Seminar Nasional Sains Data 2025 (SENADA 2025)
UPN "Veteran" Jawa Timur
E-ISSN 2808-5841
P-ISSN 2808-7283

Text Summarization Berita Online Menggunakan *Named Entity Recognition* (NER) dan *Part of Speech* (POS) *Tagging*

Mediliadita Dwi Dimastia[1], Ferdyan Paskah Dyka Wahyudi [2], Muhammad Arif Billah[3], Regita Putri Permata[4]
[1,2,3,4] *Program Studi Sains Data, Telkom University*
[1]mediliadita@student.telkomuniversity.ac.id
[2]ferdyanpaskah@student.telkomuniversity.ac.id
[3]arifhans@student.telkomuniversity.ac.id
[4]regitapermata@telkomuniversity.ac.id
Corresponding author email: regitapermata@telkomuniversity.ac.id

Abstract: *The objective of this research is to develop an automated online news summarization system using Named Entity Recognition (NER) and Part of Speech (POS) Tagging. In text summarization, NER enables the system to focus on key sections containing important entities, making the resulting summaries more informative and representative. Meanwhile, POS Tagging aids in understanding sentence structure and contextual meaning. This helps readers quickly and efficiently grasp core information, ensuring the essential message is preserved. The methods employed include data collection preparation (case folding, tokenization, stopword removal, filtering, and stemming) from CNN Indonesia news pages, as well as the use of NER and POS tagging to identify entity structures and key statements. Summary quality assessment was conducted using the BERTScore metric. The results indicate that the combination of NER and POS Tagging delivers the best performance, achieving an F1-score of 0.74—higher than using NER (0.709) or POS Tagging (0.734) separately. This combination improves precision (0.732) without compromising recall (0.7486), demonstrating its effectiveness in extracting crucial information. The study proves that integrating NER and POS tags is effective for text summarization and can serve as a solution for delivering concise and relevant information to users.*
Keywords: *Online News, Named Entity Recognition, Part of Speech Tagging, BERTScore*

Abstrak: Tujuan dari penelitian ini adalah untuk mengembangkan sistem peringkasan berita online yang dikembangkan secara otomatis menggunakan Named Entity Recognition (NER) dan *Part of Speech* (POS) *Tagging*. NER dalam peringkasan teks memungkinkan sistem untuk fokus pada bagian-bagian penting yang mengandung entitas penting, sehingga ringkasan yang dihasilkan lebih informatif dan representative. Sedangkan POS Tagging akan membantu dalam memahami struktur kalimat dan konteks makna. Ini akan membantu pembaca untuk memahami informasi inti dengan cepat dan efisien, membuat koeksistensi teks pesan penting. Metode yang digunakan termasuk mempersiapkan pengumpulan data (*case folding, tokenizing, stopword removal, filtering*, dan *stemming*) dari halaman berita CNN Indonesia, serta menggunakan NER dan pasca-tag untuk mengidentifikasi struktur entitas dan pernyataan penting. Penilaian Kualitas Ringkasan dilakukan dengan menggunakan metric BERTScore. Hasil penelitian menunjukkan bahwa kombinasi NER dan POS Tagging menghasilkan performa terbaik dengan nilai F1 sebesar 0,74 lebih tinggi dibandingkan penggunaan metode NER (0,709) atau POS Tagging (0,734) secara terpisah. Kombinasi ini meningkatkan breksion (0,732) tanpa mempengaruhi sektor modal (0,7486), dan menunjukkan efektivitasnya dalam mengumpulkan informasi penting. Studi ini membuktikan bahwa integrasi tag NER dan POS efektif dalam hal ringkasan teks pesan, dan dapat menjadi solusi untuk menyajikan kepadatan dan informasi terkait kepada pengguna.
Kata kunci: Berita Online, *Named Entity Recognition, Part of Speech Tagging, BERTScore*

I. PENDAHULUAN

Ringkasan berita adalah teks singkat yang menyajikan gagasan penting dari suatu artikel, bertujuan untuk menyampaikan inti informasi kepada pembaca[1]. Idealnya, ringkasan tidak melebihi separuh panjang berita asli dan tetap mempertahankan informasi kunci. Proses peringkasan berita melibatkan pengurangan konten tanpa menghilangkan substansi, sehingga menghasilkan representasi yang padat dari keseluruhan berita. Namun, mengekstrak kalimat-kalimat representatif membutuhkan waktu dan analisis mendalam. Sebagai contoh, kalimat akrab yang sering dijumpai dalam berita adalah kalimat utama yang dikelompokkan sebagai kalimat deskriptif dan nantinya digunakan untuk bagian dari frasa ringkasan[1].

Seiring pesatnya perkembangan teknologi berita elektronik, peringkasan berita berbasis kueri semakin penting. Teknik ini bertujuan menyajikan deskripsi singkat, terorganisir, dan dipersonalisasi dari berbagai kelompok berita, memberikan informasi lebih rinci dan spesifik dibandingkan pencarian umum di internet. Dengan demikian, pengguna dapat memperoleh informasi yang sesuai kebutuhan secara lebih efektif. Peringkasan berita otomatis terdiri dari mengambil sumber informasi, mengekstrak isi darinya, dan menyajikan konten yang paling penting kepada pengguna dalam bentuk ringkas dan sesuai dengan apa yang dibutuhkan pengguna juga [2]. Dalam penelitian ini, *Named Entity Recognition (NER)* dan *Part of Speech (POS) Tagging* dipilih karena kemampuannya dalam memodelkan peringkasan teks berita online[3].

Named Entity Recognition (NER) berperan penting dalam mengidentifikasi entitas seperti nama orang, lokasi, organisasi, dan waktu yang terdapat dalam teks berita[4]. Dengan mengklasifikasikan entitas tersebut secara otomatis, NER membantu sistem dalam mengekstraksi informasi yang relevan dan mengurangi kompleksitas dalam proses peringkasan teks [5]. Penerapan NER dalam peringkasan berita online memungkinkan sistem untuk fokus pada bagian-bagian penting yang mengandung entitas bernama, sehingga ringkasan yang dihasilkan lebih informatif dan representatif. Sementara itu, *Part of Speech* (POS) *Tagging* juga memiliki peranan yang signifikan dalam proses peringkasan teks. POS Tagging memberikan informasi tentang kategori kata dalam sebuah kalimat, seperti kata benda, kata kerja, atau kata sifat, yang membantu dalam memahami struktur kalimat dan konteks makna. Kombinasi antara POS Tagging dan NER memungkinkan sistem untuk tidak hanya mengenali entitas penting, tetapi juga memahami hubungan antar kata dalam kalimat, sehingga dapat menghasilkan ringkasan yang lebih koheren dan bermakna [7].

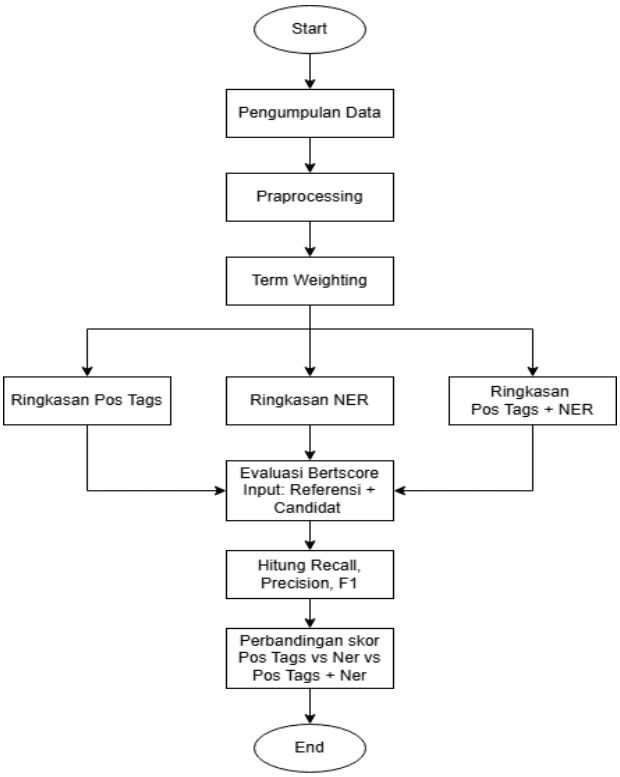
Penelitian ini bertujuan untuk menghasilkan sistem peringkasan teks berita online secara otomatis dengan memanfaatkan teknik *Named Entity Recognition* (NER) dan *Part of Speech* (POS) *Tagging*. Dengan pendekatan ini, sistem diharapkan dapat mengidentifikasi informasi penting atau inti dari suatu teks berita online, seperti nama entitas dan struktur kalimat agar memberikan ringkasan yang merepresentasikan inti dari isi berita online secara efektif atau optimal [4].

Seiring dengan perkembangan teknologi *Natural Language Processing* (NLP), berbagai metode dan algoritma telah dikembangkan untuk meningkatkan kualitas peringkasan teks otomatis[6]. Pendekatan berbasis NER dan POS Tagging terbukti efektif dalam mengekstrak informasi penting dari dokumen berita yang panjang dan kompleks. Penelitian sebelumnya menunjukkan bahwa integrasi kedua teknik ini mampu meningkatkan akurasi dalam memilih kalimat-kalimat kunci yang merepresentasikan isi berita secara optimal, sehingga sangat potensial untuk diterapkan dalam sistem peringkasan berita online saat ini [7]. Selain itu, metrik evaluasi seperti ROUGE *(Recall-Oriented Understudy for Gisting Evaluation)* telah menjadi standar untuk mengukur kualitas ringkasan otomatis. Namun, perkembangan NLP modern memperkenalkan pendekatan evaluasi berbasis *semantic similarity*, seperti BERTScore, yang mengukur kesesuaian ringkasan dengan teks sumber berdasarkan embedding kontekstual dari model BERT[8]. Metrik ini lebih unggul dalam menangkap kesamaan makna dibandingkan ROUGE yang hanya mengandalkan tumpang-tindih *n-gram.*

Dengan pendekatan ini, sistem dapat memilih kalimat yang tidak hanya mengandung entitas utama, tetapi juga memiliki struktur sintaksis yang mendukung penyampaian informasi inti secara efektif. Hasil penelitian menunjukkan bahwa penggunaan NER dan POS Tagging secara bersama-sama memberikan kontribusi signifikan dalam menghasilkan ringkasan yang lebih informatif, koheren, dan sesuai dengan makna asli berita, serta meningkatkan skor evaluasi seperti *BERTScore* dibandingkan metode konvensional [9].

II. METODE PENELITIAN

Metode penelitian ini menguraikan tahapan proses pembuatan ringkasan teks (*text summarization*) dari artikel berita online. Proses *summarization* akan dilaksanakan melalui penerapan teknik *named entity recognition* (NER) dan *part of speech* (POS) *tagging*. Adapun langkah-langkah penelitiannya meliputi:



Gambar 1. Diagram Alir Metode Penelitian

2.1. Pengumpulan Data

Dalam proses pengumpulan data, peneliti menerapkan teknik *web scraping* sebagai metode ekstraksi data guna memperoleh informasi pendukung yang relevan untuk kebutuhan pembuatan *text summarization*. Proses scraping dilakukan menggunakan aplikasi Google Colaboratory dengan sumber data berasal dari situs berita online dari CNN Indonesia. Berita online yang berhasil didapatkan dalam proses scraping ini ada lima artikel berita. Topik datanya tentang berita seputar kasus tuduhan ijazah palsu Presiden ke-7 RI. Contoh data yang telah dikumpulkan dapat dilihat pada Tabel 1.

Tabel 1. Hasil Pengumpulan Data Teks

Judul Artikel	Kalimat	Jumlah Kalimat	Jumlah Tokenisasi (Kata)
Usai Uji Labfor dan Pembanding, Bareskrim Sebut Skripsi Jokowi Identik	Bareskrim Polri menetapkan kasus dugaan ijazah palsu Presiden ketujuh RI Joko Widodo (Jokowi) yang dilaporkan Ketua Tim Pembela Ulama dan Aktivis (TPUA) Eggi Sudjana tidak memenuhi unsur pidana untuk dilanjutkan ke tahap penyidikan. Selain itu, Bareskrim juga menyatakan skripsi yang ditulis …	29	678
Jokowi Puji Penyelidikan Bareskrim soal Ijazah: Detail Sekali	Presiden ke-7 RI, Joko Widodo (Jokowi) memuji hasil penyelidikan Bareskrim Polri yang menyatakan keaslian ijazah SMA hingga S1 Fakultas Kehutanan UGM. Hasil penyelidikan tersebut diumumkan Bareskrim, Kamis (22/5) kemarin …	22	318
Bareskrim Setop Penyelidikan Laporan Dugaan	Bareskrim Polri memutuskan menghentikan penyelidikan laporan terkait dugaan kepemilikan	14	366

Ijazah Palsu Jokowi	ijazah palsu Presiden ke-7 RI Joko Widodo (Jokowi) yang dilayangkan Tim Pembela Ulama dan Aktivis (TPUA) ...		
Diperiksa soal Ijazah Palsu Jokowi, Kader PSI Bawa Dua Flashdisk	Kader PSI, Dian Sandi mengaku membawa dua buah flashdisk berisi foto hingga video untuk menghadapi pemeriksaan terkait laporan Presiden ke-7 RI Joko Widodo (Jokowi) soal dugaan ijazah palsu di Polda Metro Jaya, Rabu ...	15	313
Ketum Solmet Diperiksa Terkait Laporan soal Ijazah Jokowi	Ketua Umum kelompok relawan Solidaritas Merah Putih (Solmet) Silfester Matutina diperiksa terkait laporan ke Roy Suryo cs di Polres Metro Jakarta Selatan, Rabu (28/5). Diketahui, Roy dan tiga orang …	15	395

2.2. Text Preprocessing

Setelah melakukan scrapping terhadap berita dari internet, dilakukan serangkaian *preprocessing* untuk menyiapkan data sebelum peringkasan dan analisis lebih lanjut. Tahap ini merupakan langkah awal dalam mengubah konten berita menjadi unit-unit leksikal (token) yang siap diproses dalam tahap pembobotan. Proses ini diterapkan secara menyeluruh pada seluruh isi berita. Adapun langkah – langkah text *preprocessing* sebagai berikut[10]:

- Case folding merupakan proses transformasi uniform terhadap seluruh karakter alfabetik dalam teks menjadi format konsisten, baik uppercase maupun lowercase, untuk menstandarisasi representasi leksikal.
- Tokenizing adalah proses dekomposisi string linguistik menjadi unit-unit diskret (token) yang merepresentasikan satuan leksikal terkecil dalam sistem pemrosesan bahasa. Proses tokenisasi Kalimat dipisah menggunakan sent_tokenize dari library NLTK. Kata-kata dalam kalimat ditokenisasi menggunakan word_tokenize. Setiap teks dianalisis jumlah kalimatnya menggunakan sent_tokenize.
- Stopword removal adalah teknik penyaringan kata-kata fungsi (function words) yang memiliki distribusi frekuensi tinggi namun kontribusi semantiknya minimal dalam representasi dokumen. Dalam proses stopword Indonesia, kata-kata dalam berita yang akan dihapus ialah kata modal, kata depan, konjungsi, contohnya kata "di", "oleh", "pada", "sebuah", "karena" dan lain-lain.
- Filtering, pada tahap ini data teks dibersihkan dari kata atau karakter yang tidak sesuai. Karakter yang akan dihilangkan seperti hashtag yang dilambangkan dengan karakter "#", username yang ditandai dengan karakter "@", link URL, dan menghilangkan tanda baca seperti tanda seru (!), tanya (?), dan titik (.)
- Normalisasi Spasi yaitu menghilangkan spasi ganda, serta spasi awal dan akhir.
- Stemming untuk merubah kata bentukan menjadi kata dasar. Proses ini menggunakan Sastrawi untuk mengembalikan setiap kata ke bentuk dasarnya.
- N-gram construction adalah proses menggabungkan kata-kata yang berurutan dalam kalimat menjadi unit-unit berseri, dimana panjang seri maksimumnya adalah jumlah seluruh kata dalam kalimat tersebut dikurangi satu [11]. Hasil contstruction yang berupa bigram dapat dilihat seperti **Gambar 1.**

2.3. Part of Speech (POS) Tagging

Part of Speech Tagging (POS Tagging) merupakan teknik dalam pemrosesan bahasa alami (NLP) yang berfungsi untuk mengidentifikasi dan memberi label pada setiap kata dalam kalimat berdasarkan kelas katanya, seperti kata benda (*noun*), kata kerja (*verb*), kata sifat (*adjective*), atau kata keterangan (*adverb*). Proses ini menggunakan *Application Programming Interface (API)* khusus yang dirancang untuk Bahasa Indonesia, dalam hal ini memanfaatkan pustaka

Stanza. Stanza merupakan sebuah toolkit NLP berbasis *neural network* yang dikembangkan oleh Stanford University. Stanza menyediakan model pra-latih khusus untuk Bahasa Indonesia (id) yang dilatih pada korpus *Universal Dependencies* (UD) Indonesia, memungkinkan analisis struktur gramatikal teks dengan akurasi tinggi. Model ini mengadopsi arsitektur *bidirectional* LSTM yang dioptimalkan untuk tugas tagging sequential, dengan dukungan tokenisasi dan lemmatisasi terintegrasi. Implementasi Stanza dipilih karena kemampuannya menangani karakteristik morfologis Bahasa Indonesia, seperti prefiksasi (me-, ber-) dan konfiksasi (pe-an), yang sulit diakomodasi oleh model berbasis *rule-based* atau *statistical* tradisional. Dalam penelitian ini, POS Tagging menjadi fondasi penting untuk memahami hubungan antar kata dalam kalimat, sehingga memudahkan ekstraksi informasi inti dari suatu teks.

Penelitian oleh Pisceldo, Adriani, dan Manurung (2009) mengembangkan standar POS Tagging untuk Bahasa Indonesia dengan 25 label kelas kata. Label-label ini terbagi menjadi dua kelompok utama: 9 label untuk simbol dan tanda baca (seperti titik atau koma), serta 16 label untuk kategori kata (misalnya nomina, verba, adjektiva)[12]. Klasifikasi ini memungkinkan analisis teks yang lebih akurat dan terstruktur, khususnya dalam aplikasi seperti peringkasan otomatis, di mana pemahaman terhadap peran gramatikal kata menentukan kualitas hasil ringkasan. Implementasi POS Tagging dalam penelitian ini bertujuan untuk meningkatkan presisi sistem dalam mengidentifikasi kalimat-kalimat kunci yang merepresentasikan isi berita secara utuh.

Tabel **2**. Contoh Hasil *Part-of-Speech Tagging*

No.	Kata	Label	Artikel
0	bareskrim	JJ	1
1	polri	NN	1
2	tetap	NN	1
3	duga	NN	1
4	ijazah	NN	1
...	...	...	...

2.4. Term Weighting

Term Weighting merupakan proses perhitungan bobot setiap *term* atau kata dalam dokumen, yang bertujuan untuk mengidentifikasi tingkat kepentingan dan kontribusi masing-masing *term* dalam proses peringkasan dokumen [13]. Semakin sering suatu *term* muncul dalam dokumen tertentu namun jarang muncul dalam dokumen lainnya, maka semakin besar bobot yang diberikan terhadap *term* tersebut. Dalam proses *Term Weighting* ini, metode yang digunakan untuk pembobotan adalah TF-IDF (*Term Frequency–Inverse Document Frequency*).

2.4.1. Term Frequency−Inverse Document Frequency (TF−IDF)

TF−IDF merupakan metode statistik untuk mengevaluasi signifikansi suatu term dalam dokumen relatif terhadap seluruh koleksi dokumen. Skema pembobotan ini mengasumsikan bahwa term dengan frekuensi tinggi (stopwords) telah dieliminasi melalui proses preprocessing sebelumnya. Perhitungan TF-IDF dapat dilihat pada persamaan (1) dan (2).

$$W_{i,j} = TF \times IDF \tag{1}$$

$$IDF(t) = \log\left(\frac{N}{n_t}\right) \tag{2}$$

Dimana:

N = jumlah total berita.

n_t = jumlah berita yang mengandung kata t.

2.5. Named Entity Recognition (NER)

Named Entity Recognition (NER) adalah salah satu komponen dalam ekstraksi informasi pada *Natural Language Processing(NLP).* Teknik ini berfungsi sebagai tahap awal dalam proses ekstraksi informasi dengan mengorganisasikan teks tak terstruktur menjadi data terstruktur*Named entity recognition* digunakan untuk mengidentifikasi dan mengklasifikasi kata atau frasa ke dalam jenis entitasnya. NER dapat diimplementasikan pada *machine translation, question answering dan semantic web*[14].

Tabel 3. Contoh Label NER

No.	Label	Keterangan
1.	PERSON	Entitas yang menandai nama orang, termasuk nama depan, tengah, belakang, atau gelar.
2.	GPE	(Geo-Political Entity), Entitas geografis dengan makna politik (negara, provinsi, kota).
3.	ORG	ORGANIZATION untuk menandai nama perusahaan, lembaga, organisasi, atau institusi.
4.	FAC	FACILITY, entitas seperti bangunan, infrastruktur, atau fasilitas buatan manusia.
5.	NORP	(Nationalities/Religions/Political Groups), kelompok berdasarkan kebangsaan, agama, atau afiliasi politik.
6.	DATE	Entitas yang menunjukkan Tanggal, periode, atau waktu tertentu.
7.	EVENT	Menandai entitas acara, kejadian bersejarah, atau kegiatan.
8.	LAW	Entitas untuk nama undang-undang, peraturan, atau dokumen hukum.

Algoritma NER digunakan untuk mengenali entitas penting dalam teks seperti nama orang, organisasi, lokasi, dan waktu. Hasil dari NER membantu dalam menjaga entitas penting tetap dipertahankan dalam hasil ringkasan. Model yang digunakan adalah *en_core_web_sm* dari library spaCy. NER dilakukan pada kolom *Clean_Text_String*, dan hasil ekstraksi entitas disimpan di kolom baru bernama NER, yang berisi pasangan (entitas, label). Contoh hasil entitas:

Tabel 4. Hasil Ekstrasi Entitas NER ke dalam kolom baru

No.	Teks Artikel	NER
1	Bareskrim Polri menetapkan kasus dugaan ijazah...	[(polri, PERSON), (ri, GPE), (jokowi, ORG), (k...
2	Presiden ke-7 RI, Joko Widodo (Jokowi) memuji ...	[(presiden ri joko widodo jokowi, ORG), (polri...
3	Bareskrim Polri memutuskan menghentikan penyel...	[(polri, PERSON), (ri, GPE), (jokowi layang, P...
4	Kader PSI, Dian Sandi mengaku membawa dua buah...	[(kader psi dian, PERSON), (bawa, GPE), (kait,...
5	Ketua Umum kelompok relawan Solidaritas Merah ...	[(ketua kelompok, PERSON), (rawan, GPE), (soli...

Selanjutnya menghitung jumlah entitas (tanpa membedakan labelnya) yang muncul dalam masing-masing artikel. Hal ini bertujuan untuk melihat seberapa padat artikel tersebut dari sisi entitas yang dikenali. Contoh ringkasan entitas tanpa label per artikel:

Tabel 5. Ringkasan Entitas NER per Artikel

Artikel	Jumlah Entitas	Entitas Dominan (Top 3–5)
Artikel 1	±49	polri, ulama, jokowi, tpua eggi, skripsi
Artikel 2	±15	tim bela, ulama, polri, surakarta, sidang
Artikel 3	±28	polri, tim bela, ulama, fakultas, jokowi
Artikel 4	±22	metro jaya, dian, bawa, kuhp, jokowi
Artikel 5	±24	jakarta, tim advocate, roy, solidaritas, jokowi

Setelah melakukan proses *named entity recognition* (NER) pada setiap artikel, langkah berikutnya adalah menghitung peluang kemunculan setiap label entitas dalam masing-masing artikel. Hal ini bertujuan untuk memperoleh gambaran distribusi jenis entitas yang paling sering muncul, serta melihat fokus pembahasan yang mungkin berbeda antar artikel. Serta peluang kemunculan label NER akan menjadi inputan bobot pada proses peringkasan teks artikel. Proses penghitungan dilakukan dengan cara mengiterasi hasil NER dari setiap artikel dan mengakumulasi jumlah kemunculan tiap jenis labelnya (seperti PERSON, GPE, ORG, NORP, DATE, FAC, dll).

2.6. Evaluasi Model

Teknik evaluasi model pada penelitian ini menggunakan metode *BERTScore*. *BERTScore* (*Bidirectional Encoder Representations from Transformers Score*) adalah metode yang digunakan untuk menghitung kualitas kemiripan semantik antar token dari dua kalimat, yaitu ringkasan hasil sistem dan ringkasan referensi (*Ground Truth)*[15]. *BERTScore* akan mengukur kesamaan makna antar token dengan menggunakan representasi vektor dari model BERT dan menghitung cosine similarity antar token-token dari dua kalimat. Terdapat tiga metrik evaluasi utama *BERTScore* yang umum digunakan[16]:

- Precision : Seberapa banyak token dalam ringkasan sistem yang relevan secara semantik dengan token dalam ringkasan referensi.
- Recall : Seberapa banyak token dalam ringkasan referensi yang berhasil ditangkap secara semantik oleh ringkasan sistem.
- F1-score : metrik evaluasi yang menghitung rata-rata harmonis antara precision dan recall, berguna untuk memberikan gambaran seimbang atas kedua komponen evaluasi tersebut.

Dalam implementasi penelitian ini, BERTScore dihitung menggunakan library bert-score dalam Python[17], dengan model pralatih *bert-base-multilingual-cased* agar kompatibel dengan teks berbahasa Indonesia. Sebelum dihitung, teks ringkasan dilakukan normalisasi seperti case folding, penghapusan karakter non-alfabet, dan penghapusan spasi ganda, tanpa proses stemming atau stopword removal agar makna semantik tetap terjaga. Ringkasan referensi disusun oleh dua ahli bahasa independen yang mempertimbangkan aspek cakupan informasi, koherensi antar kalimat, dan kesesuaian isi berita. Penggunaan *BERTScore* dinilai lebih representatif dalam

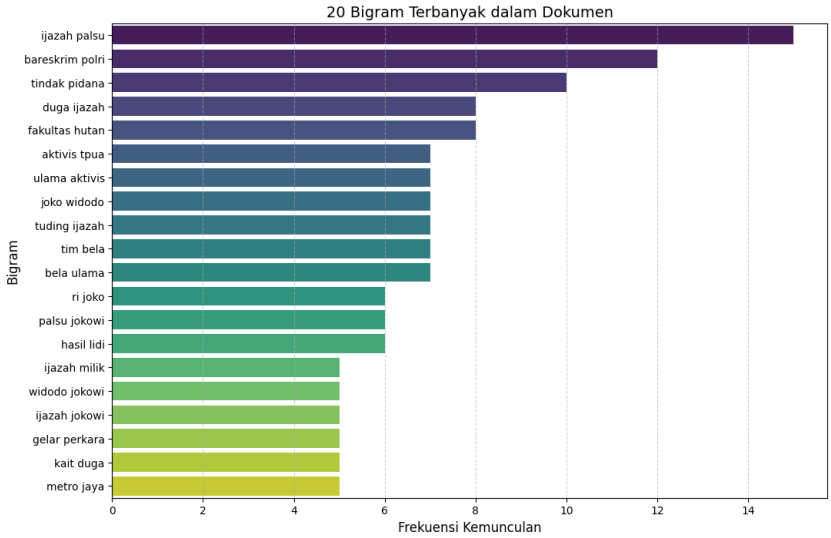
konteks ringkasan berita karena mampu menangkap kesamaan makna walaupun struktur kalimat atau pemilihan katanya berbeda [15]. Hasil evaluasi *BERTScore* diinterpretasikan dalam skala nilai 0 hingga 1, di mana nilai:

- ≥ 0.85 dianggap excellent (kesesuaian semantik tinggi),
- 0.70 – 0.85 dianggap good,
- < 0.70 dianggap perlu perbaikan.

III. HASIL DAN PEMBAHASAN

3.1. Eksplorasi Data

Hasil pengumpulan data teks dari beberapa berita kemudian dilakukan Preprocessing teks. Pada Gambar 1. menunjukkan top 20 Bigram dalam berita seperti kata "ijazah palsu", "bareskrim polri", "tindak pidana", dan "fakultas hutan".



Gambar 2. Top 20 Bigram Terbanyak dalam Berita

Gambar 3. Word Cloud Artikel Berita

Gambar 3. memperlihatkan bahwa topik utama dalam data adalah kontroversi dugaan ijazah palsu Jokowi**,** yang menjadi perhatian utama warganet atau media**.** Fokus pembicaraan

berada pada aspek pelaporan hukum, proses pemeriksaan, dan keaslian dokumen pendidikan**.** Terdapat juga keterlibatan tokoh-tokoh serta institusi hukum dalam narasi tersebut.

3.2. Part of Speech (POS) Tagging

Proses *POS (Part of Speech) tagging* dilakukan untuk mengidentifikasi jenis kata (kata benda, kata kerja, kata sifat, dsb.). Dengan informasi ini, kata-kata penting dalam kalimat dapat diidentifikasi. POS Tagging sangat membantu dalam menyaring informasi esensial dari teks berita. Fungsi POS Tagging diterapkan pada setiap token dalam Processed_Text untuk mengidentifikasi jenis katanya, misalnya:

- NN (kata benda umum),
- JJ (kata sifat),
- VB/VBD/VBP/VBZ/VBG (varian kata kerja),
- NNS (kata benda jamak), dll.

Contoh hasil POS Tagging:

Tabel 6. Hasil Pemrosesan POS_Tag

No.	Teks Artikel	POS_Tags
1	Bareskrim Polri menetapkan kasus dugaan ijazah...	[(bareskrim, JJ), (polri, NN), (tetap, NN), ...
2	Presiden ke-7 RI, Joko Widodo (Jokowi) memuji ...	[(presiden, JJ), (ri, NN), (joko, NN), (widodo...
3	Bareskrim Polri memutuskan menghentikan penyel...	[(bareskrim, JJ), (polri, NN), (putus, NN), (h...
4	Kader PSI, Dian Sandi mengaku membawa dua buah...	[(kader, NN), (psi, NN), (dian, JJ), (sandi, N...
5	Ketua Umum kelompok relawan Solidaritas Merah ...	[(ketua, NN), (kelompok, NN), (rawan, NN), ...

Data POS tagging dari seluruh artikel kemudian dikompilasi ke dalam sebuah DataFrame pos_df untuk memudahkan analisis selanjutnya. Penerapan POS Tagging dalam proses peringkasan berita memiliki manfaat signifikan, terutama dalam:

- Mengidentifikasi kalimat informatif: Kalimat yang mengandung kombinasi NN (kata benda), JJ (kata sifat), dan VB (kata kerja) cenderung memuat inti informasi.
- Peningkatan presisi dalam pemilihan kalimat ringkasan dengan mengetahui jenis kata dominan, sistem dapat memilih kalimat yang lebih padat informasi.

3.3. Named Entity Recognition (NER)

Berdasarkan **Tabel 7** membandingkan distribusi entitas bernama yang teridentifikasi dalam lima artikel berbeda. Artikel 1 mendominasi jumlah entitas PERSON dengan 25 kemunculan, menunjukkan bahwa artikel ini kemungkinan berfokus pada pembahasan tokoh tertentu, seperti profil atau wawancara. Sementara itu, entitas GPE (lokasi geografis) paling banyak muncul di Artikel 5 (8 kemunculan), mengindikasikan bahwa artikel ini membahas topik terkait wilayah atau isu internasional.

Entitas ORG (organisasi) paling menonjol di Artikel 1 dengan 22 kemunculan, yang menunjukkan bahwa artikel ini mungkin membahas topik korporasi atau lembaga, seperti merger perusahaan atau berita bisnis. Sebaliknya, Artikel 5 hanya memiliki 1 entitas ORG, berarti pembahasan organisasi hampir tidak ada. Entitas NORP (kelompok kebangsaan, agama, atau politik) hanya muncul di Artikel 3, 4, dan 5, dengan jumlah relatif sedikit,

mengisyaratkan bahwa artikel-artikel tersebut mungkin membahas keragaman budaya, isu politik, atau konflik berbasis identitas.

Tabel 7. NER dan labelnya per masing-masing artikel

Label	Article 1	Article 2	Article 3	Article 4	Article 5
PERSON	25	10	13	16	15
GPE	4	3	4	3	8
ORG	22	5	12	3	1
NORP	-	-	3	3	1
DATE	-	-	1	-	1
FAC	-	-	-	3	1

Sementara itu, entitas DATE (tanggal atau periode waktu) hanya muncul di Artikel 3 dan 5, masing-masing dengan 1 kemunculan, menunjukkan bahwa artikel-artikel ini mungkin menyebutkan momen spesifik, seperti peristiwa sejarah atau peringatan. Adapun entitas FAC (fasilitas atau infrastruktur) hanya ditemukan di Artikel 4 dan 5 dalam jumlah kecil, mengindikasikan sedikitnya pembahasan tentang bangunan atau infrastruktur tertentu. Secara keseluruhan, tabel ini membantu memahami fokus dan karakteristik masing-masing artikel berdasarkan entitas yang dominan. Data entitas yang sudah dilabeli menggunakan NER dapat dilihat pada **Tabel 8.**

Tabel 8. Named Entity Recognition Results

No.	Entity	Label	Artikel
0	polri	PERSON	1
1	ri	GPE	1
2	jokowi	ORG	1
3	ketua tim bela	PERSON	1
4	ulama	GPE	1
...	...	...	...
152	metro jaya	FAC	5
153	april	DATE	5
154	ketua tim bela	PERSON	5
155	ulama	GPE	5
156	kuhp hasut	PERSON	5

Dari hasil analisis *Named Entity Recognition* (NER) terhadap lima artikel, diperoleh informasi yang mendalam mengenai distribusi entitas bernama dalam teks. Melalui perhitungan frekuensi NER per artikel, terlihat bahwa entitas bertipe PERSON paling dominan muncul, mencerminkan bahwa artikel-artikel tersebut banyak menyoroti tokoh-tokoh atau individu tertentu yang menjadi pusat isu atau peristiwa. Disusul oleh entitas bertipe ORG (organisasi) dan GPE (entitas geopolitik), yang menunjukkan bahwa institusi dan wilayah geografis juga memiliki peran penting dalam konteks pemberitaan.

3.4. TF-IDF (Term Frequency-Inverse Document Frequency)

TF-IDF berperan sebagai metode pembobotan term untuk mengidentifikasi kata/frasa paling penting dalam dokumen berita. Integrasinya dengan NER dan POS Tagging bertujuan untuk memilih kalimat yang tidak hanya mengandung entitas bernama (hasil NER) dan kata dengan fungsi gramatikal penting (hasil POS Tagging), tetapi juga memiliki bobot semantik tinggi. Misalnya, entitas seperti "Bareskrim Polri" (label ORG) atau "Jokowi" (label PERSON) yang muncul dengan frekuensi signifikan dalam dokumen akan diberi bobot lebih besar. Demikian pula, kata dengan tag POS seperti kata benda (NN) atau kata kerja (VB)

yang memiliki bobot TF-IDF tinggi akan diprioritaskan. Perhitungan bobotnya berasal perkalian antara skor TF-IDF dengan bobot NER dan POS tag hasilnya dapat dilihat pada **Tabel 9.**

Tabel 9. Hasil Pembobotan Teks

Artikel	Kata	Label NER	Pos Tag	TF-IDF	Bobot NER	Bobot Pos Tag	Skor
1	ijazah	PERSON	NOUN	0.276	1.0	0.476	0.132
1	skripsi	ORG	NOUN	0.373	1.0	0.476	0.177
5	saudara	GPE	NOUN	0.292	1.0	0.438	0.128
...	...	...					

3.5. Summarization Text

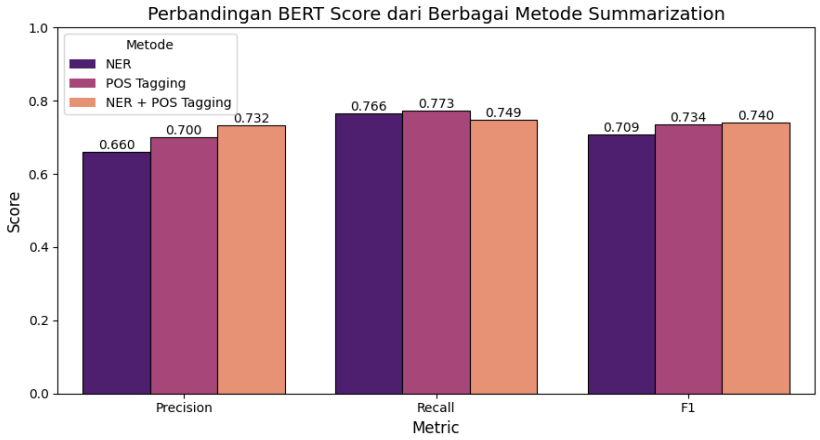
Tabel 10. Contoh Hasil Summarization Berita

Hasil Summary (Ringkasan)		
POS Tagging	**NER**	**POS Tags dan NER**
Ketua Umum kelompok relawan Solidaritas Merah Putih (Solmet) Silfester Matutina diperiksa terkait laporan ke Roy Suryo cs di Polres Metro Jakarta Selatan, Rabu . Diketahui, Roy dan tiga orang lainnya dilaporkan Tim Advocate Public Defender dari Peradi Bersatu buntut tudingan ijazah palsu Presiden ke-7 RI, Joko Widodo (Jokowi). Silfester mengungkapkan dalam pemeriksaan itu dirinya ditanya ihwal pernyataan Roy dalam sebuah acara yang menuding ijazah milik Jokowi palsu. Sebelumnya, Tim Advocate Public Defender dariPeradiBersatu melaporkan empat orang ke Polres Metro Jakarta Selatan buntut tudingan ijazah palsu Jokowi.	"Jadi saya diminta datang untuk klarifikasi dan jadi saksi layak untuk pengaduan dari Peradi Bersatu terhadap saudara RS yang menuduh ijazah Pak Jokowi palsu," kata Silfester kepada wartawan. "Pertanyaannya seputar waktu tuduhan saudara RS di salah satu program TV ya untuk Podcast, dibilang intinya bahwa saudara RS menuduh pak Jokowi ijazahnya palsu tapi saudara RS tidak mempunyai bukti-bukti atas tuduhan itu.Waktu itu saya sebagai narasumber di acara itu bersama Saudara RS juga dan ada sekitar 6 narasumber yang lainnya," tutur dia. Sementara itu, Wakil Ketua Peradi Bersatu, Lechumanan menyampaikan pemeriksaan Silfester tersebut guna melengkapi proses penyelidikan yang dilakukan kepolisian. "RS dan kawan-kawan untuk dimintai keterangan sebagai dalam rangka klarifikasi juga memang benar enggak ada peristiwa saat itu, ada enggak peristiwa atau perbuatan yang dilakukan jadi nanti semua keterangan ini dikumpulkan," ucap dia.	Silfester mengungkapkan dalam pemeriksaan itu dirinya ditanya ihwal pernyataan Roy dalam sebuah acara yang menuding ijazah milik Jokowi palsu. Sebelumnya, Tim Advocate Public Defender dari Peradi Bersatu melaporkan empat orang ke Polres Metro Jakarta Selatan buntut tudingan ijazah palsu Jokowi. Total ada empat orang yang menjadi terlapor terkait isu ijazah palsu Jokowi.

3.6. Evaluasi

Hasil summarization teks berita online dari metode NER, POS Tagging, dan gabungan keduanya kemudian akan dilakukan evaluasi menggunakan skor BERT. Hasil evaluasi dapat dilihat pada **Gambar 4.** Gambar ini membandingkan efektivitas dua metode peringkasan teks, yaitu *Part of Speech (POS) Tagging*, *Named Entity Recognition* (NER), dan gabungan kedua metode tersebut berdasarkan metrik BERT yaitu *F1 score, precision*, dan *recall*. Hasil evaluasi menunjukkan bahwa gabungan metode NER dan POS *Tagging* (NER + POS Tagging) menghasilkan performa terbaik dengan F1-score tertinggi (0.7400), diikuti oleh POS Tagging (0.7344) dan NER (0.7090). Meskipun NER memiliki recall tertinggi (0.7661), yang menunjukkan kemampuannya menangkap lebih banyak informasi dari teks sumber, presisinya lebih rendah (0.6602) dibandingkan dua metode lainnya. Hal ini mengindikasikan bahwa NER cenderung memasukkan lebih banyak konten, tetapi dengan risiko noise atau informasi kurang relevan. Sementara itu, POS Tagging mencapai keseimbangan lebih baik antara presisi (0.6999) dan recall (0.7729), menghasilkan F1-score yang lebih tinggi daripada NER.

Kombinasi metode NER dan POS Tagging berhasil menghasilkan presisi tertinggi yaitu (0.7320) dengan recall yang tetap tinggi (0.7486). Ini menunjukkan bahwa pendekatan hybrid mampu memfilter informasi tidak relevan (meningkatkan presisi) sambil mempertahankan cakupan informasi yang luas (recall tetap kompetitif). Dengan demikian, gabungan NER dan POS Tagging merupakan strategi terbaik untuk menghasilkan ringkasan yang akurat, relevan, dan komprehensif dibandingkan penggunaan metode secara terpisah.



Gambar 4. Perbandingan Skor BERT

Perbedaan skor ini mencerminkan kelebihan dan kelemahan masing-masing metode. POS Tagging unggul dalam mempertahankan koherensi dan struktur kalimat, sementara NER cenderung terbatas pada identifikasi entitas tanpa konteks utuh. Namun, keduanya saling melengkapi dimana NER membantu menyaring informasi penting (seperti nama atau lokasi), sedangkan POS Tagging memastikan ringkasan tetap gramatikal dan padat. Temuan ini sejalan dengan penelitian terkini yang merekomendasikan kombinasi kedua teknik untuk meningkatkan kualitas ringkasan otomatis, terutama untuk berita online yang kompleks.

IV. KESIMPULAN

Berdasarkan penelitian ini menunjukkan bahwa hasil evaluasi menggunakan BERTScore, metode gabungan NER dan POS Tagging memberikan performa terbaik dengan skor F1 sebesar 0.740, mengungguli metode POS Tagging saja (0.7344) dan NER saja (0.7090). Hal ini menunjukkan

bahwa integrasi kedua informasi linguistik mampu meningkatkan kualitas semantik ringkasan. Dukungan terhadap hasil ini terlihat dari distribusi tag POS dan NER yakni terdapat dominasi kata benda (NN: 779) dan adjektiva (JJ: 137) dalam struktur kalimat, yang berperan penting dalam menyampaikan makna inti, serta keberadaan entitas PERSON (79), ORG (43), dan GPE (22) yang menunjukkan banyaknya informasi penting yang bersifat entitas. Berdasarkan metode yang digunakan dalam penelitian ini, arah penelitian di masa depan dapat difokuskan pada pengembangan model evaluasi BERT yang lebih disesuaikan dengan domain atau topik tertentu (domain-specific BERT) dapat meningkatkan relevansi semantik ringkasan, terutama untuk teks-teks teknis atau ilmiah. Evaluasi manual tambahan oleh ahli bahasa juga direkomendasikan sebagai pelengkap metrik otomatis seperti BERTScore, agar dapat memperoleh gambaran yang lebih holistik terkait kualitas semantik ringkasan. Dengan demikian, pemanfaatan informasi gramatikal (POS Tagging) dan entitas bernama (NER) secara bersamaan memberikan kontribusi yang signifikan terhadap pemahaman dan peringkasan isi dokumen secara semantik.

UCAPAN TERIMA KASIH

Penulis mengucapkan terimakasih kepada Regita Putri Permata, S.Stat., M.Stat. selaku dosen pembimbing kami yang membantu dalam penulisan artikel dan pihak Universitas Telkom Surabaya sebagai universitas yang mendukung penulisan artikel penelitian ini, serta semua pihak yang membantu

REFERENSI

[1] A. S. M. Romli, *Jurnalistik online: Panduan mengelola media online*. Nuansa Cendekia, 2018.

[2] T. He, F. Li, W. Shao, J. Chen, and L. Ma, "A new feature-fusion sentence selecting strategy for query-focused multi-document summarization," *Proceedings - ALPIT 2008, 7th International Conference on Advanced Language Processing and Web Information Technology*, pp. 81–86, 2008, doi: 10.1109/ALPIT.2008.45.

[3] Roza Yuniza Putri, Widodo, and Hamidillah Ajie, "PERBANDINGAN PERINGKASAN MULTI DOKUMEN ILMIAH BERBAHASA INDONESIA MENGGUNAKAN METODE K-MEANS DAN K-NEAREST NEIGHBORS (K-NN)," *PINTER : Jurnal Pendidikan Teknik Informatika dan Komputer*, vol. 7, no. 2, pp. 19–27, Dec. 2023, doi: 10.21009/pinter.7.2.3.

[4] M. Susanty and S. Sukardi, "Perbandingan Pre-trained Word Embedding dan Embedding Layer untuk Named-Entity Recognition Bahasa Indonesia," *PETIR*, vol. 14, no. 2, pp. 247–257, Sep. 2021, doi: 10.33322/petir.v14i2.1164.

[5] M. E. Khademi and M. Fakhredanesh, "Persian Automatic Text Summarization Based on Named Entity Recognition," *Iranian Journal of Science and Technology - Transactions of Electrical Engineering*, pp. 1–12, Jul. 2020, doi: 10.1007/S40998-020-00352-2/METRICS.

[6] S. Maesaroh *et al.*, *Pembelajaran Mesin dan Kecerdasan Buatan: Teori dan Aplikasi Praktis*. Sada Kurnia Pustaka, 2024.

[7] A. P. Widyassari *et al.*, "Review of automatic text summarization techniques & methods," Apr. 01, 2022, *King Saud bin Abdulaziz University*. doi: 10.1016/j.jksuci.2020.05.006.

[8] I. IKHSAN DWI SAPUTRA, "Aplikasi Web Question Answering menggunakan Langchain OpenAI tentang Peraturan Perundang-undangan Bidang Pendidikan," *Aplikasi Web Question Answering Menggunakan Langchain OpenAI Tentang Peraturan Perundang-Undangan Bidang Pendidikan*, vol. 6, no. 1, pp. 293–306, 2024.

[9] I. Z. Khan, A. A. Sheikh, and U. Sinha, "Graph Neural Network and NER-Based Text Summarization," Feb. 2024, [Online]. Available: http://arxiv.org/abs/2402.05126

[10] C. P. Chai, "Comparison of text preprocessing methods," *Nat Lang Eng*, vol. 29, no. 3, pp. 509–553, May 2023, doi: 10.1017/S1351324922000213.

[11] M. García, S. Maldonado, and C. Vairetti, "Efficient n-gram construction for text categorization using feature selection techniques," *Intelligent Data Analysis*, vol. 25, no. 3, pp. 509–525, 2021, doi: 10.3233/IDA-205154.

[12] M. Alfian, U. L. Yuhana, and D. Siahaan, "Indonesian Part-of-Speech Tagger: A Comparative Study," *2023 10th International Conference on Advanced Informatics: Concept, Theory and Application, ICAICTA 2023*, 2023, doi: 10.1109/ICAICTA59291.2023.10390353.

[13] N. G. Yudiarta, M. Sudarma, and W. G. Ariastina, "Penerapan Metode Clustering Text Mining Untuk Pengelompokan Berita Pada Unstructured Textual Data," *Majalah Ilmiah Teknologi Elektro*, vol. 17, no. 3, p. 339, Dec. 2018, doi: 10.24843/MITE.2018.v17i03.P06.

[14] Z. Nasar, S. W. Jaffry, and M. K. Malik, "Named Entity Recognition and Relation Extraction: State-of-The-Art," *ACM Comput Surv*, vol. 54, no. 1, Jan. 2021, doi: 10.1145/3445965;JOURNAL:JOURNAL:CSUR;WGROUP:STRING:ACM.

[15] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pp. 3982–3992, Aug. 2019, doi: 10.18653/v1/d19-1410.

[16] T. Zhang, V. Kishore, F. Wu, K. Weinberger, and Y. Artzi, *BERTScore: Evaluating Text Generation with BERT*. 2019. doi: 10.48550/arXiv.1904.09675.

[17] T. Zhang, D. P. Chandrasekaran, F. Thung, and D. Lo, "Benchmarking library recognition in tweets," in *Proceedings of the 30th IEEE/ACM International Conference on Program Comprehension*, New York, NY, USA: ACM, May 2022, pp. 343–353. doi: 10.1145/3524610.3527916.