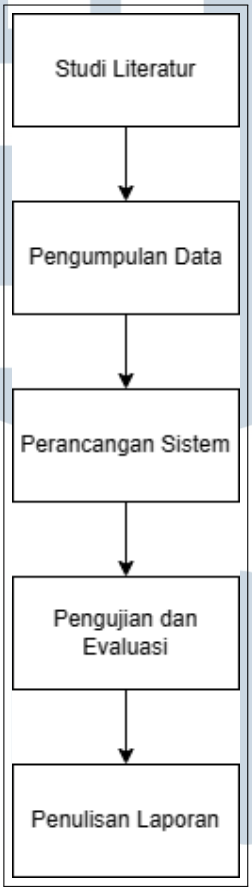
BAB 3
METODOLOGI PENELITIAN

3.1 Metodologi Penelitian

Dalam melakukan penelitian ini, terdapat metodologi penelitian atau langkah-langkah dalam mengerjakan penelitian. Berikut adalah metodologi yang akan dilakukan.



Gambar 3.1. Diagram alir metodologi penelitian

3.1.1 Studi Literatur

Studi literatur dilakukan untuk mempelajari serta memahami berbagai literatur yang sesuai dengan topik penelitian, seperti analisis sentimen, Twitter, dan *Random Forest*.

3.1.2 Pengumpulan Data

Pengumpulan data dilakukan untuk mengumpulkan data yang akan dipakai dalam penelitian ini menggunakan *library* yang bernama snscrape. Data yang akan diambil adalah data *tweets* berisikan teks yang dipublikasikan oleh user yang berhubungan dengan topik *Work from Home* (WFH).

3.1.3 Perancangan Sistem

Setelah data *tweets* diperoleh akan dilakukan perancangan sistem. Perancangan sistem akan dimulai dari tahap *text pre-processing* dengan tujuan agar pada saat membuat model latih menggunakan *Random Forest*, model akan lebih efisien dan memiliki tingkat akurasi, presisi, dan recall yang tinggi karena data yang tidak relevan sudah diproses sebelumnya. Setelah data telah dibersihkan akan dilakukan tahap pemberian label positif atau negatif pada data.

3.1.4 Pengujian dan Evaluasi

Setelah berhasil dalam membuat model latih, model akan melalui tahap pengujian dengan cara memprediksi sentimen dari data tes yang sudah disiapkan. Dilanjutkan dengan melakukan evaluasi model dengan *confusion matrix* untuk melihat tingkat akurasi, presisi, dan recall dari model yang telah dibuat.

3.1.5 Penulisan Laporan

Dilakukan penulisan laporan berdasarkan topik penelitian sebagai bentuk dokumentasi dari penelitian yang telah dilakukan. Penulisan laporan ini akan menjadi bukti bahwa penelitian ini telah dilakukan dari awal hingga selesai dan dapat menjawab rumusan masalah yang telah ditetapkan.

3.2 Perancangan Sistem

Bagian ini akan menguraikan dari masing-masing tahapan yang akan dilakukan. Berikut adalah uraian serta diagram alur dari sistem yang dibuat.

3.2.1 Gambaran Umum Sistem

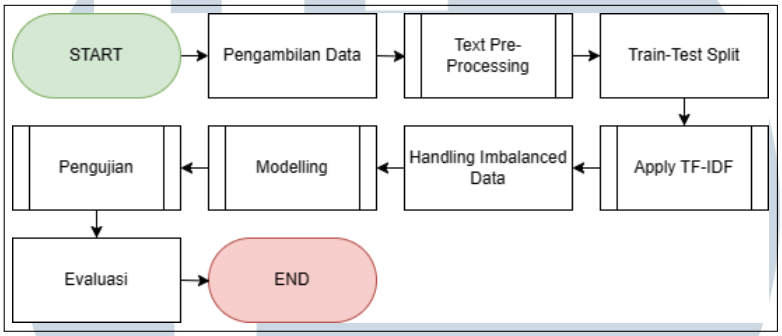
Penelitian ini membuat sistem yang dapat menganalisis sentimen tweets yang diambil dari media sosial *Twitter* dengan algoritma *Random Forest*. Pengambilan data dilakukan menggunakan *library snscrape* dengan format json. Selanjutnya data *tweets* yang telah diambil akan disimpan dalam bentuk csv agar dapat dilakukan proses berikutnya.

Data *tweets* yang telah disimpan akan diolah melalui tahap *text pre-processing*. Tahapan ini merupakan tahapan penting, karena proses ini akan membersihkan dan mempersiapkan data sebelum dilatih. Pada tahap ini akan dilakukan pembersihan data dari fitur yang tidak dipakai, seperti angka, tanda baca, *hashtag*, mengubah seluruh teks menjadi huruf kecil, dan penghapusan fitur lainnya. Selanjutnya akan dilakukan pengubahan kata slang menjadi kata baku dan menghapus teks yang tidak mengandung kata ”wfh”. Hal ini dilakukan agar data latih yang digunakan sesuai dengan topik penelitian. Setelah itu, dilakukan proses *back-translation* dari bahasa Indonesia ke bahasa Inggris, diterjemahkan kembali ke bahasa Indonesia. Selanjutnya dilakukan tahap *tokenizing* untuk memecahkan teks menjadi token dan melakukan *stopwords removal* untuk membuang kata yang tidak memiliki makna tertentu. Terakhir, dilakukan proses *stemming* untuk mengubah kata menjadi kata dasar dan menghapus teks yang panjangnya kurang dari 2 kata.

Data yang sudah bersih akan dilabel menggunakan kamus lexicon Inset. Inset merupakan kamus bahasa Indonesia, dimana setiap kata memiliki nilai positif atau negatif. *Tweets* akan dinilai setiap katanya menggunakan kamus Inset dan akan dijumlahkan. Ketika penjumlahan tersebut bernilai positif, maka *tweets* tersebut akan diberikan label positif. Sebaliknya, jika penjumlahan tersebut bernilai negatif, maka *tweets* tersebut diberikan label negatif. Setelah diberikan label, dilakukan *exploratory data analysis* untuk mendapatkan informasi yang terkandung dalam kedua sentimen tersebut, seperti visualisasi *word cloud* dan *n-gram*.

Tweets yang telah diberikan label akan dilakukan tahap *split data* dengan rasio 80% data *train* dan 20% data *test*. Selanjutnya data akan dihitung nilai TF-IDF untuk menentukan nilai dari setiap kata. Berikutnya dilakukan proses *handling imbalanced data*. Pada fase pembelajaran, algoritma pembelajaran mesin dapat terpengaruh oleh ketidakseimbangan dalam kumpulan data. Ketidakseimbangan ini mengacu pada perbedaan jumlah sampel dalam setiap kelas. Maka dari itu dilakukan penyetaraan jumlah sampel dalam setiap kelas menggunakan berbagai metode.

Setelah mendapatkan data *train* yang seimbang, dilakukan pelatihan model *Random Forest*. Kemudian akan dilakukan uji coba berdasarkan skenario yang telah dibuat. Langkah terakhir adalah mengukur kinerja model dengan *cross validation* dan mengukur hasil prediksi pada data *test* menggunakan *confusion matrix* untuk mendapatkan nilai *accuracy*, *precision*, dan *recall*.



Gambar 3.2. Diagram alir gambaran umum penelitian

3.2.2 Pengambilan Data

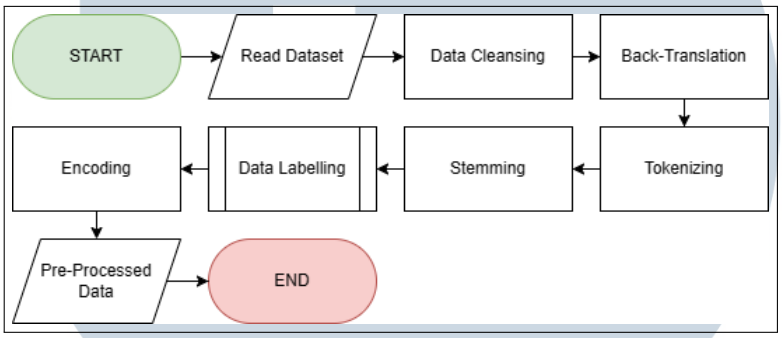
Pengambilan data dilakukan dengan bantuan *library Snscrape* dan mencari *tweets* yang mengandung kata kunci ”WFH”, ”Work From Home”, dan ”Kerja dari rumah”. *Tweets* yang diambil adalah *tweets* yang diunggah antara tanggal 1 Januari 2022 hingga 5 Maret 2023. *Data scraping* menggunakan *library Snscrape* memiliki batasan dalam mengambil data, dimana rentang tanggal hanya bisa di tahun yang sama. Selain itu tidak dapat menyaring bahasa dari *tweets*, sehingga diperlukan *library langdetect* untuk melakukan deteksi pada *tweets*.

3.2.3 Text Pre-Processing

Text pre-processing adalah tahap persiapan data sebelum data dilatih pada model *Random Forest*. Terdapat beberapa langkah yang dilakukan pada tahap ini. Seperti yang dapat dilihat pada gambar 3.3, *Text Pre-processing* dimulai dengan proses *data cleansing*, yaitu pembersihan data dari fitur yang tidak akan digunakan. Tahap pada *data cleansing* seperti, menghilangkan kolom yang tidak dipakai, menghilangkan data duplikat, menghilangkan data yang bersifat *spam*, mengekstrak *hashtag* pada *tweets*, menghilangkan beberapa fitur seperti tanda baca, *retweets*, *hashtag*, dan sebagainya, menghilangkan *tweets* yang bukan berbahasa Indonesia,

mengubah seluruh teks menjadi huruf kecil, mengganti kata slang dengan kata baku, dan menghilangkan teks jika tidak memiliki kata ”wfh”.

Setelah melalui proses *data cleansing*, *tweets* akan melalui proses *back-translation*. Hal ini digunakan untuk menghilangkan kata dalam bahasa asing dan memperbaiki kata yang memiliki kesalahan dalam pengetikan. Selain itu *back-translation* juga berguna untuk memperbanyak variasi kata dalam teks, sehingga teks memiliki kata yang lebih bervariasi.



Gambar 3.3. Diagram alir *Text Pre-Processing*

Data yang telah melalui tahap *back-translation* akan melalui tahap *tokenizing* untuk memecah kata menjadi token-token. Selanjutnya token ini akan masuk pada proses *stopwords removal*, dimana token atau kata yang tidak berarti akan dihapus. Setelah itu akan dilakukan proses *stemming*. Proses *stemming* adalah proses mengubah kata menjadi kata dasar, sehingga kata yang memiliki makna yang sama akan memiliki nilai yang sama. Selanjutnya token atau kata akan digabung menjadi kalimat dan menghilangkan data yang tidak memiliki panjang lebih dari 3 kata.

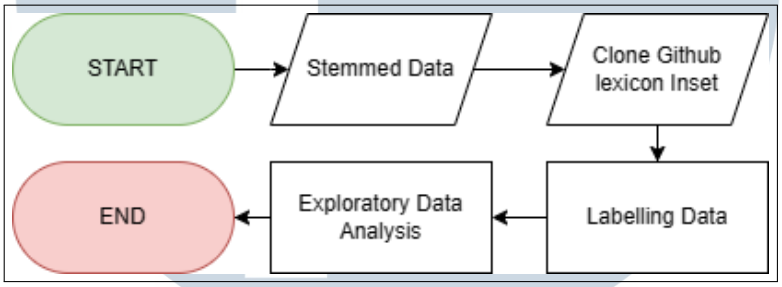
Tahap berikutnya adalah tahap *data labelling* untuk memberikan label positif atau negatif pada data. Proses *data labelling* dapat dilihat diagram alirnya pada gambar 3.4. Setelah setiap *tweet* diberi sentimen, dilakukan proses *encoding* untuk merubah sentimen positif dan negatif menjadi 1 dan 0.

3.2.4 Data Labelling

Tahap ini dilakukan untuk memberikan label pada setiap *tweets* yang telah melalui proses *text pre-processing*. Label yang diberikan adalah label positif dan negatif tergantung dengan kata yang terkandung dalam setiap teks. *Data labelling* dilakukan dengan bantuan Lexicon Inset. Diagram alir proses ini dapat dilihat pada gambar 3.4.

Seperti yang dapat dilihat pada gambar 3.4 data akan di label menggunakan kamus Inset yang telah dibuat. Kamus Inset berisikan kata dalam bahasa Indonesia dan memiliki nilai positif atau negatif sesuai dengan pengertiannya. Setiap kata akan dijumlahkan nilainya dan hasil dari penjumlahan tersebut akan menentukan sentimen dari *tweets*. Jika nilainya diatas 0, maka akan diberi label positif. Sedangkan *tweets* yang memiliki nilai dibawah 0, maka akan diberi label negatif.

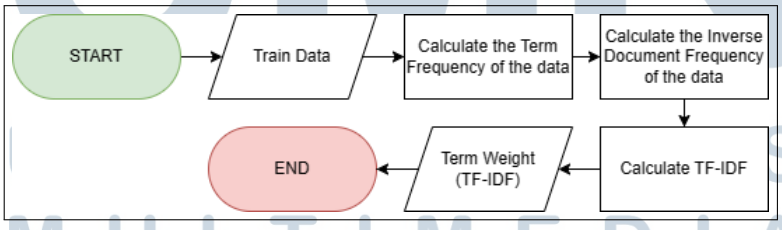
Data yang telah diberikan label akan masuk pada proses *exploratory data analysis* (EDA). Pada proses ini akan dilakukan pengeksploran data untuk mendapatkan informasi lebih detail. EDA yang dilakukan seperti melakukan visualisasi *n-gram* dan *word clouds*.



Gambar 3.4. Diagram alir *Data Labelling*

3.2.5 Apply TF-IDF

Pada gambar 3.5 menunjukan diagram alir proses *Apply TF-IDF*. TF-IDF berguna untuk memberikan bobot pada setiap *term* yang dihitung berdasarkan kemunculan *term* tersebut pada dokumen dari hasil *text pre-processing*.



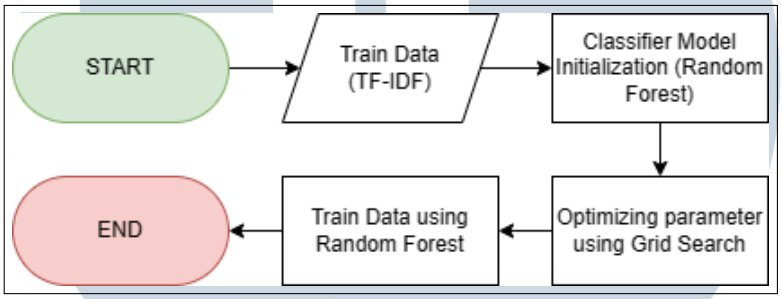
Gambar 3.5. Diagram alir *apply TF-IDF*

Langkah pertama yang dilakukan adalah menghitung nilai TF menggunakan Persamaan 2.3 untuk menghitung frekuensi kemunculan *term*. Selanjutnya dilakukan perhitungan nilai IDF menggunakan Persamaan 2.4. Setelah didapatkan nilai TF dan IDF, akan dilakukan perhitungan TF-IDF menggunakan

Persamaan 2.5. Setiap *term* akan memiliki nilai TF-IDF yang akan disimpan untuk nantinya digunakan dalam proses klasifikasi.

3.2.6 Modelling

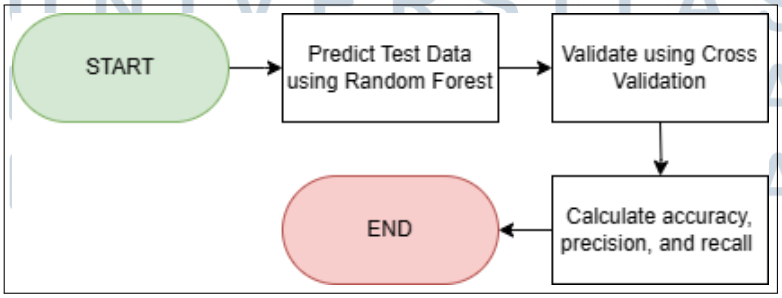
Proses *modelling* adalah proses melatih data yang telah dibersihkan dan diberikan label. Namun, sebelum melatih data tersebut, ada beberapa tahapan yang perlu dilakukan. Berikut adalah diagram alir dari tahap *modelling*.



Gambar 3.6. Diagram alir *Modelling*

Pada gambar 3.6 dapat dilihat alur dari tahap *modelling*. Dengan menggunakan nilai TF-IDF yang didapatkan pada proses sebelumnya dan data yang telah dipisahkan menjadi data *training* dan *testing*, dilakukan inisialisasi model pelatihan algoritma *Random Forest* untuk masuk pada tahap pelatihan model. Sebelum dilakukan pelatihan model menggunakan data *train*, akan dilakukan pencarian beberapa parameter dari algoritma *Random Forest* yang optimal menggunakan *Randomized Search*. Selanjutnya akan dilakukan pelatihan model menggunakan parameter tersebut dan melatih data *train* yang telah disiapkan.

3.2.7 Pengujian



Gambar 3.7. Diagram alir Pengujian

Model *Random Forest* akan melalui tahap prediksi data *test* yang telah disiapkan untuk mengklasifikasikan sentimen untuk data *test* tersebut. Selanjutnya dilakukan proses *Cross validation* yang berguna untuk menilai kinerja dan generalisasi model *machine learning*. Terakhir dilakukan perhitungan nilai *accuracy*, *precision*, dan *recall* dari model yang telah dilatih menggunakan*Confusion Matrix*.

3.2.8 Evaluasi

Tahapan terakhir dari dari implementasi algoritma *Random Forest* ini adalah evaluasi. Tahap ini berguna untuk mengevaluasi serta menganalisis dari kinerja model yang dibangun. Tujuan utama dari tahap ini merupakan untuk memperoleh pemahaman yang lebih baik mengenai keberhasilan serta kekurangan dari model yang dibangun dan dapat memberikan informasi yang berharga untuk pembaca.