BAB 1 PENDAHULUAN

1.1. Latar Belakang

Named Entity Recognition (NER) salah satu topik penting dalam *Natural Language Processing* (NLP) [1]. Topik ini biasa digunakan untuk mengidentifikasi nama entitas dalam teks yang menyerupai kategori yang telah ditentukan, seperti nama orang, Lokasi, dan organisasi [2]. *Named Entity Recognition* juga menjadi dasar dari penelitian NLP seperti Ekstraksi Relasi [3], Grafik Pengetahuan [6] dan lain lain. Saat ini penelitian *Named Entity Recognition* tersebar secara global di bidang teknik [1], Bioteknologi [3], dan bidang lainnya.

Namun, sejauh ini, dalam pemrosesan bahasa alami, terdapat tantangan yang signifikan, yaitu kesulitan untuk mengenali entitas dengan tepat. Ambiguitas serta kepekaan konteks menjadi faktor utama, karena kata-kata dapat memiliki banyak arti dan tipe entitas sering kali ambigu, sehingga mengidentifikasi jenis entitas yang benar memerlukan pertimbangan konteks sekitar dan pemahaman bahasa yang kompleks. Selain itu, NER juga menghadapi kesulitan dalam mengenali dan mengklasifikasikan entitas baru yang tidak ada dalam data pelatihan. Oleh karena itu, penting untuk terus mengkaji perkembangan model NER dengan melakukan penelitian yang cermat dan mendalam.

Banyak model yang telah dikembangkan untuk mengidentifikasi *named entity recognition*, di antaranya adalah Conditional Random Field (CRF) [7], Recurrent Neural Network (RNN) [8], Long Short-Term Memory (LSTM) [10], Bidirectional Long Short-Term Memory (BiLSTM) [10], BiLSTM-CRF [2], BERT-BiLSTM-CRF [10], dan ALBERT-BiLSTM-CRF [10]. Pada penelitian ini, dilakukan perbandingan terhadap *precision*, *recall*, dan *F1-Score* ketika menggunakan CRF, BiLSTM, dan BiLSTM CRF sebagai model dalam mengidentifikasi entitas orang. BiLSTM-CRF dipilih karena telah digunakan dalam beberapa penelitian yang berbeda [10] [11] [12]. Selain itu, karena masih banyak entitas yang

memiliki banyak arti, membuat penelitian ini penting untuk dilakukan untuk perkembangan metode NER. Oleh karena itu, penelitian ini berfokus pada ekstraksi entitas orang dalam teks bahasa Inggris. Untuk mencapai tujuan penelitian, diambil *dataset* yang sudah valid yaitu *dataset* NER CoNLL 2003 versi bahasa Inggris. Ilustrasi ekstraksi dari text dapat dilihat pada Gambar 1.1.

Gambar 1.1 Identifikasi entitas orang

1.2. Rumusan Masalah

Penelitian ini berfokus pada analisis efektivitas model CRF, BiLSTM, dan BiLSTM-CRF untuk mengidentifikasi entitas orang dalam *dataset* berbahasa Inggris. *Dataset* yang digunakan adalah dataset CoNLL 2003 versi bahasa Inggris.

1.3. Tujuan

Bedasarkan perumusan masalah sebelumnya, ditetapkan tujuan pada penelitian ini adalah menganalisis efektivitas model CRF, BiLSTM, dan BiLSTM-CRF dalam mengidentifikasi entitas orang dalam bahasa Inggris.

1.4. Batasan Masalah

Untuk memfokuskan penelitian ini, berikut adalah batasan-batasan masalah yang diterapkan:

1. *Dataset* yang digunakan dalam penelitian ini adalah CoNL 2003 versi bahasa Inggris
2. Fokus utama penelitian adalah pada identifikasi entitas orang saja, tanpa mencakup entitas lain seperti lokasi, negara, dan organisasi.

3. Penelitian ini akan mengukur dan membandingkan performa model CRF, BiLSTM, dan BiLSTM-CRF menggunakan metrik *precision*, *recall*, dan *F1-Score* untuk entitas orang.

1.5. Metode Penelitian

Metode penelitian yang digunakan dalam penelitian ini mencakup beberapa tahapan utama. Pertama, pengumpulan data dilakukan dengan menggunakan dataset CoNLL 2003 versi bahasa Inggris, yang telah terbukti valid dan reliabel untuk penelitian NER. Dataset ini mencakup entitas seperti nama orang, lokasi, dan organisasi, namun penelitian ini difokuskan pada entitas orang. Tahap berikutnya adalah pra-pemrosesan data, yang meliputi tokenisasi untuk memisahkan teks menjadi unit-unit kata, pemberian label pada setiap token sesuai anotasi yang relevan (seperti label "PER" untuk entitas orang). Setelah itu, dilakukan pembangunan model, yaitu CRF, BiLSTM, dan kombinasi BiLSTM-CRF.