BAB I

PENDAHULUAN

Part-of-speech tagging (POS), juga disebut gramatikal tagging adalah proses penugasan bagian dari tag ucapan untuk kata-kata dalam teks. Part-of-speech adalah kategori tata bahasa, umumnya termasuk kata kerja, kata benda, kata sifat, kata keterangan, dan sebagainya. Penandaan sebagian ucapan merupakan alat penting dalam banyak aplikasi pemrosesan bahasa alami seperti disambiguasi kata, penguraian, penjawaban pertanyaan, dan terjemahan mesin. *Part of Speech tagging* (POS tagging) memiliki peran penting dalam berbagai bidang pemrosesan bahasa alami *(NLP)* termasuk terjemahan mesin untuk memproses sentimen analisis. Tagset umum yang terdapat pada *Penn Treebank POS Tag* memiliki 37 tag[1]. POS Tagger ini merupakan tools yang penting dalam pemrosesan Bahasa sehingga perlu untuk dikembangkan dengan akurasi yang baik.

Indonesia adalah negara multibahasa yang luas dengan beragam budaya. Ini memiliki banyak bahasa dengan bentuk tertulis dan bahasa yang digunakan di masing-masing daerah. Dan bahasa Jawa adalah salah satu bahasa daerah di Indonesia. Penelitian bahasa Jawa ini menggunakan dataset dari media berita online. Karakteristik berita online terletak pada panjangnya kalimat. Penelitian POS tagging menggunakan bahasa Jawa juga sudah ada sebelumnya tetapi menggunakan metode basis aturan *(Rule base)*.

Dalam penelitian ini mengembangkan POS Tagging menggunakan *Conditional Random Fields* (CRF). CRF telah diterapkan pada sejumlah besar domain lain, termasuk pemrosesan teks, visi komputer, dan bioinformatika. Sejak itu, *linear-chain* CRF telah diterapkan pada banyak masalah dalam pemrosesan bahasa alami, termasuk pengenalan entitas bernama, fitur induksi untuk NER, dekomposisi dangkal, segmentasi alamat pada halaman Web [2].

Studi sebelumnya telah mengevaluasi metode CRF menggunakan tweet Indonesia di Twitter dengan hasil akurasi yang cukup bagus yaitu 71%[3]. Namun, dalam penelitian itu teks yang digunakan menggunakan teks tweet. Dalam penelitian ini, teknik yang sama untuk POS Tagger dalam bahasa Jawa akan digunakan dengan harapan bahwa hasilnya juga akan menghasilkan akurasi yang relatif baik. Adapun penelitian bahasa Jawa ini menggunakan tag set pada media

berita online. Karakteristik media berita online sangat berbeda dibandingkan dengan menggunakan tweet. Perbedaannya terletak pada panjang kalimat, di mana teks dalam tweet relatif lebih pendek atau lebih pendek. Masalah utama dari penelitian ini bagaimana metode CRF dapat memberikan akurasi paling baik untuk *POS Tag* Bahasa Jawa.