BAB I
PENDAHULUAN

1.1. Latar Belakang

Komunikasi merupakan salah satu hal paling penting yang dibutuhkan manusia sebagai makhluk sosial. Manusia berkomunikasi dengan menggunakan bahasa alami, yaitu bahasa yang secara umum digunakan baik itu dengan cara lisan, tulisan ataupun menggunakan isyarat. Secara umum komunikasi merupakan suatu proses ketika seseorang atau beberapa orang menciptakan dan menggunakan informasi agar terhubung dengan lingkungan dan orang lain (Ruben & Stewart, 2015). Dalam suatu negara, masyarakat umumnya berkomunikasi dengan menggunakan bahasa resmi negara tersebut. Ada banyak sekali bahasa yang ada di dunia salah satunya adalah Bahasa Indonesia. Bahasa Indonesia merupakan bahasa resmi negara Indonesia sebagai identitas bangsa dan lambang kebanggan nasional, yang secara luas dan umum digunakan sebagai alat komunikasi oleh 222 juta orang (Lewis, 2009). Indonesia memiliki lebih dari 742 bahasa daerah yang berbeda-beda, sehingga bahasa Indonesia merupakan bahasa pemersatu bagi masyarakat Indonesia (Lewis, 2009).

Pesatnya perkembangan teknologi pada era digital ini dan pentingnya bahasa Indonesia bagi masyarakat di negara tersebut, membuat pemrosesan terhadap bahasa Indonesia kedalam berbagai aplikasi yang membantu manusia sangat dibutuhkan. Sehingga, pengembangan sistem dan penelitian dibidang pemrosesan bahasa alami untuk bahasa Indonesia bagi masyarakat luas menjadi penting.

Pemrosesan bahasa alami atau *natural language processing* (NLP) merupakan suatu pengembangan teknik komputasi bahasa alami dalam menganalisis dan merepresentasikan teks ataupun lisan untuk mencapai pemrosesan bahasa seperti bahasa manusia (Liddy, 2001). Salah satu *task* dalam pemrosesan bahasa alami yaitu proses pelabelan kata dalam suatu kalimat berdasarkan kategori katanya, atau yang disebut dengan *part-of-speech* (PoS) *tagger*. Pemberian *tag* untuk setiap kata secara manual akan memakan banyak waktu, melelahkan dan dengan biaya yang mahal

karena diperlukannya ahli bahasa untuk memvalidasinya. Oleh karena itu, diperlukan pengembangan PoS *tagger* untuk meningkatkan waktu serta menghemat biaya. Implementasi *part-of-speech* (PoS) *tagger* sebagai dasar atau *taks* penting dalam aplikasi NLP, sudah banyak digunakan dalam berbagai perangkat lunak pengolah bahasa seperti aplikasi *grammar checker*, *speech recognition*, *question answering* dan *machine translation* (Pisceldo, Adriani, & Manurung, 2009).

Menurut Jurafsky dan Martin (2014), PoS *tagger* merupakan proses penandan *part-of-speech* untuk setiap kata pada teks kalimat masukan dalam suatu bahasa tertentu. Sebagai contoh kalimat masukan dalam bahasa Indonesia yaitu "*Saya pergi ke kebun binatang*", maka sistem PoS *tagger* secara otomatis akan memberikan *tag* kelas kata pada setiap katanya dengan pemisah berupa garis miring (/) menjadi *saya*/PRP (*personal pronouns*) *pergi*/VB (*perb*) *ke*/IN (*prepotition*) *kebun*/NN (*pommon Noun*) *binatang*/NN (*common noun*), yang dalam bahasa Indonesia *personal pronouns* berarti kata ganti orang, *verb* berarti kata kerja, *prepotition* atau kata depan dan *common noun* berarti kata benda, berdasarkan *tagset* dari penelitian yang dilakukan oleh Rashel, Luthfi, Dinakaramani, & Manurung (2014).

Salah satu metode PoS *tagger* yang telah dikembangkan (Kumar & Shekhawat, 2018) adalah PoS *tagger* dengan pendekatan *probabilistic-based* menggunakan metode *Hidden Markov Model* (HMM). *Hidden Markov Model* merupakan pengembangan dari *Markov Model* yang mengasumsikan bahwa kata secara probabilistik bergantung pada dua atau lebih kategori kata sebelumnya. Karena beberapa kata dalam bahasa Indonesia memiliki ejaan yang sama dengan kata lainnya tetapi memiliki makna berbeda (homograf), sehingga sangat mungkin sebuah kata memiliki dua atau lebih kategori kata. Sebagai contoh kata *tahu* dalam kalimat "*Saya memberi tahu pada Shinta*", kata *tahu* memiliki dua makna yang berbeda yaitu *tahu* sebagai kata benda dan *tahu* sebagai kata kerja. Sehingga, berdasarkan pada contoh kasus tersebut, algoritma *hidden markov model* akan menentukan *tag* kelas kata mana yang akan dipilih berdasarkan nilai parameter yang ada.

Brants (2000) berpendapat bahwa masalah utama PoS *tagger* disebabkan oleh adanya kata yang merupakan *out-of-vocabulary* (OOV) pada saat proses masukan. O*ut-of-vocabulary* merupakan kata yang tidak

dikenali kelas katanya oleh sistem, yang disebabkan karena kata tersebut tidak terdapat dalam *training corpus* tetapi ada dalam *testing corpus* (Muljono, Afini, & Supriyanto, 2017). Dengan *training corpus* bahasa Indonesia yang terbatas dibandingkan dengan kata dalam bahasa Indonesia yang sangat banyak, tentu sangat mungkin munculnya kata *out-of-vocabulary* (OOV). Sehingga, diperlukan suatu metode untuk dapat menyelesaikan masalah OOV tersebut.

Pada tahun 2010, Wicaksono dan Purwarianti (2010) melakukan penelitian mengenai PoS *tagger* bahasa Indonesia dengan meggunakan metode *Hidden Markov Model*, yang menghasilkan nilai akurasi sebesar 90,65% terhadap data *testing* yang mengandung 15% tingkat OOV. Sedangkan, untuk data *testing* dengan 30% tingkat OOV, menghasilkan akurasi sebesar 83,28%. Sehingga dapat disimpulkan bahwa semakin tinggi tingkat OOV yang terkandung dalam data *testing*, maka akan semakin rendah nilai akurasi yang dihasilkan oleh sistem PoS *tagger* tersebut. Pengertian akurasi itu sendiri, yaitu tingkat kedekatan pengukuran kuantitas terhadap nilai yang sebenarnya.

Salah satu bentuk kata yang paling banyak muncul sebagai OOV *word* dalam bahasa Indonesia yaitu kata yang dihasilkan dari proses morfologi yaitu proses pembentukan kata (Muljono, Afini, & Supriyanto, 2017). Proses morfologi yang dimaksud salah satunya yaitu afiksasi atau proses pembentukan kata yang memiliki imbuhan seperti kata *membantu* yang memiliki imbuhan *mem-* atau *berlari* yang memiliki imbuhan *ber-*. Kata yang memiliki imbuhan *mem-* dan *ber-* termasuk kedalam kelas kata *verba* atau kata kerja. Sehingga, dapat disimpulkan bahwa imbuhan (afiks) dapat menjadi panduan dalam proses penentuan kelas kata. Oleh karena itu, berdasarkan panduan tersebut analisis morfologi dapat menjadi solusi untuk menangani permasalahan OOV dalam sistem PoS *tagger* menggunakan HMM.

Berdasarkan pada masalah dan studi literatur diatas, maka pada penelitian ini penulis akan mengembangkan sistem PoS *tagger* bahasa Indonesia dengan pendekatan *probabilistic-based* menggunakan metode *Hidden Markov Model* (HMM) dan didukung dengan metode analisis morfologi untuk menangani permasalahan OOV.

Dengan diterapkannya algoritma *hidden markov model* dan metode analisis morfologi pada proses PoS *tagger*, diharapkan dapat meningkatkan kinerja pada *taks* NLP tersebut. Sehingga, berbagai aplikasi pengolah bahasa khususnya dalam bahasa Indonesia akan menjadi lebih baik.

1.2. Rumusan Masalah

Berdasarkan latar belakang permasalah yang telah dipaparkan pada subbab sebelumnya, maka rumusan masalah pada penelitian ini adalah:

a. Apakah metode analisis morfologi mampu menangani permasalahan OOV dalam sistem PoS *tagger* bahasa Indonesia menggunakan *Hidden Markov Model*?
b. Bagaimana sistem kerja metode analisis morfologi dalam menentukan kelas kata?
c. Bagaimana hasil kinerja sistem *PoS tagger* menggunakan algoritma *Hidden Markov Model* dengan metode analisis morfologi dibandingkan dengan tanpa analisis morfologi?

1.3. Tujuan Penelitian

Adapun tujuan dari penelitian ini berdasarkan rumusan masalah yang telah disebutkan, adalah sebagai berikut:

a. Mendesain model untuk sistem metode analisis morfologi dalam menentukan kelas kata.
b. Membangun sistem PoS *tagger* bahasa Indonesia menggunakan algoritma *hidden markov model* (HMM).
c. Mengimplementasikan metode analisis morfologi pada sistem PoS *tagger* bahasa Indonesia menggunakan *hidden markov model* (HMM).
d. Melakukan analisa terhadap hasil eksperimen yang dilakukan.

1.4. Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan manfaat dalam aspek teorits maupun praktis sebagai berikut:

a. Diharapkan dapat memberikan pengetahuan melalui pendekatan dan metode-metode yang digunakan, sehingga dapat memberikan pengetahuan dalam aspek teoritis pada bidang *natural language processing* khususnya sistem PoS *tagger* untuk bahasa Indonesia.
b. Metode analisis morfologi yang diterapkan pada peneltian ini diharapkan dapat menjadi referensi bagi peneliti lainnya dalam penanganan *out-of-vocabulary* (OOV).

1.5. Batasan Masalah

Adapun batasan masalah yang ditetatapkan dalam penelitian ini adalah sebagai berikut:

a. Bahasa yang menjadi objek utama penelitian ini adalah bahasa Indonesia.
b. Penelitian hanya dilakukan pada dataset berupa teks kalimat.
c. Sistem yang akan dirancang bekerja dalam proses pemberian *tag* kelas kata (PoS *tagger*).
d. Aturan analisis morfologi berdasarkan pada aturan morfologi dalam Bahasa Indonesia.

1.6. Sistematika Penulisan

Sistematika penyusunan penulisan skripsi ini terbagi menjadi lima bab, sesuai yang diterapkan di Universitas Pendidikan Indonesia. Bab-bab tersebut meliputi Bab I : Pendahuluan; Bab II : Tinjauan Pustaka; Bab III : Metodologi Penelitian; Bab IV : Hasil Penelitian dan Pembahasan; dan Bab V : Kesimpulan dan Saran.

BAB I PENDAHULUAN

Bab ini berisi latar belakang penelitian, rumusan masalah, batasan masalah, tujuan penelitian yang akan dilakukan, serta sistematika penulisan. Latar belakang pembahasan masalah umum yang diangkat pada penelitian dan pengambilan judul skripsi ini.

BAB II KAJIAN PUSTAKA

Bab ini berisi mengenai kajian teori dan konseop metode serta algoritma yang digunakan dalam penelitian, meliputi dasar teori mengenai Kalimat pada bahasa Indonesia, Kelas Kata dalam Bahasa Indonesia, teori mengenai *natural language processing* (NLP) mulai definisi, karakteristik, hingga contoh perangkat lunak yang ada, *part-of-speech* (PoS) *tagger, hidden markov model* (HMM) hingga pendekatan dan algoritmanya, *out-of-vocabulary* (OOV) serta analisis morfologi bahasa Indonesia yaitu mengenai afiksasi atau analisis pembentukan kata pada kata yang memiliki imbuhan.

BAB III METODOLOGI PENELITIAN

Bab ini berisi dasar teori mengenai metodologi yang digunakan untuk melakukan penelitian dan penjelasan langkah-langkah yang akan dilakukan selama penelitian. Metodologi penelitian ini meliputi pengumpulan data teks sebagai data uji, analisis, alat dan bahan penelitian, dan metode penelitian yang terdiri dari teknik pengumpulan data dan proses pengembangan perangkat lunak.

BAB IV HASIL PENELITIAN DAN PEMBAHASAN

Berisi penjelasan dari hasil penelitian yang telah dilakukan yaitu proses pengumpulan data penelitian, rancangan sistem, implementasi atau pengembangan sistem dan pengujian.

BAB V KESIMPULAN DAN SARAN

Berisi kesimpulan dan saran yang didapat dari penelitian dari mulai merumuskan masalah sampai dengan selesai.