BAB 3
METODE PENELITIAN

3.1 BAHAN DAN ALAT PENELITIAN

Bahan penelitian ini memanfaatkan data ulasan yang diambil dari *platform* Google Maps di RSU X. Data yang digunakan mencakup seluruh ulasan yang tersedia, dengan memastikan hanya ulasan yang memiliki teks (tidak kosong atau hanya berisi rating) yang diambil untuk analisis. Proses pengambilan data dilakukan menggunakan teknik *scraping* untuk mengumpulkan data ulasan dari kedua rumah sakit tersebut. *Scraping* dilakukan dengan memanfaatkan API SerpAPI untuk pengambilan hasil pencarian dari Google Maps secara otomatis. Dengan menggunakan bahasa pemrograman Python, data ulasan dikumpulkan dan disimpan untuk analisis lebih lanjut.

Alat penelitian yang digunakan adalah laptop dengan spesifikasi yang mampu mendukung pemrosesan data dan memiliki koneksi internet. Berikut adalah sistem operasi, program, aplikasi, dan *library* yang akan digunakan dalam proses penelitian ini:

1. *Operation System* Windows 10 64-bit.
2. Bahasa pemrograman Python dengan versi 3.9 keatas.
3. Google Colaboratory.
4. Visual Studio Code versi 1.72.0.
5. *Library* Python yang digunakan:
    a) Pandas versi 2.2.2 atau versi terbaru.
    b) NumPy versi 1.26.4 atau versi terbaru.
    c) Scikit-learn versi 1.6.0 atau versi terbaru.
    d) Matplotlib versi 3.8.0 atau versi terbaru.
    e) Seaborn versi 0.13.2 atau versi terbaru.
    f) Plotly versi 4.9 atau versi terbaru.
    g) Joblib versi 1.4.2 atau versi terbaru.
    h) NLTK versi 3.9.1 atau versi terbaru.

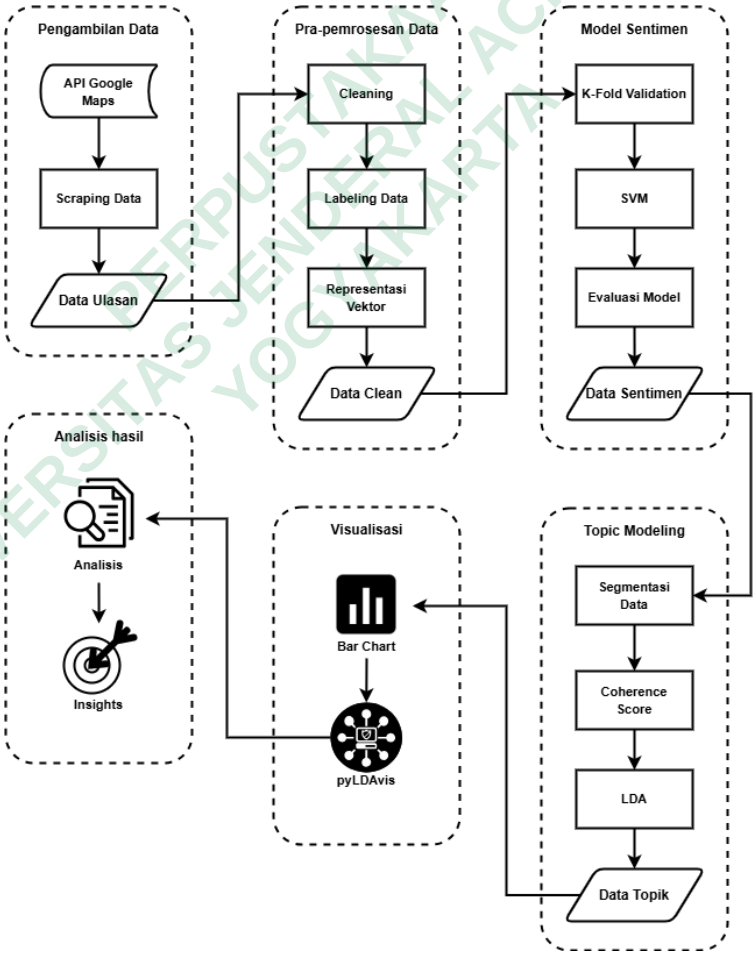
i) Sastrawi versi 1.0.1 atau versi terbaru.

j) Gensim versi 4.3.3 atau versi terbaru.

k) pyLDAvis versi 3.4.0 atau versi terbaru.

3.2 JALAN PENELITIAN

Penelitian dilakukan dengan 5 tahapan, diantaranya yaitu pengambilan data ulasan di *platform* Google Maps dengan teknik *scraping* pada lokasi RSU X Kota A dan Kota B, pra-pemrosesan data terdapat 3 subproses yaitu *cleaning*, *labeling data,* dan representasi vektor dengan TF-IDF. Kemudian, pembuatan model sentimen dengan SVM dan pembuatan topik model dengan LDA, hasil pemodelan dilakukan visualisasi dan analisis hasil. Diagram alur penelitian ditunjukkan pada **Gambar 3.1**.



Gambar 3.1 Alur Penelitian

3.2.1 Pengambilan Data

Dalam penelitian ini data diambil dari Google Maps dikedua lokasi yaitu RSU X Kota A dan RSU X Kota B. Pengambilan data menggunakan bahasa pemrograman Python dan *Application Programming Interface* (API) dari *platform* SerpAPI. SerpApi adalah layanan API yang memungkinkan pengambilan data hasil pencarian Google secara *real-time* (Pratiwi et al., 2024). Dengan menggunakan API SerpAPI, data ulasan dari kedua rumah sakit dapat diakses dan dikumpulkan secara otomatis, sehingga mempermudah proses ekstraksi informasi yang dibutuhkan.

Semua data ulasan dari RSU X di Kota A dan Kota B dikumpulkan dan disimpan dalam format *Comma Separated Value* (CSV). CSV dipilih karena memungkinkan data diolah dengan berbagai tools seperti Python, Excel, atau *software* analitik lainnya, sehingga mendukung efisiensi dalam proses eksplorasi dan pemrosesan data. Data yang telah tersimpan selanjutnya digunakan sebagai sumber utama dalam analisis lebih lanjut.

3.2.2 Pra-pemrosesan Data

Pra-pemrosesan data atau biasa disebut *preprocessing* data merupakan tahapan awal yang penting dalam analisis data yang berguna untuk membersihkan, mengubah, dan mempersiapkan data agar lebih mudah dan akurat dalam proses analisis maupun *modeling* (Daniswara & Nuryana, 2023). Data hasil *scraping* dari kedua lokasi digabungkan dengan menambahkan label lokasi, yaitu Kota A dan Kota B. Label lokasi digunakan untuk mempermudah tahap pra-pemrosesan, sehingga tidak perlu melakukan pra-pemrosesan secara terpisah untuk masing-masing data. Dalam pra-pemrosesan data terdapat 3 subproses yaitu *cleaning* yang berisi tahapan *case folding, text cleaning, normalization, tokenizing, filtering, stemming,* dan*, validation.* Subproses selanjutnya adalah *labeling data,* dan representasi vektor dengan TF-IDF.

1. *Case folding*

   Case folding akan menyetarakan huruf dalam teks menjadi huruf kecil semua.

2. *Text cleaning*

Text cleaning melakukan penghapusan kata yang tidak relevan, seperti tanda baca, *emoticon*, dan lain sebagainya.

3. *Text normalization*

   Mengubah kata-kata gaul (*slang*) menjadi bentuk baku menggunakan kamus *slangwords*.

4. *Tokenizing*

   Tokenizing akan memecah teks menjadi setiap kata.

5. *Text Filtering*

   Proses menghapus kata yang tidak penting atau tidak memiliki makna.

6. *Stemming*

   Proses transformasi kata menjadi bentuk dasarnya.

7. *Text Validation*

   Text Validation berguna untuk mengecek text hasil proses-proses sebelumnya untuk memastikan tidak ada data NaN dan data *duplicate.*

8. *Labeling Data*

   Labeling Data akan melakukan pelabelan pada setiap data dengan kategori yang sesuai. Penelitian ini menggunakan skema *average labeling*, skema ini dapat diterapkan pada ukuran dataset yang besar dengan cepat (Kharisma et al., 2023). Penerapan *average labeling* adalah rating 1, 2, dan 3 dikategorikan sebagai negatif, sedangkan rating 4 dan 5 dikategorikan sebagai positif.

9. Representasi vektor

   Representasi vektor berguna untuk mengubah kata menjadi vektor yang digunakan untuk pelatihan model *machine learning.*

Dengan tahapan pra-pemrosesan yang sistematis ini, data menjadi lebih bersih, terstruktur, dan siap digunakan untuk analisis lebih lanjut. Proses ini tidak hanya meningkatkan kualitas data tetapi juga memastikan bahwa model *machine learning* yang digunakan dapat bekerja secara lebih efisien dan akurat. Oleh karena itu, tahap pra-pemrosesan menjadi salah satu langkah krusial dalam analisis teks yang tidak boleh diabaikan, terutama dalam NLP yang membutuhkan data berkualitas tinggi untuk menghasilkan wawasan yang lebih akurat dan dapat diandalkan.

3.2.3 Model Sentimen

Setelah data dipastikan bersih saat proses pra-pemrosesan data, tahap berikutnya adalah membangun model sentimen dengan algoritma SVM. Model sentimen dengan SVM digunakan untuk mengklasifikasi data ulasan pengunjung RSU X untuk bisa mendeteksi ulasan positif atau negatif. Model Sentimen membutuhkan data *training* dan data *testing* dengan sistem *splitting* data. Untuk menentukan *splitting* data yang optimal, penulis menggunakan K-Fold *Validation*. K-Fold ditentukan jumlah *splitting* datanya yaitu 60:40, 70:30, 80:20, dan 90:10. Setelah menentukan jumlah *splitting*, K-Fold akan memproses data berdasarkan jumlah *splitting* dengan model SVM. Hasil dari K-Fold adalah *accuracy* SVM dari setiap jumlah *splitting* data. *Accuracy* tertinggi dalam *splitting* data digunakan sebagai *splitting* optimal di model SVM.

Setelah *splitting* data optimal ditentukan, selanjutnya adalah menerapkan *splitting* data optimal ke model SVM. Model SVM akan melakukan *training* atau pelatihan berdasarkan inputan data *training*. Setelah model SVM selesai melakukan pelatihan, diperlukan evaluasi model dengan data *testing* untuk melihat seberapa baik model dapat mengklasifikasikan data baru. Pada evaluasi model terdapat beberapa matriks yang digunakan seperti *Accuracy*, *Precision*, *recall*, F-1 *Score*, dan *Confusion matrix*. Setelah evaluasi model SVM menghasilkan evaluasi yang baik, hasil klasifikasi data dapat dilanjutkan ke tahap *topic modeling*.

3.2.4 Topic Modeling

Sebelum membangun topik model dengan LDA, data hasil sentimen perlu dilakukan segmentasi data. Segmentasi data dalam penelitian ini bertujuan untuk membagi data menjadi beberapa bagian. Data sentimen akan dibagi menjadi dua bagian berdasarkan label lokasi yaitu Kota A dan Kota B. Hal ini dibuat agar analisis topik lebih spesifik dari setiap rumah sakit. Setelah itu, data disegmentasi kembali berdasarkan label sentimen hasil model SVM yaitu positif dan negatif. Tujuan segmentasi kedua ini adalah agar topik keluhan dan kepuasan dapat terpisah berdasarkan data positif dan negatif. Data hasil segmentasi meliputi data sentimen di RSU X Kota A positif dan negatif, dan di RSU X Kota B positif dan negatif.

Tahap berikutnya adalah membangun model LDA untuk setiap data sentimen yang telah disegmentasi. LDA digunakan untuk mengidentifikasi topik-topik utama yang terkandung dalam ulasan berdasarkan distribusi kata. Model LDA perlu memasukkan jumlah optimal. Maka dari itu, penelitian ini menggunakan *coherence score* untuk mengevaluasi jumlah topik optimal dari setiap data. *Coherence score* tertinggi akan digunakan untuk menentukan topik dari setiap model. Hasil LDA berupa daftar topik dengan kata-kata dominan dalam data ulasan, dimana jumlah topik ditentukan berdasarkan hasil evaluasi dengan *coherence score*.

3.2.5 Visualisasi

Visualisasi data menjadi langkah penting dalam memahami distribusi dan kecenderungan topik yang muncul. Beberapa visualisasi yang digunakan dalam penelitian ini antara lain *bar charts,* dan pyLDAvis. *Bar charts* digunakan untuk menggambarkan proporsi masing-masing topik yang dihasilkan setiap model. Sehingga memudahkan dalam memahami distribusi tema dalam ulasan yang dianalisis.

PyLDAvis menjadi visualisasi penting dalam eksplorasi hasil pemodelan topik karena mampu memvisualisasikan hubungan antar topik serta kata-kata yang paling berpengaruh dalam setiap kategori. Dengan visualisasi ini, hasil analisis menjadi lebih mudah diinterpretasikan dan dapat memberikan wawasan yang lebih mendalam dalam memahami pengalaman serta sentimen pengunjung terhadap pelayanan kesehatan yang diberikan oleh RSU X.

3.2.6 Analisis Hasil

Tahap akhir dalam penelitian ini adalah analisis hasil yang bertujuan untuk memperoleh wawasan berdasarkan topik yang telah divisualisasikan dari ulasan pengunjung di RSU X. Setelah proses pra-pemrosesan dan penerapan sentimen serta pemodelan topik, dilakukan analisis terhadap kumpulan kata utama dari setiap topik untuk memahami persepsi pengunjung terhadap RSU X.

Analisis dilakukan dengan memisahkan data ulasan positif dan negatif pada masing-masing lokasi rumah sakit, yaitu RSU X Kota A dan RSU X Kota B. Setiap

kumpulan ulasan dianalisis secara terpisah berdasarkan sentimen, kemudian dilakukan ekstraksi topik menggunakan LDA. Topik-topik yang terbentuk dianalisis berdasarkan kata-kata utama untuk mengidentifikasi tema sentral dalam ulasan pengunjung. Dari hasil tersebut, dilakukan penjabaran insight untuk menggambarkan kekuatan dan kelemahan masing-masing rumah sakit berdasarkan persepsi pengunjung terhadap layanan yang diberikan.