BAB III

METODOLOGI PENELITIAN

3.1 Gambaran Umum Objek Penelitian

Objek yang diteliti adalah data *tweet* dari media sosial *Twitter*, berisikan opini masyarakat terhadap pembangunan IKN Nusantara, termasuk pembangunan infrastruktur seperti Istana Kepresidenan, Jalan Tol, dan Perkantoran Pemerintah. Berdasarkan Buku Saku Pemindahan Ibu Kota Negara, pembangunan IKN terbagi dalam beberapa tahap pembangunan, yaitu dimulai dari periode 2020-2024 sebagai pemindahan tahap awal pusat pemerintahan dari Jakarta ke IKN. Tahap awal ini berfokus pada pembangunan infrastruktur dasar seperti jalan hingga fasilitas pemerintahan seperti istana kepresidenan hingga perkantoran pemerintahan [33], [3].

Media sosial *Twitter* dipilih karena merupakan *platform* yang populer dan digunakan oleh masyarakat Indonesia [8]. Berbeda dengan *platform* sosial media seperti *Instagram* yang lebih berfokus pada konten visual seperti foto dan *video*, *Twitter* mengutamakan pesan teks singkat dengan batasan 280 karakter per postingannya [35]. Batasan ini membuat informasi yang disampaikan lebih ringkas dan terfokus karena menghasilkan teks yang lebih pendek [68]. Selain itu, *Twitter* menyediakan akses *Application Programming Interface (API)* sehingga memudahkan pengambilan data untuk kepentingan penelitian terkait topik atau isu tertentu. Penggunaan *Twitter* sebagai sumber data telah banyak digunakan dalam berbagai penelitian terkait sentimen [9], [10], [11]. Oleh karena itu, penelitian ini menggunakan *Twitter* sebagai sumber data untuk meneliti sentimen publik terhadap pembangunan IKN.

3.2 Metode Penelitian

Metode penelitian merupakan bagian penting dalam penelitian ilmiah. Terdapat beberapa jenis metode penelitian, yaitu kualitatif dan kuantitatif [69]. Penelitian ini menggunakan metode kualitatif karena akan menganalisis data tekstual mengenai sentimen publik terhadap pembangunan IKN Nusantara. Selain itu, analisis jaringan

sosial atau *Social Network Analysis* (SNA) akan dilakukan untuk mengidentifikasi hubungan antar individu dalam jaringan serta untuk menemukan akun-akun utama dalam jaringan yang berpengaruh dalam penyebaran informasi terkait pembangunan IKN.

3.3 Variabel Penelitian

Penelitian ini membagi variabel menjadi dua jenis, yaitu variabel independen dan variabel dependen.

3.3.1 Variabel Independen

Variabel independen merupakan variabel yang mempengaruhi atau memberikan dampak terhadap variabel lain [70]. Pada penelitian ini, variabel independen merupakan *tweet* terkait dengan Pembangunan IKN, Istana Kepresidenan IKN, Jalan Tol IKN, dan Perkantoran Pemerintah IKN.

3.3.2 Variabel Dependen

Variabel dependen merupakan variabel yang dipengaruhi atau memperoleh dampak dari variabel lain [70]. variabel dependen dalam penelitin ini merupakan label sentimen yang berisi klasifikasi sentimen dari data *tweet* yaitu sentimen positif dan sentimen negatif.

3.4 Teknik Analisis Data

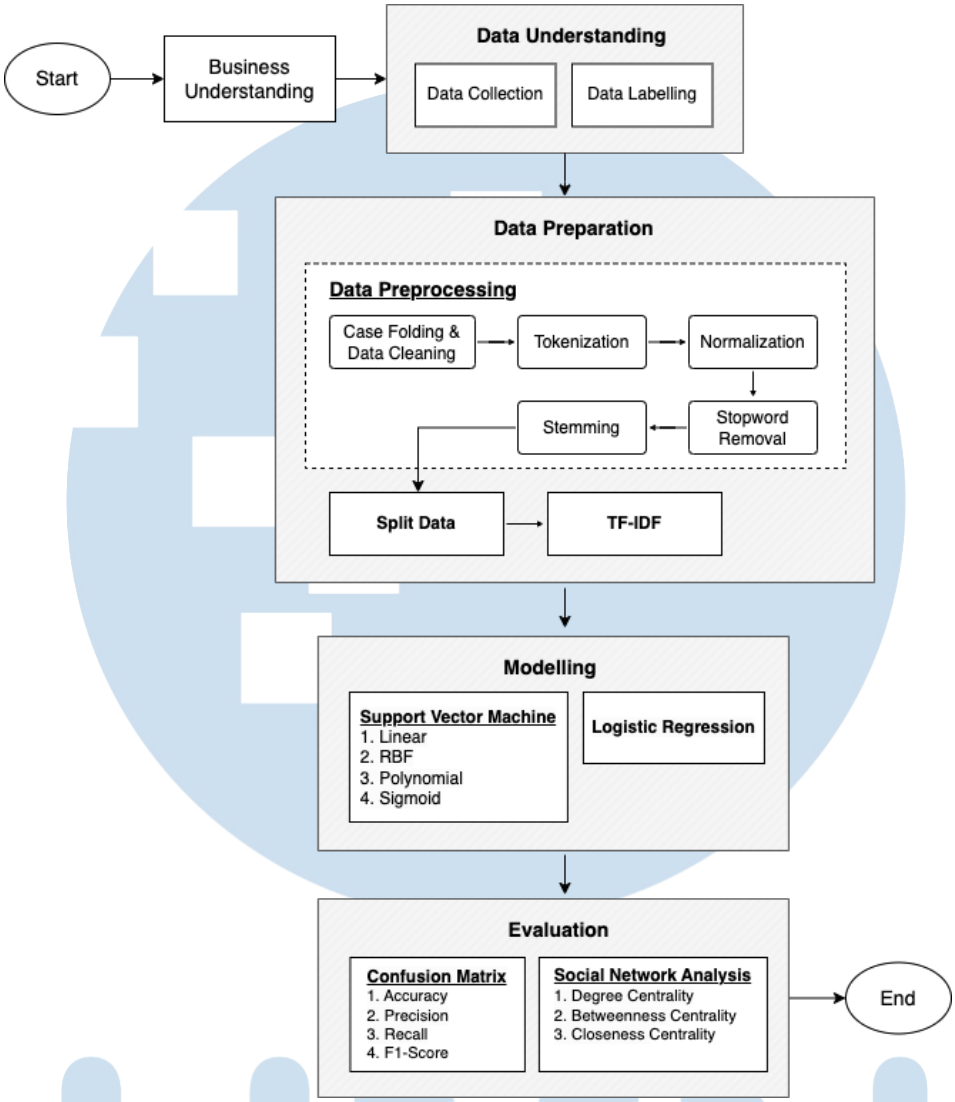
Penelitian ini mengadopsi model *Cross Industry Standard Process for Data Mining* (CRISP-DM). Tabel 3.1 berikut merupakan komparasi CRISP-DM dengan kerangka kerja lain, yaitu KDD *(Knowledge Discovery in Databases)* dan SEMMA *(Sample, Explore, Modify, Model, dan Assessment)*.

Tabel 3.1 Perbandingan *Framework Data Mining*

Framework	**KDD Process**	**SEMMA**	**CRISP-DM**
Tahapan	*Pre* KDD	-	*Business Understanding*
	Selection	*Sample*	*Data Understanding*
	Preprocessing	*Explore*	
	Transformation	*Modify*	*Data Preparation*
	Data Mining	*Model*	*Modelling*
	Interpretation/Evaluation	*Assessment*	*Evaluation*
	Post KDD	-	*Deployment*

Berdasarkan Tabel 3.1, CRISP-DM digunakan karena memiliki tahapan yang sistematis mulai dari pemahaman bisnis sebagai langkah awal sehingga membantu dalam mendapatkan pemahaman mendalam tentang sentimen publik terkait pembangunan IKN dan mengaplikasikan *insight* tersebut dalam pengambilan keputusan. Metode CRISP-DM fleksibel dan iteratif sehingga memungkinkan penyesuaian berkelanjutan sesuai dengan temuan baru dan kondisi data [71]. Berbeda dengan KDD, yang prosesnya yang lebih *linear* dan berurutan serta berfokus pada ekstraksi pengetahuan yang belum terungkap dari data yang diolah. KDD cocok untuk eksplorasi data yang tujuan utamanya adalah untuk menemukan pola dan *insight* baru tanpa batasan awal yang didefinisikan oleh kebutuhan bisnis. Sementara itu, SEMMA merupakan *framework* dengan tahapan yang paling sederhana, namun walaupun efektif dalam eksplorasi dan modifikasi data, SEMMA kurang menekankan pada tahapan pemahaman bisnis, sehingga lebih menekankan pada kualitas dan validitas model statistik.

Berikut Gambar 3.2 merupakan alur kerja penelitian ini yang dirancang berdasarkan kerangka kerja CRISP-DM.



Gambar 3.1 Alur Penelitian

3.4.1 *Business Understanding*

Penelitian ini bertujuan untuk mengetahui sentimen masyarakat terhadap pembangunan IKN yang kini memasuki fase awal yaitu pembangunan infrastruktur [3]. Selain itu, penelitian ini bertujuan untuk mengidentifikasi akun-akun yang berpengaruh dalam membentuk opini publik dan menyebarkan informasi terkait pembangunan IKN.

3.4.2 *Data Understanding*

3.4.2.1 *Data Collection*

Pengumpulan data dilakukan menggunakan *platform NodeXL,* yang merupakan *add-on* untuk *Microsoft Excel* untuk analisis jaringan sosial dan visualisasi data. *NodeXL* menyediakan akses untuk mengumpulkan data dengan memanfaatkan *API* yang tersedia dari *platform* media sosial seperti *Twitter, YouTube, Reddit, Flickr, Wikipedia* dan lainnya [65].

Pada penelitian ini, data yang dikumpulkan berasal dari *Twitter,* mencakup *tweet* yang mengandung kata kunci Pembangunan IKN, Istana Kepresidenan IKN, Jalan Tol IKN, dan Perkantoran Pemerintahan IKN. Proses pengumpulan data akan menggunakan fitur *Import from X (formerly Twitter) Search Network 3.0 (Beta)* yang tersedia di *NodeXL*. Fitur ini memungkinkan pengumpulan data postingan yang mengandung kata kunci yang telah di tentukan. Pengambilan data dilakukan selama periode tiga bulan, yaitu dari 1 November 2023 hingga 31 Januari 2024. Periode tersebut dipilih karena bertepatan dengan konteks politik Indonesia, yaitu masa kampanye pemilihan umum presiden dan wakil presiden Indonesia [72].

3.4.2.2 *Data Labelling*

Setelah mengumpulkan data, langkah berikutnya adalah pelabelan data. Proses ini melibatkan klasifikasi manual *tweet* berdasarkan sentimennya. Pada penelitian ini, data yang dikumpulkan adalah data *tweet* berbahasa Indonesia. Oleh karena itu, perlabelan manual akan dilakukan karena memungkinkan pemahaman lebih detail terhadap kalimat ambigu, sarkasme, serta istilah *slang* yang seringkali sulit dikenali oleh algoritma [73]. *Tweet* yang menyampaikan pandangan atau emosi positif, seperti dukungan atau pujian, akan diberi *label* "positif". Sebaliknya, *tweet* yang mengandung kritik, kekecewaan, atau bahasa yang kasar akan diberi *label* "negatif".

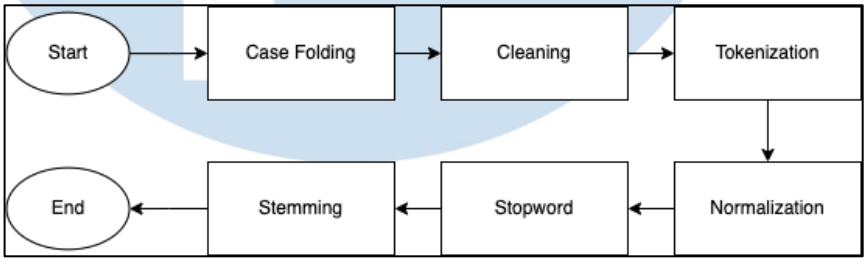
Untuk meningkatkan objektivitas dan konsistensi, proses perlabelan atau anotasi akan dilakukan oleh tiga orang anotator yang memiliki kemampuan berbahasa Indonesia yang baik dan benar serta

merupakan bagian dari masyarakat umum yang menerima informasi terkait pembangunan IKN. Proses penentuan *label* akan dilakukan berdasarkan metode mayoritas dari penilaian ketiga anotator. Sebagai contoh, jika dua dari tiga anotator memberikan *label* "negatif" pada sebuah *tweet,* sementara satu anotator lainnya memberikan *label* "positif", maka *label final* yang ditetapkan untuk *tweet* tersebut adalah "negatif" begitupun sebaliknya.

3.4.3 *Data Preparation*

3.4.3.1 *Data Preprocessing*

Tahap *preprocessing* dilakukan untuk mengubah dan memodifikasi data ke dalam format yang dibutuhkan dengan menggunakan *python.* Terdapat beberapa tahap untuk melakukan *preprocessing*, yaitu sebagai berikut.



Gambar 3.2 Diagram *Data Preprocessing*

Tahapan *preprocessing* telah diuraikan secara mendetail dalam bagian 2.2.4. Berdasarkan Gambar 3.2, proses ini terbagi menjadi enam tahapan.

1. *Case folding:* Pada penelitian ini *case folding* akan menggunakan fungsi *lower()* pada *python* untuk mengubah semua teks menjadi huruf kecil, bertujuan untuk menghilangkan perbedaan antara huruf kapital dan non-kapital [41].
2. *Cleaning*: Menggunakan ekspresi reguler *(regex)* dengan *re.sub()* untuk menghilangkan *URL, hashtag, mention,* dan karakter non-alfabet [41].

3. *Tokenization*: Teks dipecah menjadi kata-kata atau *token* [42], dengan menggunakan *nltk.tokenize.word_tokenize.*
4. *Normalization:* pengecekan dan perbaikan kesalahan penulisan, pengejaan serta mengubah singkatan atau *slang* [41], [42].
5. *Stopword Removal*: Mengeliminasi kata-kata yang tidak memiliki makna penting dan tidak memberikan pengaruh signifikan terhadap analisis teks. Pada penelitian ini, penghapusan stopword akan dilakukan menggunakan *nltk.corpus.stopwords.* Daftar *stopword* Bahasa Indonesia akan ditarik dari *corpus* tersebut untuk dieliminasi dari teks. Beberapa contoh *stopword* dalam Bahasa Indonesia yang terdapat dalam *nltk.corpus.stopwords* mencakup: "yang", "dan", "di", "dari", "untuk", "dengan", "dalam", "tidak", "ini", "itu", "adalah", "pada", "ke", "karena", "oleh", "juga", "telah", "akan", "bisa", "kami".
6. *Stemming*: Menggunakan fungsi Sastrawi, yaitu *Sastrawi.Stemmer.StemmerFactory* untuk Bahasa Indonesia, yang membantu mengubah kata berimbuhan menjadi kata dasar.

Setelah tahapan *stemming,* data dapat mengandung *list* kosong sehingga akan dihapus dan indeks akan diatur ulang untuk mempertahankan konsistensi data untuk analisis selanjutnya. Hasil *data preprocessing* kemudian diolah dengan membuat visualisasi data yang mengandung kata-kata positif dan negatif.

3.4.3.2 *Split Data*

Pada tahap ini, data dibagi dalam dua bagian, yaitu data pelatihan *(training)* dan data pengujian *(testing)*. Berdasarkan penelitian [74] mengenai analisis sentimen *review film,* penelitian melakukan eksperimen pembagian data dengan dua proporsi, yaitu 80:20 dan 90:10. Penelitian tersebut menunjukkan bahwa proporsi 80:20 memberikan akurasi yang lebih tinggi dibandingkan proporsi 90:10. Oleh karena itu, penelitian ini akan membagi data dengan

proporsi 80:20, yaitu 80% untuk data *training* dan 20% untuk data *testing*.

3.4.3.3 TF-IDF

Pada penelitian ini, implementasi TF-IDF akan dilakukan menggunakan fungsi *TfidfVectorizer()* dari *library scikit-learn* di *python*. TF-IDF memberikan bobot pada kata-kata dan mengonversi teks menjadi vektor numerik berdasarkan frekuensi kata dalam teks serta frekuensi kemunculan kata dalam seluruh dokumen [21]. Penelitian [44] menunjukkan bahwa penggunaan TF-IDF membantu dalam memproses teks sehingga menghasilkan akurasi yang lebih baik.

3.4.4 *Modeling*

Pada tahap pemodelan, setelah *preprocessing data* dan pembagian *dataset* menjadi 80% untuk *training* dan 20% untuk *testing*, serta implementasi *TF-IDF,* model dikembangkan menggunakan algoritma *Support Vector Machine* (SVM) dan *Logistic Regression* (LR). Algoritma SVM digunakan karenapa pada penelitian sebelumnya [14], [16], [18], [19], [20], [75] menunjukkan performa yang tinggi dibandingkan algoritma lain. Dalam penelitian ini, kinerja SVM akan dioptimalkan melalui pengujian nilai *parameter cost* (C) pada *kernel linear, RBF, polynomial, dan sigmoid* sebagaimana pada [53]. Sementara itu, LR dipilih berdasarkan penelitian-penelitian sebelumnya oleh [17], [21], yang menunjukkan hasil akurasi yang lebih tinggi dibandingkan SVM. Kedua algoritma ini akan dibandingkan untuk menentukan algoritma mana yang menghasilkan akurasi terbaik dalam penelitian terkait pembangunan IKN. Berikut Tabel 3.2 merupakan komparasi kelebihan dan kekurangan algoritma yang dipilih.

Tabel 3.2 Perbandingan Algoritma SVM dan LR

Perbandingan	Algoritma	
	SVM	LR
Cara Kerja	• Mencari *hyperplane* yang paling optimal untuk memisahkan kelas data yang berbeda [51].	• Mengestimasi probabilitas suatu variabel • Dalam konteks analisis sentimen, LR digunakan

	• Dalam konteks analisis sentimen, SVM digunakan untuk mengklasifikasikan teks, menjadi kategori sentimen positif atau negatif.	untuk mengklasifikasikan teks menggunakan fungsi logistik untuk menghubungkan variabel independen (fitur teks) dengan variabel dependen (kategori sentimen). Fungsi logistik, atau sigmoid, menghasilkan output antara 0 dan 1 [55].
Perbandingan	**Algoritma**	
	SVM	**LR**
Kelebihan	• Efektif menangani fitur dengan dimensi tinggi, seperti teks dalam analisis sentimen, di mana jumlah fitur (kata-kata) sangat besar • Dapat menangani berbagai jenis data dengan menggunakan fungsi kernel yang berbeda [53]	• Cepat dalam pelatihan dan membutuhkan sumber daya komputasi yang lebih sedikit. • Memberikan *output* dalam bentuk probabilitas [76] • Koefisien dalam LR dapat diinterpretasikan sebagai pengaruh relatif dari setiap fitur terhadap probabilitas hasil, yang memberikan wawasan langsung tentang faktor-faktor yang mempengaruhi sentimen
Kekurangan	• Memerlukan penyetelan parameter yang tepat, seperti *kernel* dan C untuk kinerja optimal. • Kurang efisien pada *dataset* besar karena membutuhkan komputasi yang tinggi dan waktu pelatihan yang lama [77].	• Kurang fleksibel untuk hubungan non-linear data antara fitur dan target.

Berdasarkan perbandingan pada Tabel 3.2, pemilihan SVM didasarkan pada kemampuan SVM dalam menangani data berdimensi besar dan dapat menangani berbagai jenis data melalui penggunaan fungsi *kernel* yang berbeda. Kelebihan SVM dalam hal ini menjadi pertimbangan utama, terutama dalam analisis sentimen pembangunan IKN yang melibatkan tekstual data kompleks dan variatif. Sedangkan pemilihan LR didasarkan pada kecepatannya dalam pelatihan dan kemampuannya dalam memberikan *probabilitas* output yang jelas, yang sangat bermanfaat untuk interpretasi terkait faktor-faktor yang mempengaruhi sentimen. Hal ini sangat penting dalam analisis sentimen untuk membantu pemahaman mengenai bagaimana

fitur tertentu (kata-kata dalam teks) mempengaruhi hasil (sentimen positif atau negatif).

3.4.5 *Evaluation*

Pada tahap ini akan dilakukan evaluasi kinerja model untuk mengetahui seberapa besar akurasi yang dihasilkan dari setiap model. Proses evaluasi ini akan menggunakan *confusion matrix* dengan mempertimbangkan *nilai precision, recall, f1-score,* dan *accuracy. Precision* mengukur kualitas model dalam memprediksi kelas positif dari semua data yang diprediksi positif. *Recall* merujuk pada seberapa tepat model dalam memprediksi data aktual kelas positif dari keseluruhan data positif. Selanjutnya, *f1-score* merujuk pada keseimbangan model dalam memprediksi data positif yang sebenarnya *(recall)* dan kemampuan model dalam menghindari kesalahan mengklasifikasikan data negatif sebagai positif *(precision)*. Sementara itu, *accuracy* merujuk ketepatan model dalam memprediksi prediksi data aktual.

Lebih lanjut, akan dilakukan *Social Network Analysis* (SNA) untuk mengidentifikasi akun-akun yang berpengaruh dalam diskusi pembangunan IKN dengan membuat visualisasi jaringan dan perhitungan nilai *centrality* dengan menggunakan *gephi.* Nilai *centrality* yang dihitung adalah *degree centrality, betweenness centrality dan closeness centrality* [78]. *Degree centrality* digunakan untuk melihat akun dengan keterlibatan dan popularitas tertinggi berdasarkan dengan jumlah relasi terbanyak. Akun dengan *degree centrality* yang tinggi dianggap memiliki pengaruh yang besar karena memiliki banyak hubungan langsung dengan akun lain dan menjangkau audiens yang lebih luas. *Betweenness centrality* akan mengidentifikasi akun yang sering menjadi perantara atau jembatan dalam penyebaran informasi. Sementara itu, *closeness centrality* akan mengukur seberapa dekat sebuah akun ke semua akun lain dalam jaringan, berdasarkan jarak terpendek. Nilai *closeness centrality* berkisar antara 0 sampai dengan 1, semakin mendekati 1 maka akun memiliki kedekatan dengan akun lain, sehingga berpengaruh pada kecepatan penyebaran informasi [60].