BAB III
METODE PENELITIAN

3.1 Objek Penelitian

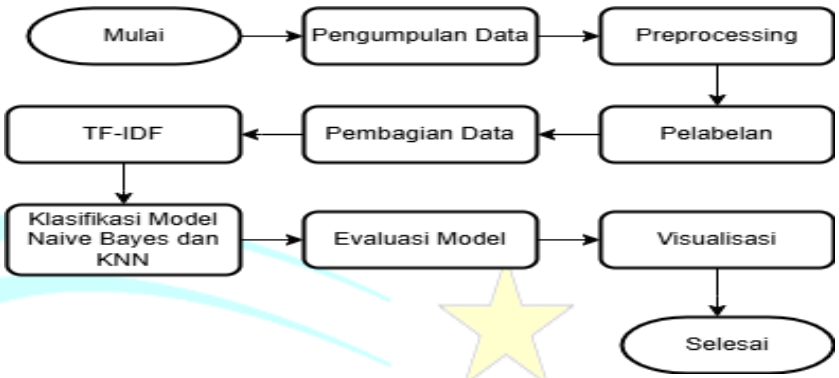
Pelaksanaan penelitian ini terdiri dari beberapa tahapan yang disusun secara sistematis untuk memperoleh hasil yang valid dan relevan. Tahapan pertama dimulai dengan identifikasi masalah, yaitu merumuskan isu yang akan diteliti berdasarkan fenomena aktual di masyarakat, dalam hal ini terkait dengan pemindahan Ibu Kota Negara (IKN). Setelah masalah dirumuskan, dilakukan studi literatur guna memperkuat landasan teori dengan menelaah referensi dari jurnal, dan penelitian terdahulu yang relevan dengan topik. Tahapan berikutnya adalah pengumpulan data, yang dilakukan melalui teknik *web crawling* dengan menggunakan pustaka *Tweet Harvest*. Data dikumpulkan dari media sosial *Twitter* dengan menggunakan kata kunci "IKN", yang mencakup berbagai atribut seperti isi tweet, nama pengguna, tanggal dan waktu unggahan, jumlah retweet dan like, serta hashtag yang menyertainya. Pengambilan data difokuskan pada tweet publik yang mengandung opini, sentimen, atau respons masyarakat terkait isu IKN.

Data yang telah dikumpulkan kemudian melalui proses pembersihan (*cleaning*) untuk menghilangkan unsur-unsur yang tidak relevan seperti simbol, tautan, spasi berlebih, atau duplikasi. Selanjutnya, data dianalisis menggunakan metode yang sesuai, seperti analisis sentimen atau klasifikasi teks, untuk mengetahui persepsi publik terhadap IKN. Tahap terakhir adalah evaluasi yang bertujuan untuk menilai kesesuaian hasil analisis dengan tujuan penelitian serta mengkaji keakuratan model yang digunakan. Setiap tahapan dalam proses ini disusun agar berjalan secara efisien, di mana beberapa kegiatan dilakukan secara bersamaan (paralel) guna menghemat waktu dan sumber daya.

3.2 Prosedur Penelitian

Terdapat beberapa langkah yang perlu diambil dalam penelitian ini. Pertama, mengumpulkan data dari X dengan memanfaatkan *python*. Kedua, melakukan tahap persiapan data, yaitu membersihkan data tersebut. Ketiga,

memberikan label pada data secara manual oleh ahli bahasa, dengan klasifikasi positif, negatif, dan netral. Terakhir, melakukan pengelompokan data dengan metode K-NN dan *Naïve Bayes*. Berikut adalah alur penelitian.



Gambar 3. 1 Alur Penelitian

3.2.1 Pengumpulan Data

Data yang digunakan dalam penelitian ini diperoleh dari media sosial Twitter (X) melalui proses *crawling* menggunakan pustaka *tweet-harvest* dengan fitur *auto token*. Pengumpulan data dilakukan dalam dua tahap, yaitu pada tanggal 15 Maret 2025 hingga 18 Maret 2025 yang menghasilkan sebanyak 810 komentar, dan pada tanggal 5 Mei 2025 hingga 7 Mei 2025 yang menghasilkan 467 komentar. Secara keseluruhan, jumlah data yang berhasil dihimpun adalah 1.277 komentar. Data tersebut disimpan dalam format CSV dan memuat informasi seperti ID tweet, *username*, waktu unggahan, isi komentar, jumlah *like*, jumlah *retweet*, serta *hashtag* yang digunakan.

Pemilihan akun pengguna Twitter dilakukan secara acak tanpa adanya kriteria khusus, sehingga data yang diperoleh mencerminkan keragaman opini dari berbagai pengguna. Pendekatan ini dimaksudkan untuk menghindari kencenderungan terhadap akun tertentu dan memastikan bahwa analisis sentimen dilakukan terhadap sampel data yang lebih objektif dan representatif.

3.2.2 *Preprocessing*

Dalam penelitian ini, tahap pra-pemrosesan dilakukan dengan beberapa langkah, yaitu: menghapus URL, angka, tanda baca, emoji, dan spasi berlebih; melakukan *case folding*, normalisasi kata tidak baku, *stopword removal*, *tokenisasi*,

serta *stemming* (Fikri Haikal et al., n.d.-b). Seluruh tahapan ini bertujuan untuk mempersiapkan data agar lebih bersih dan konsisten sebelum dilakukan analisis sentimen.

3.2.3 Pelabelan

Proses pelabelan data dilakukan dengan cara manual oleh ahli bahasa yaitu Adelya Daniyah, M.Pd yang memperhatikan konteks kalimat serta arti kata secara keseluruhan. Setiap teks diperiksa secara teliti untuk mengelompokkan sentimennya ke dalam tiga kelas, yakni positif, negatif, ataupun netral. Metode ini diterapkan untuk memastikan keakuratan label yang digunakan dalam proses pelatihan model. Berikut merupakan hasil labeling menggunakan bantuan pakar Bahasa Indonesia:

No.	TEXT	LABEL
1	lapor malaikat maut biar skalian cabut nyawa mulai jabat lengser kok bikin ramai aja mulai soal esemka bangun ikn tidak pakai uang apbn deret tipu	Negative
2	tetangga milih suami biar pindah ikn bang	Positive
3	ikn simbol harap baru perata ekonomi indonesia	Positive

Gambar 3. 2 Hasil Labeling dari Ahli Bahasa

3.2.4 Pembagian Data

Dalam penelitian ini, data terbagi atas dua bagian yakni latih maupun uji dengan perbandingannya 80% berbanding 20%. Setelah melalui tahap *preprocessing*, jumlah data yang tersedia adalah 1.112 tweet. Dari jumlah tersebut, sebanyak 889 data digunakan sebagai data latih, Sementara itu, 223 data digunakan sebagai data uji, Pembagian ini dirancang untuk melatih model secara efisien dan menilai performa algoritma *Naïve Bayes* serta KNN dalam melakukan analisis sentimen yang berhubungan dengan pemindahan Ibu Kota Negara (IKN).

3.2.5 Pembobotan TF-IDF

TF-IDF sebagai teknik yang digunakan untuk mengevaluasi signifikansi suatu kata dalam sebuah dokumen, dengan memperhatikan frekuensi kemunculannya di dalam dokumen tersebut dan frekuensi kemunculan kata tersebut di berbagai dokumen lainnya (Khoiruddin et al., n.d.).

3.2.6 Implementasi Algoritma Naïve Bayes dan KNN

Pelatihan model dilakukan menggunakan *Naïve Bayes* dan KNN untuk mengklasifikasikan sentimen terkait pemindahan IKN. *Naïve Bayes* Bekerja dengan menghitung probabilitas kata untuk setiap kelas sentimen (positif, negatif, netral) menggunakan *Teorema Bayes*, kemudian memprediksi sentimen dengan memilih kelas yang memiliki probabilitas tertinggi (Saputra et al., 2024b). Berikut adalah rumus *Naïve Bayes.*

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \quad (1)$$

Keterangan:

X	: Data dengan *class* yang belum diketahui
H	: Hipotesis bahwa data X merupakan suatu *class* spesifik
P(H|X)	: Probabilitas hipotesis H berdasarkan kondisi X (*posteriori probability*)

P(H) : Probabilitas hipotesis H (*prior probability*)

P(X|H) : Probabilitas X berdasarkan kondisi hipotesis H

P(X) : Probabilitas dari X

Sementara itu, KNN menentukan sentimen berdasarkan mayoritas K tetangga terdekat menggunakan *Euclidean Distance*, di mana model mencari kesamaan teks berdasarkan jarak antar data. Performa K-NN bergantung pada pemilihan nilai K yang optimal (Arifin et al., n.d.).

$$d = \sqrt{\sum_{i=1}^{n} (a_1 - b_1)\ 2} \quad (2)$$

Keterangan:

d : Jarak
a : Data uji/testing
b : Sample Data
I : Variable Data
n : Dimensi Data

Kedua model dievaluasi menggunakan 223 data uji dan dibandingkan berdasarkan metrik akurasi, presisi, *recall*, serta *F1-score* guna mengetahui metode yang paling optimal dalam melakukan analisis sentimen terhadap IKN.

3.2.7 Confusion Matrix

Confusion matrix adalah sebuah tabel yang digunakan untuk mengevaluasi kinerja algoritma klasifikasi dengan membandingkan hasil prediksi model terhadap label yang sebenarnya. Matriks ini sangat berguna dalam mengidentifikasi sejauh mana model mampu mengklasifikasikan data secara akurat, terutama dalam masalah klasifikasi biner. Dalam *confusion matrix* terdapat empat komponen utama, yaitu True Positive (TP), yang menunjukkan jumlah data yang benar-benar positif dan diprediksi positif; True Negative (TN), yaitu data yang benar-benar

negatif dan diprediksi negatif; False Positive (FP), yaitu data yang sebenarnya negatif namun diprediksi positif; serta False Negative (FN), yaitu data yang sebenarnya positif tetapi diprediksi negatif. Melalui keempat nilai ini, dapat dihitung metrik evaluasi seperti akurasi, presisi, recall, dan F1-score.

Pada penelitian ini, proses *confusion matrix* bertujuan untuk evaluasi model *Naïve Bayes* dan KNN dalam analisis sentimen pemindahan ibu kota negara di platform X. Berikut rumus *confusion matrix.*

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

$$F1 - Score = 2 \times \frac{precision \times recall}{precision \times recall} \quad (6)$$

3.2.8 Visualisasi

Selanjutnya, dilakukan analisis visual menggunakan *Word Cloud*, metode visualisasi ini menggambarkan berbagai kata yang paling sering muncul dalam sebuah teks, di mana kata dengan frekuensi kemunculan yang lebih tinggi akan diperbesar ukurannya, sehingga mempermudah identifikasi tema atau kata kunci utama dalam data (Tania Puspa Rahayu Sanjaya et al., 2023). Berikut merupakan hasil visualisasi.

Gambar 3. 3 Hasil Visualisasi