Penerapan Named Entity Recognition (NER) Untuk Ekstraksi Otomatis Entitas Pada Teks Berita Online Pemerintah Daerah

Rama Otari[1], Elvi Rahmi[2],

[1,2] *Program Studi Rekayasa Perangkat Lunak - Politeknik Negeri Bengkalis*

Jl. Bathin Alam, Desa Sungai Alam, Kecamatan Bengkalis, Kabupaten Bengkalis, Provinsi Riau, 28711

ramaotari28@gmail.com, elvirahmi@polbeng.ac.id

***Abstrak*— Berita pemerintah daerah memiliki peran penting sebagai sumber informasi publik, namun umumnya masih disajikan dalam bentuk teks tidak terstruktur sehingga menyulitkan proses pencarian dan analisis informasi. Penelitian ini bertujuan menerapkan metode *Named Entity Recognition* (NER) untuk mengekstraksi entitas penting sekaligus membangun sistem berbasis web yang mengintegrasikan proses scraping berita, prapemrosesan teks, ekstraksi entitas, dan penyajian hasil secara terstruktur. Metode yang digunakan memanfaatkan model *spaCy pretrained* berbasis *multilingual* dan *IndoBERT pretrained* untuk mengenali entitas PER, ORG, LOC, dan DATE pada berita Detik.com. Sistem dikembangkan menggunakan *framework Flask* dan dievaluasi dengan membandingkan hasil ekstraksi terhadap data *ground truth* menggunakan metrik *precision*, *recall*, dan F1- *score*. Hasil pengujian menunjukkan nilai *precision* sebesar 67%, *recall* sebesar 76,2%, dan F1-*score* sebesar 71,2%. Kinerja terbaik diperoleh pada entitas PER dan DATE, sementara entitas LOC dan ORG masih menunjukkan hasil yang kurang optimal. Secara keseluruhan, sistem ini dapat dimanfaatkan sebagai alat bantu awal dalam analisis berita pemerintah daerah secara lebih efisien dan sistematis.**

***Kata Kunci*— *Named Entity Recognition*, *spaCy pretrained*, *IndoBERT pretrained*, Berita Pemerintah Daerah, Ekstraksi Entitas.**

Abstract*— *Local government news is an important source of public information, but it is generally presented in unstructured text, making it difficult to search and analyze. This study aims to apply Named Entity Recognition (NER) to extract key entities and develop a web-based system that integrates news scraping, text preprocessing, entity extraction, and structured result presentation. The method utilizes a multilingual pretrained spaCy model and an IndoBERT pretrained model to recognize entities such as PER (person), ORG (organization), LOC (location), and DATE in news from Detik.com. The system was developed using the Flask framework and evaluated by comparing extraction results with ground truth data using precision, recall, and F1-score metrics. The evaluation results show a precision of 67%, recall of 76.2%, and an F1-score of 71.2%. The best performance was achieved for PER and DATE entities, while LOC and ORG entities showed less optimal results. Overall, this system can be used as an initial tool to support the analysis of local government news more efficiently and systematically.

Keywords*— *Named Entity Recognition, spaCy pretrained, IndoBERT pretrained, Local Government News, Entity Extraction

I. PENDAHULUAN

Perkembangan teknologi informasi mendorong pemerintah daerah untuk memanfaatkan media digital sebagai sarana penyampaian informasi kepada publik, salah satunya melalui berita online. Situs berita nasional seperti **Detik.com** secara rutin memuat berbagai informasi terkait aktivitas dan kebijakan pemerintahan daerah, mulai dari pelantikan pejabat, rapat DPRD, hingga penggunaan anggaran. Namun, informasi tersebut umumnya masih disajikan dalam bentuk narasi teks yang panjang dan tidak terstruktur, sehingga menyulitkan proses pencarian dan identifikasi informasi tertentu secara otomatis, seperti nama pejabat yang disebutkan, organisasi yang terlibat, lokasi yang tercantum, serta waktu yang tertulis dalam berita. Oleh karena itu, penerapan metode *Named Entity Recognition* (NER) menjadi penting untuk mengekstraksi entitas-entitas utama secara otomatis, guna mendukung proses pengolahan informasi dari teks berita pemerintah daerah secara lebih efisien dan terstruktur[1].

Untuk menyelesaikan permasalahan ini, salah satu pendekatan yang bisa digunakan adalah *Named Entity Recognition* (NER). NER merupakan bagian dari teknologi *Natural Language Processing* (NLP) yang bertugas mengenali informasi penting dalam teks, seperti nama orang, organisasi, lokasi, dan waktu. Teknologi ini sudah banyak diterapkan dalam berbagai bidang seperti berita online, media sosial, dan jurnal ilmiah. Namun, penerapan NER secara spesifik untuk berita pemerintah daerah, terutama untuk entitas penting dan umum seperti nama orang (PER), organisasi (ORG), lokasi (LOC) serta tanggal (DATE) yang memang belum banyak dikembangkan[2].

Named Entity Recognition (NER) akan diuji untuk mengetahui seberapa efektif sistem dalam mengenali entitas penting dari teks berita pemerintah daerah. Pengujian ini dilakukan dengan menggunakan tiga metrik evaluasi utama, yaitu *precision* (ketepatan), *recall* (kelengkapan), dan F1-*score* (gabungan keduanya). *Precision* mengukur ketepatan entitas yang dikenali sistem, *recall* menunjukkan seberapa banyak entitas yang berhasil ditemukan, dan F1-*score* memberikan gambaran keseluruhan performa sistem. Melalui evaluasi ini, efektivitas model dalam mengekstraksi informasi penting dari teks tidak terstruktur dapat dinilai secara objektif. Metrik *accuracy* tidak digunakan dalam evaluasi ini karena *tugas Named Entity Recognition* (NER) memiliki karakteristik data yang tidak seimbang, di mana sebagian besar token dalam teks bukan merupakan entitas. Kondisi tersebut dapat menyebabkan nilai *accuracy* menjadi tinggi meskipun sistem belum mampu mengenali entitas penting secara optimal[3].

Beberapa pendekatan teknis telah dikembangkan untuk membangun sistem NER, di antaranya adalah *spaCy* dan *IndoBERT pretrained*. *spaCy* adalah salah satu *library* NLP yang banyak digunakan karena bersifat ringan, modular, dan memiliki performa tinggi dalam ekstraksi entitas pada tugas *Named Entity Recognition* (NER)[4][5]. Di sisi lain, IndoBERT pretrained merupakan model *deep learning* berbasis *transformer* yang sudah di latih sebelumnya untuk Bahasa Indonesia. Model ini dikembangkan berdasarkan arsitektur BERT dan telah menunjukkan hasil yang sangat baik dalam tugas-tugas NLP di Indonesia[6].

Meskipun *spaCy* cepat dan efisien, library ini kurang optimal dalam mengenali entitas yang spesifik seperti nama orang (PERSON) atau jabatan di pemerintah daerah, nama instansi atau proyek pemerintah daerah (ORG/LOC), terutama jika muncul dalam format tidak standar. Untuk mengatasi hal ini, *IndoBERT* digunakan sebagai pelengkap karena mampu memahami konteks kalimat secara lebih mendalam dan mengenali entitas yang kompleks. Kombinasi kedua pendekatan ini menjadikan sistem NER bekerja lebih lengkap dan cepat, memanfaatkan kecepatan *spaCy* untuk entitas standar sekaligus kekuatan *IndoBERT* untuk entitas spesifik berbahasa indonesia, sehingga hasil ekstraksi entitas menjadi lebih informatif untuk kebutuhan pengguna. Penelitian[6] membuktikan bahwa *IndoBERT Pretrained* mampu mengungguli model-model klasik pada berbagai benchmark Bahasa Indonesia. Selain itu, studi oleh [5] yang menggunakan *spaCy* menunjukkan bahwa model ini mampu mencapai F1-*score* 94.38% dalam ekstraksi entitas dari teks media sosial. Hal ini menunjukkan bahwa baik *spaCy* maupun IndoBERT memiliki potensi besar dalam penerapan NER untuk Bahasa Indonesia.

Penelitian ini bertujuan untuk mengembangkan serta mengevaluasi pendekatan *Named Entity Recognition* (NER) dalam mengekstraksi entitas penting secara otomatis dari teks berita pemerintah daerah berbahasa Indonesia. Pendekatan ini menggabungkan keunggulan *library spaCy* serta *IndoBERT pretrained* sebagai model deep learning yang telah dilatih sebelumnya untuk Bahasa Indonesia, guna memahami kemampuan model dalam mengekstraksi entitas dari teks berita. Untuk mendukung kemudahan penggunaan, sistem ini dilengkapi dengan antarmuka berbasis web sebagai alat bantu pengguna dalam melakukan proses ekstraksi secara praktis dan terstruktur.

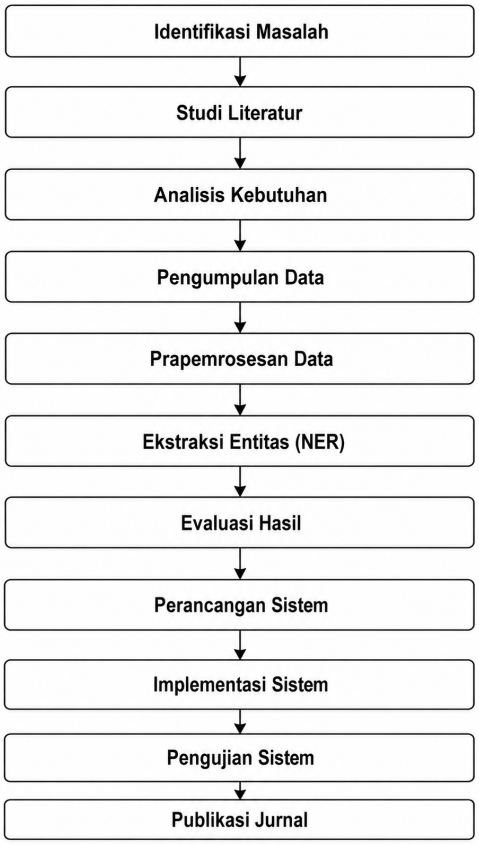
Hasil dari penelitian ini dapat memberikan solusi terhadap tantangan pengolahan informasi tidak terstruktur pada dokumen berita pemerintah, serta bermanfaat dalam mendukung otomatisasi pencarian dan pemanfaatan data secara efisien.

II. METODE PENELITIAN

Metode penelitian yang digunakan adalah model *Waterfall* karena proses pengembangan sistem dilakukan secara sistematis dan berurutan, mulai dari tahap perencanaan, analisis, perancangan, implementasi, hingga pengujian. Model ini dipilih karena sesuai untuk pengembangan sistem yang memiliki tahapan kerja yang jelas dan terstruktur sehingga memudahkan proses pembangunan sistem secara terarah[3].

A. Prosedur Penelitian

Tahapan penelitian yang dilakukan peneliti secara keseluruhan di tunjukkan pada gambar 1 berikut:



Gbr 1 prosedur penelitian

Tahapan pertama dalam penelitian ini adalah identifikasi masalah. Pada tahap ini ditemukan bahwa informasi penting

pada berita pemerintah daerah masih sulit dicari secara cepat karena data disajikan dalam bentuk teks tidak terstruktur. Pengguna harus membaca keseluruhan isi berita untuk menemukan informasi seperti nama pejabat, instansi, lokasi kegiatan, maupun tanggal pelaksanaan kegiatan. Permasalahan tersebut menyebabkan proses pencarian informasi menjadi kurang efisien.

Tahap berikutnya adalah studi literatur. Pada tahap ini peneliti mempelajari berbagai teori dan penelitian terdahulu yang berkaitan dengan *Natural Language Processing* (NLP), *Named Entity Recognition* (NER), *web scraping*, *framework Flask, library spaCy,* serta model *IndoBERT pretrained*. Studi literatur dilakukan melalui jurnal, artikel ilmiah, buku, dan dokumentasi resmi yang relevan guna mendukung proses penelitian dan pengembangan sistem.

Selanjutnya dilakukan analisis kebutuhan sistem yang terdiri dari kebutuhan fungsional dan kebutuhan nonfungsional. Kebutuhan fungsional meliputi kemampuan sistem dalam mengambil data berita secara otomatis, melakukan *preprocessing* teks, mengekstraksi entitas, menampilkan hasil ekstraksi, dan menyimpan data ke dalam file CSV/PDF. Sedangkan kebutuhan nonfungsional meliputi kebutuhan perangkat keras dan perangkat lunak seperti penggunaan bahasa pemrograman *Python*, *Visual Studio Code* sebagai code editor, *framework Flask,* serta *library* pendukung seperti *BeautifulSoup, Requests, spaCy,* dan *Transformers*.

Tahap selanjutnya adalah pengumpulan data berita menggunakan teknik *web scraping*. Proses *scraping* dilakukan pada situs berita nasional yaitu **Detik.com** dengan mengambil berita yang berkaitan dengan pemerintahan daerah berdasarkan kata kunci tertentu seperti “bupati”, “pemkab”, “sekda”, dan “gubernur”. Data yang diambil meliputi judul berita, isi berita, dan URL berita. Hasil *scraping* kemudian disimpan dalam format CSV sebagai dataset penelitian.

Setelah data diperoleh, dilakukan tahap *preprocessing* atau prapemrosesan teks. Tahap ini bertujuan untuk membersihkan data dari karakter yang tidak diperlukan seperti simbol, tag HTML, spasi berlebih, dan karakter khusus lainnya agar teks lebih mudah diproses oleh model NER.

Tahap berikutnya adalah proses ekstraksi entitas menggunakan metode *Named Entity Recognition* (NER). Pada penelitian ini digunakan kombinasi *library spaCy* dan model *IndoBERT pretrained*. *spaCy* digunakan untuk proses pengolahan teks dan pendeteksian entitas dasar, sedangkan *IndoBERT* digunakan untuk meningkatkan kemampuan identifikasi entitas berbahasa Indonesia. Entitas yang diekstraksi meliputi *PERSON*, *ORGANIZATION*, *LOCATION*, dan *DATE*. Hasil ekstraksi kemudian disimpan secara otomatis ke dalam file CSV agar dapat digunakan untuk proses analisis maupun dokumentasi data.

Tahap berikutnya adalah evaluasi hasil. Pada tahap ini dilakukan pengujian terhadap hasil ekstraksi entitas menggunakan metrik *precision, recall,* dan F1-*score* untuk mengetahui tingkat akurasi model dalam mengenali entitas pada berita.

Tahap selanjutnya adalah perancangan sistem. Pada tahap ini dilakukan perancangan antarmuka dan alur kerja sistem menggunakan *use case diagram* dan *Activity Diagram*. Sistem dirancang berbasis web menggunakan *framework Flask* sehingga pengguna dapat mengakses sistem melalui browser. Antarmuka sistem terdiri dari halaman input URL berita, proses ekstraksi entitas, dan tampilan hasil ekstraksi. Selain itu dirancang juga proses penyimpanan data hasil ekstraksi ke dalam database atau file PDF.

Setelah proses perancangan selesai, dilakukan tahap implementasi sistem. Implementasi dilakukan menggunakan bahasa pemrograman *Python* dengan *framework Flask* sebagai *backend* sistem. Proses *scraping* berita menggunakan *library Requests* dan *BeautifulSoup*, sedangkan proses ekstraksi entitas menggunakan *spaCy* dan *IndoBERT pretrained*. Sistem kemudian dijalankan menggunakan *Visual Studio Code* sebagai lingkungan pengembangan aplikasi.

Tahap berikutnya adalah pengujian sistem. Pengujian dilakukan menggunakan metode *Black Box Testing* untuk memastikan seluruh fungsi sistem berjalan dengan baik sesuai kebutuhan pengguna[7]. Pengujian dilakukan pada fitur input URL berita, proses *scraping, preprocessing* teks, ekstraksi entitas, dan penyimpanan hasil ke file PDF.

Tahap terakhir adalah dokumentasi dan publikasi hasil penelitian dalam bentuk jurnal ilmiah. Tahap ini dilakukan sebagai bentuk penyebaran hasil penelitian dan kontribusi dalam pengembangan ilmu pengetahuan khususnya pada bidang *Natural Language Processing* (NLP) dan ekstraksi informasi dari berita pemerintah daerah.

III. HASIL DAN PEMBAHASAN

A. Hasil

Bagian ini membahas hasil penerapan sistem Named Entity Recognition (NER) untuk mengekstraksi entitas penting dari teks berita online pemerintah daerah serta analisis terhadap performa sistem berdasarkan hasil pengujian yang telah dilakukan. Entitas yang menjadi fokus dalam penelitian ini meliputi nama orang (*person*), organisasi (*organization*), lokasi (*location*), dan tanggal (*date*). Proses pengujian dilakukan menggunakan data berita dari situs **Detik.com** melalui tahap scraping, prapemrosesan, serta anotasi manual sebagai data pembanding (*ground truth*).

1) Pengumpulan Data

Pengumpulan data dalam penelitian ini dilakukan secara otomatis melalui teknik web *scraping* pada situs berita Detik.com menggunakan bahasa pemrograman *Python* dengan bantuan pustaka requests dan BeautifulSoup untuk mengambil serta mengekstrak konten artikel. Peneliti menggunakan 10 kata kunci yang sering muncul dalam konteks berita pemerintah daerah, yaitu "bupati", "gubernur", "dprd", "apbd", "pemerintah daerah", "sekda", "musrenbang", "perda", "kabupaten", "provinsi"[8]. Kata kunci ini dipilih karena berkaitan langsung dengan entitas yang menjadi fokus ekstraksi, seperti nama pejabat (PER), lembaga pemerintahan (ORG), wilayah administratif (LOC), serta tanggal kegiatan

(DATE). Keempat label ini dipilih karena mengacu pada skema anotasi *Named Entity Recognition* yang digunakan dalam korpus [9] yang merupakan skema standar dalam penelitian NER. yang merupakan dataset standar untuk NER Bahasa Indonesia. Dari masing-masing kata kunci diambil 5 artikel berita yang paling relevan, sehingga terkumpul total 50 artikel yang digunakan sebagai *data uji.*

Evaluasi dilakukan dengan membandingkan hasil prediksi sistem terhadap label manual (*ground truth*), kemudian dihitung metrik *precision*, *recall*, dan F1- *score*. Meskipun jumlah artikel relatif sedikit, setiap artikel mengandung sejumlah entitas yang memadai untuk setiap kategori label, sehingga hasil evaluasi tetap representatif. Hal ini sejalan dengan temuan [10] yang menyatakan bahwa evaluasi NER dapat dilakukan secara valid meskipun data terbatas, asalkan setiap kategori entitas memiliki jumlah contoh yang cukup untuk dianalisis secara kuantitatif.

1. Prapemrosesan data

Setelah data berita dikumpulkan melalui proses scraping, tahap selanjutnya adalah preprocessing untuk membersihkan teks agar siap dianalisis oleh sistem *Named Entity Recognition* (NER). Pada tahap ini dilakukan penghapusan elemen HTML, URL, kalimat promosi, entity HTML, serta simbol yang tidak memiliki makna linguistik. Beberapa tanda baca seperti titik, koma, tanda hubung, dan garis miring tetap dipertahankan karena sering digunakan pada nama instansi atau jabatan. Selain itu, spasi berlebih juga dirapikan agar teks menjadi lebih konsisten. Tahap preprocessing tidak mencakup penghapusan *stopword* maupun stemming karena dapat mengubah bentuk asli entitas yang ingin dikenali oleh model.

2. Analisis Data/Evaluasi

Evaluasi hasil ekstraksi entitas dilakukan untuk mengetahui kemampuan model dalam mengenali dan mengklasifikasikan entitas pada teks berita pemerintah daerah. Evaluasi dilakukan dengan membandingkan hasil prediksi sistem terhadap data ground truth yang telah dianotasi secara manual. Entitas yang diuji meliputi PER, ORG, LOC, dan DATE. . Proses evaluasi dilakukan dengan menghitung metrik-metrik klasifikasi umum seperti:

1. TP (*True Positive*): Entitas yang sebenarnya ada dalam teks dan diprediksi dengan benar oleh model NER (sesuai dengan label yang sebenarnya).

2. FP *(False Positive*): Entitas yang sebenarnya ada dalam teks tetapi salah diprediksi oleh model NER (tidak sesuai dengan label yang sebenarnya).

3. FN (*False Negative*): Entitas yang sebenarnya ada dalam teks tetapi tidak diprediksi oleh model NER (tidak terdeteksi).

Berikut adalah tabel 1 rekapitulasi data hasil evaluasi:

TABEL I
HASIL PERHITUNGAN EVALUASI NER

Entitas	TP	FP	FN
PERSON	61	3	0
LOC	47	46	17
ORG	26	46	25
DATE	59	0	2
TOTAL	**193**	**95**	**42**

$$Precision = \frac{TP}{TP + FP} = \frac{193}{193 + 95} = \frac{193}{288} \text{ x } 100\% = 67\%$$

$$Recall = \frac{TP}{TP + FN} = \frac{193}{193 + 42} = \frac{193}{235} \text{ x } 100\% = 76{,}2\%$$

$$f1 - score = 2\,x\ \frac{Precision \text{ x } Recall}{Precision + Recall} = 2 \text{ x } \frac{67 \text{ x } 76{,}2}{67 + 76{,}2}$$

$$= \frac{\mathbf{10.208,4}}{\mathbf{143.2}} = 71.2\%$$

Berdasarkan hasil evaluasi ekstraksi entitas pada tabel 1, model menunjukkan kinerja yang cukup baik dengan tingkat ketepatan dan kelengkapan yang relatif seimbang. Entitas PERSON dan DATE memiliki hasil paling optimal karena mampu dikenali dengan akurat dan memiliki tingkat kesalahan yang rendah. Sementara itu, entitas LOC dan ORG masih mengalami beberapa kesalahan klasifikasi dan terdapat entitas yang tidak terdeteksi, terutama pada istilah administratif dan nama instansi pemerintahan yang kompleks. Secara keseluruhan, nilai *precision* sebesar 67%, *recall* 76,2%, dan F1-*score* 71,2% menunjukkan bahwa model memiliki performa yang cukup baik dalam mengekstraksi entitas dari teks berita pemerintah daerah, meskipun masih perlu peningkatan pada pengenalan entitas lokasi dan organisasi.

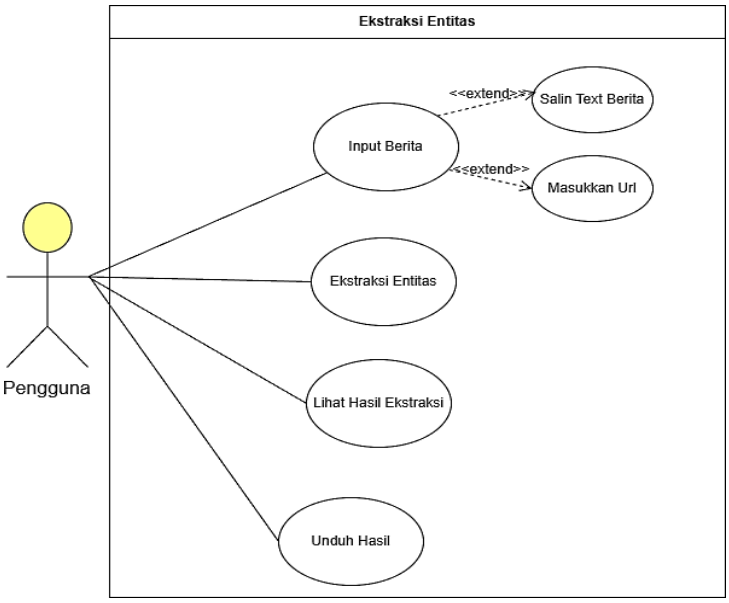
3. Perancangan

Tahap selanjutnya adalah perancangan sistem. Pada tahap ini dilakukan perancangan antarmuka dan alur kerja sistem menggunakan *use case diagram* dan *Activity Diagram*.

1. Use Case Diagram

Use Case Diagram merupakan salah satu diagram dalam *Unified Modeling Language* (UML) yang digunakan untuk menggambarkan interaksi antara pengguna (aktor) dengan sistem berdasarkan fungsi atau layanan (*use case*) yang tersedia pada sistem. Diagram ini membantu menjelaskan bagaimana pengguna menggunakan sistem serta proses apa saja yang dapat dilakukan di dalam aplikasi. Pada penelitian ini, *Use Case Diagram* digunakan untuk menggambarkan hubungan antara pengguna dengan sistem ekstraksi entitas

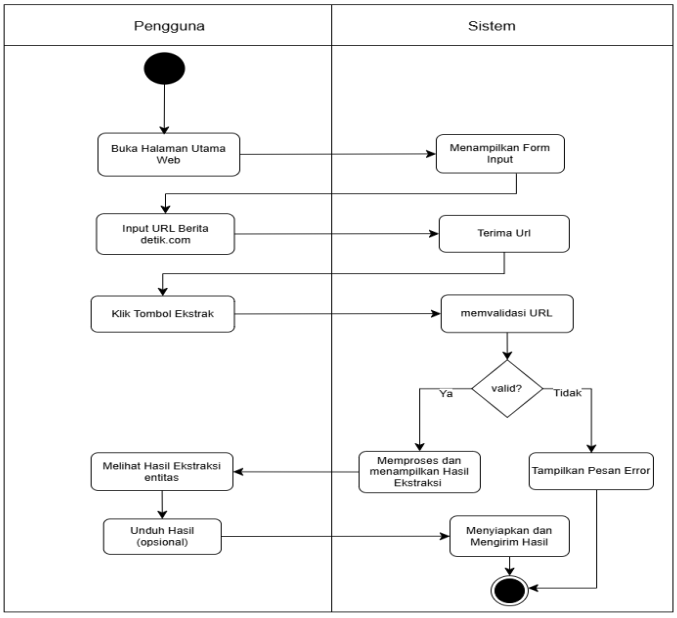
berita, seperti proses memasukkan URL berita, melakukan scraping data, menjalankan ekstraksi entitas, melihat hasil ekstraksi, dan menyimpan hasil ke dalam file PDF. Dengan adanya *Use Case Diagram*, alur kerja sistem menjadi lebih mudah dipahami sehingga memudahkan proses perancangan dan pengembangan sistem .



Gbr 2 *use case diagram*

2. *Activity Diagram*

Activity Diagram adalah salah satu jenis diagram UML (*Unified Modeling Language*) yang digunakan untuk menggambarkan alur aktivitas atau proses dalam suatu sistem, baik yang dilakukan oleh pengguna maupun sistem secara otomatis.

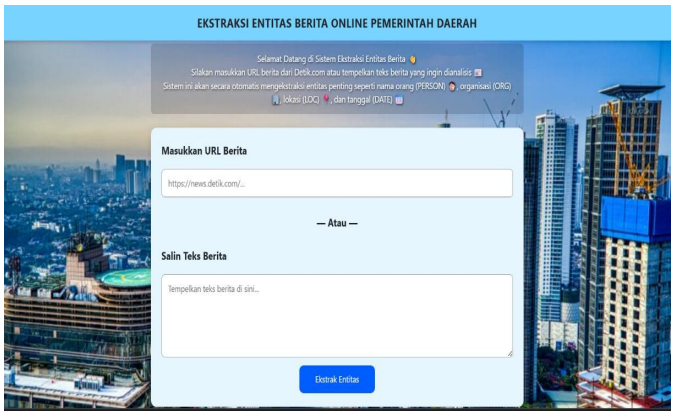


Gbr 3 *Activity Diagram*

4. *Implementasi Sistem*

Implementasi sistem dilakukan menggunakan *framework Flask*, yaitu *framework* web berbasis *Python* yang ringan, fleksibel, dan mudah digunakan dalam pengembangan aplikasi berbasis web. Sistem ini dirancang untuk melakukan proses ekstraksi entitas secara otomatis pada teks berita pemerintah daerah menggunakan metode *Named Entity Recognition* (NER) dengan kombinasi model *spaCy* dan *IndoBERT pretrained*. Selain itu, sistem juga dilengkapi dengan proses *preprocessing* teks untuk membersihkan karakter yang tidak relevan sebelum dilakukan ekstraksi entitas. Berikut merupakan tampilan sistem *website* ekstraksi entitas NER yang telah dibangun, serta pengguna juga bisa mengunduh hasil dalam bentuk *PDF*.

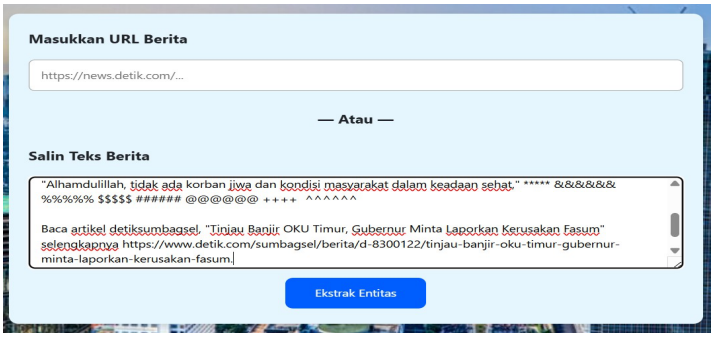
a. Tampilan *Home Page*



Gbr 4 *Home Page*

Berdasarkn gambar 4 diatas, *Flask* digunakan untuk membangun antarmuka *web* sederhana yang memungkinkan pengguna memasukkan teks berita atau pengumuman secara langsung.

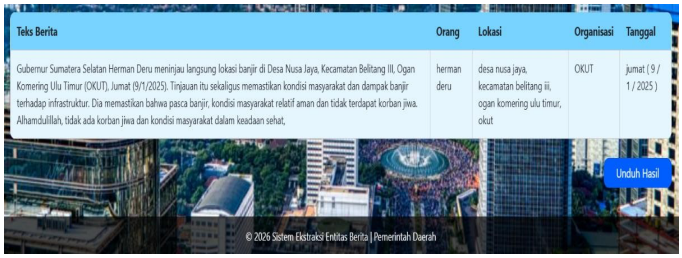
b. Tampilan Hasil Ekstraksi Entitas



Gbr 5 Mengekstrak Entitas

Pada gambar 5 di atas, Pengguna dapat memasukkan URL berita dari situs Detik.com atau menyalin langsung teks berita ke dalam kolom yang telah disediakan pada sistem. Setelah data dimasukkan, sistem akan melakukan proses preprocessing teks secara otomatis, seperti membersihkan

simbol, karakter tidak relevan, tag HTML, spasi berlebih, serta tanda baca yang tidak diperlukan agar teks menjadi lebih terstruktur dan mudah diproses. Tahap ini dilakukan untuk meningkatkan kualitas data sebelum proses ekstraksi entitas menggunakan metode *Named Entity Recognition* (NER) dilakukan oleh sistem.



Gbr 7 Hasil Ekstraksi Entitas

Selanjutnya pada gambar 7, sistem menggunakan kombinasi model *spaCy* dan *IndoBERT pretrained* untuk mengenali entitas penting seperti orang, lokasi, organisasi, dan tanggal secara otomatis setelah tombol "Ekstrak Entitas" ditekan .

IV. KESIMPULAN DAN SARAN

A. Kesimpulan

Berdasarkan hasil penelitian yang dilakukan, penerapan *Named Entity Recognition* (NER) pada teks berita pemerintah daerah berbahasa Indonesia dengan kombinasi *spaCy* dan *IndoBERT pretrained* berhasil diimplementasikan untuk mengekstraksi entitas penting dari teks tidak terstruktur. Sistem ini mampu mengenali empat jenis entitas utama, yaitu *PERSON*, *ORGANIZATION*, *LOCATION*, dan *DATE*, dari data berita yang diperoleh melalui proses web scraping. Hasil evaluasi menggunakan metrik *precision, recall*, dan F1-*score* menunjukkan kinerja yang cukup baik, dengan ***precision* sekitar 67%**, ***recall* 76,2%**, dan **F1-*score* 71,2%.** Dari hasil tersebut, entitas *PERSON* dan *DATE* memiliki performa paling optimal karena mampu dikenali dengan tingkat akurasi yang tinggi, sedangkan entitas *LOCATION* dan *ORGANIZATION* masih menjadi tantangan akibat kompleksitas penamaan wilayah dan instansi pemerintahan yang beragam. Selain itu, sistem berbasis web yang dibangun berhasil mengintegrasikan seluruh proses mulai dari scraping, prapemrosesan teks, ekstraksi entitas, hingga penyajian hasil secara terstruktur. Namun demikian, akurasi sistem masih belum optimal karena model yang digunakan masih berbasis *pretrained spaCy* dan *IndoBERT* tanpa *fine-tuning* khusus pada domain berita pemerintah daerah, sehingga sistem ini lebih tepat digunakan sebagai alat bantu analisis awal dalam pengolahan data berita.

B. Saran

Secara keseluruhan, penelitian ini membuktikan bahwa kombinasi *spaCy* dan *IndoBERT pretrained* mampu memberikan performa yang cukup baik dalam tugas ekstraksi entitas pada domain berita pemerintah daerah, khusunya untuk entitas orang dan tanggal. Disarankan agar penelitian selanjutnya melakukan penyesuaian model *Named Entity Recognition (NER)* melalui proses *fine-tuning* menggunakan data latih yang spesifik pada domain berita pemerintah daerah. Selain itu, penambahan data latih yang lebih beragam serta penerapan aturan tambahan (*rule- based filtering*) atau kamus khusus entitas pemerintahan diharapkan dapat meningkatkan ketepatan pengenalan entitas lokasi dan organisasi, sehingga sistem mampu menghasilkan ekstraksi entitas yang lebih akurat dan konsisten.

V. DAFTAR PUSTAKA

[1] N. U. Muchtar and U. M. Jember, "Analisis Kinerja Transformer Untuk Named Entity Recognition (NER) Menggunakan Indobert Pada Transkrip Video Politik Berbahasa Indonesia," vol. 10, no. October, pp. 180–199, 2025.

[2] I. Budi and R. R. Suryono, "Application of named entity recognition method for Indonesian datasets: a review," *Bull. Electr. Eng. Informatics*, vol. 12, no. 2, pp. 969–978, 2023, doi: 10.11591/eei.v12i2.4529.

[3] A. Gunawan, M. Wanda, and R. Meiyanti, "Named Entity Recognition Pada Teks Berbahasa Indonesia Menggunakan CNNs", *Prosiding Seminar Nasional Teknologi Informasi dan Teknik Informatika (SENASTIKA)*, vol. 1, no. 1, 2024.

[4] X. Schmitt, S. Kubler, E. Xavierschmittstudentunilu, J. Robert, M. Papadakis, and Y. Letraon, "A Replicable Comparison Study of NER Software :," *2019 Sixth Int. Conf. Soc. Networks Anal. Manag. Secur.*, pp. 338–343, 2019.

[5] R. M. Yanti, I. Santoso, and L. H. Suadaa, "Application of Named Entity Recognition via Twitter on SpaCy in Indonesian (Case Study : Power Failure in the Special Region of Yogyakarta)," *Indones. J. Inf. Syst.*, vol. 4, no. 1, pp. 76–86, 2021, doi: 10.24002/ijis.v4i1.4677.

[6] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," *COLING 2020 - 28th Int. Conf. Comput. Linguist. Proc. Conf.*, pp. 757–770, 2020, doi: 10.18653/v1/2020.coling-main.66.

[7] M. S. Mustaqbal, R. F. Firdaus, and H. Rahmadi, "(Studi Kasus : Aplikasi Prediksi Kelulusan SNMPTN)," vol. I, no. 3, pp. 31–36, 2015.

[8] JDIH BPK RI, "Undang-Undang Republik Indonesia tentang Pemerintahan Daerah (UU RI Nomor 23 Tahun 2014)," *Undang-Undang*, no. 1–311, pp. 1–311, 2014.

[9] R. Ljung, "Hemophilia and prophylaxis," *Pediatr. Blood Cancer*, vol. 60, no. SUPPL.1, 2013, doi: 10.1002/pbc.24340.

[10] A. Fritzler, V. Logacheva, and M. Kretov, “Few-shot classification in named entity recognition task,” *Proc. ACM Symp. Appl. Comput.*, vol. Part F1477, pp. 993–1000, 2019, doi: 10.1145/3297280.3297378.