Pendekatan Conditional Random Field Indonesian Named Entity Recognition

Proposal Proyek Pemrosesan Bahasa Alami

Oleh :

11S18022	**Devi Wahyuni Silitonga**
11S18037	**Hisar Haryanto Sinaga**
11S18046	**Simson Fransisco Panjaitan**
11S18058	**Leonardo Sianturi**
11S18066	**Jumadi Heryanto Damanik**

Institut Teknologi Del

2020/2021

DAFTAR ISI

BAB 1
PENDAHULUAN

Pada bab ini dijelaskan mengenai latar belakang topik penelitian, rumusan permasalahan penelitian, tujuan penelitian, ruang lingkup penelitian, metode penelitian dan sistematika penulisan dalam menyusun proyek ini.

1.1 Latar Belakang

Dokumen merupakan media penyajian informasi yang sangat banyak digunakan. Dokumen dapat menyimpan informasi dalam jumlah yang sangat fleksibel, dan informasi yang beragam, sesuai dengan konteks penggunaannya. Melalui dokumen, terdapat banyak informasi penting yang dapat diambil dari setiap kalimatnya, seperti nama orang, ataupun nama suatu tempat atau wilayah. Pada era modern saat ini, dokumen dapat dibuat dalam bentuk digital. Hal ini dapat mempermudah pembuatan, pengolahan informasi didalamnya, dan pendistribusian dokumen tersebut. Informasi di dalam dokumen dapat diperoleh melalui membaca dokumen tersebut secara keseluruhan, dan apabila dokumen berisi teks yang sangat panjang, maka dibutuhkan waktu yang lama untuk mendapatkan informasi yang disajikan di dalam dokumen tersebut. Untuk mengatasi masalah tersebut, diciptakan lah NER (Named Entity Recognition) yang dapat digunakan untuk memperoleh informasi secara otomatis dari sebuah dokumen teks melalui proses ekstraksi dan klasifikasi kata atau frasa secara otomatis kedalam kategori tertentu.

Penerapan NER pada proyek ini, dilakukan dengan menggunakan metode CRF, dimana akan dilakukan prediksi entitas pada *dataset* yang digunakan. Sebelum model melakukan prediksi, *dataset* akan diolah dengan metode *text preprocessing* yaitu *Lowercase,* yang akan Mengubah seluruh data menjadi huruf kecil, dan *BIO Annotation,* yang memberikan tag BIO. setelah proses text preprocessing, pembangunan model akan dilanjutkan dengan menerapkan model klasifikasi *Random Forest Classification,* dan *Conditional Random Field.* melalui proses tersebut, diharapkan model yang dibangun memiliki tingkat akurasi prediksi yang tinggi, dalam menentukan klasifikasi entitas bernama dalam bahasa Indonesia.

1.2 Tujuan

Adapun tujuan dari pengerjaan proyek ini adalah :

1. Membangun model prediksi dengan metode CRF dalam menentukan entitas bernama dalam bahasa Indonesia.

1.3 Manfaat

Manfaat yang diperoleh melalui pengerjaan proyek ini adalah :

1. Bagi Mahasiswa
   Dalam proyek ini, mahasiswa diharapkan mampu membangun model prediksi menggunakan CRF dalam menentukan entitas bernama dalam bahasa Indonesia

1.4 Ruang lingkup

Adapun ruang lingkup dari penelitian ini adalah sebagai berikut:

1. Model yang dibangun akan melakukan prediksi entitas bernama dalam bahasa Indonesia.
2. Model yang dibangun menggunakan pendekatan yang digunakan adalah Conditional Random Field (CRF).

BAB 2
Analisis

Pada bab ini dijelaskan analisis yang dilakukan terhadap data dan metode, desain pemrosesan bahasa alami yang ditampilkan dalam bentuk diagram alir, implementasi berupa kode program dan cuplikan hasil, serta 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 pada pengimplementasian pemrosesan bahasa alami.

2.1.1 Analisis Data

Dataset yang digunakan dalam proyek tersebut diperoleh dari link berikut : https://github.com/ialfina/ner-dataset-modified-dee/blob/master/singgalang/SINGGALANG.tsv. *Dataset* yang digunakan adalah SINGGALANG.tsv yang terdiri dari 1048575 *records* dan dua kolom yaitu Word dan Tag.

Tabel 1. Atribut pada *Dataset*

No.	Nama Atribut	Tipe Atribut	Deskripsi
1.	*word*	Karakter	Merupakan kata-kata dalam sebuah atau beberapa kalimat.
2.	*tag*	Kategori	Adalah tag-tag yang diberikan pada setiap kata berdasarkan sistem penandaan IOB.

Sistem IOB Tagging berisi tag dalam bentuk :

- **B** - Untuk kata bagian dalam

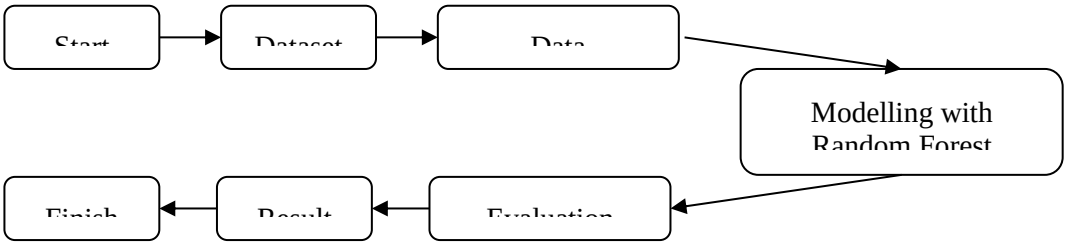
- **I** - Untuk kata-kata di dalam potongan
- **O** - Untuk kata diluar dari potongan apapun

2.1.2 Analisis Metode

2.2 Desain

Pada sub bab ini dijelaskan mengenai desain pemrosesan bahasa alami yaitu analisis pada Indonesian Named Entity Recognition yang ditampilkan dalam bentuk *flowchart* atau diagram alir seperti ditunjukkan pada gambar di bawah ini.

Gambar 1. Desain Analisis pada Indonesian Named Entity Recognition



2.2.1 *Data Preprocessing*

Data Preprocessing merupakan tahap pemrosesan awal yang dilakukan sebelum teks (*dataset*) diolah ke tahap selanjutnya. Data yang diperoleh masih dalam format yang tidak terstruktur. Selain itu, data masih dalam format *raw data* sehingga tidak memungkinkan untuk melakukan analisis pada *raw data*. Oleh karena itu perlu dilakukan *data preprocessing* untuk menghilangkan kata-kata pada teks atau dokumen yang mengandung beberapa format yang keberadaannya tidak penting dalam *text mining*.

2.2.1.1 *Data Cleaning*

Data yang diperoleh dari dataset masih memiliki beberapa *noise* yang perlu dibersihkan seperti string kosong (*incomplete data/missing value*). Beberapa cara yang dapat dilakukan untuk membersihkan data adalah sebagai berikut :

- Mengabaikan data yang hilang dengan cara menghapus data tersebut, namun cara ini tidak cukup efektif ketika terdapat banyak data yang hilang.
- Mengisi *missing value*, cara ini juga tidak cukup efektif ketika ukuran data besar dan membutuhkan waktu yang lama.
- Memperbaiki penulisan kata yang kurang tepat.
- Menghapus karakter-karakter yang bersifat *noise* pada dataset.

2.2.2 *Modelling*

2.2.2.1 *Random Forest Classifier*

Random Forest Classifier merupakan algoritma yang digunakan pada klasifikasi data dalam jumlah yang besar. Klasifikasi random forest dilakukan melalui penggabungan tree dengan melakukan training pada sampel data yang dimiliki.

Penentuan klasifikasi dengan random forest diambil berdasarkan hasil voting dari tree yang terbentuk. Pemenang dari tree yang terbentuk ditentukan dengan vote terbanyak. Pembangunan tree pada random forest sampai dengan mencapai ukuran maksimum dari pohon data. Akan tetapi, pembangunan pohon random forest tidak dilakukan pemangkasan (pruning) yang merupakan sebuah metode untuk mengurangi kompleksitas ruang.

Pembangunan dilakukan dengan penerapan metode random feature selection untuk meminimalisir kesalahan. Pembentukan tree dengan sampel data menggunakan variabel yang diambil secara acak dan menjalankan klasifikasi pada semua tree yang terbentuk. Random forest menggunakan decision tree untuk melakukan proses seleksi.

2.2.2.2 *Conditional Random Fields*

CRF adalah kelas metode pemodelan statistik yang sering diterapkan dalam pengenalan pola dan pembelajaran mesin dan digunakan untuk prediksi terstruktur. Sementara pengklasifikasi memprediksi label untuk sampel tunggal tanpa mempertimbangkan sampel. Untuk melakukannya, prediksi dimodelkan sebagai model grafis, yang mengimplementasikan ketergantungan antara prediksi. Jenis grafik apa yang digunakan tergantung pada aplikasinya. Misalnya, dalam pemrosesan bahasa alami, CRF rantai linear populer, yang mengimplementasikan dependensi sekuensial dalam prediksi. Dalam pemrosesan gambar, grafik biasanya menghubungkan lokasi ke lokasi terdekat dan/atau serupa untuk memastikan bahwa akan menerima prediksi serupa.

BAB 3
Penutup

3.1 Pembagian Tugas

Pada subbab ini dijelaskan mengenai pembagian tugas yang dilakukan oleh setiap anggota kelompok :

1. 11S18022 Devi Wahyuni Silitonga

   Mengerjakan bagian Pre Processing Lowercase dan BIO Annotation, mengerjakan pemodelan menggunakan metode Random Forest dan Conditional Random Fields beserta hasil evaluasi keakuratan dari kedua metode tersebut.

2. 11S18037 Hisar Haryanto Sinaga

   Mengerjakan bagian bagian Pre Processing Lowercase dan BIO Annotation, mengerjakan pemodelan menggunakan metode Random Forest dan Conditional Random Fields beserta hasil evaluasi keakuratan dari kedua metode tersebut.

3. 11S18046 Simson Fransisco Panjaitan

   Mengerjakan bagian bagian Pre Processing Lowercase dan BIO Annotation, mengerjakan pemodelan menggunakan metode Random Forest dan Conditional Random Fields beserta hasil evaluasi keakuratan dari kedua metode tersebut.

4. 11S18058 Leonardo Sianturi

   Mengerjakan bagian bagian Pre Processing Lowercase dan BIO Annotation, mengerjakan pemodelan menggunakan metode Random Forest dan

Conditional Random Fields beserta hasil evaluasi keakuratan dari kedua metode tersebut.

5. 11S18066 Jumadi Heryanto Damanik

Mengerjakan bagian bagian Pre Processing Lowercase dan BIO Annotation, mengerjakan pemodelan menggunakan metode Random Forest dan Conditional Random Fields beserta hasil evaluasi keakuratan dari kedua metode tersebut.

3.2 Kesimpulan

Implementasi *Conditional Random FIeld* dalam pembangunan model *Indonesian Named Entity Recognition* dapat melakukan prediksi entitas pada dataset yang digunakan. Model yang dibangun, akan memprediksi suatu entitas dengan tag *person, place,* dan *organization.* Model NER dibangun dengan mengimplementasikan *Conditional Random Field,* dan *Random Forest Classifier* dalam pengolahan *dataset* SINGGALANG. Berdasarkan hasil model yang telah dibangun, maka akurasi untuk masing masing model adalah 91% akurasi untuk model yang dibangun menggunakan CRF, dan 100% akurasi untuk model yang dibangun menggunakan *Random Forest Classifier*.