

BAB III

METODE PERANCANGAN SISTEM

Metode metode apa saja yang digunakan pada penelitian ini adalah sebagai berikut:

1. Telaah Literatur

Tahapan ini peneliti akan melakukan pencarian dan juga pendalaman dari materi-materi yang akan dibutuhkan selama proses penelitian berlangsung. Yaitu algoritma *Conditional Random Fields* dan juga *Random Forest Classifier*.

2. Analisa Kebutuhan

Pada tahap ini peneliti akan melakukan analisa pengumpulan kebutuhan apa saja yang nantinya akan dibutuhkan selama penelitian sedang berlangsung. Analisa ini akan dilakukan dari telaah literature yang telah dilakukan.

3. Pemrograman Sistem

Peneliti akan melakukan pemograman dari sistem yang akan dibuat. program akan digarap sesuai dengan kebutuhan yang telah di analisa pada tahap sebelumnya.

4. Testing Sistem

Tahap ini berfungsi sebagai tahap dimana peneliti melakukan testing untuk mengetahui apakah algoritma CRF dan juga Random Forest sudah memberikan hasil yang diinginkan atau belum.

5. Implementasi dan Evaluasi

Pada tahap ini akan dilakukan pengujian pada dataset yang berbeda dari dataset yang telah dipakai pada saat training untuk mengetahui akurasi dari hasil klasifikasi NER.

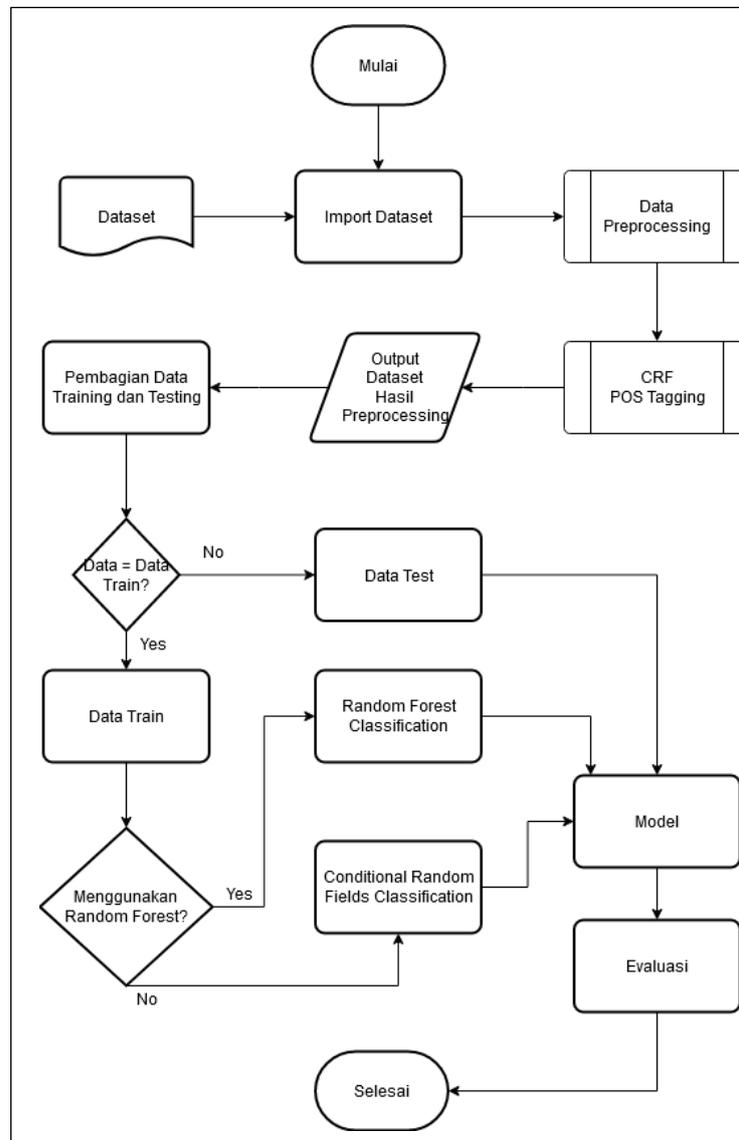
6. Konsultasi dan Penulisan

Proses konsultasi akan dilakukan secara berkala selama tahap pertama yaitu telaah literature sampai dengan implementasi dan evaluasi. Dengan tahap ini diharapkan kendala yang ditemui ketika penelitian berlangsung dapat diselesaikan dengan cepat.

3.2 Flowchart

Berikut ini adalah flowchart dari sistem, Alur dari sistem dimulai dari *preprocessing*, *conditional random fields*, sampai dengan *random forest*.

3.2.1 Flowchart Utama

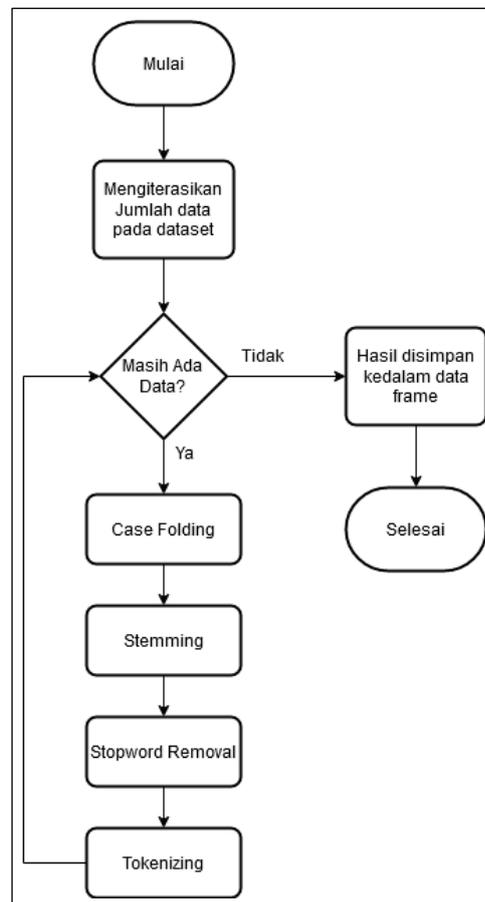


Gambar 2.1 Flowchart Utama

Gambar 3.1 menunjukkan *flowchart* utama pada penelitian, proses ini diawali dengan pengambilan atau *import* data dari dataset yang telah dilabeli dalam bentuk txt. Setelah dataset berhasil di masukkan maka proses selanjutnya yang akan dilakukan adalah proses *Text preprocessing*, setelah proses *text preprocessing* selesai selanjutnya proses CRF POS Tagging akan dilakukan. Dataset hasil pun akan didapatkan yang kemudian dibagi menjadi data testing dan data training yang

kemudian akan dilakukan *named entity extraction* yang kemudian akan dilanjutkan kedalam proses *Random Forest Classification* atau *Conditional Random Fields*. Hasil yang di dapatkan dari klasifikasi akan dilakukan evaluasi model untuk mengetahui hasil *accuracy*, *precision*, *recall*, dan *f1-score*. Proses prediksi akan di lakukan dari hasil data *training* yang telah didapatkan kepada kalimat baru yang dimasukkan.

3.2.2 Flowchart Text Preprocessing



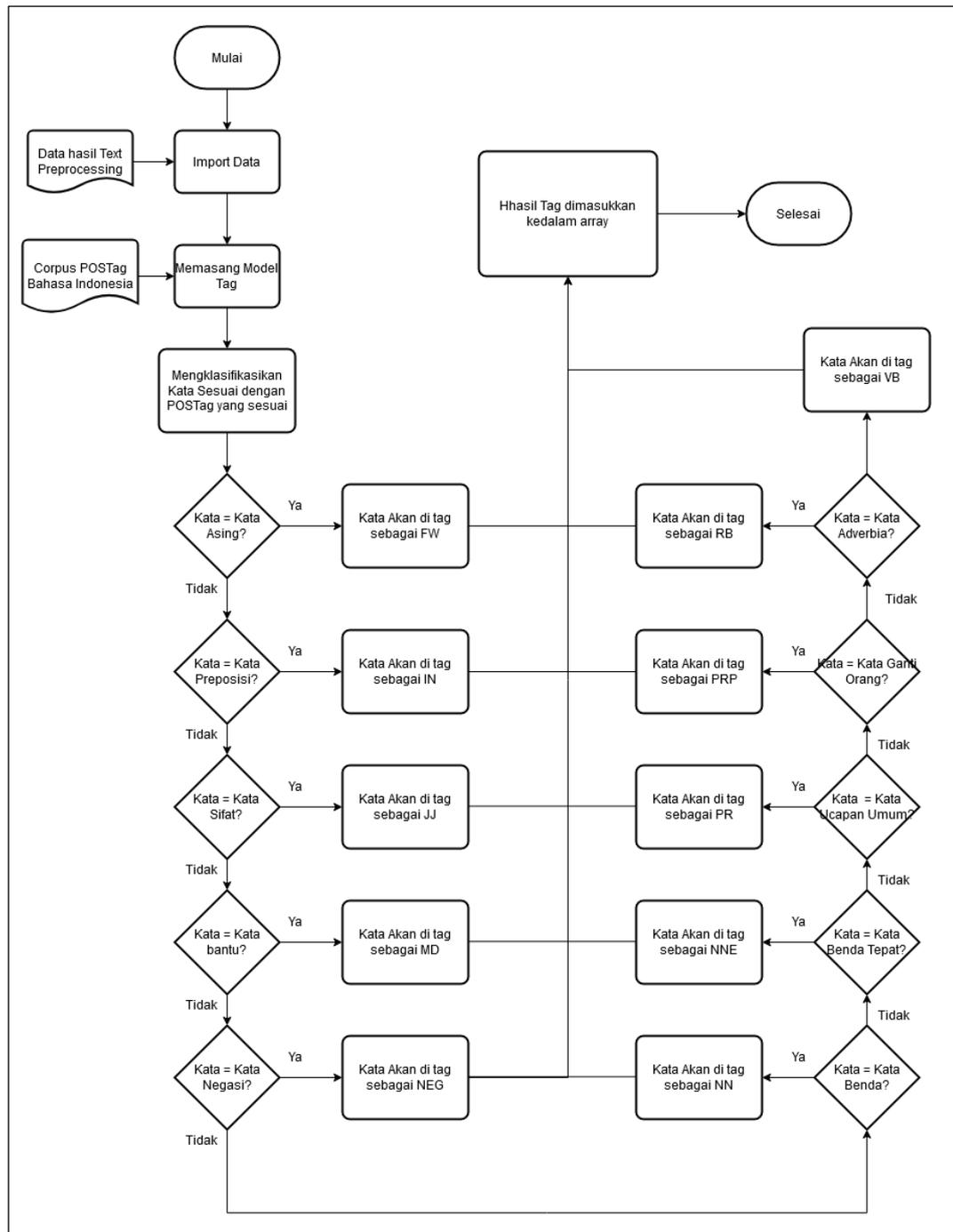
Gambar 3.2 Flowchart Text Preprocessing

Gambar 3.2 menjelaskan proses *text processing* dilakukan. Dataset yang dipakai merupakan data yang berstandar StanfordNER dimana setiap line akan terbagi dari kata yang membangun suatu kalimat dengan begitu setiap line akan di-

iterasi. Proses *text preprocessing* ini akan dibutuhkan library python yang sudah tersedia yaitu *Regex* and *Sastrawi* yang akan dipakai pada proses *noise removal*, *stemming*, dan *stopword removal*.

Tahapan pertama pada *text preprocessing* ketika data sudah dimasukkan adalah *case folding* dimana kata yang ada akan diubah dari huruf kapital menjadi huruf kecil menggunakan *function.lower()* kemudian setelah selesai tahapan selanjutnya merupakan *Noise Removal* di mana setiap kata akan dihilangkan karakter selain [a-z] dan [A-Z] sehingga karakter-karakter seperti (. , , , /, + , - ,) akan dihilangkan dari kata yang diproses. Tahapan selanjutnya dari *text preprocessing* adalah *stemming* di mana kata yang ada akan diproses dan diubah menjadi kata yang paling dasar sesuai dengan menggunakan bantuan library *Sastrawi*. *Library* yang sama akan dipakai juga dalam proses *stopword removal* di mana kata-kata henti seperti [atau, dan, tapi, namun, dll] akan dihapus karena kata-kata tersebut dianggap tidak penting dalam proses selanjutnya. Terakhir merupakan proses *tokenizing* yang memisahkan kalimat menjadi kata-kata individual.

3.2.3 Flowchart Conditional Random Fields



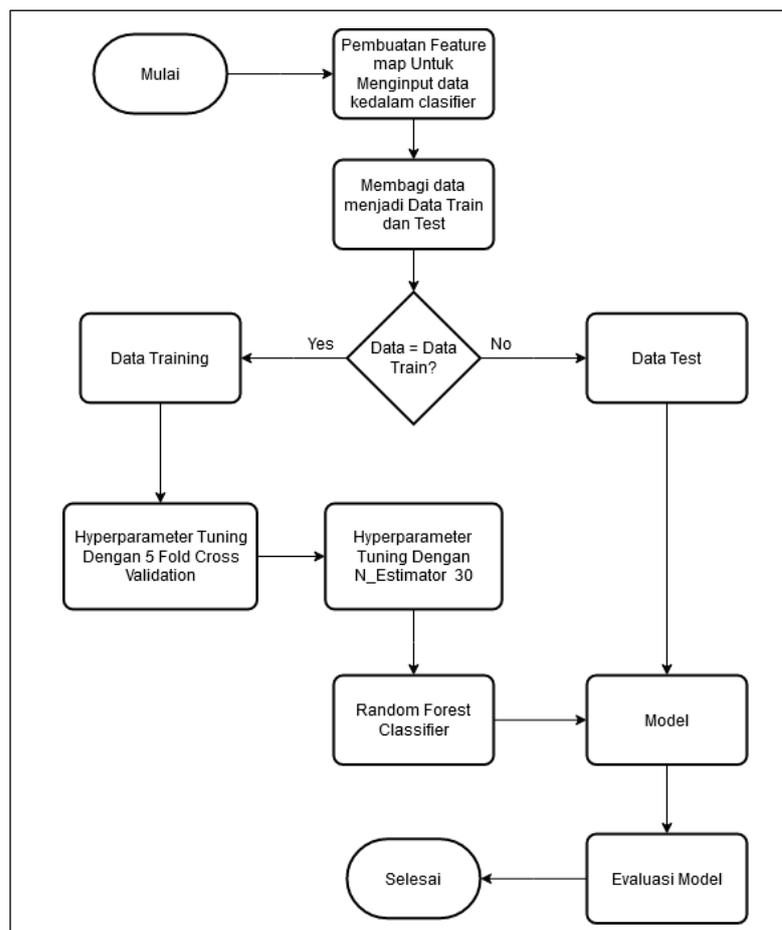
Gambar 3.3 Flowchart CRF

Gambar 3.3 merupakan penjelasan dari proses *Conditional Random Fields Tagger* dimana pada proses ini akan mengolah data yang telah di *preprocess* sebelumnya. Proses ini akan dimulai dengan membaca data hasil *preprocessing*

kemudian akan dilanjutkan dengan pemasangan model *tagger* yang didapat dari penelitian sebelumnya. Setelah model selesai dipasang kata-kata akan diproses kemudian di-*tag* sesuai dengan *tag* yang sesuai dengan kata tersebut, untuk contoh kata benda akan di-*tag* dengan *tag* NN atau kata bantu akan di-*tag* dengan *tag* MD.

Setelah proses *tag* telah selesai dilakukan maka proses selanjutnya adalah memasukkan data kedalam *dataframe* yang sesuai dengan format IOB, dimana format IOB merupakan format yang sering dipakai dalam proses *Named Entity Recognition*.

3.2.4 Flowchart Random Forest

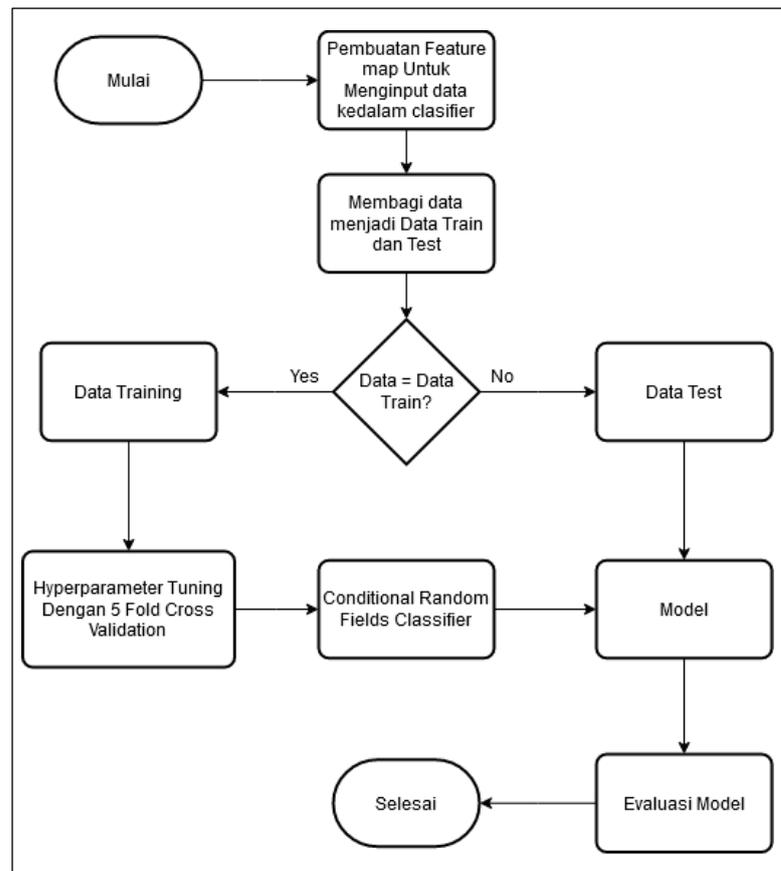


Gambar 3.4 Random Forest Flowchart

Gambar 3.4 menjelaskan bagaimana proses *random forest* akan berlangsung pada penelitian ini. Proses dimulai dengan pembuatan *feature map* yang berfungsi untuk mengekstrak kata dan juga *tag* yang berada didalam dataset. Setelah itu data tersebut akan dibagi menjadi 2 data yaitu data *testing* dan juga data *training*. Setelah itu akan dilakukan *hyperparameter tuning* dengan *n_estimators* 30, dan 5 *fold cross validation*, hal ini membuat data yang telah di input akan dibagi lagi menjadi 5 subset data dimana subset data tersebut akan dilakukan train-test. Beberapa model *decision trees* dan juga *neural network* terkadang mampu mendapatkan 100% akurasi pada data training namun akan berperforma lebih buruk pada data yang baru. metode ini diharapkan dapat membuat *random forest* untuk mengklasifikasikan secara merata pada 5 subset data dan tidak terlalu *overly optimistic* pada performanya.

Setelah data dibagi menjadi 5 subset random forest pun akan dijalankan untuk melakukan klasifikasi. Setelah klasifikasi selesai akan dilakukan evaluasi model untuk mengetahui hasil performa dari *random forest* dalam mengklasifikasikan *named entity*.

3.2.5 Flowchart Conditional Random Fields Classifier



Gambar 3.5 Conditional Random Fields Classifier Flowchart

Gambar 3.5 menjelaskan bagaimana proses *conditional random fields classifier* akan berlangsung pada penelitian ini. Proses dimulai dengan pembuatan *feature map* yang berfungsi untuk mengekstrak kata dan juga tag yang berada didalam dataset. Setelah itu kata kata tersebut akan digabungkan menjadi kalimat yang memiliki informasi-informasi tambahan dan juga tag yang berfungsi untuk memberikan konteks sebuah kalimat kedalam *classifier*. Setelah itu data kalimat tersebut akan dibagi menjadi 2 data yaitu data testing dan juga data training. Setelah itu akan dilakukan *hyperparameter tuning 5 fold cross validation*. Proses *CRF classifier* pun akan dilakukan untuk mendapatkan model yang kemudian dapat dievaluasi untuk mendapatkan hasil performa.