BAB 1
PENDAHULUAN

1.1 Latar Belakang

Seiring dengan perkembangan teknologi, banyak kemajuan yang tercipta untuk kehidupan masyarakat untuk mengakses media informasi. Salah satu teknologi yang berkembang saat ini adalah portal berita online. Dengan adanya portal berita online masyarakat dapat mengakses informasi dengan cepat isu yang saat itu terjadi.

Portal berita online masuk kedalam situs web yang paling banyak dikunjungi dengan kategori News & Media, seperti detik.com, tribunews.com, dan kompas.com yang masuk ke dalam 10 besar situs web yang paling banyak dikunjungi menurut Hootsuite dalam laporan "Digital 2020: Indonesia". Dalam Natural Language Processing (NLP), Named Entity Recognition merupakan sub-bahasan yang cukup banyak digunakan untuk penelitian. Tugas utama dari Named Entity Recognition (NER) yaitu membantu mengidentifikasi dan mendeteksi nama entitas dari suatu kata yang terdapat dalam kalimat. Sumber data yang akan digunakan yaitu berita bahasa Indonesia yang berasal dari media faktual. kata yang terdapat pada berita bahasa Indonesia dapat merujuk nama entitas orang atau lokasi atau organisasi, sehingga untuk menentukan nama entitas tersebut harus mempertimbangkan terlebih daulu dengan melihat pola disekitarnya. Namun dengan meningkatnya keragaman bahasa yang digunakan pada portal berita sehingga menimbulkan kerumitan baru pada sistem

Information Retrieval (IR) [2]. dengan melakukan crawling beberapa media Mainstream dan dilakukan analisis kemudian data hasil crawling akan disimpan pada Elasticsearch dan dibuat satu dashboard untuk melakukan analisis bertia. pada penelitian ini, peneliti akan membuat pengelolaan Named Entity Recognition menggunakan polyglot.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah disampaikan, permasalahan yang dapat dirumuskan adalah sebagai berikut:

1. Bagaimana membuat sistem yang dapat melakukan *entity recognition* yang terintegrasi dengan portal berita online?

1.3 Batasan Masalah

1. Jumlah media *online* yang digunakan sebanyak 3 portal berita yang terverifikasi administrasi dan faktual menurut Dewan Pers Indonesia yaitu Tribunnnews.com, Detik.com, dan Kompas.com.
2. Named-entity hanya dapat menggunakan 3 kategori. yaitu nama orang (*Person*), nama tempat (*Location*), dan nama organisasi (*Organization*).

1.4 Maksud dan Tujuan Penelitian

1. Membuat sistem ekstraksi *named entity recognition* yang terintegrasi dengan portal berita *online*.

1.5 Metode Penelitian

1. Melakukan pengumpulan data berita dengan melakukan *crawling* terhadap 3 portal berita *online* mainstream. *Crawling* dilakukan dengan metode *recursive crawling* dan menggunakan beberapa heuristik untuk mendeteksi apakah alamat tersebut sebuah artikel atau bukan. Kemudian hasil dari *crawling* tersebut disimpan dalam bentuk *lucene document* pada *Elasticsearch*.
2. Dari data yang sudah diperoleh dengan cara *crawling* tersebut kemudian dapat dilakukan *recognition* tiga entitas dalam sebuah berita yaitu *Person* yang menunjukan subjek pelaku, korban atau tokoh yang sedang dibicarakan dalam sebuah berita. *Organization* menunjukan sebuah nama organisasi atau kelompok pada suatu pemberitaan. *Location* menunjukan keterangan nama tempat kejadian perkara atau tempat berita tersebut berasal.
3. Membuat sebuah *REST API* untuk melakukan *recognition* agar bisa dilakukan hanya dengan memasukan konten berita yang disimpan pada *Elasticsearch*. Dan bisa diintegrasikan dengan sistem atau aplikasi lain untuk melakukan *request recognition*.
4. Kemudian melakukan pengukuran performa dari sistematika yang digunakan, dalam hal ini mengukur akurasi *polyglot* dalam melakukan *recognition*, kecepatan *response REST API*, kemampuan *crawling* sampai dengan *indexing* ke *Elasticsearch*.

1.6 Sistematika Penulisan

Adapun sistematika penulisan skripsi ini adalah sebagai berikut:

BAB I PENDAHULUAN

Bab ini menerangkan tentang latar belakang masalah, rumusan masalah, batasan masalah, maksud dan tujuan penelitian, metode penelitian dan sistematika penelitian.

BAB II LANDASAN TEORI

Berisi tinjauan pustaka dan dasar-dasar teori yang digunakan untuk membangun sistem serta membantu pengolahan berbagai macam laporan yang berkaitan langsung dengan ilmu atau masalah yang diteliti.

BAB III METODE PENELITIAN

Bab ini menjelaskan tentang analisis dan metode yang digunakan dalam melakukan *named entity recognition*, serta hal-hal yang diperlukan dalam integrasi dengan sistem *Elasticsearch*, pengukuran performa dari sistem yang digunakan beserta hasil yang telah dibuat.

BAB IV HASIL PENELITIAN DAN PEMBAHASAN

Bab ini berisikan hasil dari penetilian yang dilakukan beserta pembahasanya.

BAB V PENUTUP

Berisi kesimpulan dan saran yang penulis rangkum selama proses penelitian.