BAB 1

PENDAHULUAN

1.1 Latar Belakang

Perkembangan pada pengolahan bahasa alami di Indonesia membuat korpus bahasa Indonesia menjadi sumber yang penting sebagai data untuk melakukan pengolahan dan data penelitian, korpus sendiri merupakan kumpulan dokumen teks yang dijadikan sebagai dataset standar bagi para peneliti pengolahan bahasa alami tersebut.Bahasa memang berfungsi sebagai alat refleksi dan ekspresi terhadap nilai budaya masyarakat, melalui bahasa pula manusia memperbarui proses berpikir dan secara simultan teknologi berperan besar memperbaiki kualitas dan kuantitas pengembangan bahasa.Untuk kasus di Indonesia sumber daya data penelitian berbahasa Indonesia masih terbatas, hal ini diutarakan oleh Ketua Jurusan Teknik Informatika Fakultas Teknologi Industri Universitas Trisakti Anung B. Ariwibowo[27]. Akan tetapi literature penulisan dapat ditemukan dalam beberapa *platform* daring seperti media sosial ataupun portal berita.

Dengan banyaknya pengguna *internet* di Indonesia yang renta usia penggunanya mulai dari usia 16 tahun hingga usia 64 tahun dengan masing-masing jenis perangkat seperti *mobile phone* sebesar 96%, *smartphone* sebesar 94%, *non-smartphone mobile* phone sebesar 21%, laptop dan komputer sebesar 66%, tablet sebesar 23%, konsol gim sebesar 16%, dan *virtual reality device* sebesar 5,1%[1]. Akan tetapi data menunjukan bahwa masyarakat Indonesia memiliki presentase kepemilikan ponsel sebesar 125% lebih besar atau sebanyak 338.2 juta koneksi *intenet* melalui *mobile phone* dibandingkan dengan jumlah penduduknya[1].

Banyaknya orang yang terhubung dengan koneksi internet membuat banyak orang dapat aktif dalam media sosial dan Indonesia memiliki 160 juta pengguna

aktif media sosial[1], penggunaan aktif dalam media sosial membuat banyak masyarakat bebas dalam menuliskan segala pikiran mereka untuk menilai segala sesuatu yang terjadi dalam media sosial. Dilakukannya perbandingan *POS-Tagger* pada penelitian untuk menghasilkan aplikasi yang dapat membantu pengembangan korpus dalam bahasa Indonesia dan melihat hasil *tagger* yang dapat mengeluarkan hasil keluaran *tagger* yang optimal berdasarkan tiga metode yang digunakan yaitu *N-Gram*, *TnT*, dan *Classifier Based*. Dengan menggunakan data teks yang dapat diperoleh dari media sosial ataupun portal berita daring.

Masyarakat Indonesia yang menggunakan koneksi *internet* banyak menghabiskan waktu mereka untuk menggunakan media sosial sebagai wadah mereka dalam mencari informasi ataupun berkomentar, adapun media sosial seperti *facebook, youtube, whatsapp, instagram, twitter,* portal berita daring, dan masih banyak lainnya. Pada media sosial banyak sekali informasi dan hiburan yang dicari oleh penggunanya untuk memuaskan keinginan mereka. Membuat berbagai macam tulisan daring yang dapat diperoleh untuk dijadikan data.

Dari banyaknya data di atas, kalimat komentar yang terdapat pada media sosial ataupun tulisan informasi berita yang terdapat pada portal berita daring dapat dijadikan media pembelajaran untuk machine learning pada pengolahan teks yaitu *Natural Language Processing* dengan menjadikan kalimat-kalimat tersebut sebagai data teks yang akan dijadikan *corpus* bahasa Indonesia dan dapat digunakan untuk pengolahan teks pada *machine learning*.

Untuk *literature review* pada jurnal yang berjudul "*Investigating Bi-LSTM and CRF with POS Tag Embedding for Indonesian Named Entity Tagger*" tahun 2018 oleh Devin Hoesen dan Ayu Purwarianti, metode yang digunakan merupakan *Bi-directional long short term memory* dan *Condotional Random Field* dengan menggunakan data teks sebanyak 8400 kalimat sebagai arsitekturnya untuk membangun named entity. Pada jurnal yang berjudul "*On Part of Speech Tagger for Indonesian Language*" oleh R.Sandra Yuwana, Asri R.Yuliani dan Hilman F.Pardede menggunakan metode *unigram, hidden markov metode, TnT, Brills. Naïve Bayes* dan *Maximum Entropy*.

Pada penelitian, *dataset* yang digunakan bersumber dari *github* dan dapat diperoleh secara daring dengan cara mengunduh *dataset* tersebut, *dataset* yang digunakan dalam penelitian merupakan *dataset POS-Tag* Indonesia yaitu dataset *Indonesian_Manually_Tagged_Corpus.tsv* yang sudah diimplementasikan oleh Budi Harta dalam penelitian *POS-Tagger* menggunakan metode HMM(*Hidden Markov Models*)[28].

1.2 Rumusan Masalah

Berdasarkan latar belakang yang sudah dijelaskan berikut rumusan masalah dari penelitian ini adalah:

1. Bagaimana cara implementasi pengolahan *POS-Tag* dapat dilakukan?
2. Bagaimana hasil perbandingan yang didapatkan dari *POS-Tagging* dengan menggunakan metode *N-Gram, TnT,* dan *Classifier Based*?

1.3 Batasan Masalah

Berdasarkan latar belakang yang sudah dijelaskan berikut batasan masalah dalam penelitian ini adalah:

1. Data yang digunakan merupakan *dataset Indonesian_Manually_Tagged_Corpus.tsv* bahasa Indonesia.
2. Metode yang digunakan adalah *N-Gram, TnT,* dan *Classifier Based*.
3. Aplikasi *POS-Tag* ditampilkan dalam bentuk *GUI.*
4. Penelitian ini menggunakan *dataset* berupa teks dalam bahasa Indonesia
5. Proses melakukan *tagging* dipengaruhi oleh kalimat yang dimasukan oleh pengguna.
6. Perbedaan penggunaan *dataset* dapat mempengaruhi pada perbedaan tagging pada teks

1.4 Tujuan Penelitian

Berdasarkan latar belakang yang sudah dijelaskan tujuan dalam penelitian ini adalah:

1. Mengimplementasikan pengolahan *POS-Tag*.

2. Menghasilkan perbandingan *POS-Tag* pada metode *N-Gram*, *TnT*, dan *Classifier based*.

1.5 Manfaat Penelitian

Berdasarkan latar belakang yang sudah dijelaskan manfaat yang diberikan dalam penelitian ini adalah :

a. Manfaat Akademik
    1. Sebagai alat atau wadah untuk melatih dan mengembangkan kemampuan penulis dalam bidang programming dan data analisis
    2. Penelitian ini diharapkan menjadi inspirasi bagi peneliti lain untuk dikembangkan kembali aplikasi ini.
b. Manfaat Praktis
    1. Mengembangkan aplikasi *POS-Tagging* bahasa Indonesia.
    2. Memudahkan pengguna untuk menentukan *tagging* pada teks dengan menggunakan *dataset*

1.6 Sistematika Penulisan

BAB 1 PENDAHULUAN

Penelitian ini diangkat dari banyaknya karya tulis yang dapat dijadikan media pembelajaran untuk mengembangkan *corpus* bahasa Indonesia dengan menggunakan *tagging* pada teks.

BAB 2 TINJAUAN PUSTAKA

Bab ini membahas kajian pustaka berdasarkan sumber-sumber seperti buku dan jurnal. Kajian pustaka dalam penelitian mengenai penelitian terdahulu, *library* yang digunakan, serta teori yang akan digunakan untuk mendukung penelitian ini.

BAB 3 METODE PENELITIAN

Bab ini berisi tentang uraian kerangka pemikiran mengenai langkah-langkah dalam penelitian dan metode pengembangan aplikasi *POS-Tagging* berbahasa Indonesia.

BAB 4 HASIL DAN PEMBAHASAN

Bab ini berisi uraian hasil pengembangan aplikasi yang dicapai berdasarkan metode penelitian berupa hasil dari setiap tahapan yang dilalui serta tampilan antarmuka aplikasi akhir serta melakukan uji coba aplikasi dan pembahasan dari penelitian.

BAB 5 SIMPULAN DAN SARAN

Bab ini berisi uraian kesimpulan yang dicapai dari hasil penelitian yang telah dilakukan dan merumuskan saran atau rekomendasi yang tepat dari peneliti yang dapat digunakan dalam penelitian selanjutnya.