BAB I

PENDAHULUAN

1.1 Latar Belakang

Pada era teknologi saat ini, informasi tersedia secara melimpah dalam berbagai bidang. Kemudahan dalam menyebarkan informasi yang ditunjang oleh perkembangan *User Generate Content* (UGC) menjadi salah satu faktor penyebabnya. Pada UGC, user (pengguna) dapat menyebarkan informasi dengan mudah karena UGC menyediakan layanan untuk men-generate *content*-nya sendiri. *Content* yang dimaksud seperti mengunduh gambar, musik, video dan tulisan pada media tertentu. Salah satu media sosial yang mengkudung UGC adalah twitter yang pada akhirnya menjadi bagian dari kehidupan sehari-hari.

Twitter merupakan mikroblog atau media *sharing* informasi yang banyak digunakan dalam penyebaran informasi. Penelitian Semiocast, lembaga riset media sosial yang berpusat di Paris, Prancis, menyatakan bahwa Indonesia adalah pengguna twitter terbesar kelima di dunia dengan jumlah akun 19,5 juta (Semiocast, 2010). Selain jumlah akun, jumlah *tweet* yang dihasilkan pun terus meningkat. Beberapa referensi menyatakan kurang lebih *tweet* yang dihasilkan mencapai 400 juta per hari dengan beragam topik yang sedang hangat pada masa itu.

Ketersediaan informasi yang melimpah tersebut pada satu sisi dapat bermanfaat. Namun di sisi lain, dapat menimbulkan masalah seperti berlebihnya informasi yang diterima atau dikenal sebagai *information overload*. Kondisi ini

adalah kondisi dimana banyak informasi yang diterima tapi tidak dibutuhkan. Untuk itu diperlukan teknik dalam memilah atau mengklasifikasi informasi dari sekian banyak informasi yang disediakan. Teknik ini dikenal sebagai ektraksi informasi atau pengambilan informasi pada data tekstual. Informasi yang diambil dapat berupa *event*, entitas atau relasi pada setiap teks.

Fungsi ekstraksi informasi adalah mencari kata (token) dan entitas yang dapat mewakili isi dari data tekstual. *Named Entity Recognition* (NER) merupakan komponen dasar dari ekstraksi informasi yang bertugas untuk mengenali entitas tersebut. Entitas yang dikenali nantinya dapat dimanfaatkan sebagai metadata untuk tahap selanjutnya pada ekstraksi informasi. Selain itu, dimanfaatkan di dalam peringkasan dokumen (*summarize*), *profiling* atau *event detection*.

Metode yang digunakan dalam penelitian tentang NER pun beragam. Salah satu metode pembelajaran mesin yang digunakan dalam ekstraksi informasi adalah perceptron, yang merupakan bagian dari model *neural network*. Perceptron mempunyai kecepatan komputasi dalam mengklasifikasi objek karena menggunakan pendekatan linier yang membagi objek kedalam dua kelas. Perceptron pernah digunakan pada penelitian Ciaramita dan Altun pada tahun 2005 untuk dokumen formal dalam hal ini sebuah novel.

Berdasarkan permasalahan di atas, penelitian ini lebih difokuskan pada pengenalan entitas atau *named entity recognition* pada twitter berbahasa Indonesia menggunakan metode perceptron. Diharapkan penelitian ini dapat memberikan manfaat untuk para peneliti ektraksi informasi khususnya dalam mengekstraksi informasi pada twitter.

1.2 Rumusan Masalah

Rumusan masalah pada penelitian ini adalah:

1. Bagaimana mengembangkan sistem yang mampu mengenali entitas nama orang dan lokasi pada *tweet* berbahasa Indonesia?
2. Bagaimana algoritma perceptron dapat mengklasifikasikan entitas pada tipe entitas nama orang dan lokasi?

1.3 Tujuan Penelitian

Tujuan yang ingin dicapai pada penelitian ini adalah:

1. Membuat sistem yang mampu mengenali entitas nama orang dan lokasi pada *tweet* berbahasa Indonesia.
2. Menggunakan algoritma perceptron untuk mengklasifikasikan entitas nama orang dan lokasi pada tipe entitasnya.

1.4 Batasan Masalah

Berikut beberapa batasan masalah dari penelitian ini:

1. Penelitian lebih difokuskan pada *Named Entity Recognition*.
2. Data yang digunakan diambil dari mikroblog yaitu *tweet*.
3. Algoritma yang digunakan adalah Perceptron.
4. Kategori entitas yang diteliti hanya nama orang dan lokasi.
5. Analisa pemodelan menggunakan pemodelan berorientasi objek yaitu UML.
6. Sistem dikembangkan dengan bahasa pemograman Java.

1.5 Metodologi Penelitian

Tahapan yang akan dilalui pada skripsi ini adalah sebagai berikut:

1. **Studi Literatur**, dilakukan dengan mengkaji NER dan model perceptron dari berbagai sumber.
2. **Pengumpulan Informasi**, dilakukan dengan wawancara pada beberapa narasumber terkait dengan NER.
3. **Analisa dan Perancangan Sistem**, dilakukan analisa dan perancangan sistem NER termasuk fitur-fitur apa saja yang mempengaruhi pengenalan entitas.
4. **Implementasi Sistem**, dilakukan implementasi berdasarkan hasil analisa dan perancangan dengan menggunakan bahasa pemograman Java.
5. **Pengujian dan Evaluasi**, dilakukan pengujian pada sistem yang telah dibuat, kemudian hasilnya dievaluasi.

1.6 Sistematika Laporan

Laporan disusun secara sistematis sehingga mudah dibaca, ditelusuri, dan dievaluasi. Sistematika penulisan laporan skripsi ini terbagi menjadi lima bab sebagai berikut:

BAB I Pendahuluan

Bab ini membahas latar belakang, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, metodologi penelitian dan sistematika laporan.

BAB II Kajian Pustaka

Bab ini membahas teori-teori yang mendukung dalam penyusunan skripsi seperti NER, model perceptron dan beberapa contoh dari penelitian yang ada.

BAB III Metodologi Penelitian

Bab ini menguraikan metode yang digunakan dalam penelitian secara rinci.

BAB IV Hasil Penelitian dan Pembahasan

Bab ini menguraikan tahapan yang harus dilalui mulai dari *preprocessing* data twitter sampai sebuah entitas ditemukan di dalamnya. Tahapannya akan dijelaskan dengen rinci dan mendalam.

BAB V Kesimpulan dan Saran

Bab ini menguraikan beberapa kesimpulan dari hasil penelitian untuk menjawab rumusan masalah. Pada bagian saran, diisi rekomendasi dari penulis untuk penelitian selanjutnya.