BAB 3

METODOLOGI PENELITIAN

3.1 Metodologi Penelitian

Untuk mengukur pengaruh dari implementasi metode Recursive Feature Elmination (RFE) terhadap model Random Forest (RF) dalam mengklasifikasikan subkategori berita akan dilakukan tahapan-tahapan berikut.

1. Perumusan Masalah

Proses perumusan masalah dilakukan berdasarkan penelitian terdahulu (Fadilah, 2020) yang mendapatkan nilai rata-rata F1-Score sebesar 93,49% dengan menggunakan sejumlah 1012 fitur pada data teks masukan untuk melatih sebuah model RF. Dalam penelitian yang dilakukan akan diujikan pengaruh dari pengurangan jumlah fitur dengan menggunakan metode Recursive Feature Elimination (RFE) terhadap model RF yang telah dihasilkan pada penelitian sebelumnya. Hal ini dilakukan untuk menghasilkan model klasifikasi subkategori berita berbasis RF yang lebih efisien.

2. Studi Literatur

Pada tahapan ini aktivitas yang dilakukan adalah meninjau teori-teori yang digunakan dalam riset sebelumnya serta teori tambahan yang akan digunakan untuk meningkatkan performa model yang telah dikembangkan. Teori-teori yang dimaksudkan terdiri dari literatur-literatur terkait dengan metode-metode *text preprocessing*, algoritma Random Forest, metode ekstraksi TF- IDF, metode Recursive Feature Elimination, dan perhitungan performa model dengan menggunakan F1-Score.

3. Pengembangan Sistem

Pada tahapan ini kegiatan yang dilakukan adalah melakukan perancangan model yang telah dilengkapi dengan metode Recursive Feature Elimination untuk memperoleh model klasifikasi yang lebih efisien bila dibandingkan dengan penelitian sebelumnya.

4. Uji Coba dan Evaluasi Sistem

Pada tahapan ini akan dilakukan proses uji coba berdasarkan data yang telah tersedia dan parameterisasi model dengan metode TF- IDF dan Random Forest yang telah dilengkapi metode Recursive Feature Elimination. Melalui tahap ini dapat diketahui dampak dari pengurangan jumlah fitur terhadap performa yang dihasilkan oleh model. Evaluasi dilakukan untuk mengetahui kelebihan dan kekurangan dari cara yang digunakan dalam riset ini bila dibandingkan dengan riset sebelumnya.

5. Penarikan Kesimpulan

Pada tahap ini akan dilakukan proses penarikan kesimpulan berdasarkan eksperimen-eksperimen yang telah dilakukan dalam riset. Pada tahap ini, akan dilakukan proses penjabaran hasil berdasarkan rumusan masalah yang ingin dicapai dalam penelitian. Kemudian, melalui kesimpulan yang dihasilkan akan ditambahkan penjabaran terkait dengan peluang-peluang penelitian lanjutan yang dapat dilakukan.

6. Konsultasi Penelitian

Pada tahapan kosultasi penelitian dilakukan penyusunan laporan yang bertujuan untuk mendokumentasikan seluruh wujud proses riset serta merumuskan hasil akhir dari riset yang telah dilakukan. Proses penyusunan

laporan akan dilakukan melalui serangkaian pertemuan untuk mendiskusikan hasil penelitian bersama dengan kedua dosen pembimbing.

3.2 Pengumpulan Data

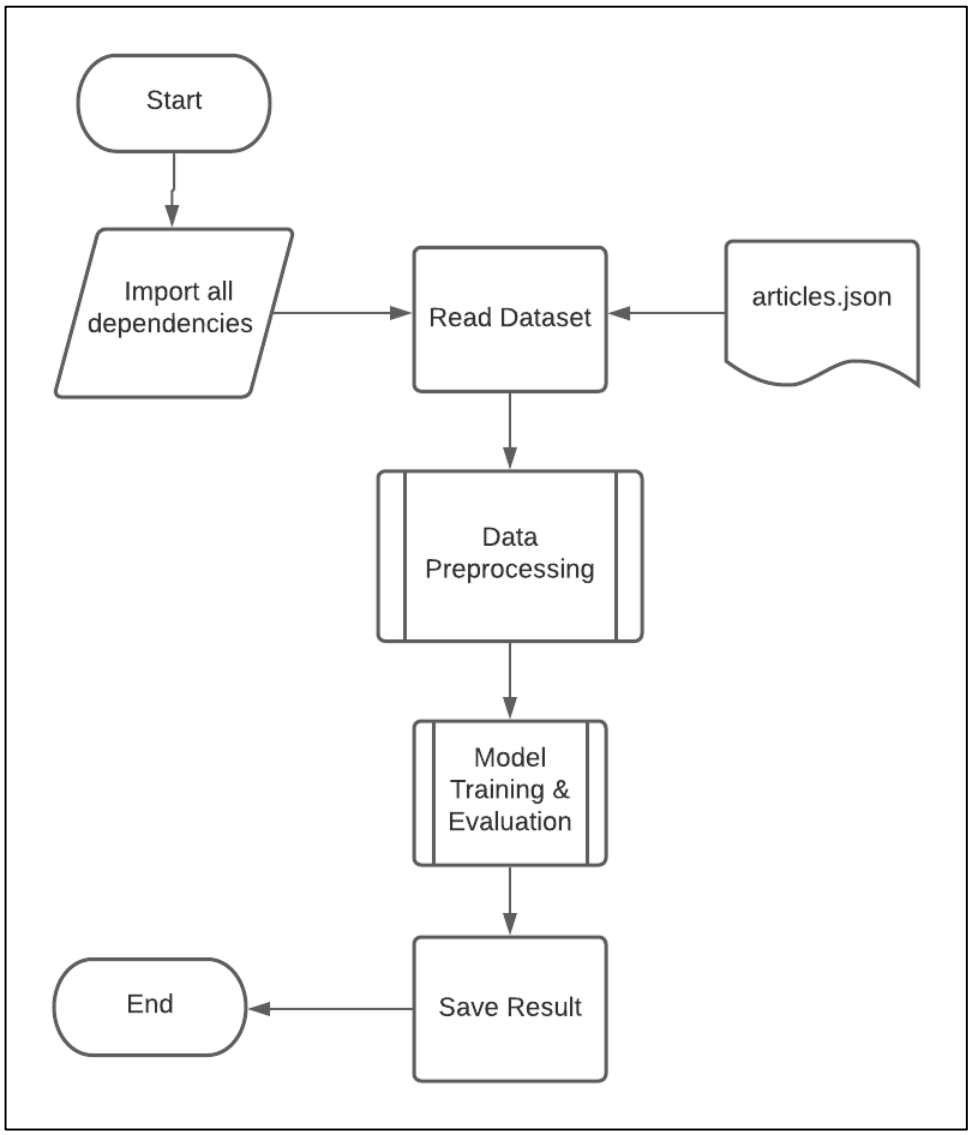
Data dalam peneilitian ini diperoleh dari penelitian sebelumnya (Fadilah, 2020). Dalam penelitian, data disimpan dalam format file Javascript Object Notation (JSON) untuk mempermudah proses pengolahan data. Data diperoleh dengan melakukan proses *web crawling* pada situs merahputih.com milik PT. Merah Putih berdasarkan izin yang telah diberikan pada penelitian sebelumnya. Dataset dari hasil proses crawling hanya terdiri dari 3 subkategori berita yang terdapat dalam kategori "Indonesiaku" yaitu kuliner, tradisi, dan travel.

3.3 Perancangan Aplikasi

Perancangan sistem mencakup dari flowchart dari proses klasifikasi dari algoritma RF, proses pemodelan, pelatihan, serta evaluasi model klasifikasi untuk subkategori berita pada kategori “Indonesiaku” di dalam *website* merahputih.com. Flowchart proses klasifikasi dengan algoritma RF akan menggambarkan bagaimana model klasifikasi dapat dibentuk dengan menggunakan algoritma RF. Di sisi lain pada proses pemodelan dan pelatihan, akan digambarkan bagaimana algoritma RFE diimplementasikan untuk mengefesiensikan model klasifikasi RF. Terakhir, pada proses evaluasi model akan dijabarkan tahapan-tahapan yang dilakukan untuk mengevaluasi keberhasilan penelitian.

3.3.1 Flowchart Klasifikasi

Dalam penelitian yang dilakukan, proses pertama yang dilakukan adalah meng-*import* seluruh *library* yang dibutuhkan dalam penelitian. Kemudian, proses dilanjutkan dengan membaca dataset yang akan digunakan dalam penelitian. Berdasarkan dataset yang telah dibaca, akan dilakukan rangkaian *text preprocessing* terhadapnya. Berdasarkan data hasil *text preprocessing* akan dilakukan proses pelatihan dan evaluasi model. Pada proses pelatihan model, model akan dibentuk dengan menggunakan parameter terbaik yang dihasilkan berdasarkan penelitian sebelumnya (Fadilah, 2020). Implementasi metode RFE sebagai metode tambahan untuk mengefesiensikan model juga akan dilakukan dalam proses pelatihan model. Kemudian, berdasarkan model yang telah dilatih akan dilakukan proses evaluasi terhadap keberhasilan model per sejumlah fitur yang digunakan untuk melatihnya. Proses diakhiri dengan menyimpan performa klasifikasi per pengurangan fitur yang telah dilakukan dalam tahap sebelumnya. Flowchart yang memberikan gambaran umum terkait dengan proses-proses yang akan dilakukan dalam penelitian ditunjukkan oleh Gambar 3.1.

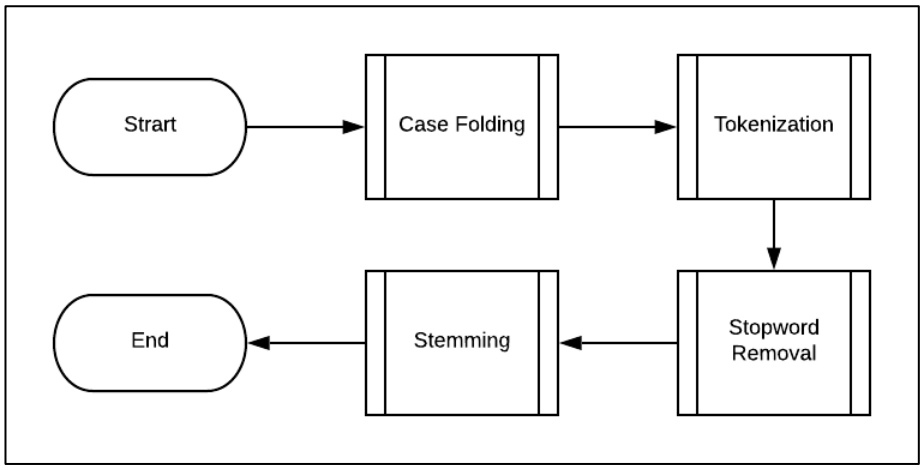


Gambar 3. 1 Flowchart Klasifikasi

3.3.2 Flowchart Text Preprocessing

Pada tahap *text processing*, proses diawali dengan proses *case folding* untuk menghilangkan karakter-karakter selain alphabet a-z dan A-Z, menghilangkan spasi yang berlebih antara kata, dan mengecilkan seluruh huruf yang menyusun teks. Kemudian akan dilakukan proses tokenization untuk mengubah sebuah teks menjadi sekumpulan list kata-kata yang menyusun teks. Berdasarkan list kata yang menyusun sebuah teks, akan dilakukan proses

stopwords removal untuk mengilangkan kata-kata muncul secara universal dan tidak mencirikan suatu kelompok dokumen. Kemudian, untuk mempermudah proses klasifikasi, maka akan dilakukan proses *stemming* untuk mengubah setiap kata yang muncul kembali ke kata dasarnya. Proses stemming dilakukan guna menghilangkan variasi kata-kata yang akan diolah. Gambaran terkait dengan rangkaian proses yang akan dilakukan ditunjukkan pada Gambar 3.2.

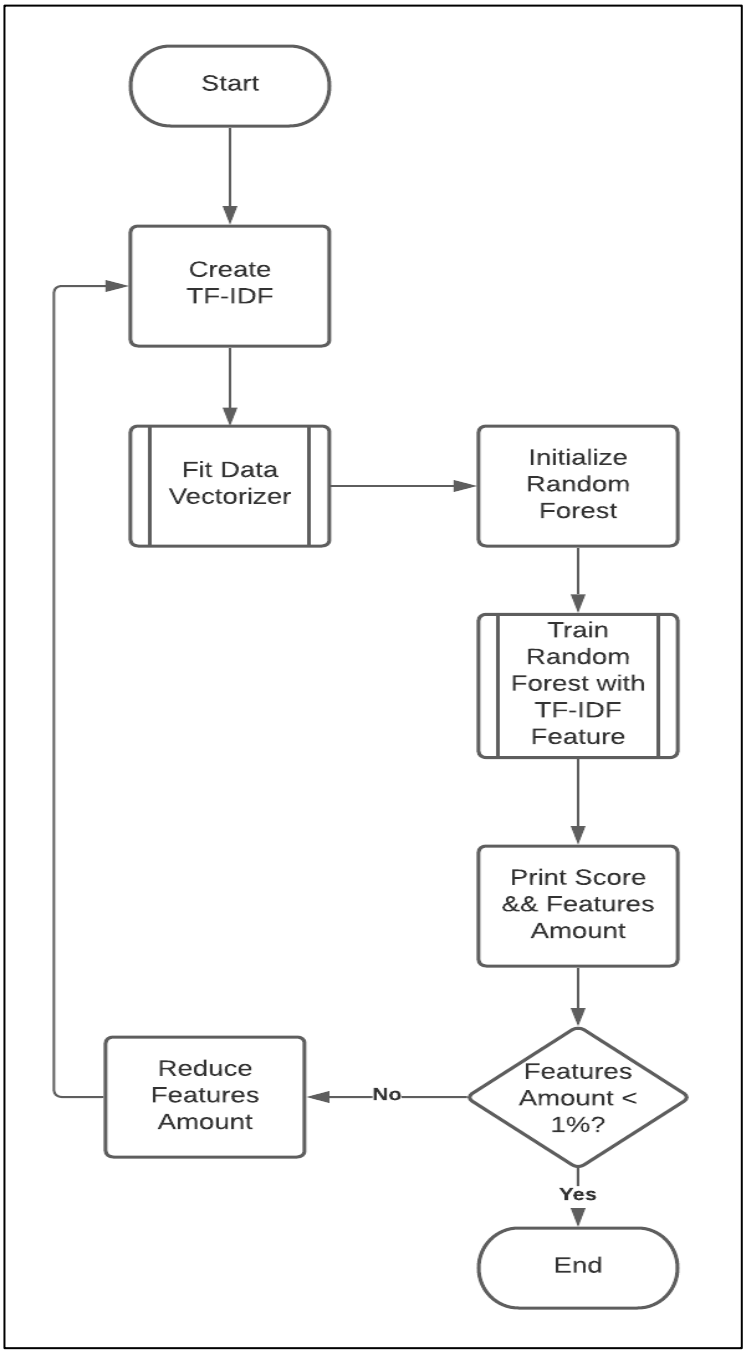


Gambar 3. 2 Flowchart Pre-processing

3.3.3 Flowchart Model Training & Evaluation

Proses pelatihan dan evaluasi model diawali dengan pembuatan model TF-IDF dan proses ekstraksi fitur berdasarkan data yang hasil tahap praproses. Kemudian, proses dilanjutkan dengan menginisialisasi model RF berdasarkan parameter terbaik pada penelitian terdahulu. Kemudian, dilakukan proses pelatihan model dengan menggunakan model TF-IDF dan RF dengan menggunakan data latih. Setelah model selesai dilatih, akan dicetak jumlah fitur yang digunakan untuk melatih model klasifikasi beserta performa yang dihasilkan oleh model. Setelah itu, jika jumlah fitur masih lebih besar dari 1% dari jumlah fitur awal yang digunakan, akan dilakukan proses pengurangan fitur dan proses

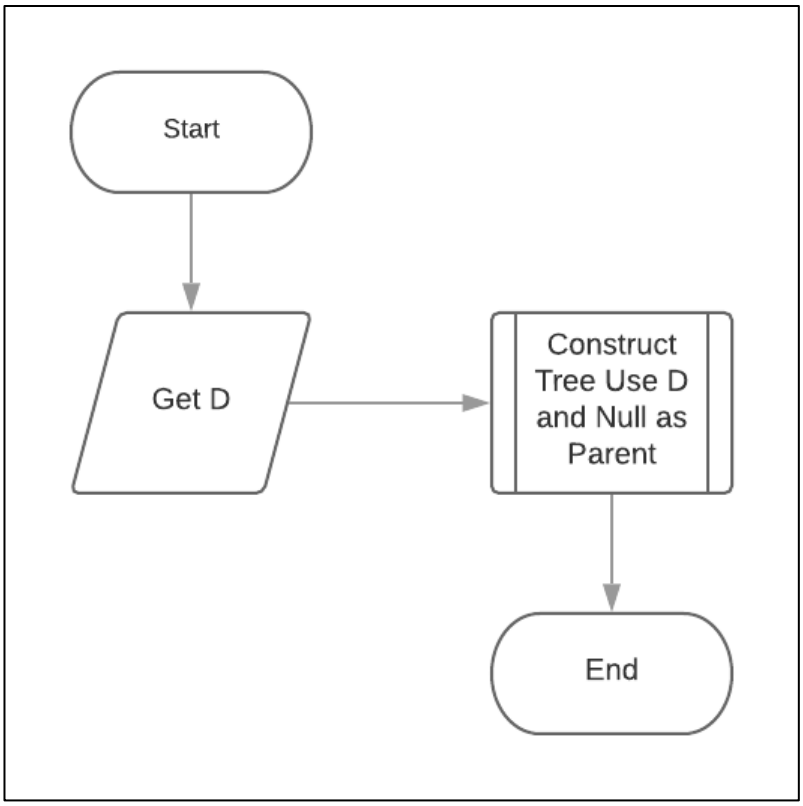
pelatihan model kembali dengan menggunakan jumlah fitur yang telah dikurangi. Proses pengurangan fitur dan pelatihan model akan terus diulangi sampai jumlah fitur kurang dari 1% dari jumlah fitur awal yang digunakan.



Gambar 3. 3 Flowchart Data Training

3.3.4 Flowchart Algoritma

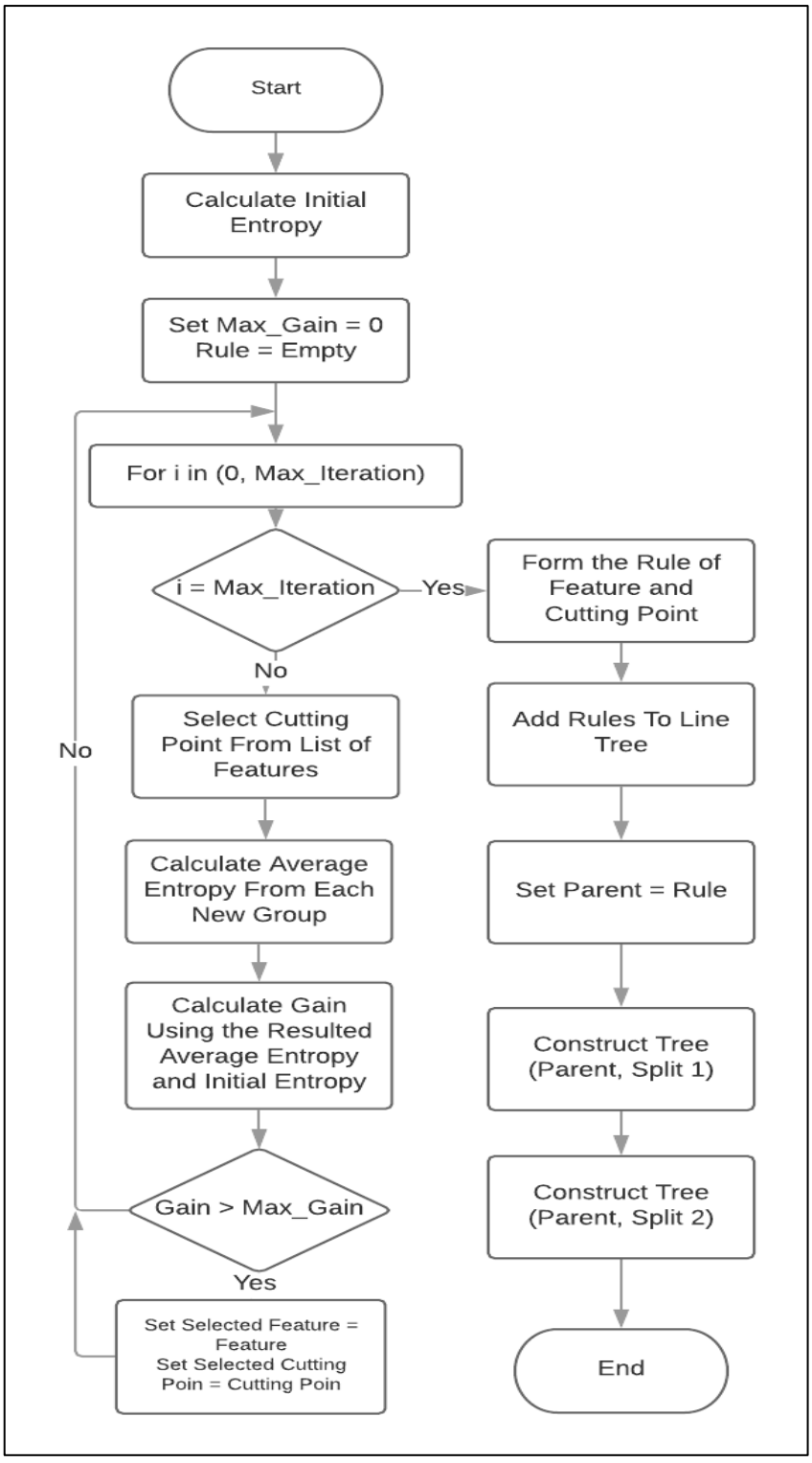
Algoritma Random Forest Classifier (RF) bekerja dengan melatih sekumpulan Decision Tree berdasarkan subdataset yang dibentuk melalui mekanisme *bagging*, maka dari itu, untuk menjelaskan proses pelatihan yang terjadi dalam algoritma RF terlebih dahulu akan dijelaskan proses pelatihan (pembentukan) sebuah Decision Tree. Tahapan proses pembentukan sebuah Decision Tree dapat dilihat pada Gambar 3.5.



Gambar 3. 4 Flowchart Decision Tree

Proses pembentukan sebuah Decision Tree dilakukan berdasarkan kumpulan dataset *D* sebagai masukan. Kemudian proses akan dilanjutkan dengan proses rekonstruksi *tree* berdasarkan dataset *D* dengan *parent node* bernilai *Null*.

Proses pembuatan *tree* atau pohon dapat dilihat pada flowchart tree pada Gambar 3.6 .



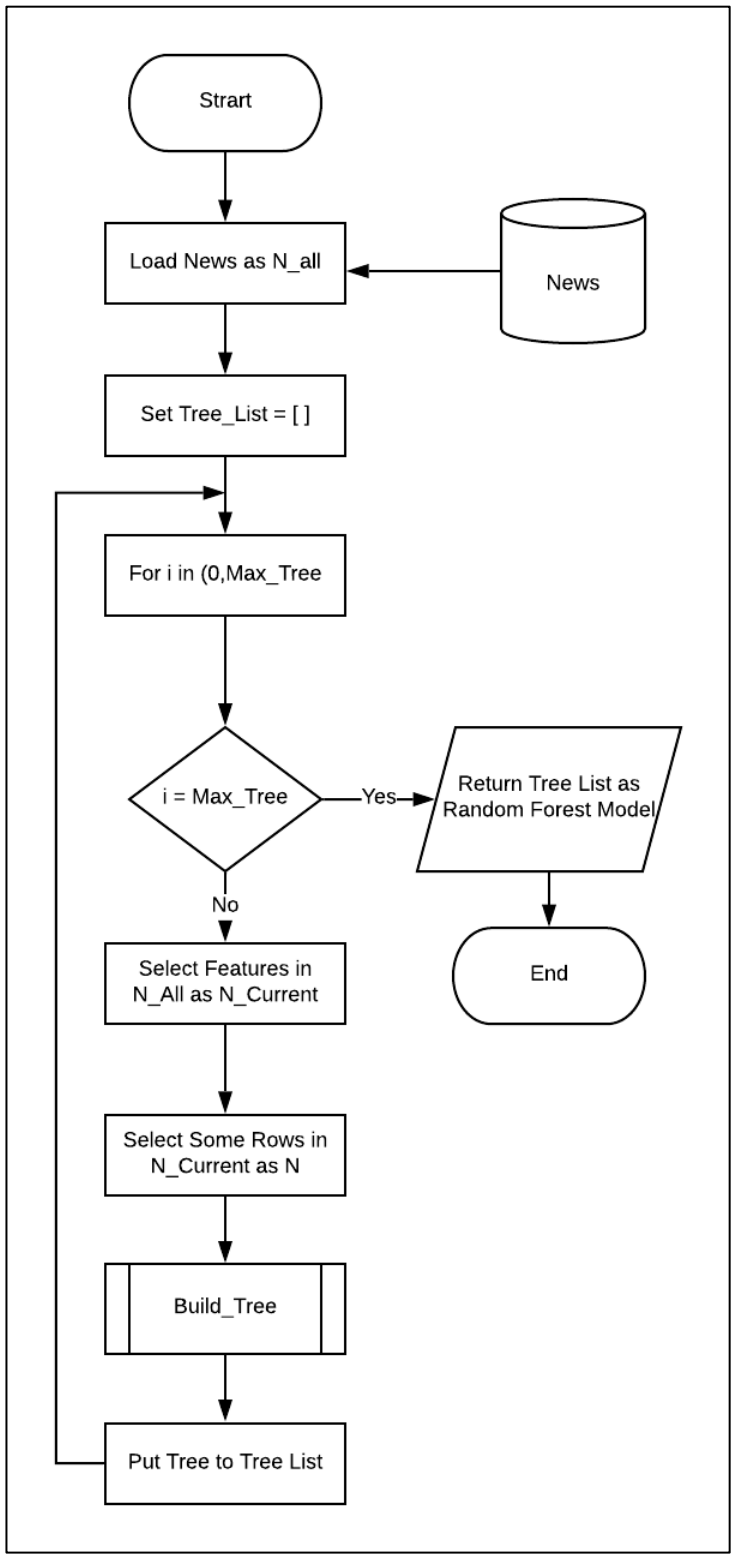
Gambar 3. 5 Flowchart Tree Construction

Proses rekonstruksi tree dimulai dengan menghitung nilai entropy mula-mula pada data masukan. Kemudian, proses dilanjutkan dengan menginisialisasi variabel Max_Gain menjadi nol dan variabel Rule menjadi "Empty". Kemudian, berdasarkan nilai parameter max_iteration yang di-*set* oleh pengguna akan dilakukan proses pengulangan yang terdiri dari tahap-tahap berikut.

1. Pilih salah satu fitur yang tersedia dalam kumpulan fitur milik dataset dan tentukan nilai *cutting point* secara acak untuk membagi dataset masukan ke dalam kumpulan subkelompok
2. Hitung nilai rata-rata entropi dari setiap subkelompok
3. Hitung nilai Gain sementara dan jika nilai Gain sementara lebih besar dari Max_Gain maka simpan nilai Selected Feature menjadi fitur yang terpilih dan set nilai *cutting point*.

Saat iterasi ke-*i* sama dengan nilai Max_Iteration maka bentuk dan sisipkan Rule sebagai node ke dalam tree. Kemudian set nilai parent node menjadi node yang berisikan Rule dan lakukan proses konstruksi tree berdasarkan setiap kelompok yang telah dibagi oleh *cutting point* berdasarkan fitur dengan nilai Gain terbesar.

Berdasarkan alur pembentukan Decision Tree yang telah dijelaskan, proses pembentukan model klasifikasi berita berdasarkan subkategori yang dimilikinya dengan menggunakan Algoritma Random Forest digambarkan pada Gambar 3.7 berikut.



Gambar 3. 6 Flowchart Algoritma Random Forest Classifier

Proses pembentukan model klasifikasi dengan menggunakan algoritma RF dimulai dengan menginisialisasi nilai N_all sebagai data artikel berita. Kemudian, berdasarkan nilai Max_Tree yang di-*set* oleh pengguna lakukan tahap-tahap berikut secara berulang.

1. Pilih sekumpulan fitur yang akan membangun sebuah Decision Tree secara acak
2. Pilih sekumpulan individu dalam dataset secara acak
3. Bangun Decision Tree berdasarkan kumpulan fitur dan individu yang telah terpilih
4. Tambahkan Decision Tree ke dalam Tree List

Proses pengulangan akan terus dilakukan hingga nilai Max_Tree tercapai. Setelah nilai Max_Tree tercapai, proses pengulangan akan dihentikan dan Tree List yang telah dihasilkan akan dikembalikan sebagai model Random Forest.