BAB 3
METODE PENELITIAN

Dalam penelitian ini menggunakan metode *Algoritma K-Means* untuk mengelompokkan sekumpulan data ke dalam beberapa kelompok (*cluster*) berdasarkan kemiripan data. Yang dimana penelitian ini membutuhkan data *tweet* yang didapatkan dari Twitter yang berkaitan dengan Yogyakarta, selanjutnya dilakukan pengolahan data berupa *preprocessing* untuk mendapatkan hasil yang diinginkan. Setelah dilakukan *preprocessing* selanjutnya *TF-IDF* untuk memberikan bobot yang lebih tinggi untuk kata-kata yang sering muncul dalam sebuah data.

3.1 BAHAN DAN ALAT PENELITIAN

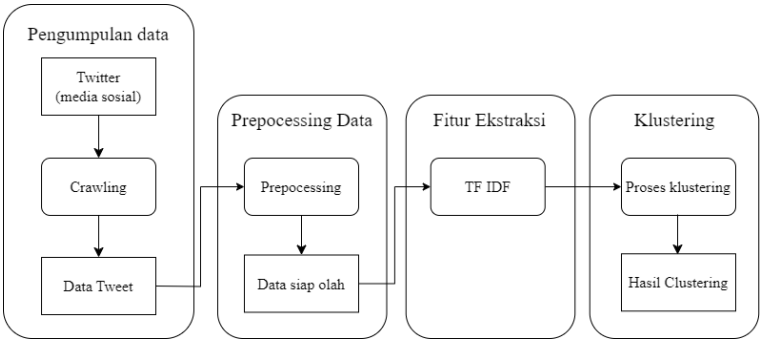
Bahan yang akan digunakan dalam penelitian ini mencakup data *tweet* yang terkait dengan Yogyakarta di media sosial Twitter.

Dalam penelitian ini, diperlukan sistem operasi dan perangkat lunak tertentu untuk menjalankan sistem informasi dan aplikasi yang dikembangkan, serta dukungan jaringan internet. Berikut adalah beberapa sistem operasi dan program aplikasi yang digunakan dalam penelitian ini yaitu :

1. Sistem Operasi: *Windows* 11
2. *Anaconda* versi 3
3. *Jupyter Notebook*
4. Microsoft Office Excel 2016

3.2 JALAN PENELITIAN

Jalan penelitian ini dilakukan berdasarkan tahapan-tahapan yang terdapat dalam metode penelitian yaitu *Algoritma K-means.* Dengan menggunakan *software* Anaconda versi 3 dan Jupyter Notebook untuk mengolah data *tweet* terkait dengan Yogyakarta dari Twitter. Adapun tahap pada jalannya penelitian ini ditunjukkan pada gambar 3 1 (Kharisma & Aesyi, 2023) :



Gambar 3. 1 Flowchart Jalan Penelitian

3.2.1 Pengumpulan Data

Pengambilan data dari media sosial Twitter berkaitan dengan topik Yogyakarta ini digunakan teknik *crawling* (Kharisma & Aesyi, 2023). Crawling adalah metode yang digunakan untuk mengumpulkan data informasi yang ada didalam web secara otomatis, pada proses ini informasi yang dikumpulkan didasarkan pada kata kunci yang telah ditentukan oleh pengguna (Saputra, 2018). Pada tahap crawling data dari *Twitter* untuk mendapatkan *tweet* dengan kata kunci "Yogyakarta" ini menggunakan *library* snscrape. Snscrape adalah sebuah *library* yang digunakan untuk mengambil data dari platform media sosial Twitter, proses ini dibuat menggunakan bahasa pemrograman python pada notebook Google Colaboratory. Data yang didapat hasil crawling data sebanyak 3006, data yang didapatkan juga data-data terbaru dimulai dari tanggal 11 Juni sampai 16 Juni 2023.

3.2.2 Preprocessing

Preprocessing adalah proses pengolahan data, dengan data yang sudah didapatkan atau data mentah yang masih perlu dilakukan eliminasi data yang tidak sesuai atau tidak relevan. Tujuannya untuk membersikan, mengubah format, dan mempersiapkan data agar sesuai dengan kebutuhan analisis atau pemrosesan selanjutnya. Adapun beberepa *library* yang dibutuhkan pada proses *preprocessing* yaitu seperti import pandas, numpy, nltk, stopword, string, re, dan sastrawi.

Setelah melakukan import library selanjutnya proses *preprocessing*, berikut merupakan tahapan dalam proses *preprocessing* data yaitu *Casefolding* untuk merubah huruf kapital menjadi huruf kecil. Cleaning data untuk membersihkan kata yang tidak penting seperti simbol, angka, *re-tweet*, dll. *Tokenizing* yaitu proses

memisahkan atau memecahkan yang awalnya berupa kalimat menjadi kata-kata berdasarkan tiap kata. *Stopwords* digunakan untuk mengurangi jumlah kata dengan cara melakukan penghapusan dan penyaringan kata contohnya yang, dan, di, dengan, dll. *Stemming* yaitu kata yang berimbuhan diubah menjadi kata asli tanpa imbuhan, sisipan, awalan, akhiran, kombinasi. *Normalization* yaitu mengubah kata yang sebelumnya merupakan singkatan dan kata tidak baku diubah menjadi kata baku.

3.2.3 Fitur Ekstraksi

Fitur ektraksi disini yaitu menggunakan *TF-IDF* (*Term frequency-Inverse Documen Frequency*), dalam *TF-IDF* ini terdapat 2 rumus yang digabungkan untuk menghitung bobot kata-kata. *TF* untuk megukur frekuensi kemunculan kata disebuah dokumen, sedangkan *IDF* untuk menghitung tingkat signifikasi suatu kata dalam dokumen tersebut (Handayani et al., 2020).

3.2.4 Klustering

Data yang selesai proses fitur ekstraksi selanjutnya mengolah data menggunakan *Algoritma K-means* untuk melakukan klustering data dalam tahap berikutnya. Berikut adalah tahap untuk menentukan jumlah *cluster* setelah melakukan tahap *preprocessing* yaitu dengan penentuan k sebagai jumlah *cluster* dengan penerapan metode *elbow method*, selanjutnya dilakukan perhitungan centroid awal untuk setiap *cluster* yang dimana jarak antara setiap objek data dengan *centroid* dihitung menggunakan metode *Euclidian Distance*. Dengan menggunakan Euclidean Distance data dapat dialokasikan ke centroid terdekat (Kharisma & Aesyi, 2023). Setelah proses pengolahan data dan klustering selesai, selanjutnya adalah analisis hasil, yang dimana tahapan ini digunakan untuk menguji kualitas kluster yang dihasilkan dari proses klustering untuk menentukan hasil perhitungan yang didapatkan.