BAB III
METODE PENELITIAN

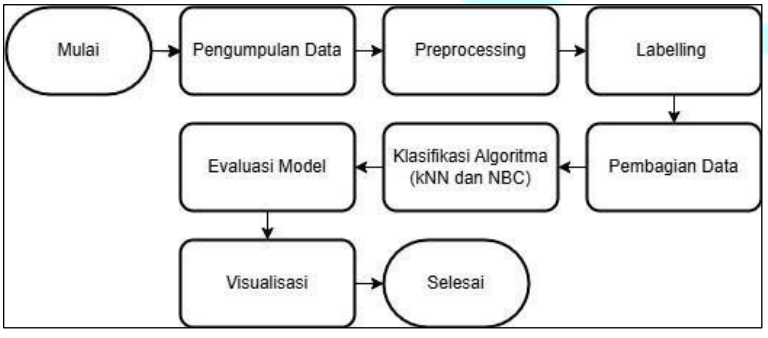
3.1 Objek Penelitian

Proses penelitian dimulai dengan pengumpulan data melalui *crawling*, kemudian dilakukan *preprocessing* untuk membersihkan data. Selanjutnya data diberi label dalam tahap *labelling* terbagi kedalam data pelatihan dan pengujian. Data pelatihan untuk melatih algoritma seperti KNN dan NBC, sedangkan pengujian digunakan untuk mengukur akurasi model. Selanjutnya, dilakukan evaluasi dan visualisasi data guna mempermudah pemahaman hasil analisis. Studi ini menerapkan metode *crawling* untuk mengumpulkan informasi dari *platform X* . Dalam proses *crawling*, peneliti menggunakan pustaka *tweet -hearvest*. *tweet -hearvest* adalah konsep Atau sarana yang digunakan untuk mengambil data dari *X*. Biasanya, alat ini dirancang untuk mengekstrak dan menyimpan *tweet* berdasarkan kata kunci, tagar, pengguna tertentu, atau kriteria lainnya.

Untuk memperoleh *tweet* yang berhubungan dengan PILKADA 2024, pencarian dilakukan melalui tagar Pilkada2024. Penelitian ini berhasil mengumpulkan total 1.203 data untuk dijadikan dataset, dataset ini disimpan dalam bentuk CSV. Rentang waktu data yang berhasil dikumpulkan adalah dari tanggal 08 Desember 2024 hingga 30 Desember 2024. Dataset ini mencakup komentar yang memiliki sentimen positif, negatif, dan netral.

3.2 Prosedur Penelitian

Seluruh prosedur dalam penelitian ini dijelaskan melalui penyajian *flowchart* dan *pseudocode* dibawah ini, untuk menggambarkan setiap tahapan secara sistematis.



Gambar 3. 1 Prosedur Penelitian

3.2.1 Pengumpulan data

Studi ini menerapkan metode *crawling* untuk mengumpulkan informasi dari *platform X*. Dalam proses *crawling*, peneliti menggunakan pustaka *tweet -hearvest*. *tweet-hearvest* adalah konsep atau sarana yang digunakan untuk mengambil data dari *X*. Biasanya, alat ini dirancang untuk mengekstrak dan menyimpan *tweet* berdasarkan kata kunci, tagar, pengguna tertentu, atau kriteria lainnya.

Untuk memperoleh *tweet* yang berhubungan dengan PILKADA 2024, pencarian dilakukan melalui tagar Pilkada2024. Penelitian ini berhasil mengumpulkan total 1.203 data untuk dijadikan dataset, dataset ini disimpan dalam bentuk CSV. Rentang waktu data yang berhasil dikumpulkan adalah dari tanggal 08 Desember 2024 hingga 30 Desember 2024.

3.2.2 *Preprocessing*

Preprocessing adalah tahap pertama sebelum teks dianalisis, yang berfungsi untuk meningkatkan keteraturan dan kebersihan data teks guna mendukung keakuratan dalam proses analisis lebih lanjut. Memainkan peran penting dalam mengatasi kemungkinan kesalahan saat mengekstrak fitur atau atribut dari teks *tweet* dapat memberikan dampak positif yang signifikan terhadap peningkatan kinerja dalam analisis sentimen.

Prapemrosesan teks adalah langkah yang krusial, karena proses ini berkontribusi langsung terhadap efektivitas algoritma yang diterapkan. Data yang lebih bersih akan mendukung performa algoritma secara maksimal (Wahyu, 2023). *Preprocessing* memiliki beberapa tahapan, diantaranya :

1. Data *Cleaning* : Langkah membersihkan data untuk menghilangkan beberapa komponen yang umum ditemukan dalam teks namun tidak mengandung arti substansial. (Agustian & Nurapriani, 2022). Proses penghapusan seperti url, penyebutan, emotikon, dan tagar(Khoiruddin et al., 2023).
2. *Case Folding:* tahap awal dalam pra-pemrosesan teks yang bertujuan untuk menyamakan format huruf dengan mengubah seluruh karakter menjadi huruf kecil. Proses ini menghilangkan perbedaan akibat penggunaan huruf besar dan kecil, sehingga menjaga konsistensi data untuk analisis lanjutan.

di *python*, langkah ini dapat dilakukan dengan fungsi .lower() yang tersedia dalam pustaka *pandas*, yang secara otomatis mengonversi semua huruf dalam teks menjadi huruf kecil. (Nevrada & Syaputra, 2025).

3. *Normalization* : Mengubah, menghapus, atau memodifikasi kata atau istilah yang disingkat (Nasution et al., 2024).
4. *Tokenizing* : tahapan pembagian teks dipisahkan ke dalam komponen-komponen yang lebih kecil. Memanfaatkan koma atau spasi untuk memisahkan elemen dalam teks(Khoiruddin et al., 2023)
5. *Stopwords Removal: Stopwords* merupakan kosakata yang umum ditemukan dalam teks, tetapi perannya dalam mendukung analisis sangat terbatas. seperti kata "dan", "atau", dan "adalah". Mengeliminasi *stopwords* dapat meningkatkan efisiensi model dengan menyederhanakan jumlah kata yang dianalisis selama proses pemodelan. (Gunawan & Aziza, 2025).
6. *Stemming* : Tahapan *stemming* langkah yang digunakan untuk mengubah kata terinfleksi ke dasarnya (Setyanto & Sasongko, 2024). Prosedur *Stemming* ini memanfaatkan *library sastrawi* (Muttaqin et al., 2024).

3.2.3 Pelabelan Data

Dalam penelitian ini, Proses pemberian label pada data dilakukan secara manual dengan melibatkan ahli bahasa yakni Adelya Daniyah, M.Pd. hal ini guna memastikan tingkat akurasi yang tinggi, konsistensi, dan validitas anotasi sentimen. Setiap data teks hasil *Preprocessing* dianalisis secara mendalam dengan mempertimbangkan struktur linguistik, makna kontekstual, serta nuansa emosional yang terkandung di dalamnya.

Ahli bahasa menerapkan pedoman klasifikasi yang telah distandardisasi, mencakup definisi operasional untuk kategori sentimen positif, negatif, dan netral, guna mengurangi subjektivitas individu selama proses anotasi. Selain memberikan label polaritas, pelabelan ini juga bertujuan mengidentifikasi ambiguitas makna dan konteks yang dapat mempengaruhi hasil klasifikasi algoritma di tahap selanjutnya, sehingga menghasilkan dataset yang berkualitas tinggi untuk kebutuhan pelatihan dan pengujian model.

Proses klasifikasi komentar dilakukan dengan menganalisis setiap teks berdasarkan aspek semantik, sintaksis, dan konteks wacana, guna mengidentifikasi kecenderungan sentimen yang tersembunyi. Komentar-komentar yang telah melewati proses *Preprocessing* selanjutnya diklasifikasikan secara terstruktur dibagi ke dalam tiga label utama: positif, negatif, serta netral. dengan berpedoman terhadap kriteria linguistik yang telah dirumuskan sebelumnya. Penekanan diberikan pada analisis makna implisit serta ekspresi emosional, sehingga setiap kategori sentimen dapat diidentifikasi secara komprehensif dan akurat. Pendekatan berbasis ahli bahasa ini diharapkan mampu menghasilkan dataset berlabel yang valid, meningkatkan kinerja model, serta memperkuat keandalan analisis sentimen berbasis machine learning dalam penelitian ini.

3.2.4 Pembagian Data

Dalam proses pembagian dataset memisahkan data berlabel jadi dua bagian, mencakup data pelatihan dan data pengujian. Pemisahan ini dilakukan melalui pendekatan berbasis persentase, di mana sekitar 80% dari seluruh dataset diperuntukkan bagi data pelatihan, sedangkan sekitar 20% lainnya digunakan untuk pengujian. Dengan cara ini, mayoritas data dapat dimanfaatkan untuk mengoptimalkan proses pembelajaran model, adapun sebagian kecil data dialokasikan sebagai data uji guna mengukur performa dan akurasi model yang telah dilatih.

3.2.5 Implementasi Algoritma *K-Nearest Neighbors*

Model *K-Nearest Neighbors (KNN)* ialah teknik dasar yang sering dimanfaatkan sebagai langkah pertama dalam mempelajari *machine learning*. Kelebihan dari *KNN* terletak pada kesederhanaannya serta kemampuannya untuk menangani data yang tidak linier atau yang belum terklasifikasi dengan jelas. Namun, kelemahan dari metode ini adalah penurunan kinerja pada dataset besar, karena algoritma perlu menghitung jarak antara setiap pasangan data pelatihan. (Fadil, 2024). Untuk menentukan label kategori atau nilai dari suatu sampel, *K-Nearest Neighbors* (KNN) mencari K Memanfaatkan data dengan jarak terdekat untuk menentukan label atau nilai yang sesuai. Pendekatan ini mengelompokkan objek berdasarkan data yang paling dekat. KNN juga menghitung jarak kedekatan

data untuk memperkirakan nilai bagi *query* baru. Metode *K-Nearest Neighbors* dapat diterapkan untuk menentukan kedekatan antar data dengan langkah-langkah berikut:

$$d = \sqrt{\sum_{i=1}^{n}(a_1 - b_1)^2} \quad (3)$$

Penjelasan:

d = nilai jarak
a = data uji
b = data sampel
i = indeks variable
n = banyaknya variabel

3.2.6 Implementasi Algoritma *Naïve Bayes Classifier*

Naive Bayes Classifier mengaplikasikan *Teorema Bayes* serta mengandalkan prinsip probabilitas dasar, dengan anggapan bahwa tiap fitur tidak saling bergantung. Algoritma ini diterapkan dalam berbagai masalah, seperti analisis sentimen, penyaringan *email spam*, pengelompokan *email* otomatis, pengurutan *email* berdasarkan prioritas, serta pengkategorian dokumen.

Metode *Naive Bayes* yang paling sesuai untuk pengklasifikasian teks atau dokumen adalah *multinomial Naive Bayes*, yang mempertimbangkan tidak hanya kemunculan kata, tetapi juga frekuensinya. Teknik ini diperkenalkan oleh ilmuwan asal *Inggris*, *Thomas Bayes*. Dibawah menunjukkan rumus klasifikasi data menggunakan *Naive Bayes* :

$$P\ (C|X) = \frac{P(X|C)\cdot P(C)}{P(X)} \quad (4$$

Keterangan:

X : data yang tidak terlabeli.

C : asumsi mengenai kategori kelas yang mungkin untuk data X .

P(C|X) : *posterior probability,* yaitu peluang hipotesis C benar setelah mempertimbangkan data X .

P(C) : *prior probability* dari hipotesis C, yakni peluang awal tanpa memperhitungkan data X .

P(X |C) : *likelihood*, yang merujuk pada kemunculan data X dengan asumsi bahwa hipotesis C adalah benar.

P(X) : *evidence*, yang mengacu pada total data X tanpa memperhitungkan kelas yang terkait.

3.2.7 Evaluasi Model

Model evaluasi adalah langkah untuk menilai performa dari sebuah model *machine learning* setelah proses pelatihan dengan data yang tersedia. Pentingnya evaluasi model terletak pada kemampuannya dalam menunjukkan seberapa baik model dapat melakukan prediksi, sekaligus membantu dalam memilih model yang paling sesuai untuk suatu permasalahan (Subagyo et al., 2024).

Beberapa metrik yang sering digunakan dalam *klasifikasi* meliputi *akurasi*, metode evaluasi ini mencakup *accuracy*, Mengukur sejauh mana model menghasilkan prediksi yang benar dengan frekuensi tertentu. *precision* , yang menentukan tingkat ketepatan model dalam mengklasifikasikan data ke pada Kelas positif; *recall*, yang menilai efektivitas model dalam mengidentifikasi seluruh kasus positif; serta *f1-score*, yang mengelompokkan *precision* serta *recall* dalam menilai secara adil kemampuan model.

Rumus yang diterapkan :

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

$$Precision = \frac{TP}{TP+FP} \quad (6)$$

$$Recall = \frac{TP}{TP+FN} \quad (7)$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision \times Recall} \quad (8)$$

Keterangan:

TP (*True Positive*): jumlah kasus positif yang berhasil diklasifikasikan secara akurat oleh model sebagai positif.

TN (*True Negative*): jumlah kasus negatif yang benar dikategorikan oleh model sebagai negatif.

FP (*False Positive*): jumlah kasus negatif yang keliru diklasifikasikan sebagai positif oleh model.

FN (*False Negative*): jumlah kasus positif yang salah diprediksi sebagai negatif oleh model.

3.2.8 Visualisasi

Setelah kedua algoritma diterapkan, dilakukan visualisasi data dengan memanfaatkan *wordcloud* dan grafik pada setiap kategori sentimen (positif, netral, dan negatif) yang diperoleh dari data latih yang telah dilatih (Khoiruddin et al., 2023).

World cloud adalah salah satu metode untuk memvisualisasikan yang menunjukkan keterkaitan antara frekuensi kata, dengan meningkatkan ukuran kata yang lebih sering digunakan secara cepat (Sanjaya et al., 2023).