**LAPORAN AKHIR**

***Indonesian Named-entity Recognition using Conditional Random Fields (CRF)***



**Pengusul:**

1. 12S17009 – Prince Ephraim Prabowo Silaban
2. 12S17055 – Joshua Koko Sarasi Manurung
3. 12S18001 – Cindy Angelia Siregar
4. 12S18003 – Citra Hutajulu
5. 12S18017 – Putri Yohana Panjaitan
6. 12S18035 – Angelina Naomi Christina Sinaga

**11S4037 – PEMROSESAN BAHASA ALAMI**

**FAKULTAS INFORMATIKA DAN TEKNIK ELEKTRO**

**INSTITUT TEKNOLOGI DEL**

**2021**

**DAFTAR ISI**

# DAFTAR GAMBAR

# DAFTAR TABEL

# BAB 1 PENDAHULUAN

Bab pendahuluan berisi penjelasan terkait latar belakang, tujuan, manfaat, dan ruang lingkup pengerjaan proyek.

**1.1 Latar Belakang**

Pada dokumen teks mengandung banyak informasi seperti nama orang, nama organisasi dan nama tempat. Pada umumnya untuk memperoleh informasi dari dalam dokumen teks secara manual dengan cara membaca seluruh isi dokumen. Apabila dokumen teks tersebut sangat panjang, maka dibutuhkan waktu yang lama untuk memperoleh informasi yang terdapat dalam dokumen tersebut. Oleh karena itu *Named Entity Recognition* dapat digunakan untuk memperoleh informasi pada dokumen teks secara otomatis dan dapat dilakukan dalam waktu yang singkat [1]. Pengenalan entitas bernama (*Named Entity Recognition* atau NER) merupakan salah satu bentuk pemrosesan bahasa alami (*Natural Language Processing* atau NLP), subbidang kecerdasan buatan (*Artificial Intelligence* atau AI) yang dapat digunakan untuk memperoleh entitas bernama seperti nama orang, organisasi dan tempat pada sebuah dokumen teks secara otomatis [2]. *Named entity recognition* (NER) diterapkan pada identifikasi entitas dalam artikel berita, algoritma pencarian agar berjalan lebih efektif dan efisien, pengembangan rekomendasi konten untuk suatu media, dan penanganan *customer support*.

Selama lima tahun terakhir, penelitian terkait NER sudah banyak dilakukan dalam berbagai metode dan tujuan yang digunakan. Metode yang digunakan dapat dikategorikan dalam dua kategori pendekatan, yaitu pendekatan *rule-based* dan pendekatan *machine learning* [3]. Metode *rule-based* menggunakan aturan yang telah didefinisikan berdasarkan pengetahuan linguistik dengan analisis yang dilakukan pada tingkatan sintaksis dan semantik. Metode tersebut memiliki keterbatasan karena aturan sintaksis dan semantik harus didefinisikan sebanyak mungkin untuk memperoleh hasil yang optimal.

Saat ini, penelitian mengenai *Named Entity Recognition* untuk Bahasa Indonesia telah banyak dilakukan. Pada Tahun 2018, William G. melakukan sebuah penelitian *Named Entity Recognition* Bahasa Indonesia dengan menggunakan algoritma *Deep learning* [4]*.* Pada penelitian tersebut, data yang digunakan berasal dari DBPedia Indonesia sebagai data berlabel dan artikel berita dari situs berita Indonesia (*kompas.com, cnnindonesia.com, tempo.com*) sebagai data tidak berlabel dengan menggunakan arsitektur *hybrid bidirectional LSTM (BLSTM).* Hasil dari penelitian tersebut dapat disimpulkan bahwa jumlah dataset yang digunakan dalam penelitian masih sedikit sehingga hasil eksperimen melaporkan bahwa kinerja Deep Learning masih cukup rendah untuk beberapa label. Oleh karena itu, pada penelitian selanjutnya dapat mengumpulkan dan membangun dataset yang lebih banyak. Dengan memiliki ukuran dataset yang besar, dapat mendukung kinerja yang lebih baik. Untuk penelitian ini juga akan menggunakan metode supervised learning untuk menghasilkan kinerja yang lebih baik.

**1.2 Tujuan**

Adapun tujuan dari pengerjaan proyek ini adalah menghasilkan model *Named Entity Recognition* Bahasa Indonesia yang digunakan untuk mempermudah proses menentukan jenis entitas pada kata atau frasa yang terdapat pada kalimat atau teks Bahasa Indonesia. Dan juga untuk mengidentifikasi / mengklasifikasi entitas yang sangat berguna saat ingin mengekstrak informasi.

**1.3 Ruang Lingkup**

Ruang lingkup yang akan dibahas dalam penelitian ini adalah:

1. Masukan pada model NER berupa Bahasa Indonesia sesuai dengan pedoman Penulisan Umum Ejaan Bahasa Indonesia (PUEBI).
2. Dataset yang digunakan pada proyek bernama dataset SINGGALANG.

# BAB 2 ISI

Pada bab ini dijelaskan analisis yang dilakukan terhadap data dan metode, desain pemrosesan bahasa alami yang ditampilkan dalam bentuk *flowchart* atau diagram alir, implementasi berupa kode program dan cuplikan hasil, dan hasil evaluasi kuantitatif terhadap implementasi pemrosesan bahasa alami yang dilakukan.

**2.1 Analisis**

Pada subbab ini dijelaskan analisis yang dilakukan terhadap data dan metode yang digunakan dalam mengimplementasikan pemrosesan bahasa alami.

**2.1.1 Analisis Data**

**2.1.2 Analisis Metode**

**2.2 Desain**

**2.3 Implementasi**

**2.4 Hasil**

# BAB 3 PENUTUP

**3.1 Pembagian Tugas dan Tanggung Jawab**

**3.2 Kesimpulan**

**3.3 Saran**

# DAFTAR PUSTAKA