# BAB 3
# METODOLOGI PENELITIAN

## 3.1 Gambaran Umum Objek Penelitian

Objek Penelitian yang akan digunakan adalah text yang akan diambil dari X yang akan dijadikan Dataset Tweets. Dataset Tweets merupakan dataset yang diambil dari platform X dengan kata kunci @KPU_ID dengan jangka waktu November 2023 sampai Februari 2024. Dataset ini berisi sebanyak 4314 data yang berisi tulisan yang diunggah oleh user X untuk menceritakan pengalaman penulis terhadap akun @KPU_ID. Dataset telah diverifikasi dan dilabel oleh pakar sebanyak 0,1% dari total jumlah dataset.

## 3.2 Metode Penelitian

Metode penelitian pada *data mining* ada 3 arsitektur yang dijadikan sebagai referensi untuk melakukan alur penelititan. 3 arsitektur tersebut bernama CRISP-DM, SEMMA, dan KDD. Untuk memperjelas perbedaan antara 3 arsitektur tersebut, ada penjelasan perbandingan pada tabel 4.3 Perbandingan KDD, SEMMA, and CRISP-DM untuk data mining.

Berdasarkan tabel 4.3 Perbandingan KDD, SEMMA, and CRISP-DM untuk data mining, dapat dilihat pada penelitian ini memiliki hubungan erat dengan arsitektur *Knowledge Discovery in Databases* karena memiliki keunggulan dalam menganalisis dataset. KDD sering dianggap cocok untuk analisis sentimen karena beberapa alasan kunci yang membuatnya sangat sesuai untuk tugas *data mining* seperti ini dibandingkan dengan metodologi lain seperti Crisp-SM dan SEMMA. Berikut adalah alasan mengapa KDD menjadi pilihan unggul untuk analisis sentimen[38]:

1. **Pemilihan Data (Data Selection)**
    - Identifikasi dan pemilihan subset data yang relevan untuk dianalisis. Pemilihan data pada penelitian ini menggunakkan *data scraping* RapidAPI. Data kemudian di label dengan kamus leksikon. Pelabelan kamus leksikon dilakukan untuk membantu penelitian dalam pelabelan dataset yang jumlah besar. Pada penelitian ini juga ingin melihat bagaimana akurasi data label hasil dari kamus leksikon.

Tabel 3.1. Perbandingan KDD, SEMMA, and CRISP-DM untuk data mining [37] [38] [39]

| Criteria | KDD | SEMMA | CRISP-DM |
|---|---|---|---|
| **Jumlah Langkah** | 5 | 5 | 6 |
| **Steps** | 1. Data Selection<br>2. Pre-processing<br>3. Transformation<br>4. Data Mining<br>5. Interpretation | 1. Sample<br>2. Explore<br>3. Modify<br>4. Model<br>5. Assess | 1. Business Understanding<br>2. Data Understanding<br>3. Data Preparation<br>4. Modeling<br>5. Evaluation<br>6. Deployment |
| **Strengths** | Comprehensive, flexible, and widely used | Specific to SAS software, but provides detailed guidance for modeling | Widely adopted, comprehensive, and well-documented |
| **Weaknesses** | Can be complex and time-consuming | Limited applicability outside of SAS software | May be perceived as rigid, depending on implementation |

2. **Pra-pemrosesan Data (Data Preprocessing)**
    - Membersihkan dan mempersiapkan data mentah untuk analisis lebih lanjut. Ini mencakup penanganan data yang hilang, penghapusan duplikasi, koreksi kesalahan, dan normalisasi data.
3. **Transformasi Data (Data Transformation)**
    - Mengubah data ke dalam format yang sesuai untuk analisis. Pada proses ini dilakukan embedding untuk mengubah kalimat menjadi angka dalam bentuk vektor. Membantu model dalam pembelajaran karena model hanya bisa memproses data berupa angka.
4. **Data Mining**
    - Menerapkan metode dan algoritma untuk mengekstraksi pola dari data. Metode yang digunakan bisa beragam, termasuk clustering,

classification, regression, dan association rule learning. Pada penelitian ini menggunnakkan LSTM, dengan mengubah hiperparameter.

5. **Evaluasi Pola (Pattern Evaluation)**

   - Mengevaluasi pola yang ditemukan untuk memastikan bahwa model relevan dan bermanfaat. Ini melibatkan pengujian validitas pola terhadap data dan mengukur performanya menggunakan *Confusion Matrix*

KDD (Knowledge Discovery in Databases) adalah proses untuk mengekstraksi pengetahuan yang bermanfaat dari data secara sistematis. Analisis sentimen adalah salah satu aplikasi di mana KDD sangat cocok karena melibatkan ekstraksi informasi berharga dari data teks untuk memahami opini, perasaan, atau sentimen yang terkandung dalam teks tersebut.

### 3.3 Teknik Pengumpulan Data

Pengumpulan data untuk penilitian ini menggunakkan teknik *Data Scrapping*. Dengan menggunakkan API RapidAPI untuk mengambil data tweets dari X. Pada penelitian ini pengumpulan data menggunakan kata kunci @KPU_ID yang merupakan akun resmi dari Komisi Pemilihan Umum. Pemilihan kata kunci dilakukan supaya mendapat semua tweets yang berhubungan dengan Komisi Pemilihan Umum. Pengambilan data dilakukan pada bulan November 2023 hingga Februari 2024 dimana isu pemilihan umum untuk 2024 dekat pelaksanaannya.

Pengunaan API dilakukan dalam *google collab*. Pengumpulan dengan metode *scrapping* dengan bantuan API RapidAPI berhasil dilakukan dengan memperoleh total 8415 data mentah. Pengumpulan data dengan scraping ini membantu penelitian dalam menghemat waktu sehingga memanfaatkan program komputer untuk membantu mengambil data terutama membutuhkan data skala besar seperti implementasi *machine learning*. Dapat dilihat dari tabel 3.2 Data pelabelan leksikon 5 data sampel yang didapatkan dari hasil *scraping*.

Data *scrapping* kemudian akan diberi pelabelan dengan kamus leksikon. Kamus leksikon adalah kumpulan kata-kata kunci beserta polaritasnya (seperti positif, negatif, netral) yang digunakan dalam analisis sentimen untuk menilai atau mengklasifikasikan sentimen dari teks. Dalam konteks analisis sentimen, kamus leksikon menyediakan referensi yang konsisten dan terstruktur untuk mengevaluasi

Tabel 3.2. Data pelabelan leksikon

| Tweet | Sentimen |
|---|---|
| @KPU_ID Berdalih dg kesalahan sirekap, sangat tidak wajar. Kelihatan ini kelompok konspirasi dan perlu diproses hukum secepatnya | neutral |
| @KPU_ID Konferensi press apa klarifikasi ketidak netralan? Masa konferensi pers gak ada tanya jawab? Persis kayak 02 yg g mau tanya jawab dengan wartawan. | positive |
| @KPU_ID kalian harus dengar keluhan rakyat | neutral |
| @KPU_ID Ahhh kirain gw pelantikan prabowo sama gibran... | neutral |
| ”@KPU_ID Jujurlah,.. Jangan sampai Ibumu menangis dan menyesal telah melahirkanmu,..” | negative |

emosi, opini, atau pandangan yang terkandung dalam sebuah teks berdasarkan kata-kata yang digunakan.

Penggunaan kamus leksikon dalam analisis sentimen penting karena menyediakan landasan konsisten dan terstruktur untuk mengevaluasi sentimen teks. Kamus ini memungkinkan identifikasi otomatis dan efisien terhadap polaritas kata-kata, memungkinkan analis untuk memahami keseluruhan sentimen sebuah teks dengan lebih baik. Manfaatnya mencakup efisiensi dalam penilaian sentimen, pembaruan dan pengembangan yang dapat disesuaikan, serta kemampuan untuk mengukur perubahan sentimen dari waktu ke waktu. Keseluruhan, penggunaan kamus leksikon membantu meningkatkan akurasi dan keterandalan analisis sentimen dalam riset.

Dari tabel 3.2 Data pelabelan leksikon bisa kita lihat data yang telah di label dengan sentiment positif, negatif dan netral. Data ini akan bermanfaat pada *supervised* machine learning untuk membantu pembuatan model machine learning pada penelitian ini.

## 3.4 Variabel penelitian

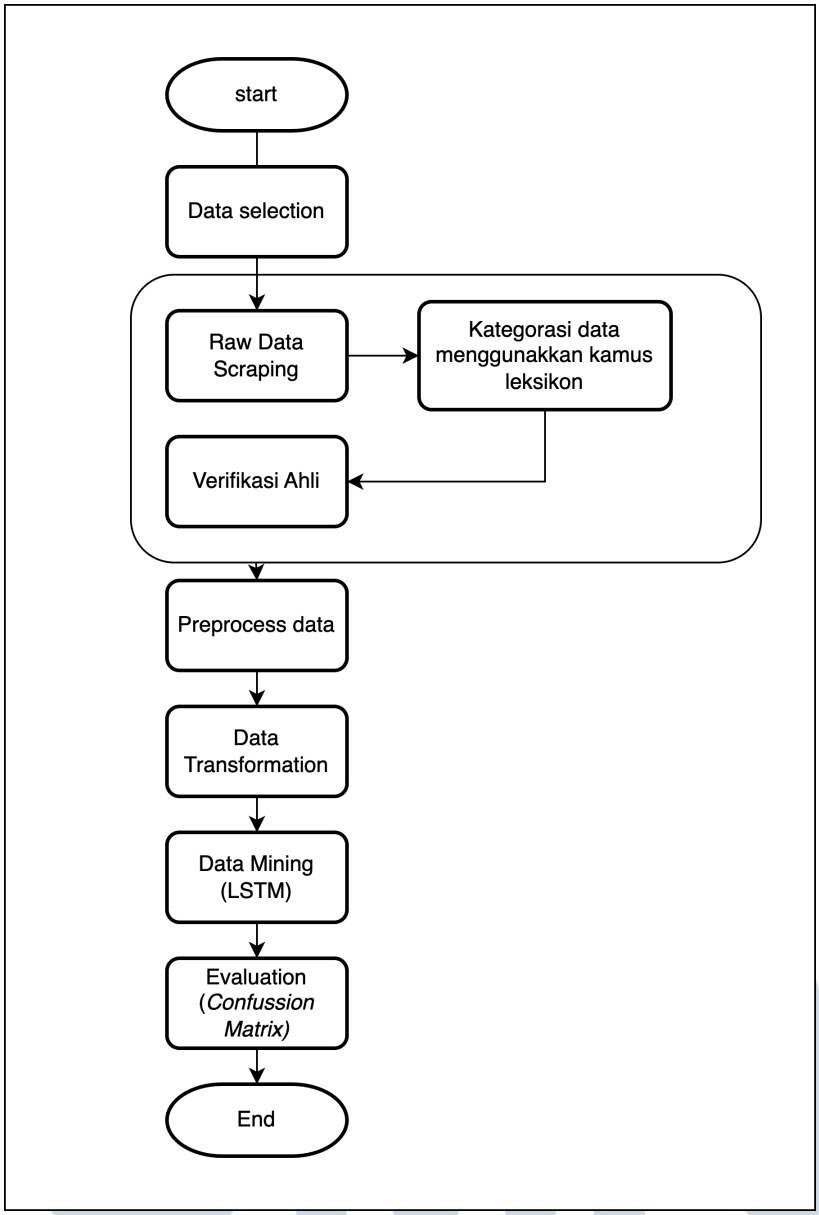
- Variabel Independen pada penelitian ini merupakan variabel yang bisa berdiri sendiri tanpa pengaruh variabel lain. Pada penelitian ini variabel independen

merupakan data tweets

- Variabel Dependen pada penelitian ini merupakan variabel yang membutuhkan variabel pendukung dari variabel independen. Variabel dependen yang diteliti merupakan algoritma LSTM yang akan dipake untuk menganalisis sentimen tweets.

## 3.5 Alur Penelitian

Alur penelitian dapat diliha pada 3.1 Flowchart Alur Penelitian akan dilakukan sesuai dengan metodologi KDD dimana akan dimulai dengan *data selection* dimana pada penelitian ini menggunakkan RapidAPI sebagai library *scraping data*. Data tersebut kemudian akan di proses dengan leksikon untuk pelabelan yang didukung oleh verfikasi ahli. Setelah di verfikasi dan dilabel data akan masuk ke proses *preprocessing* yang akan melalu tahap *case folding*, *filtering*, *stopword removal*, dan *stemming*. Data kemudian akan melewati *data transformation* yaitu tokenisasi dan *embedding* supaya model bisa membaca dengan Angka. Setelah itu akan dilakukan *data mining* yaitu klasifikasi teks menggunakkan model LSTM yang akan dirancang. Pada langkah terakhir ada langka evalusai model untuk melihat hasil dari kinerja model dan kekurangan yang model miliki.




Gambar 3.1. Flowchart Alur Penenlitian

## 3.6 Teknik Analisis Data

Teknik analisis data menggunakkan perancangan bangun model menggunakkan algoritma LSTM dengan mencari evaluasi performa dan akurasi tertinggi. Untuk mendapat akurasi dan performa dari analisis model akan dilakukan optmisasi model dengan *tuning hyperparatmeter* yaitu *lstm hidden units*, dimensi embedding, *learning rate* dan *batch size*. Penelitian juga akan dilakukan pada

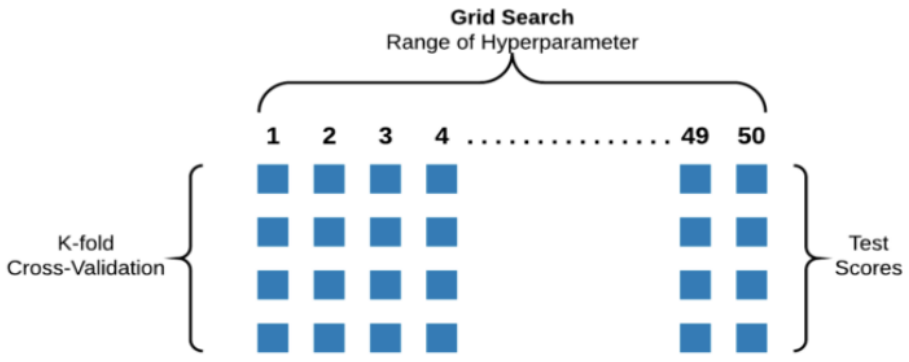
2 pembagian data testing dan training yaitu 80:20 dan 90:10. Setelah dilakukan *tuning* dan skenario-skenario yang ditentukan, akan di evaluasi dengan confussion matrix untuk mencari akurasi, presisi, *recall* dan f1-score.

### 3.6.1 Hyperparameter Tuning

*Hyperparameter tuning* digunakan untuk membangun struktur model dan tidak dapat dipelajari dari data, dan nilainya diatur sebelum proses pembelajaran dimulai. Oleh karena itu, hiperparameter mirip dengan pengaturan algoritma yang dapat disesuaikan untuk mengoptimalkan kinerja dan mencegah overfitting. Penelitian mencoba memilih satu set hiperparameter optimal untuk algoritma pembelajaran untuk meningkatkan kinerja model. Ada dua metode yang sering digunakan untuk melakukan penyetelan hiperparameter yang disebut 1) Grid Search dan 2) Random Search[40].

### 3.6.2 Grid Search

Grid search adalah metode tradisional untuk melakukan penyetelan hiperparameter. 3.2 Ilustrasi Grid Search Cross Validation menjelaskan metode ini bekerja dengan mendefinisikan subset nilai kandidat untuk setiap hiperparameter, dan melatih semua kombinasi yang mungkin dari hiperparameter tersebut. Kemudian, setiap model yang mungkin dilatih dievaluasi pada set validasi, dan konfigurasi terbaik dari hiperparameter akan dipilih pada akhirnya.




Gambar 3.2. Ilustrasi Grid Search Cross Validation

sumber: [40]

### 3.6.3 Early Stopping

*Grid search* merupakan cara *tuning hyperparameter* dengan melatih model LSTM untuk setiap kombinasi hiperparameter. *Grid search* memerlukan banyak waktu untuk melatih semua model dan memilih kombinasi hiperparameter terbaik. Salah satu cara untuk mencegah hal ini adalah dengan menggunakan Early Stopping dan Callbacks. Ide dari Early Stopping adalah melacak suatu ukuran (seperti loss validasi) dan kapan pun kriteria penghentian (seperti tidak ada peningkatan dalam nilai ukuran yang dipantau dalam beberapa langkah berturut-turut, mencapai batas yang telah ditentukan sebelumnya untuk ukuran tersebut, atau peningkatan yang telah ditentukan sebelumnya dalam ukuran tersebut) terpenuhi, dapat menghentikan proses pelatihan. Ukuran yang pada penelitian ini digunakan loss validasi karena set validasi tidak digunakan dalam proses pelatihan[41].

```
es = EarlyStopping(monitor='val_loss', mode='min', verbose=1,
    patience=5)
```

Kode 3.1: Contoh Kode Early Stopping

Seringkali, tanda pertama dari tidak ada lagi peningkatan mungkin bukan waktu terbaik untuk menghentikan pelatihan. Ini karena model mungkin mengalami fase stabil tanpa peningkatan atau bahkan sedikit memburuk sebelum menjadi jauh lebih baik. *Early sstopping* dapat mengantisipasi hal ini dengan menambahkan penundaan pada pemicu dalam hal jumlah epoch tidak melihat peningkatan. Ini dapat dilakukan dengan mengatur argumen ”patience”.

Setelah mendapat model dengan tingkat akurasi tertinggi, model akan diuji pada data yang telah diverifikasi dan label oleh ahli. Dari uji model tersebut akan dilakukan proses evaluasi dan pembahasan pada hasil dari kinerja model.