BAB I

PENDAHULUAN

1.1 Latar Belakang Masalah

Berita online merupakan salah satu bentuk informasi yang dapat selalu diakses oleh semua orang. Berita secara umum dapat diartikan sebagai laporan atau informasi yang dapat dipercaya dan sesuai dengan fakta atau realitas. Berita tradisional pada zaman dahulu disajikan melalui kertas koran, televisi, radio, dan banyak cara lain. Seiring berkembangannya zaman dan juga teknologi berita juga semakin gampang untuk diakses oleh semua orang, terutama bagi mereka yang menggunakan teknologi internet. Tahun 2019 jumlah pengguna internet pada Indonesia menurut survey yang telah dilakukan oleh Asosiasi Penyelenggara Jasa Internet Indonesia atau APJII, jumlah pengguna internet di Indonesia ada 171,17 juta jiwa atau sekitar 64,8% dari jumlah populasi Indonesia yang berjumlah 264 juta jiwa (Kompas, 2019).

Berita selalu mengandung informasi yang akan membahas suatu entitas atau objek, di mana suatu entitas itu bisa berupa seorang manusia, sebuah lembaga atau organisasi, ataupun suatu lokasi. Untuk menggali informasi yang penting dari suatu berita sesuai dengan *object* yang sedang dibahas pada artikel tersebut tentunya akan memakan waktu yang cukup lama jika dilakukan oleh manusia dengan membaca berita tersebut kata demi kata. Oleh karena itu dibutuhkanlah suatu sistem yang dapat membantu proses ini untuk menentukan nama entitas pada sebuah artikel berita.

Named Entity Recognition (NER) berfungsi untuk mengklasifikasikan setiap kata pada suatu dokumen kedalam kategori yang telah dibuat sebelumnya. NER termasuk kedalam *information extraction* yang berfungsi untuk meng-ekstrak informasi tertentu dari sebuah dokumen (Zhou dan Su, 2001). NER dapat berfungsi untuk membantu proses ini dikarenakan NER dapat membantu pembaca untuk mengetahui *object* apa yang sedang dibicarakan pada sebuah berita dengen menentukan entitas apa yang sedang dibicarakan bisa entitas berupa individu, perusahaan maupun lokasi ketiga entitas ini menjadi titik tumpuh pada penelitian ini dikarenakan pembaca dapat mengetahui berita mana yang ingin pembaca baca dengan mengetahui terlebih dahulu berita tersebut membahas tentang individu, perusahaan ataupun lokasi yang membuat mereka tertarik untuk membaca. Oleh karena itu dilakukanlah penelitian dengan menggunakan metode *Named Entity Recognition* (NER). NER dapat digunakan untuk menggali informasi penting yang terdapat pada sebuah berita dengan waktu yang lebih singkat. NER sendiri merupakan langkah paling pertama dari Natural Language Processing (NLP), NER juga merupakan langkah paling penting pada NLP. (Magge dkk., 2018) di mana pada bagian ini Named Entity Recognition akan mengekstrak informasi dari teks yang sudah ada kemudian mengolah teks tersebut menjadi data yang nantinya dapat diklasifikasikan sesuai dengan *entity relation* yang telah ditentukan.

Penelitian NER ini akan dilakukan menggunakan 2 metode yaitu metode Conditional Random Fields (CRF) yang berfungsi untuk mensegmentasikan dan melabelkan kata-kata yang terdapat pada sebuah artikel berita,dan juga metode *Random Forest Classifier* untuk mengklasifikasikan *entity relation* dari data yang

sudah disegmentasi dan dilabel oleh CRF. *Random Forest* menjadi pilihan metode klasifikasi karena random forest dapat membuat korelasi dari *feature map* yang telah dibuat dan *random forest* juga disebut sebagai *classifier* yang akurat.

Random Forest Classifier merupakan kumpulan dari *learning model* yang mengambil *deicision tree* sebagai *basic classifier*. Random Forest akan melatih beberapa *decision trees* dengan metode Bagging. Proses *bagging* merupakan proses dimana sampel akan diklasifikasikan diproses dan hasilnya akan didapatkan dari jumlah suara yang dihasilkan oleh sebuah *decision tree* (Anandarajan dkk., 2019).

Dari penjelasan latar belakang masalah, maka akan dilakukan penelitian mengenai implementasi algoritma *Conditional Random Fields* dan *Random Forest* untuk *Named Entity Recognition* pada dataset artikel berita.*Conditional Random fields* juga telah pernah dipakai didalam penelitian NER sebelumnya yang berhubungan dengan nama entitas pada dunia *biomedical* yang dapat dilihat pada peneletian Li, dkk(2015). *Random forest* memiliki toleransi yang tinggi terhadap *noise* dan anomali pada data. *Random forest* juga memiliki skalabilitas yang bagus dikarenakan *classification rules* yang dilatih dari sampel yang telah diberikan dan tidak membutuhkan pengetahuan sebelumnya dari suatu klasifikasi (Anandarajan dkk., 2019).

Penelitian ini diharapkan dapat membantu pembaca berita ataupun media penyedia berita untuk mengetahui siapa atau perusahaan apa yang dibahas pada suatu artikel berita. Penelitian ini diharapkan dapat membantu pembaca berita untuk menemukan berita yang sesuai dengan topik yang sedang mereka cari.

1.2 Rumusan Masalah

Berdasarkan uraian yang telah dilakukan pada Latar belakang maka rumusan masalah pada penelitian ini adalah sebagai berikut.

1. Bagaimana cara mengimplementasikan algoritma *Conditional Random Fields* dan *Random Forest* pada *Named Entity Recognition* pada artikel berita.

2. Bagaimana cara mengukur akurasi, precision, recall, f_1-score Conditional Random Fields dan Random Forest pada aplikasi Named Entity Recognition pada artikel berita

1.3 Batasan Masalah

Batasan masalah pada penelitian ini adalah sebagai berikut.

1. Named Entity Recognition hanya akan ditujukan kepada artikel berita yang berbahasa Indonesia dan dikelaskan menjadi 3 kelas yaitu, nama individu, lokasi, dan juga perusahaan.
2. Data yang akan diproses merupakan data artikel berita yang telah didapatkan dari rekomendasi penelitian sebelumnya.
3. Data yang digunakan sebanyak 200 kalimat judul dari artikel berita.

1.4 Tujuan Penelitian

1. Mengimplementasikan algoritma Conditional Random Fields dan Random Forest Classifier pada Named Entity Recognition pada artikel berita.

2. Mengukur akurasi, *precision, recall,* f_1-score *Conditional Random Fields* dan *Random Forest* pada aplikasi *Named Entity Recognition* pada artikel berita.

1.5 Manfaat Penelitian

Membantu mengklasifikasikan suatu berita sesuai dengan entitas/object yang sedang dibahas pada berita tersebut untuk membantu pembaca mendapatkan berita yang pembaca inginkan dengan waktu yang lebih singkat.

1.6 Sistematika Penulisan

Sistematika penulisan yang digunakan pada skripsi ini adalah sebagai berikut.

BAB I PENDAHULUAN

Bab ini berisi perihal latar belakang, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, dan juga sistematika penulisan tentang penelitiang yang dilakukan.

BAB II LANDASAN TEORI

Bab ini berisi penjelasan tetang metode dan teori-teori yang berkaitan dengan penelitian yang dilakukan. Teori-teori ataupun metode yang digunakan adalah teori tentang Named Entity Recognition, Text Preprocessing, Conditional Random Fields (CRF) dan Random Forest

BAB III METODOLOGI PENELITIAN DAN PERANCANGAN

Bab ini berisi tentang penjelasan dari metode penelitian, flowchart dan rancangan aplikasi Named Entity Recognition yang telah dibuat.

BAB IV IMPLEMENTASI DAN PENGUJIAN

Bab ini berisi tetang implementasi dan pengujian dari aplikasi Named Entity Recognition.

BAB V SIMPULAN DAN SARAN

Bab ini berisi tentang kesimpulan dari penelitian yang telah dilakukan dan saran untuk pengembangan penelitian lebih lanjut dengan penelitian yang sama.