BAB III
METODE PENELITIAN

3.1 Metode Penelitian

Dalam dunia *data mining* dan analisis data, terdapat beberapa *framework* metodologi yang memandu proses penemuan wawasan dari data. Tiga di antaranya yang paling dikenal adalah CRISP-DM, *Knowledge Discovery in Database* (KDD)[47], dan SEMMA (*Sample, Explore, Modify, Model, and Assess*)[48]. Setiap metode memiliki kelebihan dan kekurangannya masing-masing menyesuaikan dengan kebutuhan dalam penelitian. Hal ini dapat dilihat dalam tabel 3.1.

Tabel 3.1 Perbandingan Framework

Perbandingan	CRISP-DM	KDD	SEMMA
Definisi	Metodologi *data mining* yang paling populer dan diterima secara luas, bersifat komprehensif, fleksibel, dan iteratif, dirancang untuk proyek di berbagai industri.	Sebuah proses yang berfokus pada penemuan pengetahuan dari data, seringkali dari basis data yang sudah ada.	Metodologi *data mining* yang lebih ringkas dan berorientasi pada tahapan teknis inti dalam pengembangan model, dikembangkan oleh SAS Institute
Tahapan	1. Pemahaman Bisnis. 2. Pemahaman Data. 3. Persiapan Data. 4. Modeling. 5. Evaluasi. *6. Deployment.*	1. Seleksi. 2. Pra-pemrosesan. 3. Transformasi. 4. Data Mining. 5. Evaluasi. 6. Pengetahuan.	1. Sampel data. 2. Eksplorasi data. 3. Modifikasi data. 4. Pembangunan data. 5. Penilaian data.
Kelebihan	1. Sangat terstruktur dan komprehensif, mencakup seluruh siklus proyek.	Penekanan kuat pada penemuan pola dan ekstraksi pengetahuan baru dari data	1. Ringkas dan *straightforward*, fokus pada aspek teknis inti.

	2. Fleksibel dan dapat disesuaikan untuk berbagai jenis proyek data mining. 3. Berorientasi pada tujuan bisnis/penelitian yang jelas. 4. Mendukung sifat iteratif, memungkinkan perbaikan berkelanjutan di setiap tahap.		2. Efisien untuk proyek di mana data sudah relatif bersih atau tujuan sudah jelas.
Kekurangan	Bisa terasa terlalu "berat" untuk proyek yang sangat kecil atau tahap eksplorasi awal.	Cenderung lebih linier, meskipun beberapa interpretasi memungkinkannya bersifat iteratif	Kurang terstruktur dalam pemahaman bisnis dan persiapan data awal.

Berdasarkan perbandingan ketiga metode diatas, penelitian ini menggunakan salah satu metode CRISP-DM karena memiliki alur yang sesuai dengan kebutuhan peneliti mulai dari pemahaman bisnis atau permasalahan, pemahaman data, persiapan data, pemodelan data, evaluasi hingga deployment seperti pada gambar 3.1 *Framework* CRISP-DM.

Gambar 3.1 Framework CRISP-DM [49]

3.1.1 Business Understanding

Tahap awal ini berpusat pada penetapan arah proyek, mengidentifikasi problematikanya, dan merumuskan sasaran yang ingin dicapai dari proses analisis data.

3.1.2 Data Understanding

Pada fase ini, data yang relevan dikumpulkan dari beragam sumber, kemudian ditinjau dan dideskripsikan karakteristiknya. Penilaian kualitas data juga dilaksanakan guna memastikan kesiapannya untuk proses selanjutnya.

3.1.3 Data Preparation

Fase ini melibatkan serangkaian kegiatan untuk menyempurnakan data. Prosesnya mencakup penyaringan data berdasarkan kriteria

tertentu dan pembersihan elemen yang kurang valid, sehingga data yang akan diproses memiliki integritas tinggi.

3.1.4 Modeling

Dalam tahap pemodelan, fokusnya adalah pada seleksi dan pengujian teknik analisis yang paling tepat. Berbagai skenario pengujian dirancang, dan model dibangun menggunakan pendekatan yang paling relevan untuk menjawab pertanyaan penelitian.

3.1.5 Evaluation

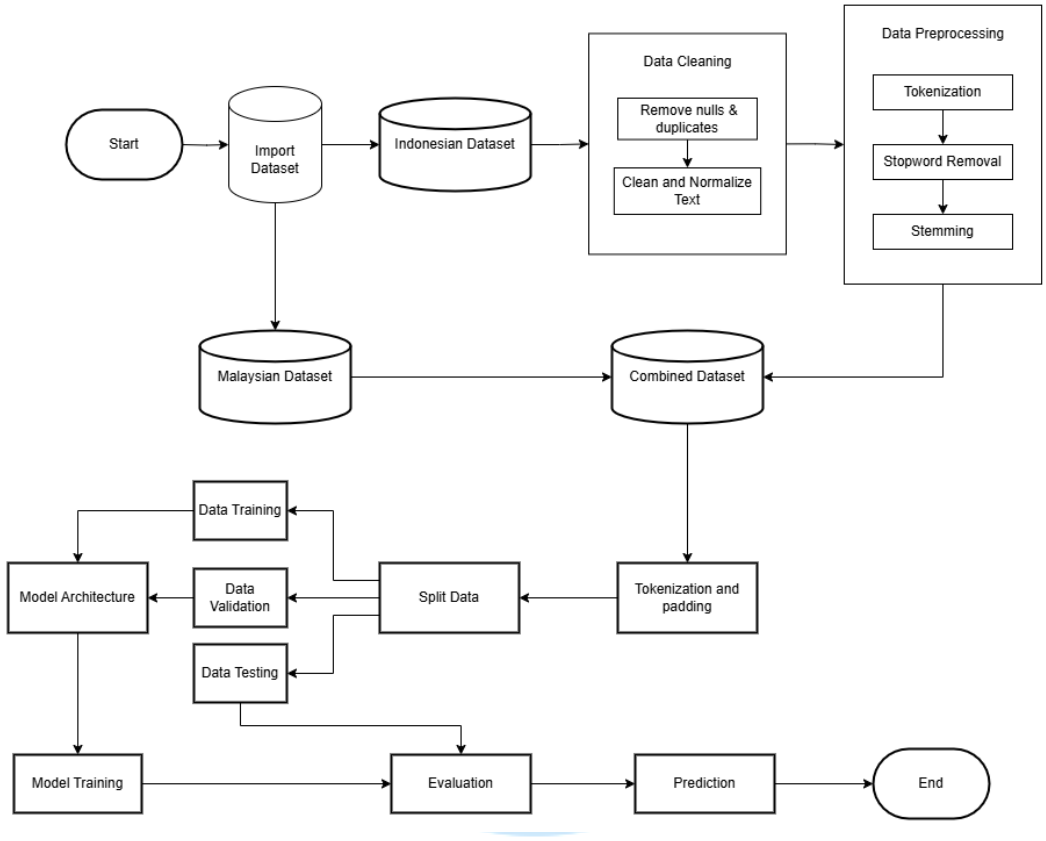
Pada tahap ini, kinerja model yang telah dikembangkan dievaluasi secara menyeluruh. Metrik penilaian yang relevan diterapkan untuk mengukur seberapa efektif model dalam mencapai tujuan penelitian yang telah ditetapkan.

3.1.6 Deployment

Tahap akhir ini mencakup dokumentasi komprehensif dari seluruh proses dan temuan penelitian. Hasil yang diperoleh dirumuskan menjadi kesimpulan, disertai rekomendasi untuk potensi implementasi di masa depan serta arah bagi riset selanjutnya.

3.2 Tahapan Penelitian

Setiap penelitian memiliki alur yang menjadi ciri khusus dari penelitian itu tersendiri. Gambar 3.2 merupakan tahapan penelitian yang dilakukan oleh peneliti, mulai dari perencanaan hingga evaluasi. Penelitian ini diawali dengan pemahaman permasalahan terkait penyebaran hoaks, yang dilanjutkan dengan pengumpulan data sekunder untuk dua bahasa yang berbeda, yaitu Indonesia dan Malaysia. Khusus untuk dataset Indonesia, dilakukan serangkaian tahap persiapan yang terdiri dari Data Cleaning (penghapusan data null & duplikat, serta normalisasi teks) dan Data Preprocessing (Tokenisasi, Stopword Removal, dan Stemming).



Gambar 3.2 Tahapan Penelitian menggunakan CRISP-DM

Setelah dataset Indonesia siap, data tersebut digabungkan dengan dataset Malaysia. Dataset Malaysia tidak melalui tahap persiapan dalam penelitian ini karena data yang diperoleh sudah dalam keadaan bersih. Dataset gabungan ini kemudian melalui proses Tokenisasi dan Padding untuk mengubah data teks menjadi format numerik yang seragam. Selanjutnya, data yang telah siap dibagi menjadi tiga bagian menggunakan metode split data, yaitu data latih, data validasi, dan data uji.

Setelah data terbagi, proses berlanjut ke tahap pemodelan. Pertama, pada tahap Model Architecture, arsitektur model Bi-LSTM dirancang dengan menentukan lapisan-lapisan yang akan digunakan, seperti lapisan Embedding, Bi-LSTM, dan Dense Layer. Kemudian, pada tahap Model Training, model dilatih menggunakan data latih, sementara kinerjanya dipantau pada setiap epoch menggunakan data validasi untuk mencegah overfitting. Tahap selanjutnya adalah Evaluation, di mana

model yang telah dilatih dievaluasi secara menyeluruh menggunakan data uji untuk mengukur performa akhirnya dengan berbagai metrik. Terakhir, model yang telah teruji dapat digunakan untuk tahap Prediction, yaitu melakukan klasifikasi pada data berita baru, yang menandai akhir dari alur penelitian ini.

3.3 Teknik Pengumpulan Data

Penelitian ini memanfaatkan data sekunder yang diperoleh dari platform Kaggle.com, bersumber dari dua koleksi data yang berbeda. Sumber pertama adalah dataset berjudul *"Indonesian Fact and Hoax Political News"*, yang diunggah oleh M. Razif Rizqullah dan Radhinansyah Hemsa Ghaida pada tahun 2023. Data ini dikumpulkan melalui *scraping* dari beberapa media berita ternama seperti CNN, Kompas, dan Tempo, serta dari platform Turnbackhoax. Dataset ini berisi total 31.332 baris data, dengan distribusi label sebagai berikut: 20.768 baris data berlabel [0] dan 9.616 baris data berlabel [1].

Sumber kedua adalah dataset berjudul *"Indonesia False News (hoax) Dataset"*, yang diunggah oleh Muhammad Ghazi Muharram. Dataset ini memiliki total 4.701 baris, di mana 4.231 baris di antaranya telah memiliki label. Distribusi label pada dataset kedua ini adalah 3.465 baris data berlabel [1] dan 766 baris data berlabel [0]. Kemudian dataset ketiga berjudul "Malay-fake-news-classification" yang didapatkan dari respositori Github. Dataset ini memiliki jumlah total 37.592 baris data dengan persebaran 25.030 data asli dan 12.562 data palsu. Sehingga total keseluruhan data yang digunakan adalah 67.935 baris.

3.4 Variabel Penelitian

Penelitian ini menggunakan dua variabel yang berbeda, yaitu variabel dependen dan variabel independen.

3.4.1 Variabel Dependen

Variabel dependen, yang juga dikenal sebagai variabel terikat, merujuk pada luaran atau efek yang diamati sebagai respons terhadap perubahan pada variabel lain, dan menjadi pusat perhatian dalam suatu

penelitian. Karakteristik utamanya adalah ketergantungannya pada variabel independen, di mana setiap modifikasi pada variabel independen akan secara proporsional memengaruhi tingkat perubahan pada variabel dependen [121]. Dalam konteks penelitian ini, variabel dependen adalah luaran dari proses deteksi berita, yang dikelompokkan ke dalam dua kategori: 0 untuk mengindikasikan bukan hoaks, dan 1 untuk hoaks. Klasifikasi ini merefleksikan status validitas berita, menjadikannya hasil pokok dari model klasifikasi yang dikonstruksi. Adanya variasi pada ciri-ciri teks berita, selaku variabel independen, diharapkan akan turut serta memodifikasi hasil deteksi yang dihasilkan.

3.4.2 Variabel Independen

Variabel independen, atau variabel bebas, didefinisikan sebagai elemen yang berfungsi sebagai pemicu perubahan atau memberikan dampak signifikan terhadap suatu kondisi atau nilai tertentu. Dalam konteks penelitian, variabel ini berperan sebagai prediktor yang dimanfaatkan untuk menghasilkan estimasi terhadap variabel dependen. Dalam studi ini, variabel independen adalah keseluruhan konten berita yang disajikan dalam format teks panjang. Teks berita ini berperan sebagai input esensial bagi model, dengan asumsi bahwa karakteristiknya secara langsung memengaruhi atau menentukan luaran dari proses deteksi hoaks. Dengan menganalisis berbagai fitur dan pola yang terdapat dalam teks berita, model akan mampu mengklasifikasikan apakah berita tersebut tergolong hoaks atau bukan.