BAB I

PENDAHULUAN

1.1 Latar Belakang

Teknologi membawa perubahan baru yang signifikan dalam aktifitas keseharian manusia. Dengan penggunaan sarana teknologi seperti komputer, *smartphone, tablet*, ataupun perangkat digital lainnya semakin memudahkan akses terhadap berbagai informasi, seperti informasi pendidikan, kesehatan, hiburan ataupun bidang lainnya. Teknologi juga mendorong peningkatan produktivitas dalam kegiatan sehari-hari melalui penggunaan aplikasi maupun perangkat lunak yang mendukung kegiatan keseharian.

Pengenalan entitas bernama atau *Named Entity Recognition* (NER) merupakan salah satu turunan ilmu yang penting dalam pemrosesan bahasa alami (*Natural Language Processing*) (Jiang, Banchs & Li, 2016). Pengertian dari entitas dalam konteks ini dapat berupa objek nyata, meliputi nama, lokasi, organisasi, dan juga objek abstrak seperti sebuah konsep dan hubungan (Cheng & Iwahara, 2018). Pengidentifikasian entitas dalam sebuah dokumen teks bertujuan untuk menyederhanakan pencarian dan analisis informasi yang didasarkan pada labelisasi entitas terhadap setiap kata dalam teks sesuai jenis entitasnya (Setiyoaji et al., 2017).

Media sosial, khususnya *Instagram* menjadi salah satu platform paling popular untuk mengekspresikan opini, berbagi pengalaman maupun berkomunikasi lintas batas (Giarsyani et al., 2020). *Caption* pada postingan *Instagram* mengandung informasi tekstual yang berpotensi mengandung entitas lokasi, baik secara eksplisit maupun implisit. Pengidentifikasian entitas dalam sebuah dokumen teks bertujuan untuk memudahkan pencarian informasi yang didasarkan pada pembagian nama entitas terhadap tiap kata yang ada dalam teks. Informasi ini dapat dimanfaatkan lebih lanjut dalam konteks analisa spasial, pemetaan sosial, maupun pengambilan keputusan berbasis lokasi.

Namun demikian, proses indentifikasi *named entity* lokasi tidaklah mudah. Karakteristik bahasa dari teks yang diperoleh pada media sosial cenderung berbahasa informal, tidak baku, singkat dan sering kali berupa campuran dari berbagai bahasa atau dialek. Hal ini menjadi kendala proses ekstraksi entitas secara akurat. Tidak semua entitas lokasi bisa diperoleh langsung dalam format yang dikenali, sehingga dibutuhkan pendekatan yang mampu mengenali pola-pola linguistik secara fleksibel dan kontekstual.

Oleh karena itu, diperlukan metode yang mampu untuk melakukan klasifikasi dan identifikasi lokasi secara efektif dalam teks. Salah satu pendekatan yang bisa diterapkan adalah algoritma *Support Vector Machine* yang dikenal efektif dalam klasifikasi data berbasis teks.

Studi terdahulu terkait klasifikasi teks telah membuktikan efektifitas algoritma SVM dalam klasifikasi teks berbahasa Indonesia untuk mengkategorikan 3 bentuk pertanyaan berbahasa Indonesia yang mendapatkan nilai akurasi mencapai 0.92%, nilai *f1-score* sebesar 0.9%, presisi 0.93% dan *recall* 0.89% (Yusliani & Syahrini, 2022).

Pada penelitian Putra (Putra & Kurniawan, 2020) dilakukan ekstraksi entitas lokasi, orang, organisasi dan detail kejadian dari berita online yang berkaitan dengan kebakaran. Masing-masing data entitas diklasifikasikan ke dalam kelas tertentu dengan label *loc, per, org,* dan *misc*, dengan metode *Bidirectional* LSTM – CNNs dipilih sebagai metode yang diterapkan. Performa yang didapat dari penelitian tersebut sebesar 75% angka akurasi, *recall* dan *presicion.* Serta berhasil menampilkan titik persebaran lokasi kebakaran hasil klasifikasi dari rentang tanggal 1 Januari 2020 sampai 20 April 2020.

Selanjutanya berdasarkan penelition Setiyoaji (Setiyoaji et al., 2017) diekstraksi entitas nama, tempat, zat dan guna yang ada pada teks tanaman obat. Menggunakan metode *Hidden Markov Model* (HMM) dan Algorima Viterbi diperoleh hasil akhir sebesar *average precision* 0.5447, *recall* 0.7402 dan nilai *f-measure* sebasar 0.5606.

Kemudian pada penelitian oleh (Dirgantara et al., 2018) dilakukan pengenalan pada 6 (enam) entitas, yaitu entitas merek, entitas tipe, entitas harga, entitas spek,

entitas n_spek dan entitas n_tag. Penelitian ini berfokus pada ekstraksi informasi seputar fitur produk ponsel pada *e-commerce*. Hasil akhir yang diperoleh dari penerapan *rule template* sebesar 97,20% tingkat akurasi. Adapun untuk *Hidden Markov Model* memberikan nilai akurasi 92,23%.

Dari penelitian-penelitian yang telah dilaksanakan sebelumnya, penulis tertarik untuk menggunakan SVM. Algoritma ini adalah salah satu bagian dari metode *machine learning* pada kelas *supervised learning* yang sering digunakan untuk pengklasifikasian data (Giarsyani et al., 2020). SVM adalah algoritma dengan performa yang baik untuk digunakan dalam pengkasifikasian teks (Purnamawan, 2015).

Penelitian berfokus pada analisis teks postingan *Instagram*, dengan entitas lokasi sebagai analisa utama dalam pendekatan *Named Entity Recognition* (NER). Setiap teks postingan *Instagram* akan diklasifikasikan ke dalam dua kelas, yaitu *Contains_Location* dan *No_Location*, berdasarkan keberadaan entitas lokasi yang ada dalam teks. Untuk mendukung proses klasifikasi ini, model SVM digunakan dan dilatih menggunakan data *caption* yang telah dilabelisasi sebelumnya.

Dari latar belakang yang telah dijabarkan diatas, penulis tertarik membuat penelitian yang judul "***Named Entity* Lokasi dalam Teks Postingan *Instagram* Menggunakan Algoritma *Support Vector Machine*"** yang bertujuan agar model mampu mengidentifikasikan dan mengklasifikasikan teks secara otomatis berdasarkan ada atau tidaknya informasi lokasi yang terkandung dalam teks postingan *Instagram*, yang kedepannya bisa dikembangkan untuk pendeteksian entitas lokasi yang lebih spesifik.

1.2 Rumusan Masalah

Dari penjabaran latar belakang di atas rumusan masalah yang akan dibahas yaitu:

1. Bagaimana mengimplementasikan *Named Entity Recognition rule-based* dengan Algoritma *Support Vector Machine* dalam mengidentifikasikan dan mengklasifikasikan entitas lokasi pada teks postingan *Instagram*?
2. Bagaimana bentuk pola yang diberikan pada NER *rule-based* agar bisa mengekstraksi entitas lokasi eksplisit pada teks postingan *Instagram*?

3. Bagaimana evaluasi kinerja yang dihasilkan dari implementasi NER *rule-based* dengan Algoritma *Support Vector Machine* dalam mengidentifikasikan dan mengklasifikasikan entitas lokasi pada teks postingan *Instagram*?

1.3 Batasan Masalah

Berdasarkan latar belakang dan rumusan masalah di atas, penulis memberikan batasan-batasan masalah sebagai berikut:

1. Data diperoleh dari teks postingan *Instagram* yang memuat deskripsi postingan.
2. Data yang dianalisis adalah teks berbahasa Indonesia, sehingga sistem tidak dirancang untuk mendeteksi entitas lokasi dalam bahasa lain seperti bahasa Inggris atau bahasa daerah.
3. Data diambil sejumlah 400 data postingan, dengan pembagian 70% data testing dan 30% data uji.
4. Jenis kata lokasi yang diklasifikasikan berupa lokasi eksplisit, seperti provinsi, kabupaten/kota, dan nama tempat yang terdaftar resmi secara geografis/administratif.
5. Diagram yang digunakan dalam perancangan berupa Diagram Konteks, DFD, dan *Table Relation*.
6. Sistem akan dibangun dengan menggunakan bahasa pemrograman *Ruby* dan *framework Ruby on Rails*.
7. Evaluasi model hanya dilakukan pada data uji terbatas yang diambil dari hasil *crawling* dan belum mencakup analisis terhadap data besar secara *real-time*.

1.4 Tujuan Penelitian

Tujuan yang ingin dicapai pada penelitian ini yaitu:

1. Mengimplementasikan *Named Entity Recognition rule-based* dan Algoritma *Support Vector Machine* dalam pengidentifikasian dan klasifikasi entitas lokasi pada teks postingan *Instagram*.
2. Menerapkan pola identifikasi kata lokasi pada NER *rule-based* untuk mengekstraksi entitas lokasi eksplisit.

3. Menganalisa tingkat akurasi yang diperoleh dari hasil implementasi NER *rule-based* dengan Algoritma *Support Vector Machine* dalam mengidentifikasikan dan mengklasifikasikan entitas lokasi pada teks postingan *Instagram*.

1.5 Manfaat Penelitian

Manfaat dari penelitian ini adalah sebagai berikut:

1. Menghasilkan Sistem Informasi yang bisa mengidentifikasikan dan mengklasifikasikan entitas lokasi dari data postingan teks *Instagram*.
2. Sebagai sarana bagi penulis untuk dapat menerapkan pengetahuan selama menempuh pendidikan studi Teknik Informatika di Universitas Malikussaleh.
3. Sebagai bahan referensi dan tambahan ilmu pengetahuan tentang penelitian terkait *Named Entity Recognition* (NER) dan klasifikasi data tekstual bagi para pembaca.