BAB III
METODOLOGI PENELITIAN

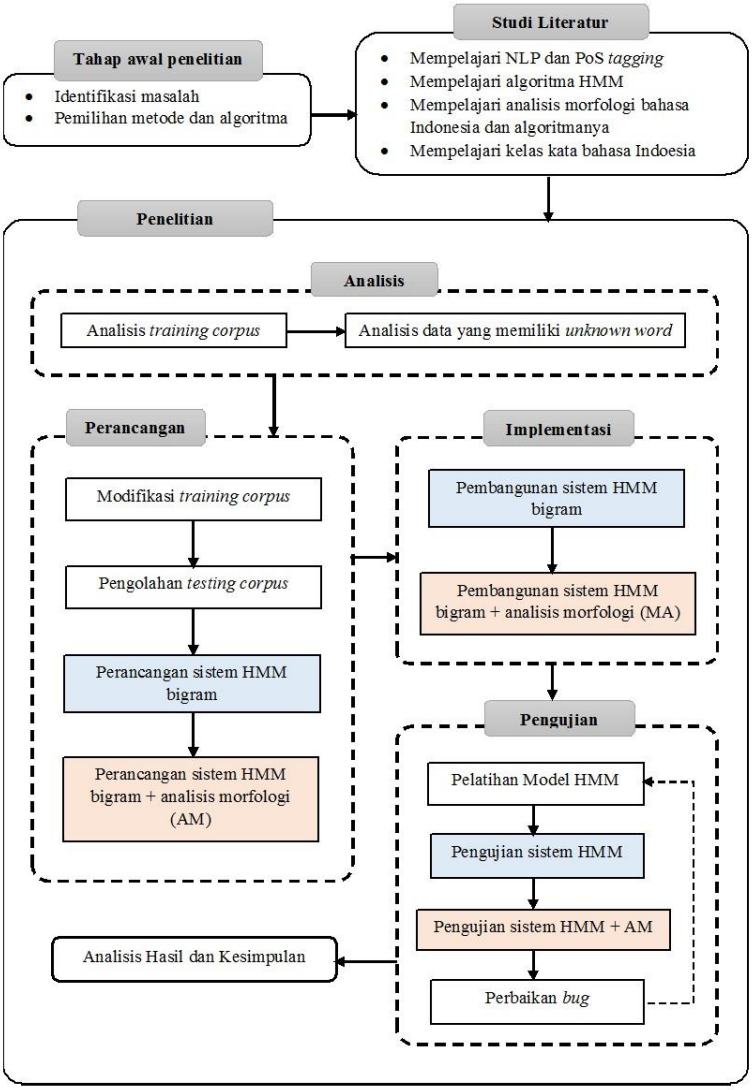
3.1. Desain Penelitian

Tujuan akhir penelitian ini, yaitu untuk mengetahui hasil kinerja sistem *PoS tagging* Bahasa Indonesia dengan menggunakan *Hidden Markov Model* (HMM) dan sistem yang menggunakan HMM ditambah dengan metode analisis morfologi (AM) untuk mengatasi *out-of-vocabulary* (*OOV).* Maka, dibutuhkannya desain penelitian yang berisi tahapan atau prosedur yang akan dilakukan selama penelitian sebagai panduan untuk mempermudah proses penelitian. Adapun gambaran umum desain penelitian pada penelitian ini ditunjukan pada Gambar 3.1.

Sebelum melakukan penelitian, dibutuhkan persiapan untuk mendukung dan mempermudah proses penelitian. Berikut ini beberapa persiapan yang dibutuhkan sebelum mulai melakukan penelitian:

a) Menentukan sumber data yang akan digunakan, seperti kumpulan berita dalam bahasa Indonesia dari media masa *online* dan dari penelitian yang pernah dilakukan sebelumnya.
b) Mengumpulkan data yang dibutuhkan, yaitu data yang berisi teks dalam bahasa Indonesia yang akan digunakan sebagai *testing corpus* dan mengumpulkan *training corpus* yang berisikan teks beserta *PoS tag*-nya.
c) Menetukan bahasa pemrograman yang akan digunakan dalam proses pembangunan sistem. Dalam penelitian ini, sistem akan dibangun menggunakan bahasa pemrograman *python*.
d) Menentukan alat dan bahan penelitian. Alat yang dimaksud yaitu perangkat keras (*Hardware*) dan perangkat lunak (*Software*) untuk mendukung proses analisis, perancangan, implementasi atau pembangunan dan pengujian.

Gambar 3.1 Desain Penelitian



A. Tahap Awal

Pada tahap ini merupakan tahap penentuan bahan untuk dijadikan objek utama penelitian. Mengidentifikasi masalah merupakan tahap paling awal yang perlu dilakukan, penyelesaian dari masalah yang ditentukan akan menjadi tujuan utama penelitian.

Pada penelitian ini, masalah yang yang akan diteliti yaitu semakin banyaknya aplikasi pengolah teks dengan berbagai bahasa, menjadikan salah satu *taks Natural Language Processing* (NLP) yaitu *Part of Speech* (PoS) *tagger* menjadi penting. *Hidden Markov Model* (HMM) merupakan salah satu algoritima pada PoS *tagging* dengan berbasis probabilistik. Namun, berdasarkan beberapa penelitian HMM memiliki kekurangan yaitu masalah pada OOV *words* yang muncul, maka diperlukan suatu metode untuk menangani masalah tersebut. Pada penelitian ini, metode yang akan digunakan untuk menangani masalah tersebut yaitu metode analisis morfologi.

B. Studi Literatur

Sebelum melakukan penelitian, diperlukannya studi literatur yang bertujuan untuk mempelajari dan memahami teori-teori yang berkaitan dengan objek utama penelitian. Adapun teori-teori yang dipelajari pada penelitian ini yaitu PoS *tagger* yang merupakan salah satu *taks* dalam NLP, kalimat dalam bahasa indonesia, macam-macam kelas kata atau *part-of-speech* pada bahasa Indonesia, algoritma HMM untuk menentukan *tag* terbaik dalam suatu teks beserta contoh dan prosesnya, analisis morfologi bahasa Indonesia khususnya afiksasi dan reduplikasi, dan algoritma analisis morfologi. Teori-teori tersebut diperoleh dari berbagai sumber seperti buku, jurnal nasional maupun internasional, artiker, informasi dari situs internet dan sumber ilmiah lainnya. Adapun dokumentasi studi literatur yang telah dilakukan ditunjukan pada bab sebelumnya.

C. Analisis

Analisis dilakukan pada *train corpus* dan data *testing* yang memiliki *out-of-vocabulary words*. *Train corpus* merupakan data latih yang berisi kumpulan kalimat dalam bahasa Indonesia yang telah diberi *tag* secara manual, sedangkan *testing corpus* yaitu korpus yang berisi

kumpulan kalimat yang akan menjadi bahan pengujian. Pada penelitian ini, sistem akan menggunakan *train corpus* dari penelitian Rashel dkk (2014). Bagian yang dianalisis dari *corpus* tersebut yaitu bentuk data, *bug* dan *tag*-nya. Sedangkan pada data *testing* yang memiliki *oov words* akan dianalisis bentuk dan ciri-cirinya. Analisis tersebut akan menjadi referensi pada data uji berisikan kalimat bahasa Indonesia yang didapat dari media massa *online*. Kemudian akan ditentukan jumlah *token* secara keseluruhan dan jumlah *token* yang merupakan *oov* dan disesuaikan jumlahnya dengan kebutuhan penelitian. Kumpulan data yang sudah diproses akan menjadi *testing corpus* yang akan digunakan pada tahap pengujian. Hasil dari analisis juga akan menjadi referensi pada proses perancangan sistem yang akan dibangun khususnya pada rancangan aliran data.

D. Perancangan

Pada tahap perancangan, dilakukan proses pembuatan rancangan atau rencana dari kerangka kerja sistem secara keseluruhan yang akan dibangun, dimulai dari merancang arsitektur sistem; aliran data; hingga *output* sistem. Dua sistem utama yang akan dirancang yaitu rancangan sistem PoS *tagger* menggunakan HMM *bigram* saja dan HMM *bigram* ditambah metode analisis morfologi. Namun, setelah merancang sistem HMM bigram, terlebih dulu dilakukan proses modifikasi *training corpus* dan pengumpulan serta pengolahan *testing corpus*, sebelum merancangan kedua sistem tersebut.

E. Implementasi

Dokumentasi rancangan sistem yang telah dibuat harus dimengerti oleh mesin, dalam hal ini yaitu komputer, maka rancangan tersebut harus diubah kedalam bentuk yang dapat dimengerti oleh komputer melalui proses *coding.* Pada penelitian ini, sistem yang telah dirancang akan diubah ke dalam bahasa komputer, dengan menggunakan bahasa pemrograman *python*.

F. Pengujian

Sistem yang telah dibangun harus diujikan terlebih dahulu, dengan tujuan agar sistem dapat sesuai dengan rancangan yang telah dibuat. Selain pengujian algoritma yang digunakan pada sistem HMM bigram

dan analisis morfologi untuk menghilang *bug*, pengujian dalam hal ini juga mencakup pengujian yang dilakukan dengan menggunakan *testing corpus* sebagai objek utama penelitian. Kemudian, hasil akurasi dari kedua sistem akan dianalisis untuk mengetahui bagaimana peningkatan akurasi pada proses PoS *tagging* menggunakan metode analisis morfologi yang mengatasi OOV *words*.

3.2. Metode Penelitian

Dalam metode penelitian dijabarkan tahapan-tahapan yang dilakukan dalam penelitian. Metodologi penelitian terdiri dari beberapa tahapan yang terkait secara sistematis. Tahapan ini diperlukan untuk memudahkan dalam melakukan penelitian. Tahapan yang dilakukan dalam penelitian ini dijelaskan pada subbab berikut.

3.2.1 Pengumpulan Data

Pada penelitian ini penulis mencari data dan informasi yang akurat mengenai penelitian yang akan dilakukan, yang dapat menjadi referensi dan menunjang proses penelitian. Ada dua teknik pengumpulan data yang diterapkan dalam penelitian ini, yaitu:

a. Eksplorasi dan Studi Literatur

   Eksplorasi dan studi literatur dilakukan untuk menyelesaikan permasalahan pada penelitian yang akan diteliti serta mendapatkan dasar-dasar teori-teori yang ada. Teori-teori yang dikumpulkan bersumber dari buku, *website*, artikel, jurnal nasional maupun internasional dan sumber ilmiah lainnya yang mendukung dalam penelitian ini.

b. Wawancara

   Wawancara dilakukan kepada beberapa narasumber yang ahli dibidang bahasa Indonesia dan beberapa ahli lainnya terkait penelitian ini.

3.2.2 Pengembangan Perangkat Lunak

Dalam tahap ini metode pengembangan perangkat lunak yang digunakan adalah pendekatan terstruktur yaitu model sekuensial linier

atau model *waterfall*. Model sekuensial linier mengusulkan sebuah pendekatan pengembangan perangkat lunak yang sistematik dan sekuensial mulai dari sistem level dan terus maju dari *analysis*, *design*, *coding*, *testing* dan *maintenance* (Pressman, 2010).

Gambar 3.2 Model *Waterfall*

a. *Analysis* (Analisis), pada tahap ini kebutuhan sistem yang diperlukan dalam pembangunan perangkat lunak dirumuskan dan dianalisis dari mulai sampai dengan selesai. Kebutuhan pada penelitian ini yaitu korpus berupa *training corpus* dan *testing corpus*.
b. *Design* (Desain), berdasarkan hasil analisis, kemudian akan dibuat rancangan sistem yang melibatkan identifikasi dan menjadi gambaran dasar serta referensi pada tahap implementasi.
c. *Coding* (Implementasi), setelah desain atau raancangan sudah dibuat kemudian masuk ke tahap implementasi. Tahap ini adalah tahap mengolah data menjadi informasi yang berupa kode program dengan menerapkan algoritma *Hidden Markov Model bigram* dalam proses *tagging* yang kemudian akan mengimplementasikan HMM menggunakna analisis morfologi bahasa Indonesia untuk menangani OOV.
d. *Testing* (Pengujian), pengujian sistem digunakan untuk mengetahui apakah perangkat lunak yang dibangun masih terdapat error atau tidak. Jika ada yang error maka sistem akan kembali

diperbaiki. Pengujian juga dilakukan untuk ekperimen terkait objek utama penenlitian.

e. *Maintenance* (Perawatan), perawatan pada perangkat lunak untuk perbaikan atau pengembangan. Ini merupakan kelebihan model *waterfall* karena pengembang dapat kembali ketahap sebelumnya, meskipun pengembang sudah sampai pada tahap akhir.

3.3. Alat dan Bahan Penelitian

Untuk menunjang penelitian yang akan dilakukan, maka diperlukanlah alat dan bahan untuk mendapatkan hasil yang baik dan terstruktur.

3.3.1. Alat Penelitian

Dalam proses pengumpulan data dan pembangunan sistem dalam penelitian ini dilaksanakan dengan menggunakan beberapa alat bantu berupa perangkat keras (*hardware*) dan perangkat lunak (*software*) sebagai berikut.

a. Perangkat keras (*Hardware*), yang terdiri dari :
   1. Processor Intel(R) Core(TM) i5-4300U CPU @ 1.90GHz 2.49GHz
   2. RAM 4 GB
   3. *Harddisk* 320 GB
   4. *Mouse*
   5. *Keyboard*

b. Perangkat lunak (*Software*), yang terdiri dari :
   1. Sistem Operasi (SO) Microsoft Windows 10 64 bit
   2. Python 3.6.4 (64-bit)
   3. Python IDLE *shell* dan *editor*
   4. Jupyter notebook
   5. *Text editor* Notepad dan Sublime Text 3
   6. Google Chrome
   7. Microsoft Office Excel 2016

3.3.2. Bahan Penelitian

Bahan penelitian menggunakan dua korpus sebagai *input* sistem yaitu *train corpus* berupa kumpulan kalimat bahasa Indonesia yang telah

diberi *tag* secara manual pada setiap *token*-nya dan *testing corpus* berisi kumpulan kalimat bahasa Indonesia tanpa adanya *tag*, yang telah dimodifikasi dengan 30% tingkat kandungan OOV *words* didalamnya. Kedua korpus disimpan dalam format berkas *tab separated value* (.tsv) dengan bentuk satu kalimat perbaris. Sementara *output* sistem berupa kalimat bahasa Indonesia yang ada dalam *testing corpus* beserta PoS *tag*-nya.