BAB III
METODOLOGI PENELITIAN

3.1 Gambaran Umum Objek Penelitian

Objek penelitian yang dilakukan dalam penelitian ini berfokus pada Prediksi Pemilihan Presiden RI Tahun 2024. Pemilihan umum merupakan proses masyarakat Indonesia memilih seseorang untuk mengisi jabatan politik. [62]. Pada proses pemilihan presiden biasanya opini-opini masyarakat terhadap calon presiden dan wakil presiden yang dibagikan di sosial media salah satunya adalah X. X sering digunakan sebagai tempat di mana diskusi publik terjadi secara langsung, termasuk debat politik dan kampanye pemilihan. Dengan demikian, X memiliki potensi besar untuk memberikan wawasan yang berharga tentang sentimen masyarakat terhadap calon presiden dan isu-isu politik terkini. Data yang digunakan dalam penelitian ini berasal dari hasil *scraping* X dengan berbagai *keyword*, yang dilakukan melalui *platform Google Colab* dengan bantuan bahasa pemrograman *Python*. Periode pengumpulan data berlangsung dari November 2023 hingga Januari 2024, sesuai dengan periode kampanye yang dilakukan oleh calon presiden dan wakil presiden.

3.2 Metode Penelitian

Metode penelitian terdapat beberapa jenis, metode yang paling sering digunakan adalah metode kualitatif dan kuantitatif. Metode kualitatif adalah metode yang menggunakan pengamatan, wawancara. Metode kuantitatif adalah metode yang menggunakan analisis statistik terhadap data angka [63]. Penelitian ini menggunakan metode kuantitatif karena menggunakan dataset yang berasal dari sosial media X. Dataset ini diperoleh dengan cara *scraping data*. Pada penelitian dengan metode kuantitatif akan menggunakan perbandingan algoritma *Naive Bayes, Random Forest,* dan *Support Vector Machines (SVM)*.

Dalam data *minning* terdapat beberapa metode yang bisa digunakan yaitu metode *Cross Industry Standard Process (CRISP-DM), Knowledge Discovery* in *Database (KDD),* dan *SEMMA (Sample, Explore, Modify, Model, and Assess)* [59]. Berikut merupakan tahapan dari masing-masing metode. [64]

Tabel 3. 1 Tahapan Metode *KDD, SEMMA, CRISP-DM*

KDD	SEMMA	CRISPDM
Pre KDD	-	Business Understanding
Selection	Sample	Data Understanding
Pro processing	Explore	
Transformation	Modify	Data Preparation
Data Mining	Model	Modeling
Interpretation/Evaluation	Assesment	Evaluation
Post KDD	-	Deployment

Pada ketiga metode tersebut masing-masing memiliki kelebihan dan kekurangan. Berikut merupakan kelebihan dan kekurangan dari metode *Cross Industry Standard Process (CRISP-DM)*, *Knowledge Discovery in Database (KDD)*, dan *SEMMA (Sample, Explore, Modify, Model, and Assess)*. [65][66][67]

Tabel 3. 2 Perbandingan *KDD, SEMMA, CRISP-DM*

Metode	Kelebihan	Kekurangan
CRISP-DM	Metode yang sering digunakan di lingkungan industri, karena kelebihannya dalam mengatasi berbagai masalah dalam proyek-proyek *data mining*.	Dapat terasa kompleks untuk proyek kecil.
KDD	- KDD dapat Meningkatkan pengambilan Keputusan yang lebih baik. - KDD dapat digunakan untuk mendeteksi aktivitas penipuan dengan mengidentifikasi pola dan anomali. - KDD dapat digunakan untuk membangun model prediktif yang dapat meramalkan tren dan pola masa depan.	- KDD dapat meningkatkan masalah privasi karena melibatkan pengumpulan dan analisis data dalam jumlah besar, yang dapat mencakup informasi sensitif tentang individu. - Proses KDD sangat bergantung pada kualitas data, jika data tidak akurat atau konsisten, hasilnya bisa menyesatkan
SEMMA	SEMMA dirancang oleh SAS Institute dan diterapkan pada kumpulan data besar untuk menemukan pola yang sebelumnya tidak teridentifikasi, sehingga dapat dimanfaatkan sebagai keunggulan dalam dunia bisnis.	SEMMA dirancang bersamaan dengan sebuah aplikasi atau alat dari SAS yang dikenal sebagai Enterprise Miner, sehingga penggunaannya menjadi terbatas apabila digabungkan dengan alat penambangan data yang berbeda atau dalam konteks proses bisnis secara keseluruhan.

Berdasarkan Tabel 3.1 dan Tabel 3.2 dapat disimpulkan bahwa *CRISP-DM* menawarkan prosedur standar untuk penambangan data yang bisa

diintegrasikan dalam strategi penyelesaian masalah secara umum di bidang bisnis atau riset. Dibandingkan dengan metodologi penambangan data lainnya, *CRISP-DM* lebih komprehensif dan terdokumentasi secara rinci. Selain itu, dari hasil *survey* mendapatkan bahwa *CRISP-DM* adalah metode *data mining* yang paling populer. Hasil survey dari KDNuggets memperlihatkan bahwa *CRISP-DM* di tahun 2014 mencapai lebih dari 40% dan pada tahun 2020 *CRISP-DM* mencapai lebih dari 45%[68]. Penelitian ini akan menggunakan teknik atau metode *CRISP-DM (Cross Industry Standard Process for Data mining)* karena metode tersebut paling popular digunakan dan *CRISP-DM* merupakan sebuah standarisasi untuk melakukan proses *data mining* agar menghasilkan solusi pemecahan masalah dalam suatu perusahaan atau organisasi [36]. *CRISP-DM* memiliki enam tahapan dalam proses *data mining* yaitu, *Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, Deployment* [37] Dalam melakukan penelitian ini, menggunakan *tools* google colab dengan menggunakan bahasa pemrograman *Python*.

3.3 Variabel Penelitian

Penelitian ini menggunakan dataset yang berasal sosial media X. Dataset ini diperoleh dengan cara *scraping data X*. Pada variabel penelitian terdapat dua variabel, yaitu variabel independen dan variabel dependen.

3.3.1 Variabel Independen

Variabel independen adalah variabel yang tidak terikat atau bebas yang dapat memberikan dampak terhadap variabel lain. Pada penelitian ini, *Variable* yang digunakan untuk klasifikasi adalah *Variable full_text*. *Variable* yang digunakan untuk metode social network analysis adalah *Variable username, location, favorite_count, quote_count, reply_count,* dan *retweet.*

1. *full_text*: berisi teks lengkap dari tweet yang diposting oleh akun.
2. *username*: nama pengguna dari akun yang men-tweet.
3. *location*: lokasi atau daerah dari akun pengguna yang men-tweet (bisa kosong atau tidak terisi).

4. *favorite_count*: jumlah tweet yang difavoritkan oleh pengguna lain.
5. *quote_count*: jumlah tweet yang di-quote oleh pengguna lain.
6. *reply_count*: jumlah tweet yang mendapatkan balasan dari pengguna lain.
7. *retweet*: jumlah tweet yang di-retweet oleh pengguna lain.

3.3.2 Variabel Dependen

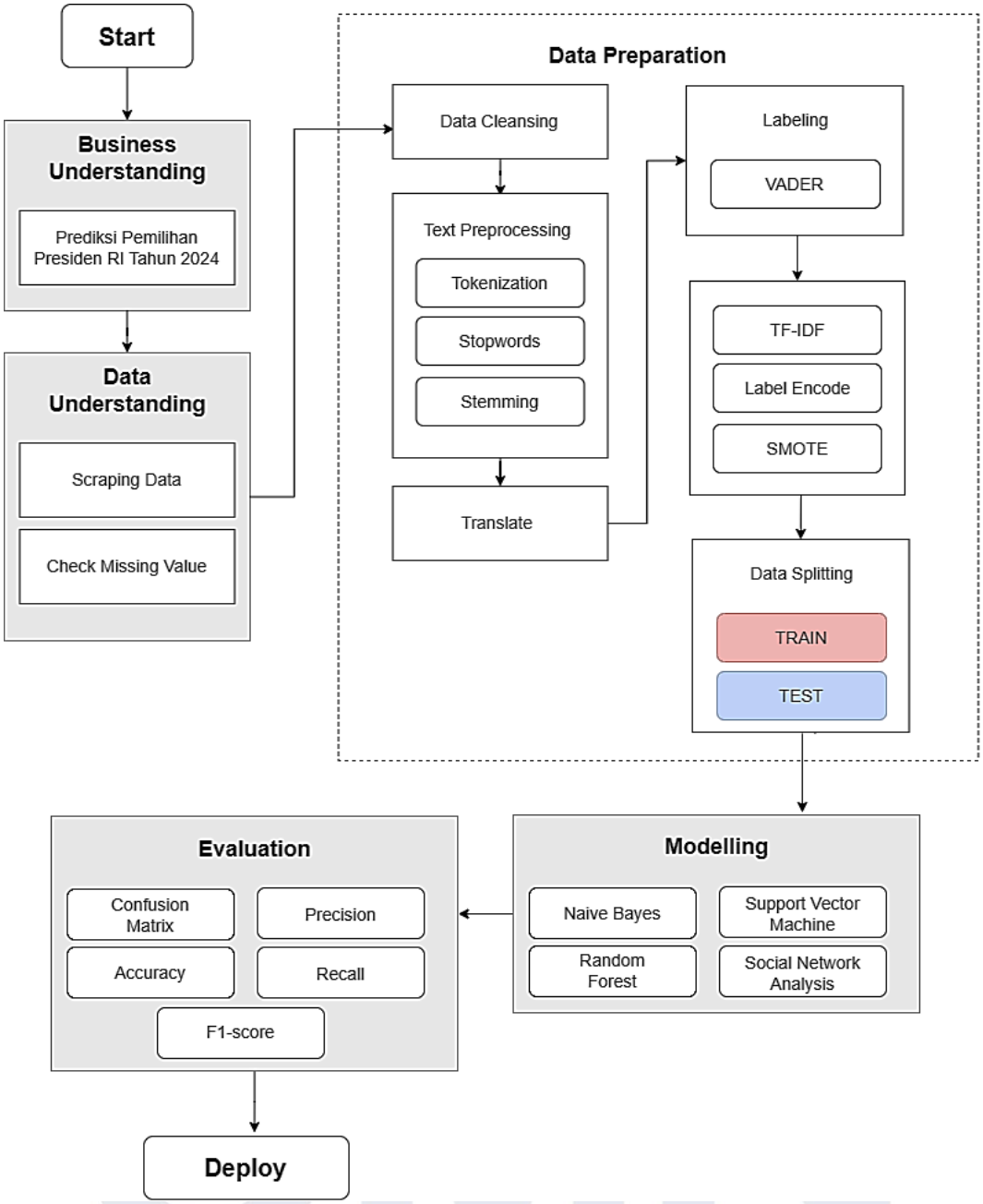
Variabel Dependen adalah variabel yang terikat atau dipengaruhi oleh variabel independent. Pada penelitian ini *Variable* dependen untuk analisis sentimen terdapat pada *variable* sentimen (positif, negatif, netral), sedangkan untuk *variable* dependen *Social Network Analysis* adalah *degree centrality.*

3.4 Teknik Pengumpulan Data

Teknik pengumpulan data memiliki beberapa teknik yaitu, data primer, data sekunder, dan data tersier. Data primer adalah data yang dikumpulkan sendiri atau data asli, data sekunder adalah data yang didapatkan dari berbagai sumber studi literatur, dan data tersier adalah data penunjang yang didapatkan dari ensiklopedia atau sumber lain yang memiliki masalah yang dapat diteliti dan di analisis[63]. Penelitian ini menggunakan data primer yang mana data dikumpulkan dari hasil *scraping* media sosial X dengan periode selama November 2023 hingga Januari 2024.

3.5 Teknik Analisis Data

Dalam penelitian ini, akan menerapkan model *CRISP-DM (Cross Industry Standard Process for Data mining).* Berdasarkan gambar 3.1 terdapat enam tahapan dalam *CRISP-DM* yang akan dilakukan pada penelitian ini dengan penjelasan sebagai berikut.



Gambar 3 1 Teknik Analisis Data

3.5.1 *Business Understanding*

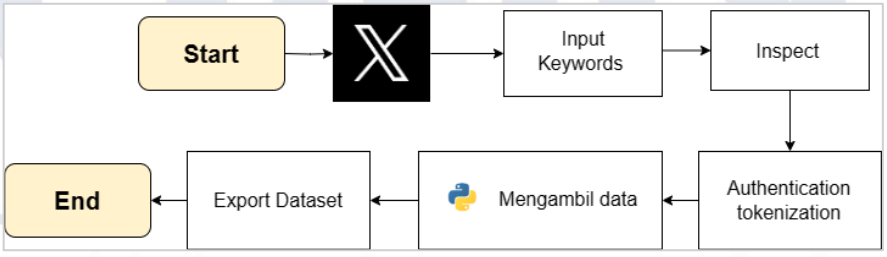
Pada tahapan ini bertujuan untuk menentukan tujuan dari projek dalam ruang lingkup bisnis. Pada penelitian ini tujuan utama adalah untuk Prediksi Pemilihan Presiden RI Tahun 2024. Pada masa kampanye pemilihan presiden banyak berita-berita yang tersebar luas yang dapat berdampak pada opini masyarakat. Prediksi Pemilihan Presiden RI Tahun 2024 dilakukan untuk melihat sentimen dari opini masyarakat terhadap masing-masing calon presiden dan wakil presiden. Penelitian ini akan membandingkan algoritma *Naive Bayes*, *Random Forest*, dan *Support Vector Machines (SVM)* dan metode *social network analysis (SNA)* dengan bantuan *framework CRISP-DM*.

3.5.2 *Data Understanding*

Pada tahapan ini merupakan tahapan untuk memahami data-data yang digunakan untuk menganalisis data tersebut.

3.5.2.1 Scraping Data

Proses pengumpulan data dilakukan melalui teknik *scraping* data menggunakan bantuan bahasa pemrograman *Python* dari *platform* media sosial X. Data diambil dalam rentang periode dari November 2023 hingga Januari 2024. Proses *scraping* data dilakukan untuk mengumpulkan informasi yang relevan dari *platform* X, termasuk teks-*tweets* dan informasi pengguna, sehingga dapat memberikan gambaran yang komprehensif tentang percakapan yang terjadi selama periode yang ditentukan.



Gambar 3 2 Proses *scraping data*

Gambar 3.2 menggambarkan proses pengumpulan data *tweet* atau *scraping data* untuk analisis, mulai dari awal hingga data siap digunakan. Proses dimulai dengan menentukan kata kunci dan mendapatkan token

otentikasi yang dibutuhkan untuk mengakses X API. Setelah itu, menggunakan bahasa pemrograman *Python*, data *tweet* dikumpulkan berdasarkan kata kunci yang telah ditentukan berdasarkan *keyword* yang pernah *trending* dan nama dari masing-masing pasangan calon. Data yang terkumpul kemudian disimpan ke dalam sebuah dataset, yang merupakan kumpulan dari *tweet* yang relevan untuk analisis. Langkah terakhir dalam proses ini adalah mengekspor dataset tersebut, membuatnya tersedia untuk dianalisis lebih lanjut.

3.5.2.2 Check Missing Value

Check missing value merupakan salah satu langkah penting dalam tahap *Data Understanding*. Proses ini berguna untuk mengevaluasi apakah terdapat data yang memiliki nilai *null* atau kosong dalam dataset. Dengan melakukan *check missing value*, dapat mengidentifikasi seberapa banyak data yang hilang. Langkah ini juga membantu dalam menentukan strategi untuk mengatasi nilai-nilai yang hilang, seperti penggantian nilai *null* dengan nilai yang sesuai atau penghapusan baris data yang tidak lengkap.

3.5.3 *Data Preparation*

Pada tahapan ini, dilakukan proses pembersihan data. Beberapa proses yang dilakukan pada tahap *text preprocessing* adalah c*ase folding, tokenizing, stemming,* dan *stopword removal* [65]. Tahapan dalam *data preparation* meliputi *data cleansing, preprocessing, data translation, labeling, TF-IDF, label encoding, SMOTE,* dan *data splitting.*

3.5.3.1 Data Cleansing

Pada data cleansing akan dilakukan pembersihan data. Tahapan ini akan melakukan *remove url, remove hashtag, remove html tags & code, remove username,* angka, tanda baca, emoji, dan spasi ekstra, *lowercase.* Tahapan ini dilakukan agar dataset yang digunakan menjadi bersih dan siap dilakukan untuk tahap *preprocessing.*

3.5.3.2 Preprocessing

Text Preprocessing adalah proses pembersihan teks ke dalam bentuk yang dapat diprediksi dan dianalisis untuk tugas tertentu. Tujuan utama dari *preprocessing* teks adalah untuk memecah teks menjadi bentuk yang dapat dicerna oleh algoritma pembelajaran mesin [30]. Pada bagian *text preprocessing* akan dilakukan *tokenization, remove stopwords,* dan *stemming.* Tokenisasi merupakan proses untuk membagi teks menjadi unit yang lebih kecil, yaitu, token, mungkin pada saat yang sama membuang karakter tertentu, seperti tanda baca. Setelah melakukan tokenisasi, tahapan selanjutnya adalah *remove stopwords. Stop-words* adalah kata-kata yang umum digunakan dalam suatu bahasa. *Stop-word* dihilangkan dari teks sehingga kita dapat berkonsentrasi pada kata-kata yang lebih penting. Tahapan terakhir yang dilakukan dalam *text preprocessing* adalah *stemming. Stemming* adalah proses dasar berbasis aturan untuk menghilangkan bentuk infleksional dari sebuah token. Token diubah menjadi bentuk akarnya. Misalnya kata 'bermasalah' diubah menjadi 'masalah' setelah dilakukan *stemming* [30][31].

3.5.3.3 Data Translation

Pada tahapan ini akan melakukan translate data dari bahasa Indonesia menjadi bahasa inggris dengan bantuan *library deep translator*. Data yang di translate merupakan data yang telah dilakukan pembersihan data dan *preprocessing data*. Tahapan ini dilakukan untuk proses *labeling* data menggunakan library *VADER.*

3.5.3.4 Labeling

Pada proses *labeling* data akan menggunakan bantuan dari *library VADER (Valance Aware Dictionary Sentiment Reasoner).* VADER akan mengevaluasi setiap teks dengan mencetak skor positif, negatif, atau netral, dan semua skor akan dijumlahkan untuk menghasilkan nilai *compound score*. *Compound score* adalah matriks yang memperhitungkan semua skor yang dinormalisasi dari rentang -1 hingga +1. Jika nilai

komposit lebih besar dari 0,05, maka dianggap positif; jika kurang dari -0,05, dianggap negatif; dan jika antara -0,05 dan 0,05, dianggap netral.

3.5.3.5 TF-IDF

TF-IDF (Term Frequency-Inverse Document Frequency) adalah metode ekstraksi fitur yang umum digunakan dalam klasifikasi teks berbasis ruang vektor. *TF-IDF* memberikan bobot pada setiap kata yang muncul dan menghitung nilai invers di dalam kalimat. *TF-IDF* dilakukan untuk mengonversi teks menjadi vektor numerik berdasarkan frekuensi kata dalam teks tersebut dan juga frekuensi kemunculan kata tersebut dalam seluruh dataset [32].

3.5.3.6 Label Encode

Label Encoding adalah teknik pengkodean yang digunakan untuk mengubah label kategori menjadi label numerik. Label sentimen positif, negatif, netral diubah menjadi label numerik dengan keterangan ‘negatif: 0’, ‘netral:1’, dan ‘positif: 2’. *Label encode* ini dilakukan setelah proses *TF-IDF*.

3.5.3.7 SMOTE

Dataset yang dihasilkan memiliki ketidakseimbangan data dari hasil *labeling*. Pada tahapan ini akan dilakukan keseimbangan data menggunakan teknik *SMOTE*. Hal pertama yang dilakukan dalam teknik *SMOTE* adalah identifikasi kelas mayoritas (dominan) dan kelas minoritas (kurang dominan) dalam dataset hasil *labeling* menggunakan *VADER*. Kemudian, menyesuaikan parameter agar menghasilkan jumlah sampel sintetis yang sama dengan kelas mayoritas.

3.5.3.8 Data Split

Pada proses ini, data dibagi menjadi dua bagian, yaitu *data training* dan *data testing*. *Data training* digunakan untuk melatih model, seperti mencari parameter terbaik untuk algo yang digunakan, sedangkan *data testing* digunakan untuk memvalidasi atau mengevaluasi kinerja model. Dari penelitian [69] tentang analisis sentimen kanjuruhan, peneliti

melakukan pembagian *ratio* dengan membandingkan tiga *ratio*, yaitu 60:40, 70:30, dan 80:20. Dari hasil penelitian tersebut menunjukkan bahwa ratio 70:30 memberikan hasil akurasi yang tinggi.

3.5.4 *Modeling*

Pada tahapan ini, akan dilakukan analisis dari pemodelan dengan algoritma klasifikasi yang sudah dipilih, yaitu *Naive Bayes, Random Forest,* dan *Support Vector Machines (SVM),* serta metode *Social Network Analysis (SNA)* untuk prediksi Pemilihan Presiden RI Tahun 2024. Pemodelan dengan ketiga algoritma ini akan dilakukan secara terpisah, dan akan dilakukan perbandingan dari ketiga algoritma klasifikasi. Dalam *Social Network Analysis (SNA),* dipilih untuk menggunakan *degree centrality* karena metrik ini secara efektif dapat mengidentifikasi akun-akun yang paling berpengaruh dalam jaringan social [42]. *Degree centrality* memberikan wawasan yang penting tentang distribusi dan pengaruh informasi dalam jaringan, sehingga dapat memberikan pemahaman yang mendalam tentang dinamika politik dan opini publik dalam konteks Pemilihan Presiden RI 2024. *Proses modeling* ini akan menggunakan *tools Google Colab* dengan bahasa pemrograman *Python*.

3.5.5 *Evaluation*

Pada tahapan ini akan dilakukan penilaian atas performa dari model yang sudah dibuat. Tahapan ini akan menampilkan *confusion matrix, akurasi, precision, recall, F1-score* dari model algoritma *Naïve Bayes, Random Forest*, dan *Support Vector Machine*. Setelah itu, akan dilakukan perbandingan performa dari ketiga algoritma tersebut.

3.5.6 *Deployment*

Pada tahapan ini merupakan tahapan akhir yang akan melakukan implementasi terhadap hasil yang didapatkan. Penelitian ini akan menggunakan visualisasi data berupa *dashboard* yang berisi hasil analisis data Prediksi Pemilihan Presiden RI Tahun 2024 dan akan diimplementasikan ke dalam *website*.