BAB 1
PENDAHULUAN

1.1 Latar Belakang Masalah

Named Entity Recognition (NER) atau pengenalan entitas bernama adalah salah satu bagian atau tugas dari *natural language processing (nlp).* Tujuan dari *NER* adalah untuk mengidentifikasi atau mengklasifikasi sebuah entitas misalnya nama orang, organisasi, waktu, lokasi dan sesuatu entitas lain dalam sebuah teks yang sangat berguna dalam kasus ekstraksi informasi [1].

Penelitian tentang *NER* sudah dilakukan di Bahasa Indonesia. Salah satunya adalah penelitian yang menggunakan metode *HMM* dalam kasus pembangkit pertanyaan otomatis [2] dengan hasil akurasi yaitu 42,54%. Rendahnya akurasi pada penelitian tersebut disebabkan oleh adanya kata-kata ambigu yang tidak terdeteksi.

Sementara itu, dalam metode pembelajaran mesin terdapat metode yang terbukti mendapatkan pencapaian performa paling tinggi*(state-of-the-art)* dalam kasus *NER* [3], yaitu *Bidirectional LSTM-CRF*. *Bidirectional LSTM* menggabungkan konteks sebelumnya dan konteks setelahnya dengan memproses data dari dua arah [1] yang selanjutnya diklasifikasi menggunakan *CRF* [3]. Pada penelitian yang dilakukan oleh Guillaume Lample, dkk [3] mengkomparasikan dua *neural* arsitektur dalam mengatasi *NER*, yaitu *Stack-LSTM* (*S-LSTM*) dan *Bidirectional LSTM-CRF*. Hasil dari komparasi pada dataset bahasa Inggris, *S-LSTM* mendapatkan akurasi 90,33% dan *Bidirectional LSTM-CRF* mendapatkan akurasi 90.94%. Pada penelitian lain yang dilakukan oleh T Anh Le, dkk [1] metode *Bidirectional LSTM-CRF* mendapatkan akurasi 87,17% dibandingkan dengan metode NeuroNER yang mendapatkan akurasi 85.37% untuk dataset Gareev's. Dari penelitian-penelitian yang telah dilakukan terbukti bahwa metode *Bidirectional LSTM-CRF* tersebut memiliki akurasi yang cukup tinggi. Oleh karena itu, dalam penelitian ini akan digunakan metode *Bidirectional LSTM-CRF* untuk menangani kasus pada teks bahasa Indonesia

Berdasarkan Uraian tersebut maka pada penelitian ini akan dilakukan Implementasi metode *Bidirectional LSTM-CRF* pada kasus Named Entity Recognition dalam teks bahasa Indonesia.

1.2 Identifikasi Masalah

Berdasarkan penjelasan yang telah diuraikan pada latar belakang, maka identifikasi masalah dalam penelitian ini adalah sebagai berikut

1. Rendahnya tingkat akurasi dari penelitian sebelumnya [2].
2. Belum adanya penelitian untuk membuktikan keefektifan metode *Bidirectional LSTM-CRF* dalam kasus *named entity recognition* pada teks berbahasa Indonesia

1.3 Maksud dan Tujuan

Maksud dari penelitian ini adalah untuk membangun sebuah sistem NER (*Named Entity Recognition*) dengan menggunakan metode *Bidirectional LSTM-CRF* pada teks bahasa Indonesia. Adapun tujuan dari penelitian ini adalah sebagai berikut :

1. Mengetahui akurasi dari metode *Bidirectional LSTM-CRF* dalam menangani kasus *NER* pada teks bahasa Indonesia.
2. Membuktikan keefektifan metode *Bidirectional LSTM-CRF* dalam menangani kasus *NER* pada teks bahasa Indonesia.

1.4 Batasan Masalah

Berikut batasan masalah yang meliputi data masukan, proses, dan data keluaran.

1. Data Masukan

   Artikel berita politik berbahasa Indonesia yang bersumber dari penelitian yang dilakukan oleh Rusliani dengan format file *.txt [2].

2. Proses
   a. Ekstraksi Fitur yang digunakan adalah sebanyak 6 fitur (InitCap, AllCap, AllLower, Digits, Containts Digits, Punctuation) [4] .

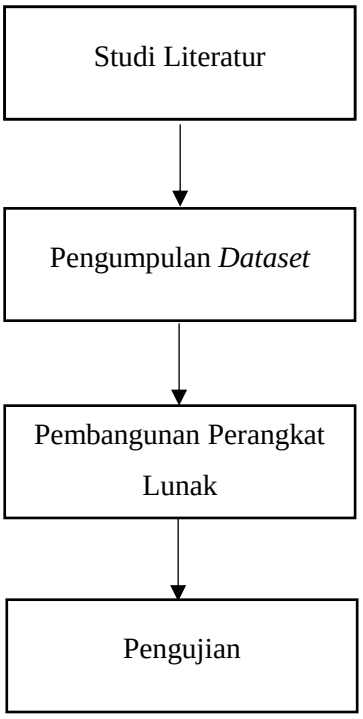
b. Nama entitas kelas yang dikenali adalah person, location, time, organization, quantity dan other [2] [5].

3. Data Keluaran

   Kata dan nama entitas kelas

1.5 Metodologi Penelitian

Metodologi penelitian yang digunakan adalah metode Deskriptif [6]. Metode ini digunakan karena penelitian ini menggunakan metode dalam *named entity recognition* berdasarkan hasil yang didapat dari penelitian yang dilakukan sebelumnya untuk mendapatkan hal-hal apa saja yang dilakukan dalam penelitian ini. Berikut adalah skema yang digunakan pada penelitian ini.



Gambar 1.1 Alur Penelitian

1.5.1 Studi Literatur

Dalam penelitian ini hal pertama yang dilakukan adalah mengidentifikasi masalah. Untuk melakukan identifikasi masalah maka dilakukan studi literatur. Studi literatur pada penelitian ini adalah mempelajari literatur-literatur seperti buku,

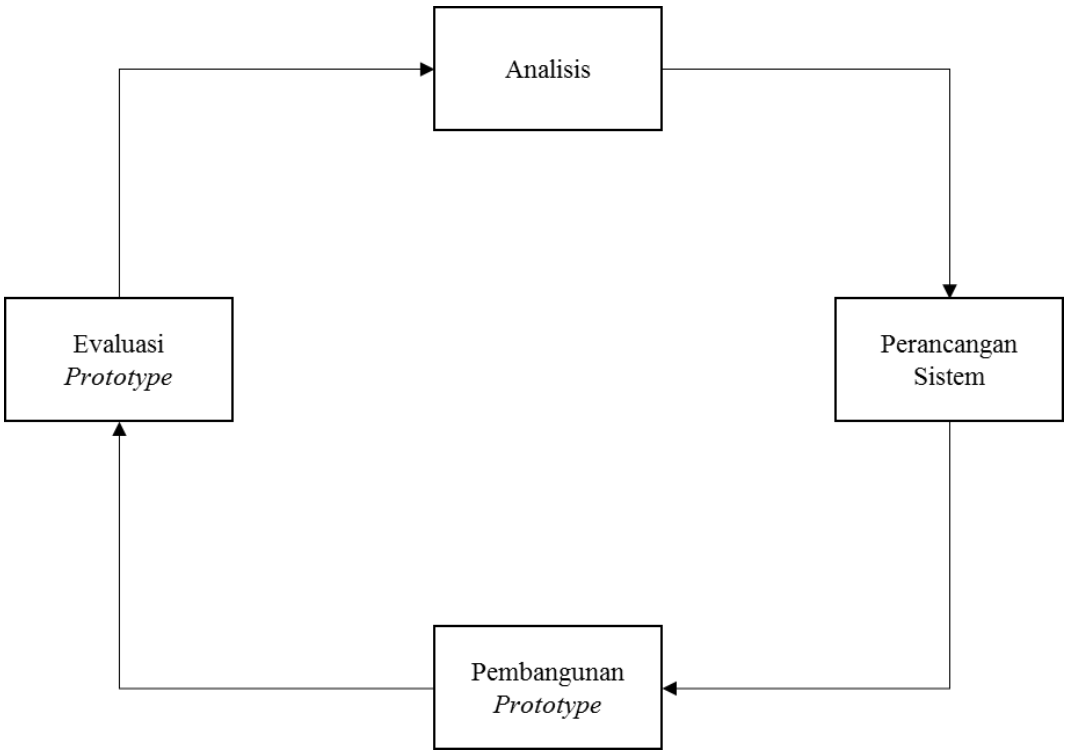
jurnal, artikel ilmiah, dan *website* yang berhubungan dengan pengenalan entitas bernama dan *Bidirectional LSTM-CRF*.

1.5.2 Pengumpulan Dataset

Dalam menggunakan algoritma *Bidirectional LSTM-CRF*, maka dibutuhkan suatu *dataset*. *Dataset* yang digunakan pada penelitian ini berupa artikel berita politik yang sudah diberi tag. *Dataset* tersebut didapatkan dari penelitian yang dilakukan Rusliani [2].

1.5.3 Metode Pembangunan Perangkat Lunak

Metode pembangunan perangkat lunak dalam penelitian ini adalah menggunakan model *prototype* [7].



Gambar 1.2 Model *Prototype* [7]

1. Analisis Metode

Pada tahapan ini dilakukan analisis kebutuhan dan analisis proses metode *bidirectional LSTM-CRF* dalam mengenali entitas bernama. Setelah

kebutuhan metode dan elemen simulator terkumpul maka dibuat perancangan simulator yang akan dibangun.

2. Implementasi Metode

Proses yang dilakukan dalam tahapan ini adalah membangun simulator pengenalan entitas berdasarkan analisis metode yang telah dilakukan. Simulator dibangun menggunakan bahasa pemograman *python* dan teks editor sebagai tempat menyimpan data pelatihan dan pengujian.

3. Pengujian

Setelah simulator dibangun tahap selanjutnya adalah melakukan pengujian hasil implementasi *bidirectional LSTM-CRF* untuk mengenali entitas bernama pada teks bahasa Indonesia. Pengujian ini difokuskan pada pengujian akurasi data.

4. Penarikan kesimpulan

Di tahap ini dilakukan penarikan kesimpulan berdasarkan hasil pengujian implementasi metode *bidirectional LSTM-CRF* untuk mengenali entitas bernama pada teks bahasa Indonesia.

1.5.4 Pengujian

Pengujian dilakukan terhadap fungsionalitas sistem ekstraksi informasi dengan menggunakan metode *Black-Box,* dan nilai akurasi menggunakan perhitungan akurasi.

1.6 Sistematika Penulisan

Sistematika penulisan laporan akhir penelitian ini disusun untuk memberikan gambaran umum tentang penelitian yang dilakukan. Sistematika penulisan tugas akhir ini adalah sebagai berikut:

BAB 1 PENDAHULUAN

Pada bab ini berisi penjelasan mengenai latar belakang, rumusan masalah, maksud dan tujuan, batasan masalah, metodologi penelitian, dan sistematika penulisan untuk memberikan gambaran urutan pemahaman dalam menyajikan laporan ini.

BAB 2 TINJAUAN PUSTAKA

Pada bab ini membahas landasan teori yang digunakan untuk menganalisis masalah dan mengolah data penelitian yaitu teori mengenai *named entity recognition*, metode *Recurrent Neural Network* (*RNN*), metode *Long Short Term Memory* (*LSTM*), *bidirectional LSTM*, metode *Conditional Random Field* (*CRF*), kombinasi *bidirectional LSTM* dan *CRF*, dan teori-teori pembangun simulator.

BAB 3 ANALISIS DAN PERANCANGAN

Pada bab ini berisi tentang analisis masalah, analisis sistem yang berisi analisis data masukan dan analisis proses, analisis kebutuhan sistem yang berisi analisis kebutuhan non fungsional dan analisis kebutuhan fungsional, perancangan antarmuka dan jaringan semantik.

BAB 4 IMPLEMENTASI DAN PENGUJIAN

Pada bab ini menjelaskan mengenai proses implementasi pengujian metode *bidirectional LSTM-CRF* dalam *named entity recognition* pada teks bahasa Indonesia yang meliputi penginputan data sampai dengan hasil klasifikasi, serta pengujian dan pengukuran terhadap klasifikasi yang dihasilkan oleh proses pengenalan entitas.

BAB 5 KESIMPULAN DAN SARAN

Pada bab ini menjelaskan tentang hasil yang diperoleh dari penelitian terhadap metode *bidirectional LSTM-CRF* dalam *NER* pada teks bahasa Indonesia berdasarkan maksud dan tujuan yang telah ditentukan serta saran untuk pengembangan *NER* selanjutnya.