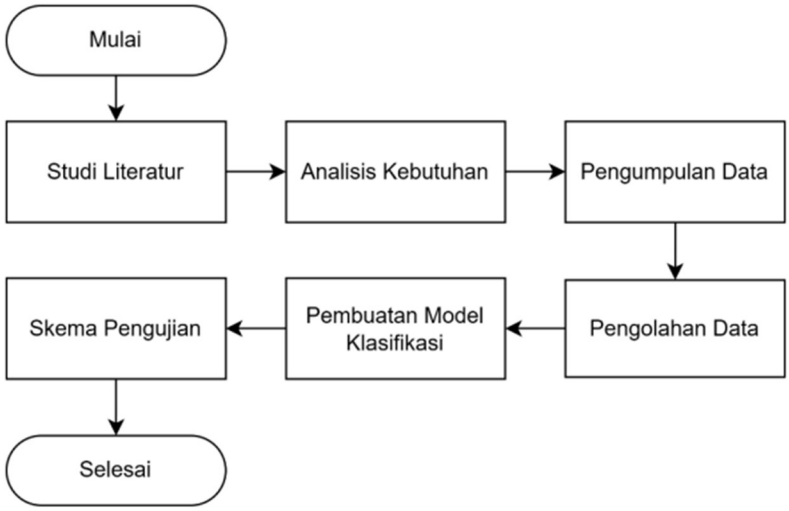
BAB II
METODE PENELITIAN

Pada bab Metodologi, penelitian ini menjelaskan langkah-langkah yang digunakan selama penelitian analisis sentimen. Diagram alir digunakan untuk menggambarkan terjadinya suatu proses sehingga mempermudah memahami proses yang sedang terjadi. Penelitian ini dapat dikategorikan dalam tipe penelitian implementatif karena mengimplementasi algoritma Random Forest untuk mengetahui sentimen masyarakat terhadap isu perceraian di Indonesia.

Langkah-langkah metodologi yang dilakukan untuk mencapai tujuan dari penelitian ini digambarkan melalui diagram alir di bawah ini.



Gambar II.1 Langkah-langkah Penelitian

Tahapan penelitian ini dimulai dengan melakukan studi literatur untuk mengidentifikasi permasalahan yang hendak diselesaikan menggunakan teknologi. Kemudian dilajutkan dengan analisis kebutuhan selama penelitian, lalu melakukan pengumpulan data berdasarkan topik yang diangkat. Pada tahap pembuatan model terlebih dahulu dilakukan preprocessing data yang kemudian dilanjutkan dengan pembuatan model klasifikasi menggunakan algoritma machine learning. Di akhir penelitian akan dilakukan pengujian terhadap hasil model berdasarkan rancangan skenario.

2.1 Studi Literatur

Studi literatur dilakukan dengan mengumpulkan sumber informasi yang relevan berdasarkan topik yang diangkat. Sumber referensi yang digunakan sebagai dasar acuan penelitian meliputi buku, skripsi, penelitian, jurnal, karya ilmiah, tesis, internet. Topik-topik referensi yang digunakan berkaitan dengan analisis sentimen, penggunaan algoritma Random Forest, dan topik lainnya yang mendukung tujuan penelitian ini.

2.2 Analisis Kebutuhan

Pada tahap analisis kebutuhan ini meliputi analisis kebutuhan data dan kebutuhan software & hardware agar penelitian berjalan sesuai harapan.

2.3 Kebutuhan Data

Data yang dibutuhkan dalam penelitian ini adalah data tweet pada media sosial Twitter yang berhubungan dengan isu perceraian di indonesia, dengan batasan data yang diambil mulai dari tanggal 29 November 2023 sampai 13 Desember 2023.

2.4 Kebutuhan Hardware dan Software

Pada penelitian ini diperlukan perangkat software dan hardware yang dapat menunjang keberhasilan penelitian. penelitian ini menggunakan Laptop Asus Vivobook M1403QA dengan kapasitas RAM sebesar 16GB dan CPU mencapai 8 core. Sedangkan software yang akan digunakan meliputi web browser untuk menjalankan program, jupyter notebook, dan bahasa pemrograman Python.

2.5 Pengumpulan Data

Tahap pengumpulan data bertujuan mendapatkan data mentah sebanyak mungkin dengan memperhatikan kata kunci yang digunakan dan batasan waktu. Tahapan ini meliputi penjelasan sumber data dan rancangan awal struktur data.

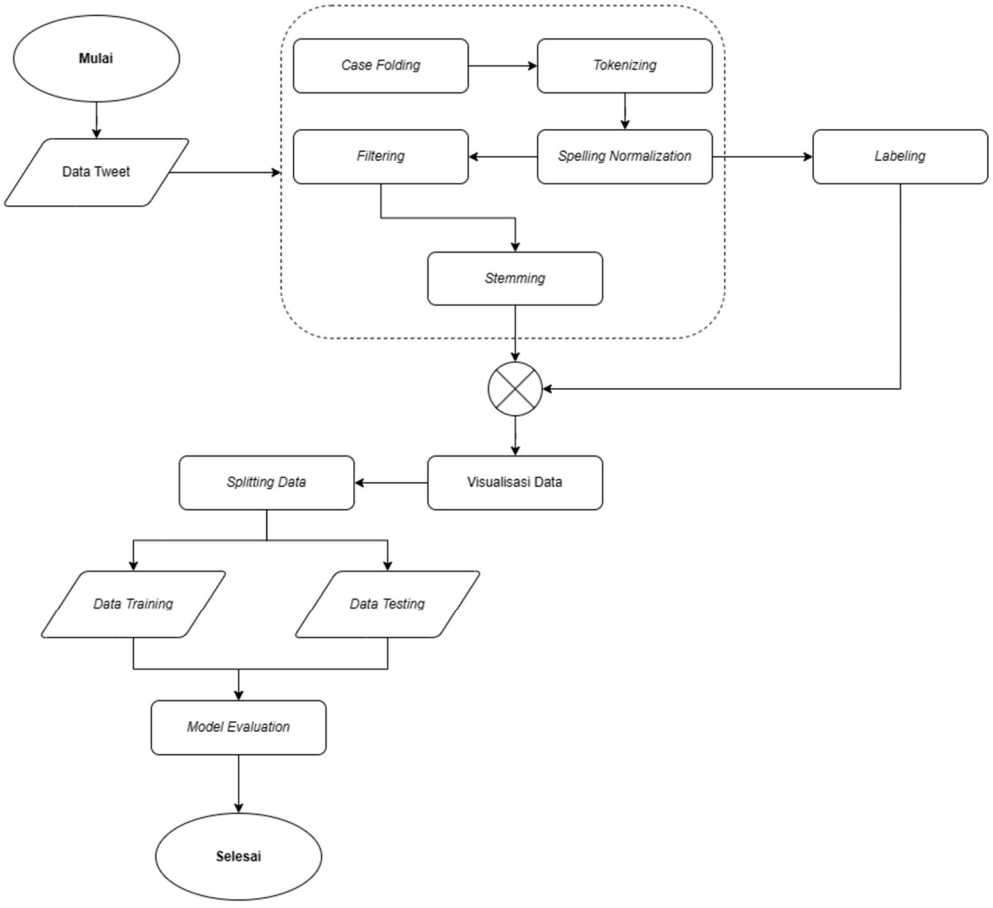
2.5.1 Sumber Data

Penelitian ini menggunakan sumber data sekunder yang diperoleh dari media sosial Twitter melalui TwitterAPI dengan mengambil seluruh data tweet yang berkaitan dengan opini atau pendapat masyarakat tentang isu perceraian di Indonesia. Data tweet diambil dengan ketententuan

berbahasa Indonesia dan menggunakan kata kunci “perceraian”. Data tweet yang diambil adalah tweet yang dibuat sejak tanggal 29 Januari 2023 sampai 13 Desember 2023.

2.6 Pengolahan Data

Data set tweet pada Twitter yang telah dikumpulkan mengenai isu perceraian diolah menggunakan software Python 3.11 (64-bit). Sebelum dilakukan klasifikasi menggunakan algoritma machine learning, terlebih dahulu dilakukan preprocessing guna menghilangkan beberapa permasalahan yang bisa mengganggu tahap pemrosesan data menggunakan algoritma klasifikasi. Selanjutnya dilakukan pelabelan secara otomatis untuk mengetahui apakah sentimen tweet termasuk positif, netral, atau negatif. Dilanjutkan dengan melakukan visualisasi terhadap dataset dan perhitungan menggunakan metode feature extraction. Langkah-langkah preprocessing data dijelaskan menggunakan diagram alir seperti pada Gambar 3.2

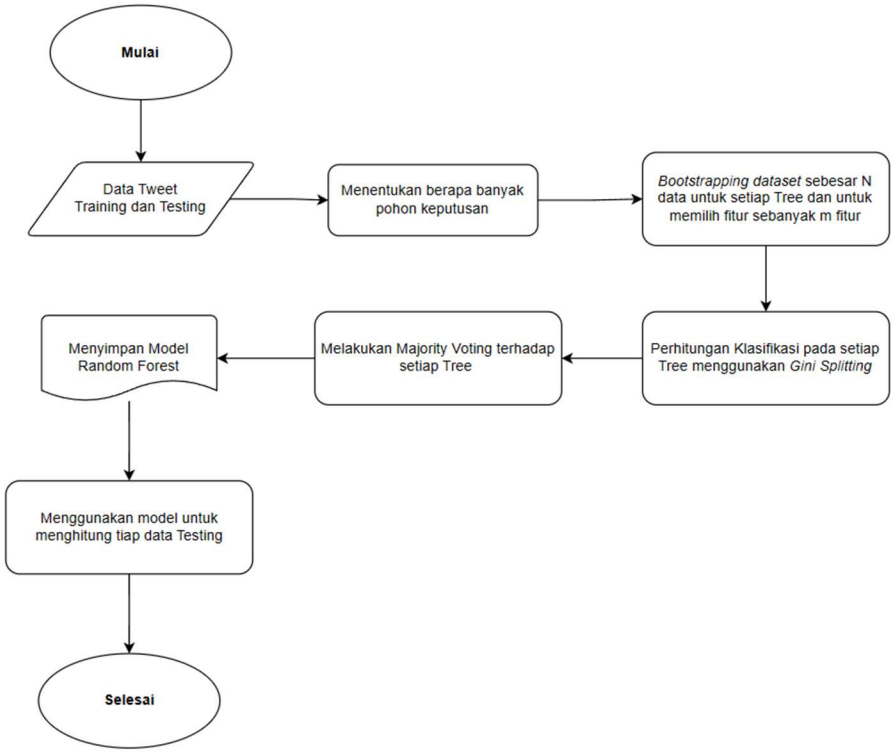


Gambar II.2 Diagram alir metode pengolahan data

1. Menggunakan data tweet yang didapatkan melalaui crawling dan scraping pada platform media sosial Twitter melalui fasilitas terbuka Twitter API (Application Programming Interface) berdasarkan topik permasalahan.
2. Data Preprocessing
    a. Melakukan case folding, yaitu mengubah semua kata dengan huruf besar (upper case) dalam dokumen menjadi huruf kecil (lower case) seluruhnya serta menghilangkan tanda baca, link, nomor, dan character.
    b. Melakukan tokenizing yang bertujuan memecah kalimat menjadi kumpulan kata yang biasa disebut token.
    c. Melakukan spelling normalization untuk mengubah data tweet yang mengandung kata tidak baku menjadi kata baku. Melakukan filtering atau stopword removal adalah tahapan dengan menghilangkan kata yang dianggap tidak penting (stoplist) dan menyimpan kata yang bermakna penting (wordlist).
    d. Melakukan stemming, yaitu mengonversi kata yang berimbuhan menjadi bentuk kata dasar.
3. Melabeli setiap data tweet dengan label positif, negatif, dan netral menggunakan package Python yaitu transformers dengan model fine-tuned Indonesian RoBERTa
4. Melakukan visualisasi data tweet berupa grafik dan word cloud menggunakan package matplotlib, seaborn dan wordcloud yang tersedia di Python.
5. Membagi data tweet menjadi data training dan testing dengan rasio tertentu yang nantinya digunakan sebagai masukkan pada tahap pembuatan model klasifikasi.
6. Melakukan evaluasi model untuk menilai seberapa baik model dalam memprediksi sentiment pada dataset isu perceraian.

2.7 Pembuatan Model Klasifikasi

Setelah dataset dilakukan preprocessing dan dilakukan lebeling selanjutkan dilakukan klasifikasi menggunakan algoritma machine learning Random Forest guna mengetahui performa model. Setelah itu dipilih model terbaik untuk mengklasifikasi data tweet baru. Langkah-langkah klasifikasi data dijelaskan melalui diagram alir seperti pada Gambar 3.3.



Gambar II.3 Diagram alir metode klasifikasi

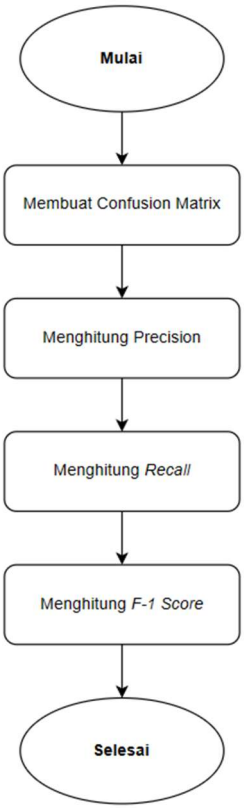
1. Menggunakan data tweet hasil ekstraksi fitur meliputi data training dan data testing yang dipanggil dari file data berformat pickle.
2. Membuat model klasifikasi Random Forest
    a. Menentukan banyak tree untuk setiap model pohon klasifikasi dari $b = 1$ sampai B sehingga untuk setiap tree dapat dituliskan T_b
    b. Membuat sampel dataset untuk setiap tree T_b dari N total data training dengan cara bootstrapping atau diacak
    c. Menentukan variabel m yaitu banyak fitur atau atribut secara acak sebagai splitting point sebanyak $\sqrt{p}$ total fitur.
    d. Melakukan perhitungan klasifikasi untuk setiap tree T_b diawali dengan menentukan root node dengan mencari kriteria pemisah yang memiliki nilai gini-split paling kecil pada data training.
    e. Mengulangi perhitungan gini-split untuk membentuk internal node sampai data dari sebuah node memiliki kelas homogen atau sudah tidak terdapat atribut beserta

kriteria pemisahan, proses ini dilakukan pada setiap T_b. Jadi setiap T_b memiliki internal node pemisah dan kedalaman yang berbeda-beda.

f. Melakukan majority voting terhadap hasil klasifikasi model setiap Tb sehingga didapatkan hasil akhir klasifikasi yaitu kelas prediksi. Proses ini disebut *debugging*.

2.8 Model Evaluation

Tahapan skema evaluasi model meliputi beberapa Langkah yang digambarkan melalui diagram alir di bawah ini.



Gambar II.4 Diagram Alir Evaluasi Model

1. Confusion matrix adalah tabel yang digunakan untuk menggambarkan kinerja model klasifikasi pada set data pengujian, membandingkan prediksi model dengan nilai sebenarnya. Matriks ini terdiri dari empat sel: True Positive (TP) yaitu jumlah kasus yang benar diprediksi sebagai positif, True Negative (TN) yaitu jumlah kasus yang benar diprediksi sebagai negatif, False Positive (FP) yaitu jumlah kasus yang seharusnya negatif, tetapi diprediksi sebagai positif, dan False Negative (FN) yaitu jumlah kasus yang seharusnya positif, tetapi diprediksi sebagai negatif.

2. Precision (presisi) adalah ukuran dari sejauh mana model benar dalam memprediksi kasus positif. Precision memberikan informasi tentang akurasi prediksi positif model dan mengukur tingkat kepercayaan terhadap kelas positif.
3. Recall (recall atau sensitivitas) adalah ukuran dari sejauh mana model mampu menemukan semua kasus positif yang seharusnya ditemukan. Recall memberikan pandangan tentang seberapa baik model dapat mengidentifikasi semua kasus positif yang ada.
4. F1 score adalah metrik yang menggabungkan precision dan recall menjadi satu angka yang menyediakan ukuran komprehensif tentang kinerja model. F1 score berguna ketika terdapat trade-off antara precision dan recall; misalnya, ketika keseimbangan antara kedua metrik itu penting.