BAB III

METODOLOGI PENELITIAN

3.1 Gambaran Umum Objek Penelitian

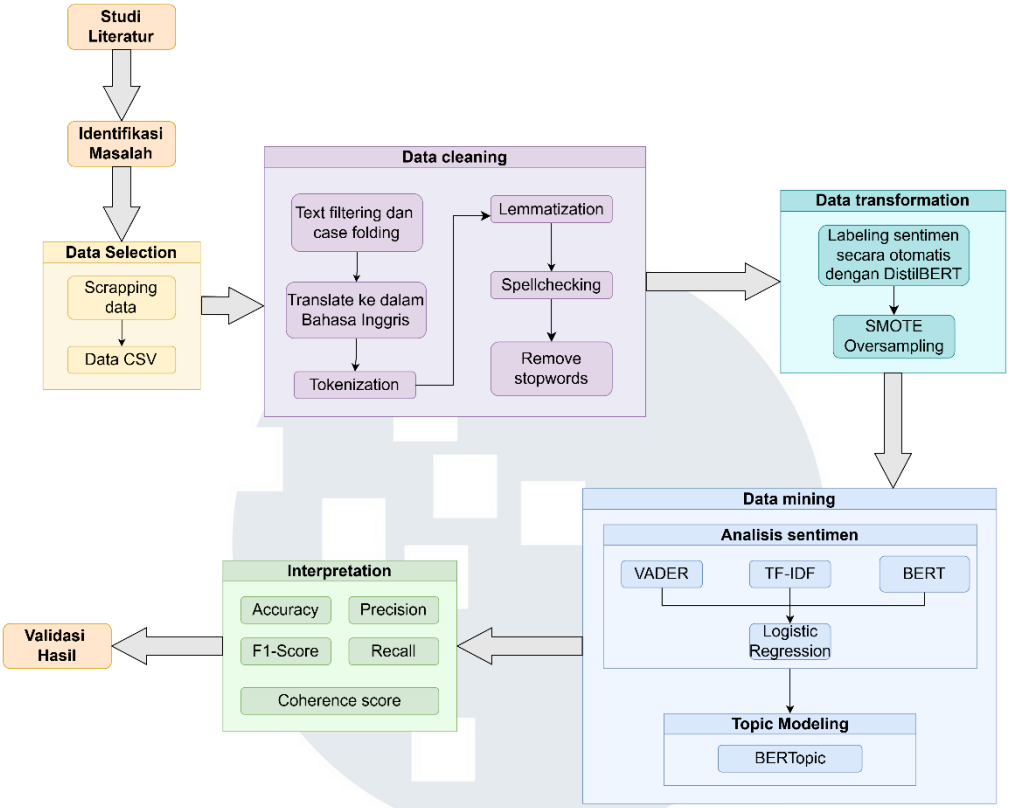
Penelitian ini menganalisis data berupa *tweets* dari media sosial X sebagai objek utama untuk memahami dinamika respon dan opini publik. *Platform* X memiliki volume interaksi pengguna yang tinggi karena sifatnya yang *micro-blogging* sehingga memungkinkan eksplorasi yang lebih mendalam. Data diambil dari seluruh *tweet* berbahasa Indonesia dengan kata kunci "Coretax" yang diunggah pada periode 1 Januari 2025 hingga 28 Februari 2025. Periode tersebut menjadi pilihan yang tepat karena relevansinya dalam merepresentasikan reaksi masyarakat dua bulan pertama pasca peluncuran Coretax. *Dataset* yang telah dikumpulkan terdiri dari 23.427 baris dan 15 atribut. Tabel 3.1 menjelaskan informasi lebih lanjut mengenai informasi atribut pada *dataset* dari media sosial X.

Tabel 3.1 Deskripsi atribut *dataset* objek penelitian

Atribut	Tipe data	Deskripsi
conversation_id_str	string	ID *tweet* utama yang memulai percakapan
created_at	timestamp	Tanggal penulisan *tweet*
favorite_count	integer	Jumlah *like* pada *tweet*
full_text	string	Isi konten *tweet*
id_str	string	ID *tweet* individual
image_url	string	URL gambar yang dilampirkan pada *tweet*
in_reply_to_screen_name	string	*Mention* pertama atau *username* pemilik tweet
lang	string	Bahasa yang digunakan pada *tweet*
quote_count	integer	Jumlah pengguna yang mengutip *tweet*
reply_count	integer	Jumlah komentar pada *tweet*
retweet_count	integer	Jumlah *retweet*
tweet_url	string	URL *tweet*
user_id_str	string	ID unik pengguna
username	string	*Username* yang dibuat pengguna

3.2 Metode Penelitian

3.2.1 Alur Penelitian



Gambar 3.1 Alur penelitian

Penelitian ini mengimplementasikan kerangka kerja KDD sebagai landasan metodologi untuk memastikan proses analisis data yang terstruktur dan berorientasi tujuan. KDD dipilih karena pendekatan iteratif yang memfasilitasi adaptasi terhadap data *tweet* dengan volume besar. Selain itu penelitian ini bertujuan untuk mendapatkan pemahaman baru yang berdasarkan dari data *tweets* berupa sentimen dan topik. Dataset yang dikumpulkan cenderung tidak beraturan dan terdiri dari *tweets* yang cukup beragam sehingga proses *mining* ditekankan pada tahap *cleaning* dan *transformation* untuk menghindari bias. Metode *data mining* KDD pada penelitian ini terdiri lima tahap pada Gambar 3.1 yang diuraikan sebagai berikut.

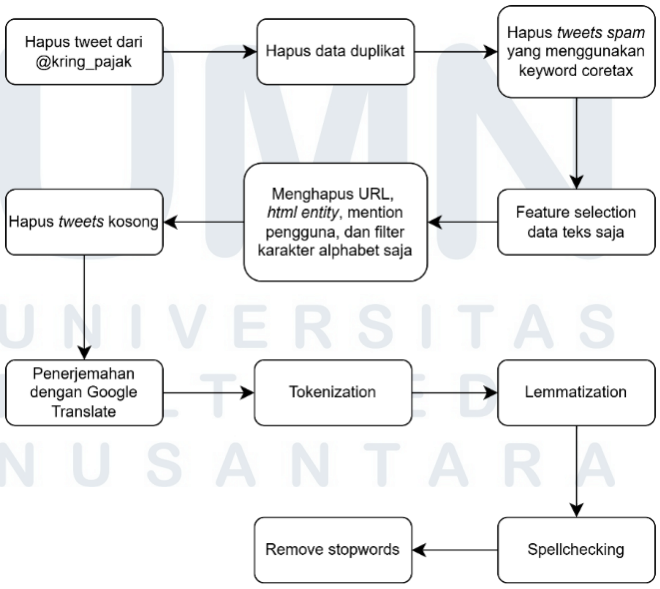
1. *Data Selection*

Tahap ini berfokus pada ekstraksi data dari sumber yang relevan untuk proses analisis. Salah satu tujuan penelitian ini untuk mengidentifikasi sentimen dan topik yang dibahas mengenai Coretax pada media sosial X. Dengan pertimbangan tersebut maka dibutuhkan data *tweets* yang diunggah dengan kata kunci "Coretax." Untuk menjaga relevansi, data yang diambil hanya dalam rentang 1 Januari hingga 28 Februari 2025. Data yang berhasil dikumpulkan kemudian disimpan dalam format CSV. Contoh *tweets* dapat dilihat pada Tabel 3.2.

Tabel 3.2 Hasil *data selection*

Tweets yang membahas Coretax
fitur preview faktur pajak tuh sangat2 membantu banget tauuu. plis tambahin di Coretax min @kring_pajak
Ternyata Coretax punya misi terselubung membuat orang-orang akunting jadi jago javascript. Inilah industri 4.0 yang sebenarnya. Kalo di manufaktur belum minimal buat industri keuangan dulu lah ya
Coretax udh ditahap bikin konsultan kantor w kyk gini terakhir dia sampe ngomong cabut aja tuhannnn nyawa saya Coretax bikin gila TAPI GUE YANG INPUT AJA BENERAN GILA JADINYA

2. *Data Cleaning*



Gambar 3.2 Alur *data cleaning*

Data cleaning merupakan tahap paling penting dalam alur KDD. Tahap ini mengidentifikasi dan memperbaiki *error* pada data untuk memastikan data yang akurat, utuh, dan konsisten. Gambar 3.2 menunjukkan alur proses melakukan *data cleaning* yang diuraikan sebagai berikut.

a. Menghapus *tweets* dari @kring_pajak (*customer service* layanan perpajakan) untuk menjaga hasil sentimen dan ekstraksi topik dalam merepresentasikan persepsi publik.
b. Menghapus data *tweets* duplikat untuk mengurangi redundansi.
c. Menghapus *tweets* tidak relevan terkait Coretax atau spam yang menggunakan *keyword* "Coretax." Proses ini dilakukan secara otomatis menggunakan *rule* berupa kata-kata yang terindikasi spam seperti pada Lampiran F.
d. *Feature selection* dengan memilih atribut *full_text* sebagai basis analisis sentimen dan pemodelan topik dalam penelitian ini.
e. *Text filtering* untuk menghapus URL, HTML entity, *mention* pengguna untuk memastikan atribut *full_text* hanya terdiri dari karakter alfabet saja.
f. Menghapus *tweets* kosong umumnya karena *tweet* hanya berisi gambar atau video tanpa *caption*.
g. Penerjemahan *tweets* dengan Google Translate melalui *library* deep-translator karena leksikon VADER dioptimalkan dalam bahasa Inggris.
h. *Tokenization* dengan fungsi split.
i. *Lemmatization* dengan POS tagging WordNetLemmatizer pada *library* NLTK.
j. *Spellchecking* dengan *library* Hunspell untuk memastikan kata-kata dari proses *lemmatization* tetap sesuai.
k. *Remove stopwords* untuk menghilangkan kata-kata tanpa makna yang serung muncul dengan menggunakan *library* NLTK.

3. *Data Transformation*

Data Transformation tidak kalah pentingnya dengan *data cleaning* karena perannya untuk menyiapkan data bersih tersebut ke dalam struktur yang sesuai untuk kebutuhan analisis. Pada penelitian ini, transformasi diawali dengan pelabelan sentimen otomatis dengan model DistilBERT yang dapat menangkap hubungan kontekstual dalam sebuah teks. Pemilihan metode ini didasari dengan performa model tersebut yang sangat tinggi pada penelitian [22]. Proses pelabelan sentimen dilakukan melalui *library* Transformers yang tersedia pada *platform* Hugging Face dengan memetakan kelas sentimen pada positif dan negatif dengan model distilbert-base-uncased-finetuned-sst-2-english untuk bahasa Inggris. Parameter yang digunakan dalam model ini dicantumkan pada Tabel 3.3 untuk menghasilkan label sentimen dan skor probabilitas terhadap label sentimen tersebut.

Tabel 3.3 Parameter dalam model DistilBERT

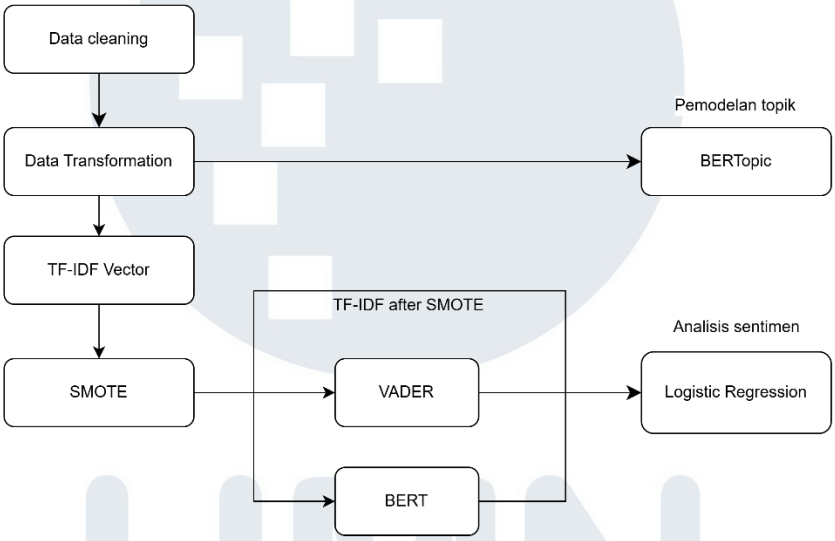
Parameter	***Nilai***
vocab_size	30.522
max_position_embeddings	512
n_layers	6
n_heads	12
dim	768
hidden_dim	3.072
dropout	0,1
attention_dropout	0,1

Hasil pelabelan otomatis dengan DistilBERT menghasilkan data yang siap untuk dianalisis lebih lanjut, namun seringkali dihadapkan pada masalah ketidakseimbangan kelas. Untuk mengatasi distribusi kelas yang tidak merata, diterapkan teknik SMOTE untuk menghasilkan sampel sintetik pada kelas minoritas melalui interpolasi linier antar *k-nearest neighbors* dalam *feature space*. Penelitian ini menggunakan metodologi Kristiyanti et al. [50] dengan menerapkan SMOTE sebelum membagi data ke dalam *training* dan *testing* untuk mencapai performa paling tinggi.

Dengan teknik tersebut, jumlah sampel data pada kelas sentimen positif dan negatif menjadi seimbang sehingga siap digunakan untuk tahap berikutnya.

4. *Data Mining*

Data mining merupakan inti dari seluruh proses dalam KDD. Tahap ini bertujuan ekstraksi pola dan *insight* dari data yang telah diproses sebelumnya. Hasil yang diharapkan pada *data mining* ialah *output* model yang digunakan untuk menganlisis data. Pada konteks penelitian ini, *data mining* mencakup dua tugas, diantaranya analisis sentimen dan pemodelan topik dengan alur yang ditunjukkan pada Gambar 3.3.



Gambar 3.3 Alur analisis sentimen dan pemodelan topik dalam *data mining*

Analisis sentimen dilakukan dengan menggunakan model *hybrid* VADER, TF-IDF, BERT dan Logistic Regression. VADER menyediakan skor sentimen berbasis leksikon, TF-IDF menghitung bobot kata dalam *corpus*, dan BERT untuk memproses representasi semantik dengan arsitektur *transformer*. Hasil ketiganya kemudian dijadikan sebagai *input vector* dalam Logistic Regression. Pengujian model *hybrid* dilakukan dalam dua skenario yaitu dengan SMOTE dan tanpa SMOTE. Selain itu,

penelitian ini membandingkan empat *splitting ratio* dalam *classifier* Logistic Regression yaitu 60:40, 70:30, 80:20, dan 90:10. Metode pengujian ini dilakukan untuk membandingkan performa model *hybrid* dalam melakukan klasifikasi sentimen dalam konteks Coretax.

Pemodelan topik menggunakan BERTopic mengelompokkan *tweets* ke dalam topik berdasarkan kesamaan semantik kata-kata pada data. BERTopic memanfaatkan *sentence transformers* untuk menghasilkan *text embedding*, dilanjutkan dengan *dimension reduction* menggunakan UMAP dan *clustering* dengan HDBSCAN. Pendekatan dengan metode ini memfasilitasi identifikasi topik tanpa menentukan jumlah *cluster*. Hasil analisisis sentimen dan pemodelan topik diintegrasikan untuk mengetahui sentimen yang ada terhadap Coretax dan topik yang dinbahas pada masing-masing sentimen.

5. *Interpretation/Evaluation*

Setelah model berhasil melakukan klasifikasi sentimen, selanjutnya dilakukan evaluasi dengan confusion matrix. Hasil ini mempertimbangkan metrik seperti *accuracy*, *recall*, *precision*, dan *f1-score*. Sedangkan pada pemodelan topik, evaluasi dilakukan dengan metrik c_v *coherence score*.

3.3 Variabel Penelitian

Dalam penelitian ini terdapat dua jenis variabel yaitu variabel independen dan variabel dependen. Variabel independen merupakan variabel bebas yang dapat mempengaruhi variabel dependen. Variabel dependen merupakan variabel terikat yang dipengaruhi oleh variabel independen. Kedua variabel ini terdapat pada masing-masing analisis sentimen dan pemodelan topik.

3.3.1 Analisis Sentimen

Variabel independen dalam analisis sentimen yaitu data *tweets* di X yang membahas sistem Coretax. Data tersebut mencakup teks *tweet* yang telah melalui tahap *preprocessing*. Sedangkan variabel dependen dalam analisis

sentimen merupakan hasil *labeling* oleh VADER. *Label* sentimen yang diberikan diantaranya positif dan negatif.

3.3.2 Pemodelan topik

Variabel independen dalam pemodelan topik yaitu data *tweets* X yang telah diberikan label oleh metode DistilBERT. Data tersebut mencakup teks *tweet* maupun dalam bentuk token dan label sentimennya. Sedangkan variabel dependen pada pemodelan topik merupakan topik hasil ekstraksi pada masing-masing sentimen.

3.4 Teknik Pengumpulan Data

Pengumpulan data dilakukan dengan teknik *scraping*. *Scraping* merupakan proses pengumpulan data dari sumber tertentu secara otomatis. Saat ini API Developer pada media sosial X versi tidak berbayar sudah tidak mendukung *scraping* data. Data dikumpulkan dengan bantuan *command-line tool* bernama Tweet Harvest yang dikembangkan oleh Helmi Satria.

3.4.1 Tweet Harvest

Command-line tool ini mengadopsi pendekatan *browser automation* melalui *framework* Playwright untuk meniru perilaku manusia dalam mengakses sebuah situs web. Tweet Harvest menggunakan auth token pada cookies untuk mengekstraksi data dari *user interface* web X [51]. Cara kerja dari *tool* ini terdiri dalam tiga tahap. Tahap pertama dimulai dengan menginisialisasi parameter *query* untuk melakukan pencarian di X. Sintaks *query* ini menyesuaikan dokumentasi *platform* X seperti kata kunci, tagar, rentang waktu, atau *username* spesifik. Tahap kedua dilanjutkan mengatur batas maksimum *tweet* yang diambil. Tahap ketiga dilakukan dengan menjalankan *tools* ini melakukan *crawling* hingga data dapat di-*import* ke dalam bentuk CSV.

3.5 Teknik Analisis Data

Penelitian analisis sentimen dan pemodelan topik terkait sistem Coretax di media sosial X menggunakan bahasa pemrograman Python. Keputusan tersebut

diambil atas dasar pertimbangan implementasi dan ketersediaan *library* dengan cakupan luas dan beragam dalam memproses NLP dan analisis data. *Library* seperti Pandas dan NLTK digunakan untuk memfasilitasi tahap *preprocessing* hingga proses *modeling*. Proses analisis ini dilakukan pada *platform* Google Colab yang menyediakan fitur Python Notebook dengan *runtime* secara *cloud-based* Google. Keunggulan utama Google Colab terletak pada kemampuannya menjalankan kode dengan performa tinggi. Penggunaan IDE tersebut dapat mendukung proses penelitian tidak bergantung pada performa komputasi pada *local environment*.