BAB III
METODE PENELITIAN

3.1 Objek Penelitian

Objek penelitian ini berupa tanggapan pengguna media sosial X terhadap kebijakan pemerintah mengenai kenaikan Pajak Pertambahan Nilai (PPN) menjadi 12%. Penelitian ini akan mengidentifikasi sentimen pengguna, baik dalam bentuk tanggapan negatif, netral, maupun positif, yang diunggah dalam bentuk teks di platform media sosial tersebut. Fokus penelitian diarahkan pada analisis sentimen untuk memahami persepsi masyarakat terhadap kebijakan tersebut, sesuai dengan rumusan masalah dan tujuan penelitian yang telah ditentukan.

Lokasi penelitian dilakukan secara daring dengan memanfaatkan platform media sosial X sebagai sumber data utama. Media sosial X dipilih karena penggunaannya yang luas di Indonesia.

Waktu penelitian dilakukan mulai dari Oktober 2024 sampai dengan April 2025.

Tabel 3.1 Waktu Penelitian

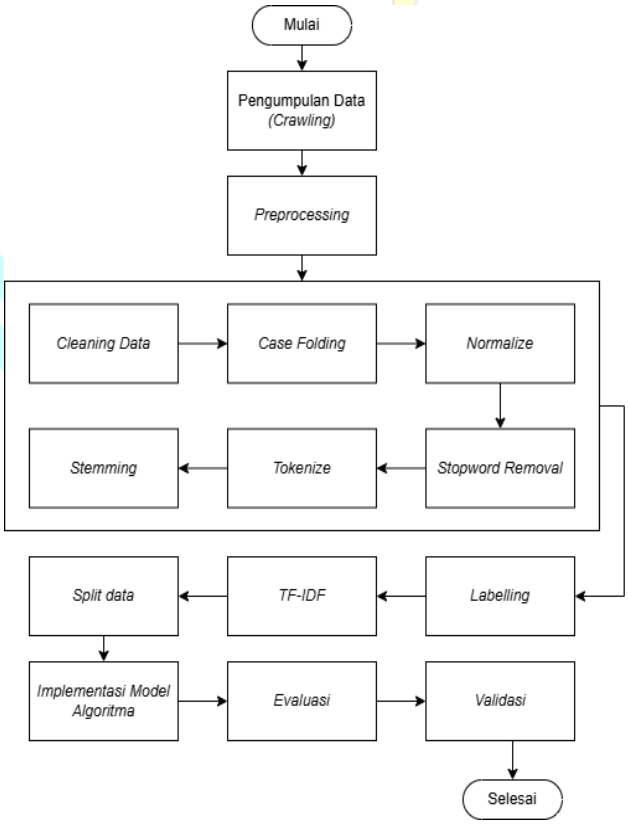
No	Jenis Kegiatan	2024			2025			
		10	11	12	1	2	3	4
1	Pembuatan Proposal							
2	Pengumpulan data							
3	Menganalisis data							
4	Pembuatan model ML							
5	Penyusunan laporan akhir							

Lingkup batasan penelitian ini hanya mencakup unggahan teks dalam bahasa Indonesia yang relevan dengan topik penelitian dan diunggah dalam kurun waktu tertentu, yaitu dimulai dari bulan November hingga bulan Desember. Unggahan dalam bentuk non-teks, seperti gambar atau video, tidak akan dianalisis.

Tahapan penelitian yang akan dilakukan pada studi ini. Penelitian dimulai dengan tahap pengumpulan data menggunakan teknik *crawling* untuk mengakses data dari media sosial atau sumber lainnya. Data yang diperoleh kemudian melalui proses *preprocessing* untuk meningkatkan kualitas data sebelum dianalisis. Proses

preprocessing mencakup beberapa langkah, yaitu *cleaning* data untuk menghapus data yang tidak relevan, *case folding* untuk mengonversi teks menjadi huruf kecil, *normalizing* untuk menyelaraskan format teks, *stopword removal* untuk menghilangkan kata-kata umum yang tidak memiliki pengaruh signifikan, *tokenizing* untuk memecah teks menjadi kata-kata, serta *stemming* untuk mengubah kata menjadi bentuk dasarnya.

Setelah *preprocessing* selesai, data akan dibagi menjadi dua bagian, yaitu data latih dan data uji, untuk keperluan pelatihan dan pengujian model. Selanjutnya, data latih digunakan dalam proses implementasi model algoritma *machine learning* untuk membangun model yang dapat menganalisis sentimen. Tahapan terakhir adalah pengujian model untuk mengevaluasi kinerja dan akurasi model yang telah dikembangkan.



Gambar 3.1 Alur Penelitian

3.2 Pengumpulan Data

Dataset yang digunakan dalam penelitian ini awalnya terdiri dari 1.871 data, yang diperoleh menggunakan Google Colab dengan bantuan library Python bernama tweet-harvest. Data tersebut dikumpulkan dari media sosial X selama

periode 29 Oktober 2024 hingga 17 Desember 2024, dengan kata kunci “PPN 12%”.

conversation_id_str	created_at	favorite_count	full_text
1818297575273120197	Wed Jul 31 10:33:10 +0000 2024	0	@Carbaroxabane @RodriChen Kan katanya ada prog...
1818594224226308162	Wed Jul 31 10:27:05 +0000 2024	0	Makan Bergizi Gratis Prabowo Belum Diputuskan ...
1818592489391243607	Wed Jul 31 10:20:11 +0000 2024	0	Bukan Program Pemerintah Pusat Begini Antusias...
1818589849538199685	Wed Jul 31 10:09:42 +0000 2024	0	lagi budgeting trs nyadar ternyata tdk memenuh...

Gambar 3.2 Contoh data

3.3 *Preprocessing Text*

Preprocessing merupakan Langkah yang dilakukan untuk membersihkan data dan memperbaiki data yang rusak. Biasanya data yang di dapatkan dari hasil *crawling* tidak terstruktur dengan baik dan mengandung banyak karakter. Tujuan dilakukannya adalah untuk memperbaikinya. Sehingga didapatkan data yang sudah terstruktur untuk dilakukan pada saat melakukan pemodelan. Ada beberapa tahap dalam melakukan *preprocessing*, seperti *cleaning, case folding, normalize, stopword removal, tokenize, stemming.*

	full_text	created_at	username
0	@txtdrimedia Ppn 12% KOCAK	Sun Dec 15 14:48:26 +0000 2024	puphy
1	@txtdrimedia pajak motor nambah 2 ppn naek 12%...	Sun Dec 15 14:45:01 +0000 2024	lippetew
2	OJK: PPN 12% Berdampak Temporer https://t.co/K...	Sun Dec 15 14:42:20 +0000 2024	InvestorID
3	@aHeartShaker oh iyaya ppn naik jadi 12% . .	Sun Dec 15 14:41:34 +0000 2024	posnie
4	Colek Prabowo soal Wacana Kenaikan PPN 12 Pers...	Sun Dec 15 14:40:26 +0000 2024	OposisiCerdas

Gambar 3.3 contoh dataset sebelum dilakukan *preprocessing*

3.3.1 *Cleaning*

Cleaning merupakan proses pembersihan data teks agar lebih siap untuk dianalisis. Cleaning bertujuan untuk menghilangkan elemen-elemen yang tidak

relevan atau dapat mengganggu proses analisis, seperti simbol, angka, tanda baca, URL, dan karakter khusus. Proses ini penting karena data yang diambil dari media sosial seringkali mengandung banyak "noise" yang tidak berkontribusi pada pemahaman makna teks.

3.3.2 *Case Folding*

Case folding dalam analisis sentimen adalah proses standarisasi teks dengan mengubah semua huruf menjadi huruf kecil. Tujuan utamanya adalah untuk menghilangkan perbedaan yang disebabkan oleh penggunaan huruf kapital, sehingga analisis menjadi lebih konsisten.

3.3.3 *Normalize*

Proses ini bertujuan mengubah variasi kata atau frasa yang berbeda menjadi bentuk standar sehingga memudahkan pengolahan lebih lanjut, terutama dalam analisis teks yang membutuhkan konsistensi data. Variasi dalam teks dapat muncul dari penggunaan singkatan, ejaan tidak baku, penulisan slang, atau kesalahan ketik yang sering ditemukan pada data teks mentah, terutama di media sosial atau sumber teks tidak terstruktur lainnya.

3.3.4 *Stopword Removal*

Stopword removal dilakukan untuk menghilangkan kata-kata umum (*stopwords*) yang tidak memiliki makna signifikan dalam analisis teks, seperti "dan", "yang", atau "dalam". Kata-kata ini sering kali muncul dalam teks dengan frekuensi tinggi namun tidak memberikan informasi yang relevan untuk proses pengolahan atau analisis lebih lanjut. Proses ini bertujuan untuk meningkatkan efisiensi pengolahan teks dengan memastikan bahwa hanya kata-kata yang bermakna atau relevan yang dipertahankan.

3.3.5 *Tokenize*

Tokenization digunakan untuk memecah teks menjadi unit-unit kecil yang disebut token. Token biasanya berupa kata, frasa, atau bahkan karakter, *tokenize* digunakan untuk mengubah data teks menjadi format yang lebih mudah dikelola oleh algoritma *machine learning.*

3.3.6 *Stemming*

Stemming untuk mengubah kata-kata dalam teks ke bentuk dasarnya (*root word* atau *stem*) dengan menghapus imbuhan seperti awalan, akhiran, atau sisipan. Dalam analisis sentimen, *stemming* bertujuan untuk menyederhanakan representasi kata agar berbagai bentuk kata yang memiliki makna serupa dapat dianggap sama oleh model.

3.4 *Labelling*

Labelling merupakan tahap penting dalam analisis sentimen, di mana tanggapan masyarakat di media sosial X dikategorikan menjadi tiga kategori utama, yaitu positif, netral, dan negatif. Proses ini bertujuan untuk memberi label pada setiap entri atau komentar yang ada berdasarkan sentimen yang terkandung dalam teks tersebut. Kategori positif diberikan kepada tanggapan yang menunjukkan perasaan atau pandangan yang mendukung atau menyukai suatu topik atau kebijakan, sedangkan kategori negatif diberikan untuk tanggapan yang menunjukkan ketidaksetujuan, penolakan, atau ketidakpuasan. Kategori netral digunakan untuk tanggapan yang tidak menunjukkan sentimen yang jelas, baik mendukung maupun menentang, dan cenderung bersifat informasional atau netral.

3.5 TF-IDF

TF-IDF digunakan untuk mengekstraksi fitur dari teks yang telah diproses. Proses ini menghasilkan bobot untuk setiap kata, yang menunjukkan seberapa penting kata tersebut dalam dokumen tertentu dibandingkan dengan keseluruhan koleksi dokumen. Hasil ekstraksi ini berupa matriks yang digunakan sebagai masukan untuk analisis selanjutnya.

3.6 Pembagian Dataset

Dataset yang digunakan dalam penelitian ini berjumlah 1.806 data, yang kemudian dibagi menjadi dua bagian utama. 1.000 data pertama digunakan untuk pelatihan dan pengujian awal model menggunakan skenario 80:20. Sedangkan 806 data kedua digunakan sebagai validasi akhir untuk menguji performa model terhadap data baru yang belum pernah diproses sebelumnya

3.7 Klasifikasi

Pada penelitian ini, tiga algoritma machine learning digunakan untuk analisis sentimen terhadap isu kenaikan Pajak Pertambahan Nilai (PPN) sebesar 12% pada media sosial X, yaitu *Support Vector Machine* (SVM), *random forest*, dan *decision tree.*

3.7.1 *Support Vector Machine* (SVM)

SVM bekerja dengan cara mencari *hyperplane* yang memaksimalkan margin antara dua kelas data yang berbeda. Keunggulan SVM adalah kemampuannya untuk menangani data yang tidak terpisahkan secara linear dengan menggunakan kernel trick, sehingga memungkinkan pengklasifikasian yang lebih baik pada dataset yang kompleks, seperti teks dalam analisis sentiment.

Hyperplane : $f(x) = w.x + b$ (7)

w : Vektor bobot.

x : Vektor fitur (representasi teks).

b : bias (offset dari *hyperplane*).

$f(x)$: skor keputusan untuk menentukan kelas.

Penelitian ini digunakan kernel linear karena data analisis teks memiliki dimensi fitur yang tinggi, seperti pada representasi TF-IDF atau *Word Embedding*. Kernel linear dipilih karena efisien secara komputasi, sederhana, dan meminimalkan risiko overfitting, terutama pada data berdimensi tinggi di mana pemisahan linear sering cukup efektif. Selain itu, kernel ini sesuai untuk analisis sentimen yang berfokus pada akurasi tanpa menambah kompleksitas model.

3.7.2 *Random Forest*

Random Forest menggabungkan keputusan dari berbagai pohon untuk membuat keputusan yang lebih stabil dan akurat. Salah satu keuntungan utama menggunakan *Random Forest* yaitu analisis sentimen adalah kemampuannya untuk mengurangi *overfitting.*

$\gamma = mode\{h_1(x), h_2(x), \dots, h_m(x)\}$ (8)

γ : Prediksi akhir (kelas sentimen).

$h_1(x)$: Prediksi dari pohon ke-i.

m : Jumlah pohon dalam Random Forest.

mode : Fungsi untuk memilih nilai yang paling sering muncul (mayoritas voting).

Jumlah pohon dalam random forest, yang dikenal sebagai n_estimators, diatur menjadi 100, yang merupakan nilai default dalam pustaka Scikit-learn. Jumlah ini dipilih karena memberikan keseimbangan antara akurasi dan efisiensi komputasi. Dengan 100 pohon, algoritma mampu menghasilkan prediksi yang stabil tanpa memerlukan waktu komputasi yang terlalu lama.

3.7.3 *Decision Tree*

Decision Tree digunakan untuk membangun model klasifikasi dengan cara membagi data berdasarkan fitur yang paling informatif. Setiap cabang pada pohon keputusan mewakili suatu keputusan atau pembagian berdasarkan suatu fitur, sementara setiap daun menunjukkan kelas hasil klasifikasi. *Decision tree* mudah dipahami dan diinterpretasikan, serta efektif dalam menangani data dengan struktur yang jelas.

$$\text{Gini impurity} = 1 - \sum_{i=1}^{c} p_i^2 \qquad (9)$$

Gini impurity digunakan pada studi ini karena mampu menghitung ketidakmurnian suatu simpul dengan efisien. Kriteria ini mengukur probabilitas suatu data salah klasifikasi jika dipilih secara acak, sehingga membantu dalam membagi data dengan akurasi tinggi.

3.8 Evaluasi Model

Pada tahap evaluasi *confusion matrix* digunakan untuk mengukur kinerja model klasifikasi algoritma SVM, *random forest*, dan *decision tree*. *Confusion Matrix* sendiri merupakan metode yang digunakan untuk mengukur performa model klasifikasi dengan cara membandingkan prediksi model terhadap data actual.

3.9 Pengujian

Pengujian akhir menggunakan 806 data yang tersisa, yang tidak digunakan dalam tahap pelatihan maupun pengujian awal. Validasi ini bertujuan untuk menguji model terhadap data baru yang belum pernah diproses sebelumnya guna mengukur tingkat generalisasi model dalam menangani data yang tidak terlihat selama pelatihan. Dengan demikian, metode ini memastikan bahwa model tidak hanya bekerja optimal pada data latih, tetapi juga mampu mengklasifikasikan sentimen secara akurat pada data yang benar-benar baru.