

BAB I

PENDAHULUAN

1.1 Latar Belakang

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.

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.

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.

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 menggunakan pendekatan berbasis aturan dan statistik. 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. 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 [1] membuktikan bahwa *IndoBERT Pretrained* mampu mengungguli model-model klasik pada berbagai *benchmark* Bahasa Indonesia. Selain itu, studi oleh [2] 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.

1.2 Rumusan Masalah

Berdasarkan latar belakang tersebut, rumusan masalah dalam penelitian ini dapat dirangkum sebagai berikut: penelitian ini berfokus pada bagaimana menerapkan pendekatan *Named Entity Recognition* (NER) dengan memanfaatkan kombinasi model *spaCy* dan *IndoBERT pretrained* untuk mengekstraksi secara otomatis entitas penting maupun entitas umum dari teks berita pemerintah daerah yang bersifat tidak terstruktur. Selain itu, penelitian ini juga mengkaji sejauh mana kinerja dan efektivitas sistem yang dibangun dalam melakukan proses ekstraksi entitas berdasarkan metrik evaluasi *precision*, *recall*, dan *F1-score*. Selanjutnya, penelitian ini menyoroti bagaimana merancang sebuah sistem berbasis web sederhana yang mampu menampilkan hasil ekstraksi entitas tersebut secara otomatis, terstruktur, dan mudah diakses sebagai alat bantu analisis informasi dari berita *online* pemerintah daerah.

1.3 Batasan Masalah

Adapun batasan masalah dari penelitian ini hanya berfokus pada sejauh mana hasil ekstraksi entitas otomatis teks pemerintah daerah menggunakan NER dengan menerapkan pendekatan *spaCy* dan model *IndoBERT pretrained*. Data diambil hanya dari situs Detik.com untuk menjaga konsistensi struktur.

Penelitian ini tidak bertujuan melatih model dari awal, melainkan mengukur performa model NER *spaCy* dan *IndoBERT* yang telah dilatih sebelumnya (*pretrained*).

1.4 Tujuan

Penelitian ini bertujuan untuk menerapkan *Named Entity Recognition* (NER) menggunakan pendekatan *spaCy* dan model *IndoBERT pretrained*. *SpaCy* dan *IndoBERT Pretarined* diterapkan untuk membantu mengambil dan menampilkan informasi penting yang tertulis di dalam berita pemerintah daerah secara otomatis, sehingga pengguna dapat melihat isi utama berita dengan lebih cepat dan rapi. Selain itu, penelitian ini juga bertujuan untuk mengevaluasi performa hasil ekstraksi entitas menggunakan metrik evaluasi seperti *precision*, *recall*, dan *F1-score* guna mengetahui tingkat ketepatan dan kelengkapan model dalam mengenali entitas. Hasil ekstraksi entitas tersebut dimanfaatkan untuk menyajikan ringkasan isi berita berdasarkan informasi utama yang tertulis di dalam teks, tanpa melakukan penafsiran makna atau opini. Sebagai bentuk implementasi, sistem berbasis web sederhana dirancang untuk menampilkan hasil ekstraksi entitas secara terstruktur dan mudah diakses, yang diharapkan dapat mendukung upaya digitalisasi informasi dalam konteks pemerintahan

1.5 Manfaat

Penelitian ini diharapkan memberikan manfaat berupa pendekatan praktis dalam pengolahan teks berbahasa Indonesia melalui kombinasi *spaCy* dan *IndoBERT pretrained* untuk kebutuhan ekstraksi informasi dari berita pemerintah daerah. Metode ini dapat menjadi referensi dalam pengembangan sistem informasi berbasis NER yang mampu menangani dokumen tidak terstruktur, khususnya di

lingkungan administrasi publik dan pemerintahan. Selain itu, sistem web yang dihasilkan dapat digunakan sebagai alat bantu dalam mengotomatisasi dan memvisualisasikan hasil ekstraksi entitas penting, sehingga mendukung efisiensi dan integrasi informasi dalam sistem digital pemerintahan.