BAB III

METODOLOGI PENELITIAN

3.1 Gambaran Umum Objek Penelitian

Data pengaduan dari fitur LAKSA (Layanan Aspirasi Kotak Saran Anda) dalam aplikasi Tangerang LIVE menjadi objek utama penelitian ini. Tangerang LIVE merupakan *super-app* yang dikembangkan oleh pemerintah Kota Tangerang sebagai wujud dukungan akan implementasi *smart city*. Tidak hanya sebagai dukungan, aplikasi Tangerang LIVE sangat bermanfaat bagi masyarakat karena mengintegrasikan berbagai layanan publik digital ke dalam satu aplikasi layaknya portal akses yang memudahkan masyarakat dalam mengurus banyak administrasi.

Fitur LAKSA (Layanan Aspirasi Kotak Saran Anda) merupakan jembatan komunikasi warga Kota Tangerang dan Pemerintah Kota Tangerang yang memfasilitasi pelaporan keluhan, penyampaian aspirasi, dan permintaan layanan darurat terkait infrastruktur dan fasilitas publik. Berdasarkan kegunaannya, fitur LAKSA menghasilkan data berupa teks tidak terstruktur yang mencakup berbagai kategori permasalahan, mulai dari kerusakan jalan, penerangan jalan umum, tumpukan sampah, dan pelayanan administrasi kependudukan. Tingginya volume interaksi masyarakat pada fitur ini menjadikannya sumber data yang kaya yang berujung menimbulkan asumsi internal bahwa sebagian besar isi laporan pada fitur LAKSA termasuk dalam kategori negatif tanpa adanya bukti valid.

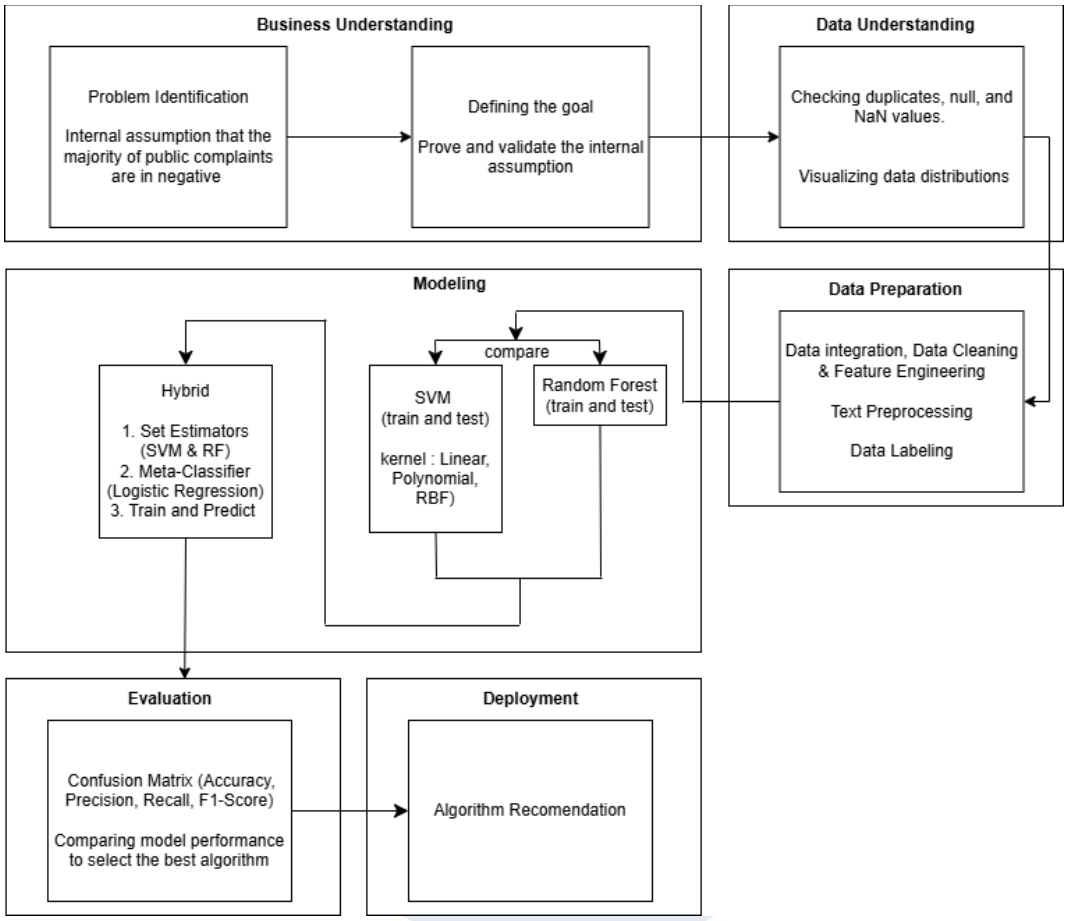
Dalam penelitian analisis sentimen yang kini dilakukan, data yang digunakan merupakan data primer yang diperoleh secara resmi melalui kerja sama dengan Dinas Komunikasi dan Informatika (Diskominfo) Kota Tangerang. Cakupan data penelitian dibatasi pada laporan pengaduan yang masuk selama periode empat tahun terakhir yakni mulai dari tahun 2021 hingga tahun 2024. Pemilihan rentang waktu dilakukan untuk memastikan tren isu yang dikumpulkan bukan merupakan isu yang sudah terbelakang dan memastikan volume data cukup representatif untuk melatih dan menguji model *machine learning*. Fokus analisis pada objek ini adalah membuktikan asumsi sebagian besar laporan pengaduan yang masuk berkategori negatif.

3.2 Metode Penelitian

Alur penelitian ini dirancang secara sistematis dengan mengadopsi metode CRISP-DM (Cross-Industry Standard Process for Data Mining). CRISP-DM menyediakan panduan langkah demi langkah yang terstruktur untuk menjalankan penelitian analisis sentimen ini, mulai dari definisi masalah hingga evaluasi solusi. Proses penelitian dengan CRISP-DM diawali dengan tahap business understanding untuk mengidentifikasi kebutuhan Diskominfo Kota Tangerang dalam menganalisis pengaduan masyarakat secara objektif. Selanjutnya, tahap data understanding dan data preparation yang mencakup pengumpulan dan pemahaman data pengaduan dari fitur LAKSA aplikasi Tangerang Live dan dilanjutkan dengan proses text preprocessing untuk membersihkan dan mengubah data teks mentah menjadi format yang siap dianalisis oleh model. Pada tahap modeling, penelitian ini akan membangun tiga arsitektur model yang berbeda: model murni *Random Forest*, model murni *Support Vector Machine* (SVM), dan sebuah model hibrida (hybrid model) yang dirancang untuk mengkombinasikan kekuatan dari kedua algoritma tersebut. Kemudian, ketiga model tersebut (RF murni, SVM murni, dan model hibrida) akan dibandingkan pada tahap evaluation. Kinerja setiap model akan dievaluasi secara ketat menggunakan metrik seperti akurasi, presisi, dan *recall* untuk menentukan arsitektur model mana yang kinerjanya paling unggul dan optimal untuk dataset ini. Tahap terakhir, *deployment*, dalam konteks penelitian ini diwujudkan dalam bentuk rekomendasi arsitektur model terbaik kepada Diskominfo. Dengan pendekatan CRISP-DM yang terstruktur ini, penelitian diharapkan dapat menghasilkan analisis yang valid dan solusi model klasifikasi paling akurat untuk meningkatkan kualitas layanan publik..

3.2.1 Alur Penelitian

Penelitian yang dilakukan dalam beberapa tahapan dibuat dan dibentuk kedalam sebuah alur penelitian untuk mempermudah proses penelitian yang dibuat sebagai berikut :



Gambar 3.1 Alur Penelitian Sentimen Analisis

Alur penelitian ini dirancang secara sistematis dengan mengadopsi metode CRISP-DM (Cross-Industry Standard Process for Data Mining). Kerangka kerja ini menyediakan panduan langkah demi langkah yang terstruktur untuk menjalankan proyek analisis sentimen ini, mulai dari definisi masalah hingga evaluasi solusi. Berikut adalah penjabaran detail dari setiap fase yang diadaptasi untuk penelitian ini, sesuai dengan alur pada Gambar 3.1

Diawali dengan fase *Business Understanding* yang bertujuan untuk menyelaraskan aspek teknis data mining dengan kebutuhan strategis instansi pengelola. Pada tahap *Problem Identification*, penelitian berfokus pada adanya asumsi internal di lingkungan Diskominfo Kota Tangerang yang menyatakan bahwa mayoritas pengaduan masyarakat pada fitur LAKSA cenderung bersifat negatif. Namun, karena asumsi tersebut belum didukung oleh bukti empiris yang terukur, maka langkah selanjutnya adalah *Defining the Goal*. Tujuan

utama yang ditetapkan adalah untuk membuktikan dan memvalidasi asumsi internal tersebut secara objektif. Melalui klasifikasi sentimen, penelitian ini berupaya memberikan landasan data yang valid bagi pemerintah kota dalam mengevaluasi efektivitas layanan publik dan merespons kebutuhan masyarakat secara lebih akurat.

Setelah tujuan ditetapkan, dilakukan tahap *Data Understanding* untuk memahami karakteristik dan kualitas dataset yang akan digunakan. Fase ini melibatkan proses *Checking duplicates, null, and NaN values* untuk mengidentifikasi keberadaan data yang tidak lengkap atau berulang yang berpotensi menimbulkan bias pada hasil klasifikasi. Selanjutnya, dilakukan *Visualizing data distributions* guna melihat profil data secara makro, seperti tren jumlah pengaduan per tahun yang menunjukkan penurunan bertahap dari tahun 2021 hingga 2024. Eksplorasi awal ini sangat penting sebagai fondasi dasar sebelum data memasuki tahap pemrosesan yang lebih kompleks, guna memastikan bahwa data yang digunakan benar-benar representatif terhadap kondisi lapangan.

Tahap Data Preparation merupakan fase yang paling intensif karena melibatkan transformasi data mentah menjadi format yang siap untuk diolah oleh mesin. Proses ini diawali dengan *Data Integration*, *Data Cleaning*, dan *Feature Engineering* untuk merapikan struktur data serta menghapus informasi yang tidak relevan. Kemudian, dilakukan *Text Preprocessing* yang meliputi pembersihan karakter, *tokenization*, hingga *stemming* untuk mereduksi kata menjadi bentuk dasarnya. Setelah teks bersih, dilakukan proses Data Labeling. Sebagaimana dilakukan dalam jurnal pertama penelitian terdahulu, proses pelabelan ini dilaksanakan secara manual melalui validasi pegawai pemerintahan untuk memastikan setiap data terklasifikasi ke dalam kategori yang tepat sesuai dengan konteks aslinya. Pendekatan manual ini krusial untuk menghasilkan *ground truth* yang berkualitas tinggi, sehingga model dapat mempelajari pola data secara lebih presisi dibandingkan hanya mengandalkan deteksi otomatis yang rentan terhadap ambiguitas bahasa.

Fase *Modeling* dalam penelitian ini merupakan tahap krusial yang mengintegrasikan pendekatan pembelajaran mesin tunggal dan metode ansambel (*ensemble learning*) untuk mencapai performansi klasifikasi yang optimal. Strategi pertama dimulai dengan membangun model tunggal sebagai *baseline*, yakni *Support Vector Machine* (SVM) dan *Random Forest*. Pada pengembangan SVM, peneliti melakukan eksperimen menggunakan tiga variasi *kernel* utama, yaitu *Linear*, *Polynomial*, dan *RBF*, untuk menguji kompleksitas hubungan antar data. Penggunaan ketiga kernel ini memungkinkan membandingkan efektivitas pemisahan data, mulai dari pendekatan linear yang sederhana hingga transformasi non-linear yang lebih kompleks, guna memastikan model mampu menangkap pola pengaduan masyarakat yang beragam baik dalam struktur teks yang sederhana maupun yang memiliki keterikatan semantik yang rumit. Secara paralel, algoritma *Random Forest* dilatih untuk menangani variansi data melalui sekumpulan pohon keputusan yang bekerja secara kolektif dalam menentukan label klasifikasi. Pada model ini, parameter *n_estimators* ditetapkan sebesar 100, yang merujuk pada standar praktik umum dalam literatur penelitian terdahulu di mana angka tersebut dianggap sebagai nilai estimator utama yang mampu menyeimbangkan performansi model dengan efisiensi komputasi.

Kebaharuan teknis utama dalam penelitian ini terletak pada perancangan arsitektur Hybrid Stacking yang menggabungkan kekuatan prediktif dari kedua algoritma tersebut. Arsitektur ini disusun dalam dua tingkatan (*level*): pada Level 0 (*Base-Estimators*), SVM dan *Random Forest* dilatih secara independen untuk mengenali pola-pola bahasa dalam teks pengaduan LAKSA. Hasil prediksi (berupa probabilitas atau label) dari kedua model dasar ini kemudian dikumpulkan dan digunakan sebagai fitur masukan baru untuk Level 1 (*Meta-Classifier*). Pada tingkatan kedua ini, digunakan algoritma *Logistic Regression* sebagai *meta-learner* yang bertugas mempelajari bobot kontribusi dari masing-masing *base-estimator*.

Penggunaan *Logistic Regression* sebagai *meta-classifier* bertujuan untuk meminimalkan risiko *overfitting* dan melakukan koreksi terhadap kesalahan

prediksi yang mungkin dilakukan oleh SVM atau *Random Forest* secara individual. Dengan mekanisme *Stacking* ini, model Hybrid tidak hanya sekadar mengambil keputusan mayoritas, melainkan secara cerdas mengombinasikan keunggulan SVM dalam menangani data teks berdimensi tinggi dan kemampuan *Random Forest* dalam menjaga stabilitas prediksi. Integrasi teknis ini diharapkan mampu menghasilkan model klasifikasi yang lebih superior dan tangguh (*robust*) dalam memvalidasi polaritas sentimen masyarakat secara akurat.

Pada tahap *Evaluation*, seluruh model yang telah dibangun diuji performanya menggunakan data uji yang tidak pernah dilihat sebelumnya oleh model. Penilaian dilakukan secara kuantitatif melalui instrumen *Confusion Matrix* yang menghasilkan parameter ukur berupa *Accuracy, Precision, Recall, dan F1-Score*. Melalui parameter tersebut, peneliti melakukan *Comparing model performance* untuk membedah keunggulan dan kelemahan masing-masing algoritma, terutama dalam menangani data yang ambigu atau kelas yang tidak seimbang. Evaluasi ini memastikan bahwa algoritma yang terpilih nantinya bukan hanya unggul secara angka akurasi, tetapi juga handal dalam mengenali sentimen negatif dan positif secara objektif sesuai dengan kebenaran label (*ground truth*).

Tahap akhir dari kerangka kerja CRISP-DM adalah *Deployment*, yang dalam konteks penelitian ini diwujudkan melalui *Algorithm Recommendation*. Hasil evaluasi yang menunjukkan performa terbaik diserahkan kepada pihak Diskominfo Kota Tangerang sebagai rekomendasi teknis solusi klasifikasi otomatis. Penyerahan model ini berfungsi sebagai alat pendukung keputusan bagi pemerintah dalam memonitor persepsi publik secara *real-time*. Dengan tersedianya model yang telah divalidasi, pihak instansi dapat melakukan pemetaan isu secara lebih cepat, menjawab asumsi internal yang selama ini ada, dan menjadikannya dasar dalam perumusan kebijakan pelayanan publik di masa mendatang.

3.3 Teknik Pengumpulan Data

Data yang digunakan dalam penelitian merupakan data primer dalam bentuk teks laporan pengaduan. Teks laporan pengaduan yang digunakan merupakan data laporan pengaduan yang masuk ke fitur Layanan Aspirasi Kotak Saran Anda (LAKSA) pada aplikasi Tangerang LIVE dan diperoleh secara langsung dari Dinas Komunikasi dan Informatika (Diskominfo) Kota Tangerang selaku mitra dalam penelitian yang dilakukan. Proses pengumpulan data dilakukan dengan bantuan dan koordinasi langsung dari pihak Dinas Komunikasi dan Informatika (Diskominfo) Kota Tangerang sehingga peran dan keterlibatan Diskominfo bersifat krusial. Hal ini dikarenakan data pengaduan pada fitur LAKSA merupakan data internal Pemerintah Kota Tangerang yang tersimpan dalam *database* dan memiliki akses terbatas dimana hanya bisa diakses oleh pegawai berwenang. Dalam tahap pengumpulan data, pihak Diskominfo membantu melakukan ekstraksi data mentah dari *database*.

Data mentah yang dibantu ekstrak pihak Diskominfo diberikan dalam file dengan format XLSX. Data mentah yang dimanfaatkan dalam penelitian terhitung sejak 1 Januari 2021 sampai dengan 30 November 2024 dengan total 10.586 laporan pengaduan. Penelitian yang dilakukan menggunakan data sensus yakni keseluruhan total 10.586 pengaduan akan digunakan dan dianalisis. Penggunaan keseluruhan data bertujuan untuk memastikan analisis sentimen dan perbandingan model memiliki representasi data yang paling lengkap dan komprehensif. Hal tersebut juga meminimalisir adanya *sampling bias* dan memaksimalkan potensi penghasilan wawasan akurat mengenai sentimen publik yang masuk.

3.4 Teknik Analisis Data

Penelitian analisis sentimen yang dilakukan menerapkan beberapa tahapan analisis data yang sistematis, bertujuan untuk mengolah data mentah dari fitur LAKSA, membangun, membandingkan, dan mengevaluasi berbagai arsitektur model klasifikasi sentimen, hingga menentukan arsitektur model paling optimal sebagai rekomendasi. Prosedur analisis data yang dirancang untuk mendukung tujuan penelitian adalah sebagai berikut :

3.4.1 *Data Preparation* dan *Data Processing*

Tahap pertama teknik analisis data merupakan fondasi dari penelitian dan bertujuan untuk mengubah data mentah menjadi data yang bersih dan siap untuk pemodelan. Proses *data preparation dan preprocessing* diawali dengan Persiapan Data yang mencakup eksplorasi awal untuk memahami struktur data (*info, head*), konversi tipe data, penghapusan kolom yang tidak relevan ('id_user', 'nik', 'nama', 'jenis_aduan'), penanganan data duplikat dan nilai yang hilang (drop NaN), serta pembersihan data anomali seperti tanggal yang tidak logis. Pada tahap ini, dilakukan juga Rekayasa Fitur (*Feature Engineering*) dengan membuat kolom baru 'durasi_proses' untuk menghitung waktu respons pengaduan.

3.4.2 Text Preprocessing

Tahap *text preprocessing* merupakan tahapan krusial pada kolom 'isi_pengaduan'. Tahapan kedua meliputi *text standardization* (pembersihan tanda baca dan *pattern* via regex), *lowercasing* (semua huruf pada kolom isi_pengaduan sama), *word tokenization* (pengubahan teks menjadi daftar kata), *stopword removal* (penghapusan kata umum maupun kustomisasi), serta *lemmatization* dan *stemming* untuk menormalisasi kata ke bentuk dasarnya. Terakhir, data yang bersih ini akan melalui proses *Pseudo-Labeling* untuk memberikan label sentimen (positif, negatif, dan netral) secara terprogram sebagai input untuk model *supervised learning*.

3.4.3 Modeling

Data yang telah melalui tahap pra-pemrosesan teks kemudian akan ditransformasi menjadi representasi vektor numerik menggunakan metode pembobotan kata TF-IDF (Term Frequency-Inverse Document Frequency). Sesuai dengan tujuan penelitian, akan dibangun dan dibandingkan beberapa arsitektur model.

1. Dilakukan perbandingan model murni. Model *Random Forest* (RF) akan dibangun. Secara terpisah, model *Support Vector Machine* (SVM) akan

diuji menggunakan tiga *kernel* yang berbeda (Linear, Polynomial, dan RBF) untuk menemukan konfigurasi SVM yang paling optimal.

2. Sebuah model hibrida (hybrid model) akan dirancang dan dibangun. Model ini akan mengkombinasikan arsitektur SVM (menggunakan kernel terbaik dari hasil perbandingan sebelumnya) dan *Random Forest*, dengan tujuan menguji apakah pendekatan gabungan ini mampu melampaui kinerja model murni.

Seluruh proses *training* model ini dilakukan dalam bahasa pemrograman Python dengan memanfaatkan pustaka (*library*) Scikit-learn.

3.4.4 Evaluasi Model

Seluruh arsitektur model yang telah dilatih (varian *kernel* SVM, RF murni, dan model hibrida) kemudian akan dievaluasi kinerjanya secara ketat menggunakan data uji (*test set*). Analisis performa akan didasarkan pada metrik-metrik standar yang berasal dari *Confusion Matrix*, yang mencakup *Accuracy* (Akurasi), *Precision* (Presisi), *Recall* (Perolehan), dan *F1-Score*. Hasil evaluasi kuantitatif ini akan menjadi dasar utama untuk perbandingan objektif, yang bertujuan menentukan arsitektur model (murni atau hibrida) mana yang paling unggul, akurat, dan andal dalam mengklasifikasikan sentimen pada data pengaduan fitur LAKSA.