 0,60 guna menyaring prediksi dengan tingkat keyakinan rendah. Untuk entitas temporal, sistem mengintegrasikan mekanisme *fallback* berbasis *regular expression* (regex) untuk meningkatkan nilai recall, yang mencakup format tanggal numerik (dd/mm/yyyy), tanggal dengan nama bulan, serta ekspresi temporal lengkap termasuk zona waktu WIB, WITA, dan WIT.

Kombinasi antara model NER berbasis BERT, strategi agregasi *span*, dan *pipeline post-processing* yang diusulkan menghasilkan peningkatan akurasi dalam mengidentifikasi entitas-entitas penting pada teks berita. Melalui pendekatan ini, sistem tidak hanya mampu mengenali entitas dengan tingkat presisi yang lebih tinggi, tetapi juga dapat beradaptasi secara lebih efektif terhadap keragaman format teks yang umum ditemukan pada artikel berita daring.

E. Format Keluaran

Sistem yang dikembangkan menghasilkan dua jenis keluaran utama. Pertama, laporan ringkas per artikel yang disajikan dalam format yang mudah dibaca oleh manusia. Laporan ini mencakup informasi *header* (judul, sumber, dan tanggal publikasi), bagian "*ABSTRACTIVE SUMMARY*", daftar entitas yang diekstraksi (*PERSON*, *ORGANIZATION*, *LOCATION*, dan *DATE*), serta ringkasan statistik yang meliputi jumlah kata asli, jumlah kata hasil ringkasan, dan rasio kompresi. Struktur ini memungkinkan pembaca untuk dengan cepat memahami inti konten setiap artikel berita beserta informasi kontekstual utamanya secara ringkas dan terorganisasi dengan baik.

Kedua, korpus terintegrasi dalam format terstruktur yang terdiri atas blok metadata JSON, *delimiter* ---*CONTENT*---, serta teks artikel lengkap. Format keluaran ini dirancang agar mudah diparsing untuk keperluan analisis lanjutan secara otomatis, integrasi ke dalam representasi tabular atau berkas CSV, serta pemrosesan data lebih lanjut oleh sistem lain[18]. Dengan demikian, keluaran yang dihasilkan tidak hanya mendukung keterbacaan oleh manusia, tetapi juga memenuhi kebutuhan pemrosesan berorientasi mesin.

Eksekusi *pipeline* dikelola melalui skrip orkestrasi berbasis shell yang menjalankan siklus berikut: (i) pembersihan artefak sebelumnya, (ii) pengambilan berita melalui RSS feed (dengan penggunaan delimiter unik dan mekanisme *safe appending*), (iii) proses peringkasan (serta NER apabila diaktifkan), dan (iv) penulisan keluaran akhir. Orkestrasi ini memungkinkan eksekusi yang berurutan dan dapat direproduksi tanpa memerlukan pemanggilan manual terhadap masing-masing skrip[19].

F. Lingkungan Python

Implementasi sistem dikembangkan menggunakan Python 3 dengan memanfaatkan beberapa pustaka utama. Pustaka Transformers[20] dan PyTorch digunakan untuk memuat serta menjalankan model mT5 pada tugas peringkasan dan model BERT pada klasifikasi *Named Entity Recognition* (NER). Akuisisi data berita melalui RSS feed difasilitasi oleh pustaka feedparser dan requests untuk permintaan HTTP, sedangkan ekstraksi konten utama artikel dilakukan menggunakan trafilatura[21], dengan HTML parser digunakan sebagai mekanisme cadangan (*fallback*) apabila diperlukan.

Deteksi perangkat dilakukan secara otomatis. Apabila GPU tersedia, sistem mengaktifkan eksekusi berbasis GPU dengan presisi 16-bit (float16) untuk mengurangi konsumsi memori dan mempercepat waktu pemrosesan. Jika GPU tidak terdeteksi, eksekusi secara otomatis dialihkan ke CPU dengan menggunakan presisi bilangan pecahan 32-bit (float32)[22]. Dengan konfigurasi ini, sistem dapat beroperasi secara fleksibel pada berbagai lingkungan komputasi, baik pada pengembangan lokal maupun pada server yang dilengkapi dengan akselerasi GPU.

G. Prosedur Evaluasi

Kinerja komponen peringkasan dievaluasi menggunakan dua pendekatan utama[23]. Pertama, performa sistem diukur menggunakan metrik ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*)[24], [25], khususnya ROUGE-1 (unigram), ROUGE-2 (bigram), dan ROUGE-L (*longest common subsequence*), dengan skor F1 digunakan sebagai kriteria evaluasi utama. Nilai ROUGE dihitung dengan membandingkan ringkasan yang dihasilkan sistem dengan ringkasan rujukan yang disusun oleh manusia (*gold standard*).

Untuk menjamin kualitas dan objektivitas, korpus rujukan disusun secara cermat. Sebanyak 100 artikel berita diringkas oleh 20 anotator yang berasal dari latar belakang beragam, termasuk jurnalis, mahasiswa, dan guru sekolah, guna meminimalkan bias individu. Ringkasan yang disusun secara manual tersebut selanjutnya melalui proses verifikasi untuk menghilangkan kesalahan tipografis, sehingga dihasilkan dataset rujukan yang bersih dan andal.

Kedua, sistem menghitung persentase ringkasan, yang didefinisikan sebagai rasio antara jumlah kata pada ringkasan dan jumlah kata pada teks asli. Hasil kuantitatif lengkap dari kedua metode evaluasi tersebut, termasuk statistik persentase ringkasan dan skor ROUGE, disajikan secara rinci pada Bab 3.

H. Tingkat Reduksi Ringkasan

Tingkat reduksi digunakan untuk mengukur efisiensi pemampatan teks, yaitu sejauh mana panjang teks asli dapat dipersingkat tanpa menghilangkan informasi esensial. Tingkat reduksi dihitung menggunakan rumus berikut:

$$Tingkat\ Reduksi\ (\%) = \left(1 = \frac{S}{D}\right)\ x\ 100 \qquad (1)$$

D menyatakan jumlah kata pada dokumen asli dan S menyatakan jumlah kata pada ringkasan hasil sistem

I. Evaluasi ROUGE

Kualitas ringkasan abstraktif dievaluasi menggunakan metrik ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*), yang mengukur tingkat kemiripan antara ringkasan sistem dan ringkasan rujukan. ROUGE-N dihitung berdasarkan tumpang tindih n-gram, sebagaimana ditunjukkan pada rumus berikut.

$$ROUGE - N = \frac{\sum_{g \in Ref} min(Count_{sys}(g), Count_{ref}(g))}{\sum_{g \in Ref} Count_{Ref}(g)}\ x\ 1 \qquad (2)$$

ROUGE mengukur kesamaan berdasarkan *Longest Common Subsequence* (LCS) antara ringkasan sistem dan ringkasan rujukan, yang merefleksikan kesesuaian struktur kalimat.

J. Kontribusi Dataset

Sebagai bagian dari penelitian ini, penulis membangun sebuah dataset berita berbahasa Indonesia yang mencakup teks asli, ringkasan rujukan yang disusun oleh manusia, serta anotasi entitas bernama. Dataset ini disediakan secara terbuka melalui platform Zenodo untuk mendukung pemanfaatannya sebagai benchmark dalam penelitian NLP di Indonesia.

Kontribusi dataset ini memberikan dua manfaat utama. Pertama, dataset ini menyediakan sumber data terstandar yang dapat digunakan untuk membandingkan berbagai pendekatan peringkasan dan ekstraksi entitas secara objektif. Kedua, dataset ini memperkaya ekosistem sumber daya NLP berbahasa Indonesia yang hingga saat ini masih relatif terbatas. Dengan tersedianya dataset ini, penelitian selanjutnya diharapkan dapat mengevaluasi pendekatan-pendekatan baru secara lebih konsisten serta mendorong pengembangan model yang lebih adaptif terhadap karakteristik bahasa Indonesia.

III. HASIL DAN PEMBAHASAN

Bab ini menyajikan hasil pelaksanaan alur kerja yang diusulkan beserta analisis kinerja sistem. Evaluasi dilakukan menggunakan dua pendekatan yang saling melengkapi, yaitu analisis kualitatif yang menilai koherensi ringkasan yang dihasilkan serta ketepatan identifikasi entitas pada contoh kasus terpilih, dan analisis kuantitatif yang mengukur kinerja sistem melalui indikator rasio kompresi dan metrik ROUGE.

Selain itu, pembahasan mencakup interpretasi terhadap hasil yang diperoleh, identifikasi keterbatasan sistem, serta eksplorasi potensi pengembangan di masa depan guna meningkatkan keandalan dan penerapan sistem yang diusulkan dalam skenario dunia nyata.

A. Gambaran Umum Hasil

Sistem yang diimplementasikan berhasil memproses 100 artikel berita yang dikumpulkan dari portal berita daring Egindo. Setiap artikel melewati seluruh tahapan alur kerja, mulai dari akuisisi data, peringkasan abstraktif menggunakan model mT5, hingga ekstraksi entitas bernama menggunakan model BERT multibahasa, dan diakhiri dengan penyajian hasil dalam format terstruktur.

Keluaran akhir disajikan dalam dua format utama. Pertama, laporan tekstual yang memuat metadata artikel (judul, sumber, dan tanggal publikasi), ringkasan abstraktif, daftar entitas yang dikelompokkan (*PERSON*, *ORGANIZATION*, *LOCATION*, dan *DATE*), serta statistik pada tingkat kata yang mencakup jumlah kata teks asli, jumlah kata hasil ringkasan, dan rasio kompresi. Format ini dirancang agar mudah dibaca oleh manusia serta memberikan gambaran yang komprehensif mengenai setiap artikel.

Kedua, hasil juga disusun dalam bentuk data tabular terstruktur, di mana setiap baris merepresentasikan satu artikel berita dan setiap kolom merepresentasikan atribut data tertentu (misalnya judul, ringkasan, entitas, dan

metadata terkait). Format ini memungkinkan analisis data secara agregat serta memfasilitasi integrasi dengan berbagai alat analitik maupun platform visualisasi data.

B. Analisis Studi Kasus

Untuk menilai kualitas hasil secara lebih mendalam, dilakukan analisis studi kasus. Sebagai contoh ilustratif, dipilih salah satu artikel berita yang telah diproses oleh sistem. Artikel yang dipilih berjudul "Ayuso Defeats Romo to Win a Stage of the Vuelta a España".

TABEL 1
CONTOH HASIL PERINGKASAN.

Komponen	Hasil yang Diperoleh dari Sistem
Cuplikan Teks Asli	Turin | EGINDO.com – Pembalap Spanyol Juan Ayuso (UEA Team Emirates-XRG) mengungguli rekan senegaranya Javier Romo (Movistar) untuk memenangkan etape ke-12 Vuelta a Espana pada hari Kamis, kemenangan etape keduanya di balapan tahun ini, sementara pembalap Denmark Jonas Vingegaard mempertahankan posisi puncak klasemen. Ayuso, yang menang etape ketujuh sendirian, ditemani Romo di 25 kilometer terakhir dari perjalanan sejauh 144,9 km dari Laredo ke Los Corrales de Buelna dan menyalip Romo tepat waktu di tikungan terakhir. "Setelah hari yang berat, entahlah, dan saya harus bermain dengan strategi," kata Ayuso, menambahkan bahwa mobil timnya telah menginstruksikannya untuk berkendara secara taktis dan membuat Romo yakin bahwa ia perlu "menarik lebih jauh" jika ingin berpeluang memenangkan etape tersebut. "Saya harus membuatnya sedikit gugup," katanya. Pembalap Prancis Brieuc Rolland (Groupama-FDJ) tertinggal 13 detik di posisi ketiga setelah sempat mengancam di kilometer terakhir, sementara Vingegaard (Visma-Lease a Bike) finis dengan aman di peleton yang menyusul lebih dari enam menit kemudian. Tidak ada perubahan di puncak klasemen umum, dengan Vingegaard mempertahankan keunggulan 50 detiknya atas Joao Almeida (UEA Team Emirates-XRG) dari Portugal, sementara pembalap Inggris Tom Pidcock (Q36.5 Pro Cycling Team) tertinggal enam detik di posisi ketiga. Ayuso menjalani Vuelta yang penuh peristiwa sejauh ini. Selain dua kemenangan etapenya, yang memulai sebagai salah satu pembalap yang diperkirakan akan menantang Vingegaard, pembalap berusia 22 tahun ini juga kehilangan sebagian waktunya setelah tertinggal di tanjakan. Keluarnya pebalap Spanyol itu dari Tim Emirates-XRG UEA diumumkan pada hari istirahat hari Senin, yang membuat Ayuso geram dan mengkritik keras waktu pengumuman tersebut keesokan harinya. Di pertengahan jalan, kelompok breakaway yang terdiri dari lebih dari 50 pebalap unggul lebih dari tiga menit dari peloton, dan ketika kelompok terdepan semakin tersisih, Ayuso mencoba untuk lolos sendiri di tanjakan terakhir. Romo mengikutinya, dan dengan Rolland masih mengejar, pasangan terdepan memulai permainan kejar-kejaran di kilometer terakhir, dan Ayuso-lah yang bangkit dari ketertinggalan di tikungan terakhir sebelum garis finis untuk mengungguli Romo. Pebalap runner-up itu sampai harus memukul setangnya karena frustrasi, hampir saja meraih kemenangan etape Grand Tour pertamanya. "Ini bukan sesuatu yang benar-benar saya nikmati, tidak bekerja sama sepenuhnya, tetapi terkadang kita harus bermain cerdas," kata Ayuso. "Dan itulah yang saya lakukan di final dan sprint." Setelah drama sehari sebelumnya di Basque Country, di mana para demonstran pro-Palestina mengakhiri etape lebih awal tanpa pemenang, etape di Cantabria berlangsung lebih santai. Para pemimpin klasemen umum tampak menahan diri untuk etape ke-13 hari Jumat yang akan menjadi salah satu etape terberat dalam balapan tahun ini. Etape pegunungan sepanjang 202,7 km menanti, dengan semua tanjakan yang akan datang di bagian akhir di mana para pembalap akan menaklukkan tiga tanjakan kategori satu, termasuk tanjakan brutal hingga finis di Angliru, tempat Vuelta dapat dimenangkan atau dikalahkan. Sumber : CNA/SL
Ringkasan Abstraktif	Pembalap Spanyol Juan Ayuso mengungguli rekan senegaranya Javier Romo untuk memenangkan etapa ke-12 Vuelta a Espana pada hari Kamis, kemenangan fase keduanya di balapan tahun ini, sementara Denmark Jonas Vingegaard mempertahankan posisi puncak klasemen.
Entitas (NER)	**ORANG :** Ayuso, Vingegaard, Brieuc Rolland, Javier Romo, Joao Almeida, Jonas Vingegaard, Juan Ayuso, Romo, Tom Pidcock **ORGANISASI :** XRG, UEA Team Emirates, FDJ, Groupama, Movistar, Q36. 5 Pro Cycling Team, Tim Emirates, UEA, Visma - Lease a Bike **LOKASI :** Romo, Denmark, Laredo, Los Corrales de Buelna, Portugal, Spanyol, Turin
Statistik	**Jumlah Kata Asli** : 448 **Jumlah Kata Rangkuman** : 34 **Persentase Ringkasan** : 7.59%

Analisis komprehensif yang disajikan pada Tabel 1 menggambarkan hasil peringkasan abstraktif dan ekstraksi entitas bernama. Ringkasan yang dihasilkan bersifat koheren, dengan berhasil mereduksi teks asli dari 448 kata menjadi 34 kata (7,59%) tanpa menghilangkan informasi esensial. Terkait *Named Entity Recognition* (NER), model berhasil mengidentifikasi sebagian besar entitas dengan benar pada kategori *PERSON*, *ORGANIZATION*, dan

LOCATION. Namun, ditemukan satu kesalahan klasifikasi, yaitu entitas "Romo" yang teridentifikasi sebagai *LOCATION*, padahal dalam konteks kalimat entitas tersebut merujuk pada seorang pesepeda bernama Javier Romo. Kesalahan klasifikasi ini disebabkan oleh ambiguitas leksikal, karena Romo juga merupakan nama lokasi geografis yang nyata (Pulau Romo di Denmark). Dalam data pelatihan model, istilah Romo kemungkinan lebih sering muncul dalam konteks lokasi, sehingga model mengembangkan bias statistik[26]. Akibatnya, ketika menghadapi token yang ambigu tersebut, model multibahasa lebih mengandalkan distribusi probabilitas yang telah dipelajari[27], [28], [29] dibandingkan dengan konteks tingkat kalimat (misalnya frasa "Ayuso defeated Romo"), yang seharusnya mengindikasikan bahwa Romo merujuk pada individu, namun justru terdeteksi sebagai lokasi.

Selain kasus Romo, selama eksperimen juga ditemukan beberapa kesalahan NER lain dengan pola serupa. Sebagai contoh, nama "Xi" dalam rujukan kepada Presiden Tiongkok Xi Jinping sebagian terdeteksi sebagai lokasi karena token Xi sering muncul dalam konteks geografis. Demikian pula, petenis Aryna Sabalenka diklasifikasikan secara keliru sebagai lokasi meskipun konteks secara jelas menunjukkan bahwa entitas tersebut adalah seorang individu. Kasus serupa juga terjadi pada Naomi Osaka, yang salah dikategorikan sebagai organisasi, padahal merujuk pada individu, karena istilah Osaka lebih umum diasosiasikan dengan nama kota di Jepang.

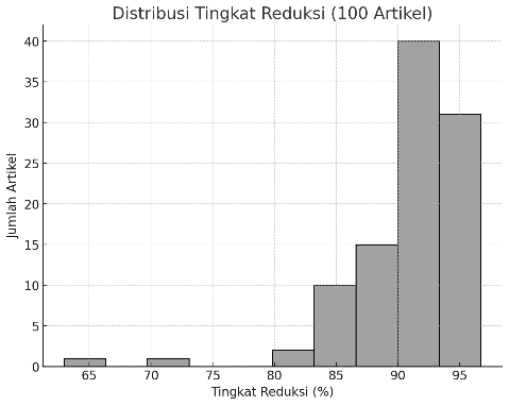
Pola umum yang mendasari kesalahan-kesalahan tersebut adalah ambiguitas leksikal, di mana satu token dapat merujuk pada individu, lokasi, atau tipe entitas lainnya. Model NER cenderung sangat bergantung pada frekuensi kemunculan token dalam data pelatihan, sehingga rentan terhadap kesalahan klasifikasi ketika suatu istilah lebih dominan pada kategori entitas yang berbeda. Untuk mengatasi keterbatasan ini, diperlukan integrasi modul *entity linking* dan *entity normalization* agar konteks tingkat kalimat dapat dimanfaatkan secara lebih optimal serta kesalahan klasifikasi dapat diminimalkan.

C. Analisis Kuantitatif Hasil

Analisis kuantitatif difokuskan pada statistik peringkasan yang dihasilkan oleh sistem, khususnya tingkat reduksi. Ukuran ini merepresentasikan derajat pengurangan jumlah kata dari teks asli ke ringkasan yang dihasilkan.

TABEL 2
STATISTIK TINGKAT REDUKSI RINGKASAN ABSTRAKTIF PADA 10 BERITA (CONTOH REPRESENTATIF)

Judul Berita	Kata Asli	Kata Rangkuman	Tingkat Reduksi (%)
fed siap memangkas suku bunga seiring pasar tenaga kerja melemah	1070	230	78,50%
bank sentral australia kaji dampak ai terhadap ekonomi	760	270	64,47%
ekoteologi: jalan spiritualitas baru menjaga bumi	490	420	14,29%
the fed umumkan konferensi pada bulan oktober tentang inovasi pembayaran	540	210	61,11%
sinner kalahkan auger - aliassime untuk melaju ke final us open lawan alcaraz	1230	420	65,85%
kejati sumut usut korupsi lahan ptpn i, ada tiga lokasi citraland terlibat	490	290	40,82%
ekspor karet alam sumut pada juli 2025 terkoreksi, mengalami penurunan	890	260	70,79%
sk on korsel teken perjanjian pasokan baterai ess dengan flatiron energy di as	700	320	54,29%
puluhan negara di dunia warganya tidak dipungut pajak penghasilan, ada negara tetangga indonesia	520	240	53,85%
refleksi antara aspirasi dan anarkis: hak berekspresi dalam demokrasi	710	280	60,56%

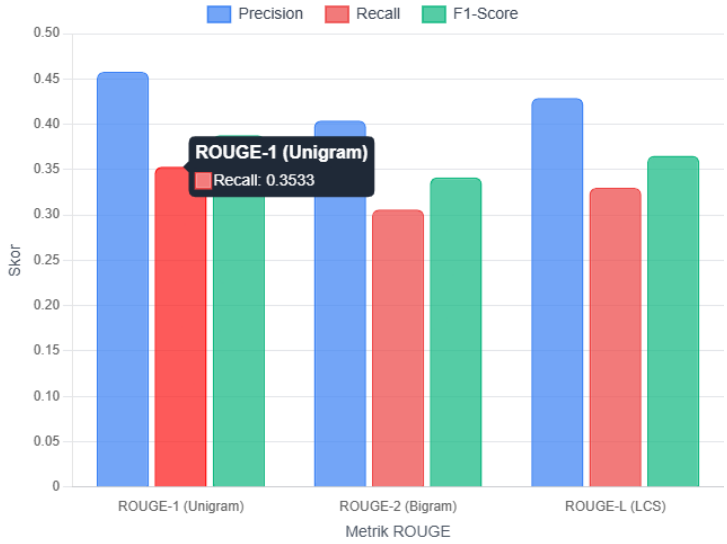


Gambar 2. Distribusi tingkat reduksi ringkasan abstraktif pada 100 berita.

Berdasarkan Tabel 2, untuk sampel 10 artikel berita, panjang ringkasan rata-rata adalah 39,73% dari teks asli (dengan tingkat reduksi sebesar 60,27%). Untuk keseluruhan 100 artikel, panjang ringkasan rata-rata adalah 37,53% (dengan tingkat reduksi sebesar 62,47%). Hasil ini menunjukkan bahwa sistem mampu menghasilkan ringkasan yang ringkas dan konsisten, sekaligus tetap mempertahankan sebagian besar informasi esensial dari artikel berita asli. Untuk memvalidasi ringkasan yang dihasilkan oleh sistem, dilakukan perbandingan terhadap standar rujukan (*gold standard*) yang terdiri atas ringkasan yang disusun secara manual. Ringkasan rujukan tersebut disiapkan oleh peringkas yang memiliki keahlian dan pengalaman dalam menganalisis serta merangkum inti informasi suatu teks. Kompetensi ini memastikan bahwa setiap ringkasan manual disusun dengan pemahaman konteks yang mendalam, kemampuan mengidentifikasi informasi kunci, serta keterampilan menyajikannya secara ringkas dan efektif. Keterlibatan para peringkas ahli ini bertujuan untuk menjamin kualitas dan objektivitas ringkasan manual sebagai tolok ukur yang valid.

```
============================================================
HASIL VALIDASI STATISTIK PERBANDINGAN RANGKUMAN
============================================================
                          Rangkuman AI Rangkuman Manusia
Metrik
Rata-rata Jumlah Kata          29.1800          109.4400
Std Dev Jumlah Kata             5.7303           35.2670
Rata-rata Rasio Kompresi (%)    9.0944           32.8858
Std Dev Rasio Kompresi (%)      4.9011           14.9703
Rata-rata ROUGE-1 (F1)          0.3879                 -
Rata-rata ROUGE-2 (F1)          0.3413                 -
Rata-rata ROUGE-L (F1)          0.3653                 -
============================================================
```

Gambar 3. Hasil evaluasi 100 berita menggunakan ROUGE dibandingkan dengan ringkasan manual



Gambar 4. Grafik rata-rata nilai ROUGE 100 berita

Untuk menilai kualitas ringkasan yang dihasilkan, digunakan evaluasi ROUGE, yang mengkaji tiga aspek, yaitu tumpang tindih unigram (ROUGE-1), tumpang tindih bigram (ROUGE-2), serta *longest common subsequence* dalam struktur kalimat (ROUGE-L) seperti yang ditampilkan di gambar 3. Salah satu temuan awal dari hasil evaluasi menunjukkan bahwa nilai presisi secara konsisten lebih tinggi dibandingkan recall seperti pada gambar 4. Secara sederhana, hal ini menunjukkan bahwa sistem memiliki tingkat ketepatan yang tinggi dalam memilih kata-kata yang penting dan relevan. Namun, karena sifat ringkasan yang sangat padat, sebagian informasi yang dianggap penting oleh manusia dapat terlewatkan.

Keterbatasan ini dapat dipahami, khususnya dalam konteks artikel berita. Tujuan utama dari sebuah ringkasan berita adalah menyajikan esensi informasi secara cepat dan akurat, sehingga pembaca dapat memperoleh gambaran umum serta memutuskan apakah perlu menelusuri artikel secara lengkap, di mana detail informasi tetap tersedia. Berdasarkan skor yang diperoleh, ROUGE-1 mencapai nilai tertinggi sebesar 0,3879, diikuti oleh ROUGE-L sebesar 0,3653, sementara ROUGE-2 memperoleh nilai terendah sebesar 0,3413. Urutan skor ini memberikan gambaran yang jelas bahwa sistem sangat efektif dalam mengidentifikasi kata-kata kunci dalam artikel berita. Tantangan utama muncul pada tahap penyusunan kata-kata tersebut menjadi frasa dan kalimat yang koheren. Nilai ROUGE-2 yang relatif rendah mengindikasikan bahwa sistem masih memerlukan pengembangan lebih lanjut dalam membangun struktur frasa dan kalimat yang lebih alami serta menyerupai gaya bahasa manusia.

D. Pembahasan

Hasil analisis kualitatif dan kuantitatif menunjukkan bahwa pipeline yang dirancang telah berfungsi sesuai dengan tujuan penelitian. Integrasi antara peringkasan abstraktif dan ekstraksi entitas bernama terbukti mampu menghasilkan keluaran yang kaya informasi sekaligus efisien. Ringkasan yang dihasilkan tidak hanya ringkas—dengan rata-rata panjang sekitar 37,53% dari teks asli (tingkat reduksi 62,47%)—tetapi juga memiliki kepadatan semantik yang tinggi, sehingga mampu merepresentasikan esensi berita secara efektif. Hasil evaluasi ROUGE (Gambar 2, Gambar 3) mengonfirmasi temuan ini, di mana nilai presisi lebih tinggi dibandingkan recall. Hal tersebut menunjukkan bahwa sistem cenderung menghasilkan ringkasan yang relevan, meskipun relatif singkat, sehingga menyebabkan beberapa detail informasi terlewatkan.

Di sisi lain, daftar entitas yang diekstraksi memberikan informasi kontekstual terkait unsur "siapa, apa, dan di mana" dalam suatu peristiwa. Namun demikian, beberapa keterbatasan penting berhasil diidentifikasi sebagai berikut:

1) Ambiguitas Semantik pada NER. Sebagai contoh, entitas "Romo" salah diklasifikasikan sebagai lokasi, padahal dalam konteks berita merujuk pada individu (Javier Romo). Kesalahan ini mencerminkan adanya bias dalam data pelatihan serta ambiguitas linguistik. Bahkan bagi manusia, interpretasi suatu kata dapat keliru tanpa konteks yang memadai; oleh karena itu, kesalahan semacam ini masih dapat dipahami dalam konteks NLP.
2) Redundansi Entitas. Entitas yang merujuk pada objek yang sama masih muncul secara terpisah[30], [31], [32], misalnya "Ayuso" dan "Juan Ayuso", atau pengulangan nama negara (contoh: "Philippines Philippines").
3) Ketergantungan pada Kualitas Sumber. Teks berita yang mengandung noise atau format yang tidak baku berpotensi menurunkan kualitas hasil peringkasan maupun keluaran NER[3], [33].

Kesalahan utama sistem sebagian besar dipengaruhi oleh ambiguitas semantik, seperti pada kasus “Romo” yang diklasifikasikan sebagai lokasi meskipun konteks kalimat menunjukkan bahwa entitas tersebut adalah seorang individu. Pola kesalahan lain yang sering muncul adalah redundansi entitas, seperti kemunculan terpisah antara “Ayuso” dan “Juan Ayuso”. Fenomena-fenomena ini mengindikasikan bahwa model masih lebih mengandalkan pola statistik dibandingkan pemahaman konteks secara menyeluruh. Analisis kesalahan ini menegaskan pentingnya integrasi modul entity linking dan entity normalization agar sistem dapat bekerja secara lebih konsisten dan akurat dalam memproses teks berita berbahasa Indonesia.

TABEL 3
PERBANDINGAN SKOR ROUGE PIPELINE VS METODE TRADISIONAL

Model	ROUGE-1 (F1)	ROUGE-2 (F1)	ROUGE-L (F1)
Pipeline (mT5)	**0.473**	**0.417**	**0.446**
Tradisional (mT5)	0.0049	0.0000	0.0045
Tradisional (TextRank)	0.0234	0.0000	0.0207

Hasil yang ditampilkan pada Tabel 3 menunjukkan bahwa di antara metode tradisional, CRF memiliki kinerja yang lebih baik dibandingkan BERT (Micro-F1 = 0,704 vs. 0,598). Tanpa konfigurasi yang tepat, CRF sering kali unggul karena aturan pelabelan sekuensinya (BIO) yang mampu menjaga konsistensi serta mengurangi kesalahan batas entitas. Sebaliknya, BERT dalam konfigurasi standar (vanilla BERT) mengevaluasi token secara individual, sehingga lebih rentan memecah entitas yang terdiri dari beberapa kata[34], [35]. Pada dataset berukuran kecil dan tidak seimbang (misalnya didominasi oleh entitas LOC), CRF cenderung lebih stabil dan menghasilkan skor Micro-F1 yang lebih tinggi.

Namun demikian, ketika dioptimalkan dalam pipeline yang diusulkan, kinerja model berbasis BERT mampu melampaui CRF dengan capaian skor Micro-F1 yang mendekati 0,70. Peningkatan ini menunjukkan bahwa kelemahan Transformer pada tugas pelabelan sekuens dapat dikompensasi melalui optimasi alur kerja, khususnya dengan penerapan teknik agregasi span untuk memperbaiki fragmentasi entitas akibat tokenisasi subword. Selain itu, integrasi mekanisme regex *fallback*, terutama untuk entitas temporal, berkontribusi dalam meningkatkan nilai recall tanpa mengorbankan presisi secara signifikan. Temuan ini menegaskan bahwa performa model NLP tidak hanya ditentukan oleh arsitektur inti, tetapi juga oleh strategi post-processing dan desain *pipeline* secara keseluruhan, sehingga pendekatan terintegrasi menjadi kunci dalam mengungguli metode klasik pada konteks pemrosesan berita berbahasa Indonesia.

Berdasarkan perbandingan yang ditunjukkan pada dua tabel sebelumnya, dapat disimpulkan bahwa pipeline mT5–BERT mengungguli pendekatan tradisional (TextRank–CRF). Pada tugas peringkasan, pipeline ini mencapai skor ROUGE-1 F1 sebesar 0,473 (Tabel 4), yang lebih tinggi dibandingkan metode tradisional. Pada tugas NER, pipeline berbasis BERT juga menunjukkan kinerja yang lebih baik dibandingkan CRF, dengan skor Micro-F1 yang lebih tinggi sebagai hasil dari penerapan agregasi span dan regex fallback. Temuan ini menegaskan bahwa optimasi alur kerja (workflow optimization) berperan signifikan dalam meningkatkan kualitas peringkasan maupun kinerja ekstraksi entitas.

IV. SIMPULAN

Penelitian ini berhasil mengimplementasikan *pipeline* NLP otomatis *end-to-end* yang mengintegrasikan akuisisi berita daring, peringkasan abstraktif, dan ekstraksi entitas bernama (NER) untuk teks berbahasa Indonesia secara konsisten dan terotomatisasi. Integrasi model Transformer menunjukkan kinerja yang efektif, di mana mT5 mampu menghasilkan ringkasan yang koheren dan informatif, sementara model berbasis BERT dapat mengekstraksi entitas kunci dengan tingkat akurasi yang memadai dalam konteks berita. Secara kuantitatif, sistem menghasilkan ringkasan dengan panjang rata-rata 37,53% dari teks asli atau tingkat reduksi sebesar 62,47%, yang menandakan efisiensi pemampatan konten tanpa kehilangan informasi esensial. Hasil evaluasi ROUGE menunjukkan nilai presisi yang lebih tinggi dibandingkan recall, mengindikasikan bahwa ringkasan yang dihasilkan cenderung relevan, ringkas, dan berfokus pada informasi inti. Meskipun demikian, masih ditemukan keterbatasan berupa kesalahan klasifikasi entitas akibat ambiguitas semantik serta redundansi variasi penulisan entitas, serta keterbatasan cakupan data yang hanya berasal dari satu portal berita. Penelitian selanjutnya disarankan untuk memperluas sumber data berita dengan izin resmi agar hasil penelitian dapat digeneralisasi dengan lebih baik, serta melakukan evaluasi kinerja secara formal menggunakan metrik standar, yaitu ROUGE untuk peringkasan dan *Precision, Recall*, serta F1-score untuk NER. Peningkatan akurasi dan konsistensi sistem dapat dicapai melalui *fine-tuning* model NER pada korpus berita berbahasa Indonesia yang teranotasi, disertai penerapan normalisasi dan *entity linking* untuk mengurangi ambiguitas dan redundansi entitas. Selain itu, pengembangan fitur lanjutan seperti analisis sentimen, pemodelan topik, serta integrasi

ke dalam layanan berbasis API diharapkan dapat meningkatkan skalabilitas dan nilai aplikatif *pipeline* dalam mendukung pemantauan media, intelijen informasi, dan aplikasi analitik berita berskala besar.

UCAPAN TERIMAKASIH

Penulis mengucapkan terima kasih kepada PT. Ekonomi Dunia Indonesia, selaku pengelola portal berita EGINDO, atas ketersediaan akses terhadap konten berita publik yang digunakan sebagai sumber data dalam penelitian ini. Seluruh data dimanfaatkan semata-mata untuk keperluan akademik dan non-komersial. Penulis menyatakan tidak terdapat konflik kepentingan serta menegaskan bahwa tidak ada sponsor maupun pihak eksternal yang memengaruhi perancangan penelitian, pengumpulan dan analisis data, interpretasi hasil, maupun keputusan untuk memublikasikan penelitian ini.

DAFTAR PUSTAKA

[1] A. Muharom and N. Rukhviyanti, "Development of Web-Based Multimedia Learning for Grade 3 Elementary School Mathematics," *INOVTEK Polbeng - Seri Informatika*, vol. 10, no. 2, pp. 1142–1152, Jul. 2025, doi: 10.35314/sj1qng08.
[2] K. Aggarwal, "A Review of Text Summarization Techniques Using NLP," *Computational Intelligence and Machine Learning*, vol. 4, no. 2, Oct. 2023.
[3] K. Bagla, A. Kumar, S. Gupta, and A. Gupta, "Noisy Text Data: Achilles' Heel of Popular Transformer Based NLP Models," Oct. 2021.
[4] D. Nagalavi and M. Hanumanthappa, "The NLP Techniques for Automatic Multi-article News Summarization Based on Abstract Meaning Representation," in *Proceedings of the Conference*, 2019, pp. 253–260.
[5] S. Sistla, "*Named Entity Recognition* : A Deep Dive," *Journal of Artificial Intelligence & Cloud Computing*, vol. 3, no. 6, pp. 1–5, Dec. 2024, doi: 10.47363/JAICC/2024(3)409.
[6] A. V. Patil, "Identifying specific details from text to populate databases and generate summaries using *Named Entity Recognition* ," *INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT*, vol. 08, no. 05, pp. 1–5, May 2024, doi: 10.55041/IJSREM33111.
[7] S. Mhatre and L. Ragha, "Implementing Extractive Summarization Methods on Extractive Datasets," in *5th International Conference on Data Intelligence and Cognitive Informatics (ICDICI)*, 2024, pp. 578–854.
[8] G. Cheirmpos, S. A. Tabatabaei, E. Kanoulas, and G. Tsatsaronis, "Benchmarking *Named Entity Recognition* Approaches for Extracting Research Infrastructure Information from Text," in *Proceedings of the Conference*, 2024, pp. 131–141.
[9] F. M. Apriansyah, T. I. Ramadhan, C. R. Hidayat, and A. K. Wijaya, "Perbandingan IndoBERT dan IndoRoBERTa Untuk Analisis Sentimen Pada Film Dokumenter Dirty Vote," *Jurnal Informatika: Jurnal Pengembangan IT*, vol. 10, no. 3, pp. 593–605, Jul. 2025, doi: 10.30591/jpit.v10i3.8607.
[10] N. Widaningsih, N. Windiyanti, and N. Rukhviyanti, "Web-based Inventory Information System using Agile Scrum Method at CV Tunggal Putra Jaya," *SISTEMASI*, vol. 14, no. 3, p. 1471, May 2025, doi: 10.32520/stmsi.v14i3.5253.
[11] D. N. Pryatama and N. Rukhviyanti, "Rancang Bangun Aplikasi Stok Barang dengan QRcode Menggunakan Metode Waterfall dan Framwork Laravel pada Konveksi Sfgiandra," *JURNAL KRIDATAMA SAINS DAN TEKNOLOGI*, vol. 7, no. 01, pp. 71–89, Feb. 2025, doi: 10.53863/kst.v7i01.1488.
[12] K. V. Benedict and N. Rukhviyanti, "Analysis of the Classification of Data on the Launch of Apple Mobile Phone Prices in China and Pakistan Using the Decision Tree Algorithm in Python Programming," *Eduvest - Journal of Universal Studies*, vol. 5, no. 9, pp. 10534–10546, Sep. 2025, doi: 10.59188/eduvest.v5i9.51409.
[13] H. A. Holiel, N. Mohamed, A. Ahmed, and W. Medhat, "English-Arabic Text Translation and Abstractive Summarization Using Transformers," in *20th ACS/IEEE International Conference on Computer Systems and Applications*, 2023, pp. 1–8.
[14] C. Meister, T. Vieira, and R. Cotterell, "Best-First Beam Search," *Trans. Assoc. Comput. Linguist.*, vol. 8, pp. 795–809, 2020, doi: 10.1162/tacl_a_00342.
[15] N. Ott, R. Horst, and R. Dörner, "Towards Reducing Latency Using Beam Search in an Interactive Conversational Speech Agent," in *IEEE Gaming, Entertainment, and Media Conference (GEM)*, 2024, pp. 1–6.
[16] S. Lemons, C. Linares López, R. C. Holte, and W. Ruml, "Beam Search: Faster and Monotonic," in *International Conference on Automated Planning and Scheduling (ICAPS)*, 2022, pp. 222–230.
[17] M. W. A. Pramana, D. P. S. Putri, and I. K. A. Purnawan, "Comparison of IndoBERT and Bi-LSTM Models for Indonesian Law Violation Text Classification," *Jurnal Informatika: Jurnal Pengembangan IT*, vol. 10, no. 4, pp. 1033–1043, Sep. 2025, doi: 10.30591/jpit.v10i4.8795.
[18] Asro Asro, Meisa Monica, Novi Rukhviyanti, and M. Yusron, "Analisis Literatur Review Perencanaan Strategi Sistem Informasi Menggunakan Metode Pieces Framework," *KRESNA: Jurnal Riset dan Pengabdian Masyarakat*, vol. 4, no. 2, pp. 161–169, Nov. 2024, doi: 10.36080/kresna.v4i2.182.
[19] D. Zatnika and N. Rukhviyanti, "Penerapan Metode Forward Chaining pada Sistem Pakar Rekomendasi Mobil Second dari Aspek Penghasilan Kerja," *Jurnal Penelitian Inovatif*, vol. 4, no. 4, pp. 2463–2476, Dec. 2024, doi: 10.54082/jupin.759.
[20] T. Wolf and others, "Transformers: State-of-the-Art Natural Language Processing," in *EMNLP System Demonstrations*, 2020, pp. 38–45. doi: 10.18653/v1/2020.emnlp-demos.6.
[21] A. Barbaresi, "Trafilatura: A Web Scraping Library and Command-Line Tool for Text Discovery and Extraction," in *ACL-IJCNLP System Demonstrations*, 2021, pp. 122–131. doi: 10.18653/v1/2021.acl-demo.15.
[22] H. Tan, R. Yan, L. Yang, L. Huang, L. Xiao, and Q. Yang, "Efficient Multiple-Precision and Mixed-Precision Floating-Point Fused Multiply-Accumulate Unit for HPC and AI Applications," in *Proceedings of the Conference*, 2023, pp. 642–659.
[23] Susan Juli Safitri, Gelar Alam Ramdhaniawan, Asro Asro, and Novi Rukhviyanti, "Analisis Literatur Review Perencanaan Strategi Sistem Informasi Menggunakan Metode Metode Five Competitive Force Pada CV. Bio Chitosan Indonesia," *Bridge : Jurnal publikasi Sistem Informasi dan Telekomunikasi*, vol. 2, no. 4, pp. 319–327, Sep. 2024, doi: 10.62951/bridge.v2i4.263.
[24] T. Hasan and others, "XL-Sum: Large-Scale Multilingual Abstractive Summarization for 44 Languages," in *Findings of ACL-IJCNLP*, 2021, pp. 4693–4703. doi: 10.18653/v1/2021.findings-acl.413.
[25] A. Al-Numai and A. Azmi, "LEMMA-ROUGE: An Evaluation Metric for Arabic Abstractive Text Summarization," *Indonesian Journal of Computer Science*, vol. 12, no. 2, pp. 470–481, 2023, doi: 10.33022/ijcs.v12i2.330.
[26] D. Arias, "Statistical Bias," in *Translational Sports Medicine*, Elsevier, 2023, pp. 163–164.

[27] N. Campolungo, T. Pasini, D. Emelin, and R. Navigli, "Reducing Disambiguation Biases in NMT by Leveraging Explicit Word Sense Information," in *NAACL-HLT*, 2022, pp. 4824–4838.
[28] T. Blevins and L. Zettlemoyer, "Moving Down the Long Tail of Word Sense Disambiguation with Gloss Informed Bi-encoders," in *ACL*, 2020, pp. 1006–1017.
[29] H. S. Yoon and others, "SMSMix: Sense-Maintained Sentence Mixup for Word Sense Disambiguation," in *EMNLP Findings*, 2022, pp. 1493–1502.
[30] H. Bast, M. Hertel, and N. Prange, "A Fair and In-Depth Evaluation of Existing End-to-End Entity Linking Systems," in *EMNLP*, 2023, pp. 6659–6672. doi: 10.18653/v1/2023.emnlp-main.414.
[31] K. Zaporojets, J. Deleu, Y. Jiang, T. Demeester, and C. Develder, "Towards Consistent Document-level Entity Linking," in *ACL Short Papers*, 2022, pp. 778–784. doi: 10.18653/v1/2022.acl-short.98.
[32] S. Dash, G. Rossiello, N. Mihindukulasooriya, S. Bagchi, and A. Gliozzo, "Open Knowledge Graphs Canonicalization using Variational Autoencoders," in *EMNLP*, 2021, pp. 10379–10394. doi: 10.18653/v1/2021.emnlp-main.808.
[33] A. Hamdi, A. Jean-Caurant, N. Sidère, M. Coustaty, and A. Doucet, "Assessing and Minimizing the Impact of OCR Quality on *Named Entity Recognition* ," in *Proceedings of the Conference*, 2020, pp. 87–101.
[34] Y. Cao and A. Yusup, "Chinese Electronic Medical Record *Named Entity Recognition* based on BERT-WWM-IDCNN-CRF," in *Dependable Systems and Their Applications (DSA)*, 2022, pp. 582–589.
[35] J. C.-W. Lin, J. M.-T. Wu, Y. Shao, M. Pirouz, and B. Zhang, "A Latent Variable CRF Model for Labeling Prediction," in *Proceedings of the Conference*, 2019, pp. 68–78.