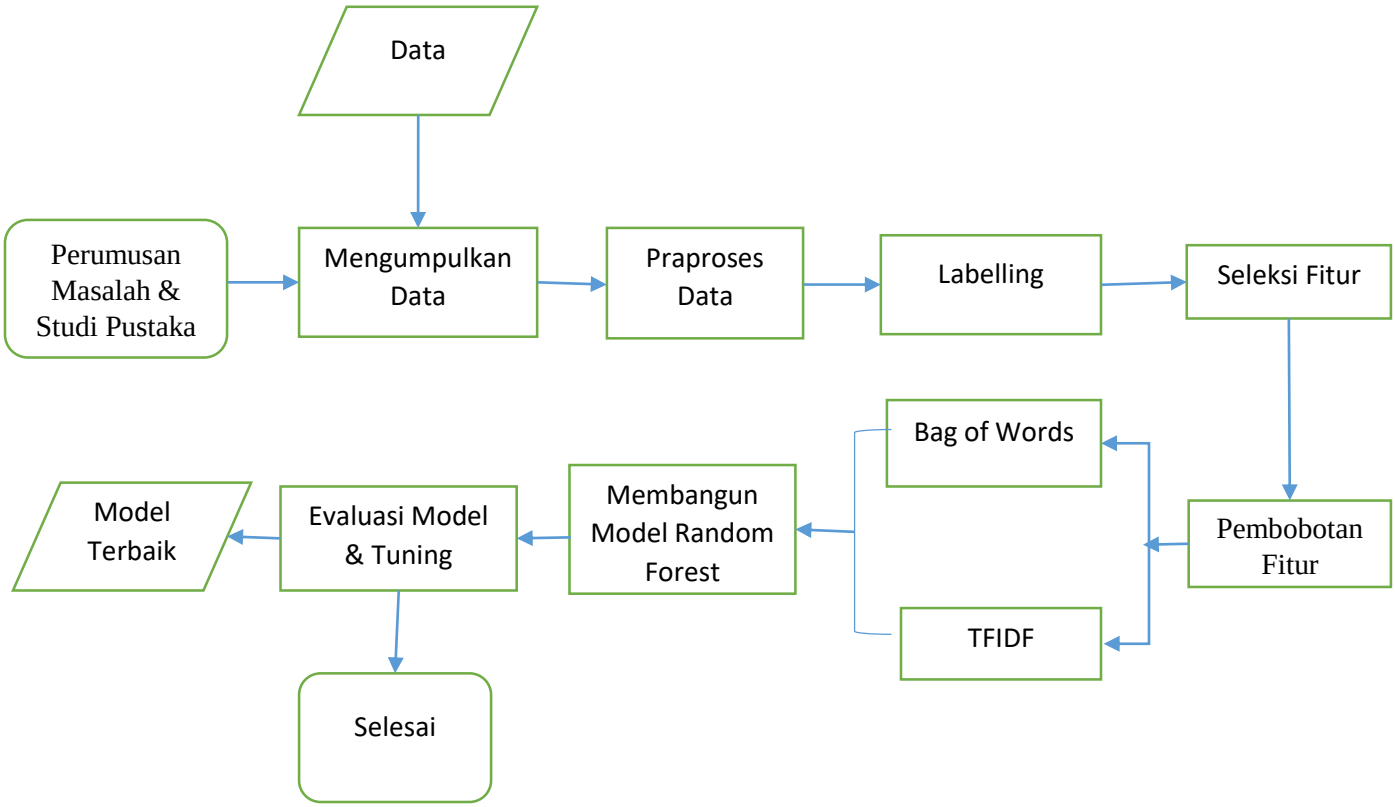
BAB III

METODOLOGI PENELITIAN

Pada gambar berikut ini, ditampilkan ilustrasi tentang alur penelitian yang akan digunakan dalam penelitian ini, memberikan gambaran visual tentang langkah-langkah yang akan dilakukan pada penelitian ini.



Gambar 3. 1 Alur Penelitian

3.1 Sumber Data

Data penelitian yang digunakan dalam penelitian ini adalah data *tweet* berbahasa Indonesia yang diperoleh dari *platform* media sosial *Twitter*. Pengumpulan data dilakukan dengan menggunakan kata kunci yang relevan, yaitu "Kampus Merdeka", "Merdeka Belajar",

"MSIB", dan "MBKM". Kata kunci ini digunakan sebagai filter pencarian untuk mengambil *tweet* yang berhubungan dengan topik penelitian.

Proses pengumpulan data dilakukan dalam rentang waktu mulai dari bulan April 2023 hingga Juni 2023. Selama periode ini, *tweet-tweet* yang mengandung kata kunci tersebut diambil dan disimpan. Data diambil secara terpisah untuk setiap kata kunci yang digunakan, dan kemudian data-data tersebut digabungkan menjadi satu dataset yang lengkap.

Data yang diambil dan digunakan dalam penelitian ini berbentuk teks, yaitu teks tweet berbahasa Indonesia. Setiap data merupakan representasi dari satu *tweet* yang mengandung informasi terkait dengan topik yang sedang diteliti. Data teks ini akan menjadi dasar untuk melakukan analisis sentimen dan pembangunan model klasifikasi pada penelitian ini.

Tabel 3. 1 Contoh Data *Tweet*

No	*Tweet*
1	cape ga sih..... dibantai project uas, blm nyiapin cv buat mbkm, sempro 😭😭😭 help ya Allah semoga masih waras bismillah menuju calon orang sukses 😊
2	aaaa capek banget mbkm
3	Sehubungan dengan tingginya antusiasme terhadap Program Merdeka Belajar - Magang Merdeka (MBKM), Departemen Pengembangan Karier BEM FH UI 2023 dan Departemen Advokasi BEM FH UI 2023 akan mengadakan sosialisasi mengenai MBKM.
4	Melihat banyaknya mahasiswa yang antusias mengikuti program-program dari Kampus Merdeka, URC kembali hadir menjadi wadah bagi mahasiswa yang tertarik dan ingin mengikuti Program Kampus Merdeka dengan mengadakan Sosialisasi Kampus Merdeka dengan tema "Get to Know MBKM to Explore
5	kuliah tinggal lulusnya malah dipaksa mbkm 4 bulan dipuskesmas wkwkwk
6	Untung gw konversi dari MBKM ahaha jadi santuyyy

3.2 Tahap Penelitian

Ada empat tahapan utama dalam alur penelitian ini, yaitu tahap awal penelitian, tahap pemrosesan data, tahap klasifikasi teks dan tahap analisis hasil.

3.2.1 Tahapan Awal

Tahap awal atau tahap perumusan terdiri dari tiga proses yang akan dilakukan sebagai berikut:

1. Perumusan Masalah

   Pada tahap pertama alur penelitian, langkah awal yang dilakukan adalah proses perumusan masalah yang menjadi latar belakang penelitian ini. Penelitian ini berfokus pada klasifikasi sentiment program MBKM yang terjadi pada media sosial *Twitter*, yang dapat dikategorikan menjadi dua kategori, yaitu sentimen positif dan sentimen negatif. Proses perumusan masalah ini bertujuan untuk mengidentifikasi masalah yang ingin dipecahkan dan memberikan arah penelitian yang jelas. Langkah selanjutnya adalah memulai penelitian dengan mengkaji literatur yang relevan dengan masalah penelitian dan mengumpulkan data.

2. Studi Pustaka

   Pada tahap ini, peneliti melakukan pengumpulan dan pemahaman yang mendalam terhadap berbagai teori, perhitungan, dan pembahasan yang berkaitan dengan penelitian yang sedang dilakukan. Sumber literatur yang digunakan mencakup jurnal ilmiah, buku, dan artikel penelitian yang menjadi referensi utama dalam mengembangkan pemahaman yang komprehensif tentang topik penelitian.

   Dalam proses pengumpulan literatur, peneliti memfokuskan pada subjek teori yang relevan dengan penelitian, seperti teknik preprocessing teks untuk membersihkan dan mempersiapkan data teks sebelum dilakukan analisis, penggunaan metode *Bag of Words* dan *TF-IDF* sebagai teknik pembobotan fitur dalam analisis teks, algoritma *Random Forest* yang digunakan untuk melakukan klasifikasi data, serta penelitian sebelumnya yang telah dilakukan dalam bidang yang serupa dan relevan dengan topik penelitian yang sedang dijalankan.

   Dengan mempelajari berbagai teori, perhitungan, dan pembahasan yang ada, peneliti dapat memperoleh pemahaman yang mendalam tentang konsep dan metodologi yang diperlukan dalam menjalankan penelitian ini. Peneliti juga dapat memanfaatkan temuan dan kesimpulan dari penelitian sebelumnya untuk mengembangkan landasan teoritis yang kuat.

3. Pengumpulan Data

   Pada tahap ini, dilakukan pengumpulan data yang akan menjadi dataset dalam penelitian ini. Sumber data yang digunakan berasal dari media sosial *Twitter*, peneliti

menggunakan kata kunci yang telah ditentukan, seperti "mbkm", "kampus merdeka", dan "merdeka belajar". Kata kunci ini berfungsi sebagai filter pencarian, sehingga data yang terkumpul berkaitan dengan topik yang sedang diteliti.. Proses pengumpulan data dilakukan melalui crawling menggunakan *Twitter* API. API ini memungkinkan akses ke data publik yang tersedia di *platform Twitter*. Dengan menggunakan API, peneliti dapat mengambil data *tweet* yang mengandung kata kunci yang telah ditentukan sebelumnya. Pengumpulan data dilakukan dalam rentang waktu tertentu, yaitu mulai dari tanggal 14 April 2023 hingga 12 Juni 2023, untuk memperoleh data yang relevan dan terkini.

3.2.2 Praproses Data

Data yang terkumpul dalam penelitian ini merupakan data teks yang bersifat tidak terstruktur. Untuk dapat mengolah data tersebut secara komputasi, diperlukan tahapan praproses data yang bertujuan untuk mengubahnya menjadi data yang terstruktur. Tahapan praproses data meliputi *data cleaning, case folding, stopword removal,* dan *stemming.* Tahapan preprocessing akan dijelaskan sebagai berikut :

1. *Data Cleaning*

   Proses ini bertujuan untuk menghilangkan elemen-elemen yang kurang relevan dalam data, seperti hashtag, username, URL, serta tanda baca seperti koma, titik, dan lain-lain. Dengan menghapus elemen-elemen tersebut, data menjadi lebih fokus pada teks utama dari setiap *tweet*, sehingga memudahkan analisis dan pengolahan lebih lanjut.

2. *Case Folding*

   Pada tahap ini, setiap karakter dalam teks diubah menjadi huruf kecil sehingga semua kata memiliki format yang seragam. Misalnya, kata-kata yang awalnya ditulis dalam huruf kapital atau campuran huruf kapital dan kecil, akan diubah menjadi huruf kecil sepenuhnya.

3. *Stemming*

   Proses ini dilakukan untuk mengubah setiap kata menjadi kata dasar dengan menghapus imbuhan di awal dan akhir. Dalam bahasa Indonesia, terdapat banyak kata

yang memiliki imbuhan seperti awalan, akhiran, atau sisipan. Misalnya, kata "mengambil" dapat dipotong imbuhan "meng-" sehingga menjadi kata dasar "ambil". Proses ini dilakukan untuk mengurangi variasi kata yang sebenarnya memiliki makna yang sama, sehingga memudahkan pemrosesan dan analisis lebih lanjut.

4. *Stopword Removal*

   Proses ini dilakukan untuk mengahapus kata – kata umum yang tidak memiliki makna penting atau kontribusi signifikan dalam teks. Kata-kata umum ini biasanya terdiri dari kata hubung, kata bantu, atau kata-kata yang sering muncul namun tidak memberikan informasi yang berarti dalam teks.

5. *Tokenizing*

   Proses ini dilakukan untuk memecah teks atau kalimat menjadi unit – unit yang lebih kecil yang disebut token. Proses tokenisasi membantu dalam mempersiapkan teks untuk analisis lebih lanjut, seperti pembobotan fitur atau ekstraksi fitur.

3.2.3 *Data Labelling*

Setelah melalui tahap praproses data, langkah selanjutnya adalah melakukan pelabelan pada dataset untuk mengelompokkan teks atau kalimat ke dalam sentimen positif atau negatif. Pelabelan ini dilakukan dengan menggunakan kamus atau daftar kata yang memiliki konotasi positif dan kamus atau daftar kata yang memiliki konotasi negatif.

Proses pelabelan dilakukan dengan memperhitungkan jumlah kata positif dan jumlah kata negatif yang terdapat dalam suatu teks. Untuk menentukan kelas sentimen, akan dilakukan perhitungan kata positif dan jumlah kata negatif dalam teks tersebut. Jika jumlah kata positif lebih banyak daripada kata negatif, maka teks tersebut akan dilabeli sebagai sentimen positif. Sebaliknya, jika jumlah kata negatif lebih banyak daripada kata positif, maka teks tersebut akan dilabeli sebagai sentimen negatif.

Metode pelabelan ini memungkinkan untuk mengklasifikasikan teks berdasarkan sentimen yang terkandung dalamnya, dengan memanfaatkan kamus kata positif dan kata negatif. Dengan menggunakan pendekatan ini, dapat diidentifikasi sentimen yang terkandung dalam teks secara relatif, berdasarkan frekuensi kata-kata yang terkait dengan konotasi positif dan negatif.

3.2.4 Seleksi Fitur

Setelah data melalui tahap praproses, langkah selanjutnya adalah melakukan seleksi fitur pada data untuk mengeliminasi fitur yang kurang relevan. Pada penelitian ini, akan diggunakan metode seleksi fitur yang dikenal sebagai metode *Chi-Square.*

Metode *Chi-Square* akan menerapkan suatu rumus perhitungan yang menghasilkan nilai statistik *Chi-Square*. Nilai ini dapat digunakan untuk menguji hubungan antara suatu term atau kata dengan kelas yang ada. Dalam konteks penelitian ini, digunakan metode *Chi-Square* untuk menentukan sejauh mana suatu term atau kata berhubungan dengan kelasnya.

Proses perhitungan *Chi-Square* didasarkan pada tabel kontingensi yang berisi nilai frekuensi suatu kata terhadap kelasnya. Tabel ini digunakan untuk menghitung frekuensi observasi yang terjadi di dalam data dan membandingkannya dengan ekspektasi frekuensi yang diharapkan jika tidak ada hubungan antara term atau kata dengan kelasnya.

Dengan menggunakan metode *Chi-Square* dalam seleksi fitur, dapat mengidentifikasi dan mempertahankan fitur-fitur yang memiliki hubungan yang signifikan dengan kelas, sementara fitur-fitur yang memiliki hubungan yang lemah atau tidak signifikan dapat dieliminasi. Hal ini akan membantu memfokuskan perhatian pada fitur-fitur yang lebih informatif dan relevan dalam proses klasifikasi.

Dengan demikian, langkah seleksi fitur menggunakan metode *Chi-Square* setelah tahap praproses akan membantu memperbaiki kualitas data dengan menghilangkan fitur yang kurang relevan dan mempertahankan fitur-fitur yang memiliki hubungan yang signifikan dengan kelas, sehingga dapat meningkatkan performa model klasifikasi yang akan dibangun selanjutnya.

3.2.5 Pembobotan Fitur

Setelah data telah melewati proses praproses dan penyaringan fitur menggunakan metode $chi - square$ (χ^2), langkah berikutnya adalah melakukan pembobotan fitur sebelum proses klasifikasi dilakukan. Pembobotan fitur bertujuan untuk memberikan nilai numerik pada setiap fitur yang ada dalam data teks, sehingga fitur-fitur tersebut dapat digunakan sebagai input untuk algoritma *machine learning* yang memerlukan data dalam bentuk numerik.

Dalam penelitian ini, akan digunakan dua metode pembobotan fitur yang umum digunakan, yaitu *Bag of Words* dan *TF-IDF*. Metode *Bag of Words* akan menghitung frekuensi kemunculan kata-kata dalam teks dan memberikan bobot berdasarkan frekuensi tersebut. Sementara itu, metode *TF-IDF* (*Term Frequency-Inverse Document Frequency*) akan memberikan bobot yang lebih signifikan pada kata-kata yang jarang muncul tetapi penting dalam konteks dokumen.

3.2.6 Klasifikasi Teks

Setelah dilakukan pembobotan fitur menggunakan metode *Bag of Words* (*BoW*) dan *Term Frequency-Inverse Document Frequency* (*TF-IDF*), langkah selanjutnya dalam proses ini adalah membangun model klasifikasi menggunakan algoritma *Random Forest*. Algoritma *Random Forest* adalah salah satu algoritma *machine learning* yang populer dalam tipe *supervised learning*.

Proses membangun model klasifikasi melibatkan pembagian data menjadi dua bagian, yaitu *data training* dan *data testing*. *Data training* akan digunakan untuk melatih (*train*) model klasifikasi, sedangkan *data testing* akan digunakan untuk menguji (*test*) performa model yang telah dilatih menggunakan *data training*. Dalam penelitian ini, proporsi yang digunakan untuk pembagian *data training* dan *data testing* adalah 90% untuk *data training* dan 10% untuk *data testing*. Artinya, 90% dari seluruh data yang tersedia akan digunakan untuk melatih model klasifikasi, sementara 10% sisanya akan digunakan untuk menguji performa model yang telah dilatih. Pembagian data dengan proporsi seperti ini penting untuk memastikan bahwa model klasifikasi telah belajar dengan cukup baik dari *data training* yang representatif dan kemudian dapat diuji dengan *data testing* yang tidak pernah dilihat oleh model sebelumnya. Dengan menggunakan proporsi yang sesuai, dapat diukur seberapa baik model dapat melakukan prediksi atau klasifikasi pada data yang baru dan belum pernah dilihat sebelumnya.

Selanjutnya, performa model klasifikasi dapat dievaluasi menggunakan berbagai metrik evaluasi seperti akurasi, presisi, *recall*, dan *f-measure*. Metrik-metrik ini memberikan informasi tentang sejauh mana model klasifikasi mampu mengklasifikasikan data dengan benar dan menganalisis performa model secara lebih rinci.

Dengan membangun dan menguji model klasifikasi menggunakan algoritma *Random Forest* serta melakukan pembagian *data training* dan *data testing* dengan proporsi yang tepat, penelitian ini diharapkan dapat menghasilkan model klasifikasi yang akurat dan dapat digunakan untuk melakukan prediksi pada data teks terkait sentimen program MBKM.

3.2.7 Analisis Hasil

Pada tahap ini, dilakukan pengujian metode dan analisis terhadap model klasifikas menggunakan algoritma *Random Forest* yang telah dibangun. Pengujian ini bertujuan untuk mengevaluasi performa model klasifikasi teks menggunakan algoritma *Random Forest* dengan metode pembobotan atau ekstraksi fitur *Bag of Words* dan *TF-IDF*.

Dalam penelitian ini, akan dibandingkan performa model klasifikasi pada dua skenario pengujian yang berbeda. Pertama, data akan dibagi menjadi dua bagian dengan rasio splitting 90:10, di mana 90% data digunakan untuk pelatihan model dan 10% data digunakan untuk pengujian. Kedua, akan digunakan metode *k-fold cross* validation dengan k=10, di mana data akan dibagi menjadi 10 bagian yang seimbang, dan setiap bagian akan digunakan sebagai data pengujian secara bergantian, sementara bagian lainnya digunakan sebagai data pelatihan.

Berikut adalah skenario pengujian yang akan dilakukan dalam penelitian ini, yang mencakup kedua rasio pembagian data dan metode cross validation:

1. *Splitting data training* dan *data testing* 90:10

Tabel 3. 2 Skenario Pengujian Model dengan *Splitting Data* 90:10

	Bag of Words	*TF-IDF*
Akurasi		
Presisi		
Recall		
F1-Score		

2. *10-fold cross validation*

Tabel 3. 3 Skenario Pengujian Model dengan *10-Fold Cross Validation*

	Akurasi		Presisi		*Recall*		*F1-Score*	
Iterasi	*BoW*	*TF-IDF*	*BoW*	*TF-IDF*	*BoW*	*TF-IDF*	*BoW*	*TF-IDF*
1								
2								
3								
4								
5								
6								
7								
8								
9								