BAB I

PENDAHULUAN

1.1 Latar Belakang

Media daring atau biasa disebut media online adalah media atau saluran komunikasi yang tersaji secara online di suatu situs *web* di *internet*. Secara bahasa (KBBI), media adalah alat atau sarana komunikasi seperti koran, majalah, radio, televisi, film, poster dan spanduk. Media juga berarti perantara atau penghubung. Sedangkan daring merupakan singkatan dari dalam jaringan yang berarti terhubung dengan jejaring komputer, *internet* dan sebagainya. Seiring dengan perkembanganya, media daring semakin diminati oleh berbagai kalangan, sehingga mulai bermunculan penelitian-penelitian berupa penggalian informasi (*data mining*) dari sebaran berita yang ada. Seperti yang saat ini sedang dikembangkan oleh salah satu media daring di Indonesia, *Beritagar.id*, yang menggali informasi dengan mengolah data dari sebaran berita-berita yang ada di seluruh Indonesia.

Dalam suatu pemberitaan, selalu mengandung unsur *5W+1H*. Dengan mengetahui elemen-elemen tersebut, maka pembaca berita dapat mengetahui secara keseluruhan isi dari berita yang disampaikan. Elemen-elemen tersebut merupakan entitas yang dapat diekstrak, dan kemudian dapat digunakan untuk membuat sebuah mesing yang mampu mengenali entitas atau biasa disebut dengan *Named Entity Recognition (NER)*. Ada beberapa metode yang dapat menyelesaikan proses *NER*, salah satu metode yang terkenal adalah *Hidden Markov Model (HMM)* yang merupakan pengembangan model statistic dari *Markov Model*. *HMM* dalam *NER* berfungsi untuk menggabungkan peluang gabungan pada proses *POS Tagging*.

Beberapa penelitian sebelumnya yang berkaitan dengan pemanfaatan *HMM* untuk *POS Tagging* menghasilkan akurasi terbaik 92,7% [3], sedangkan untuk *NER* memiliki akurasi rata-rata kurang dari 82,9% [5].

Dengan beberapa penjelasan diatas, peneliti bermaksud membuat sebuah aplikasi berbasis *web* media berita daring dengan implementasi *machine learning* yang memiliki kemampuan ekstraksi entitas yang muncul pada suatu berita. Proses ekstraksi dimulai dengan mengumpulkan banyak data berupa berita dari berbagai sumber, kemudian dilakukan proses *cleaning data*, dimana nantinya data tersebut akan dijadikan sebagai bahan *training* atau pembelajaran bagi mesin untuk dapat menghasilkan *output* berupa entitas-entitas hasil dari ekstraksi. Pembelajaran pada mesin dilakukan dengan pembelajaran terarah (*supervised learning*), yaitu dengan cara memberikan data hasil anotasi kepada mesin, untuk kemudian mesin akan melakukan *training* dari data yang telah diberikan, dan akan dilakukan evaluasi kembali untuk melihat hasil dari proses tersebut. Selanjutnya, diharapkan mesin dapat melakukan ekstraksi entitas dengan hasil yang baik, dalam arti mesin mampu mengenali entitas dari setiap berita dengan akurat dan waktu yang sangat cepat.

1.2 Rumusan Penelitian

Perumusan dalam pembuatan penelitian ini adalah "Bagaimana menghasilkan sebuah aplikasi berbasis *web* dengan implementasi *machine learning* yang mampu mengekstrak entitas dan mampu mengenali dengan tepat entitas yang muncul pada suatu pemberitaan di media daring?"

1.3 Batasan Penelitian

Batasan dalam pembuatan penelitian ini:

1. Bahasa pemrograman yang digunakan adalah *Python 3.6*;
2. Input data yang digunakan adalah berita yang diambil dari *Google News Indonesia*;
3. *NER* memanfaatkan *tools spaCy* dan *Prodigy*.
4. Proses mengumpulkan, *cleaning*, *training* dan *testing* data dilakukan secara terpisah.

1.4 Tujuan Penelitian

Tujuan utama dari penelitian ini adalah untuk membuat sebuah aplikasi berbasis *web* dengan implementasi *machine learning* yang dapat belajar dan memiliki kemampuan untuk melakukan ekstraksi entitas dari suatu berita di media daring dan juga mampu mengenali jenis dari entitas tersebut.

1.5 Manfaat Penelitian

Manfaat yang diharapkan dari penelitian ini antara lain:

1. Menemukan metode untuk meningkatkan akurasi *HMM POS Tag* dan *NER* Bahasa Indonesia;
2. Menambah pengetahuan dalam *machine learning* dan *text mining*;
3. Menerapkan ilmu pengetahuan yang telah didapat selama kuliah.

1.6 Metodologi Penelitian

Metodologi penelitian dibagi menjadi 2 tahap, Tahapan Pengumpulan Data, dan Tahapan Penelitian.

1.6.1 Tahapan Pengumpulan Data

Tahapan-tahapan yang dilakukan dalam pengumpulan data untuk penelitian ini adalah:

1.6.1.1 Menentukan Sumber

Pada tahapan ini akan dilakukan pemilihan untuk sumber berita yang baik, dengan kriteria baik adalah sumber berita yang memiliki akses yang cepat, dan memiliki konten berita yang lengkap.

1.6.1.2 Mengumpulkan Data

Pada tahap ini akan dibuat sebuah robot untuk mengunduh data dari sumber yang telah ditentukan sebelumnya.

1.6.2 Tahapan Penelitian

Tahapan-tahapan yang dilakukan dalam pembuatan penelitian ini adalah:

1.6.2.1 Studi Pustaka

1. Dengan mencari literatur, jurnal atau *paper* yang berkaitan dengan *Text Mining*, *HMM*, *POS Tagging*, dan *NER*;
2. Melakukan *sharing* dengan beberapa orang yang sudah ahli di bidang *Text Mining* dan *Entity Extraction*.

1.6.2.2 Analisis Sistem

Analisis sistem dilakukan dengan menuliskan kebutuhan fungsional dan non-fungsional dari sistem ini.

1.6.2.3 Perancangan

Melakukan percobaan-percobaan dengan kode program, mulai dari proses menyiapkan data, hingga pemilihan metode-metode yang tepat agar mendapatkan hasil yang maksimal dan dapat diimplementasikan dengan baik.

1.6.2.4 Implementasi

Mengimplementasikan hasil dari proses perancangan sebelumnya, dalam bentuk pembuatan kode program dengan bahasa pemrograman *Python 3.6* dan *framework Gunicorn* untuk dapat membuat sebuah aplikasi berbasis *web* guna memvisualisasikan hasil dari *Named Entity Recognition (NER)*.

1.6.2.5 Pengujian

Menguji program, yaitu dengan memberikan *corpus* baru pada mesin untuk mendeteksi akurasi dari *HMM Post Tagger* dan menggunakan tabel *Confussion Matrix* untuk mengukur *Precision, Accuracy, F1 Score, Recall* dan *Loss* untuk pengujian *NER*.

1.7 Sistematika Penulisan

Laporan penelitian ini akan disusun kedalam 5 bab, masing-masing bab tersebut akan disusun dan diurutkan sebagai berikut:

BAB I PENDAHULUAN

Pada bab ini akan diuraikan mengenai latar belakang masalah, perumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, metode pengumpulan data, dan sistematika penulisan penelitian.

BAB II LANDASAN TEORI

Pada bab ini akan diuraikan tentang teori-teori yang digunakan sebagai dasar dalam penulisan penelitian ini. Mencakup tinjauan pustaka yang meliputi pengertian *Text Mining, Post Tagging, HMM, dan NER.*

BAB III ANALISIS DAN PERANCANGAN

Pada bab ini akan diuraikan tentang perancangan aplikasi, model, serta cara kerja yang dilakukan dalam studi kasus ektrasi entitas berita daring di Indonesia.

BAB IV IMPLEMENTASI DAN PEMBAHASAN

Pada bab ini akan dibahas ektraksi entitas dari berita daring mulai dari *pre-processing*, *training*, *testing*, dan implementasi hasil dalam bentuk aplikasi berbasis *web*.

BAB V PENUTUP

Bab ini berisi kesimpulan yang didapat dari hasil analisis, perancangan dan implementasi dalam bentuk aplikasi serta saran untuk pengembangan yang lebih baik.