BAB I
PENDAHULUAN

1.1 Latar Belakang

Di era komputer dan internet, generasi data terus berkembang. Informasi yang dikumpulkan berasal dari berbagai sumber, termasuk artikel berita, dokumen, dan teks media sosial. Oleh karena itu, teknik yang efektif untuk mengorganisir dan menginterpretasikan data tekstual sangat dibutuhkan. Komponen penting dari analisis teks *named entity recognition* (*NER*). Teknologi pemrosesan bahasa alami yang disebut *NER* digunakan untuk mengekstrak data dari sumber teks yang tidak terstruktur termasuk email, blog, dan surat kabar. Tidak mungkin meremehkan pentingnya *NER* dalam bidang *Natural Language Processing* (*NLP)*. Dalam industri bisnis dan teknologi, NER membantu pengguna untuk menggabungkan fitur-fitur seperti asisten virtual yang lebih cerdas, sistem rekomendasi yang tepat, dan pencarian informasi yang efektif. Misalnya, asisten virtual dapat memahami dan bereaksi terhadap permintaan pengguna secara lebih efektif jika mereka mengetahui nama entitas dalam perintah pengguna.

Penelitian ini menggunakan tiga *dataset*: *Groningen Meaning Bank* (GMB), *Europeana Newspaper*, dan *DBpedia* [1]. Salah satu algoritma yang digunakan adalah kombinasi dari *Bidirectional Long Short-Term Memory* dan *Conditional Random Field* (*Bi-LSTM-CRF*), yang memanfaatkan mekanisme *bidirectional long sequence processing* untuk memahami semantik dari konteks teks dan *Conditional Random Field* untuk memastikan kemampuan pengenalan model [2]. Pada penelitian sebelumnya pada *dataset CCKS-2019* dan *dataset Second Affiliated Hospital of Soochow University* (SAHSU) mencapai *F1 Score* sekitar 75%, sedangkan penelitian lain dengan *dataset* teks paten *Traditional Chinese Medicine* (*TCM*) mendapatkan *F1 Score* 94.48% [3][4]. Selain *Bidirectional Long Short-Term Memory* dan *Conditional Random Field* (*BI-LSTM-CRF*) ada juga *Embeddings from Language Models* (*ELMo*) yang merupakan *embedding* kata yang dikontekstualisasikan. *Embeddings from*

Language Models (*ELMo*) telah terbukti memberikan peningkatan akurasi yang sebanding dalam berbagai tugas *NLP* tingkat kalimat [5][6]. Penelitian sebelumnya dengan *dataset English CoNLL* 2003 memperoleh F1-Score sebesar 95,56%, sementara penelitian lain dengan *dataset* korpus Biomedical Named Entity Recognition (B-NER) memperoleh F1 Score sebesar 69% [7][8].

Dari pembahasan di atas, penelitian ini membandingkan dua algoritma *Named Entity Recognition* (*NER*) yang berbasiskan *Bidirectional Long Short-Term Memory dan Conditional Random Field* (*Bi-LSTM-CRF*), serta *Embeddings* dari *Language Model* (*ELMo)* untuk tiga jenis *dataset* bahasa. Penelitian ini bertujuan untuk mengetahui perbandingan performa antara dua algoritma yang memiliki perbedaan kemampuan dalam melakukan NER pada *dataset* yang sama sehingga dapat diketahui apakah masing-masing algoritma memberikan hasil yang lebih baik atau lebih buruk. Penelitian ini diharapkan dapat memberikan wawasan tentang bagaimana kedua algoritma tersebut dibandingkan dalam hal melakukan *Named Entity Recognition* (*NER*). Kontribusi utama dari penelitian ini adalah untuk memandu saran berdasarkan informasi empiris tentang algoritma mana yang berkinerja lebih baik NER, oleh karena itu, penelitian ini dapat bermanfaat dalam aplikasi pemrosesan bahasa alami yang berfokus pada bisnis, teknologi, atau sektor yang berhubungan dengan kesehatan.

1.2 Rumusan Masalah

Berikut rumusan masalah pada penelitian ini, antara lain:

1. Bagaimana kinerja algoritma *Bidirectional Long Short-Term Memory dan Conditional Random Field* (*Bi-LSTM-CRF*) dan *Embeddings* dari *Language Model (ELMo)* dalam *NER*?
2. Berapa F1-Score algoritma *Bidirectional Long Short-Term Memory dan Conditional Random Field* (*Bi-LSTM-CRF*) dan *Embeddings* dari *Language Model (ELMo)* dalam *NER*?

1.3 Batasan Masalah

Berikut batasan masalah yang digunakan dalam penelitian ini, antara lain:

1. Penelitian ini hanya akan menggunakan dataset *Groningen Meaning Bank* (GMB), *Europeana Newspaper*, dan *DBpedia.*

2. Penelitian ini hanya akan menggunakan algoritma *Bidirectional Long Short-Term Memory dan Conditional Random Field* (*Bi-LSTM-CRF*) dan *Embeddings* dari *Language Model (ELMo)* untuk *NER.*

3. Penelitian ini tidak akan menggunakan model tuning untuk meningkatkan kinerja algoritma.

4. Penelitian ini dijalankan dengan laptop Thinkpad x250 dan menggunakan versi free pada google collaboratory.

1.4 Tujuan Penelitian

Penelitian ini bertujuan untuk melakukan perbandingan dua algoritma untuk melakukan *NER* menggunakan dataset *roningen Meaning Bank* (GMB), *Europeana Newspaper*, dan *DBpedia.*

1.5 Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan wawasan tentang bagaimana kedua algoritma tersebut dibandingkan dalam hal melakukan *Named Entity Recognition* (*NER*).

1.6 Sistematika Penulisan

BAB I PENDAHULUAN

Bab ini mengkaji perihal latar belakang, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian dan sistematika penulisan.

BAB II TINJAUAN PUSTAKA

Bab ini memberikan pemahaman awal tentang masalah yang akan diteliti berdasarkan kajian pustaka yang relevan pada penelitian sebelumnya. Kajian ini mencakup studi literatur dan dasar teori penelitian.

BAB III METODE PENELITIAN

Bab ini memberikan gambaran umum objek penelitian, penjelasan tentang proses alur penelitian serta alat dan bahan yang digunakan untuk menunjang penelitian.

BAB IV HASIL DAN PEMBAHASAN

Bab ini membahas mengenai hasil uji coba terhadap model yang digunakan, kemudian analisis hasil penelitian tersebut dituangkan dalam bentuk laporan penelitian.

BAB V PENUTUP

Bab ini menyampaikan kesimpulan dari keseluruhan analisis penelitian dan memberikan saran penelitian kedepannya yang berkaitan dengan kekurangan penelitian ini sehingga dapat mendorong pengembangan penelitian lebih lanjut.