BAB III

METODOLOGI PENELITIAN

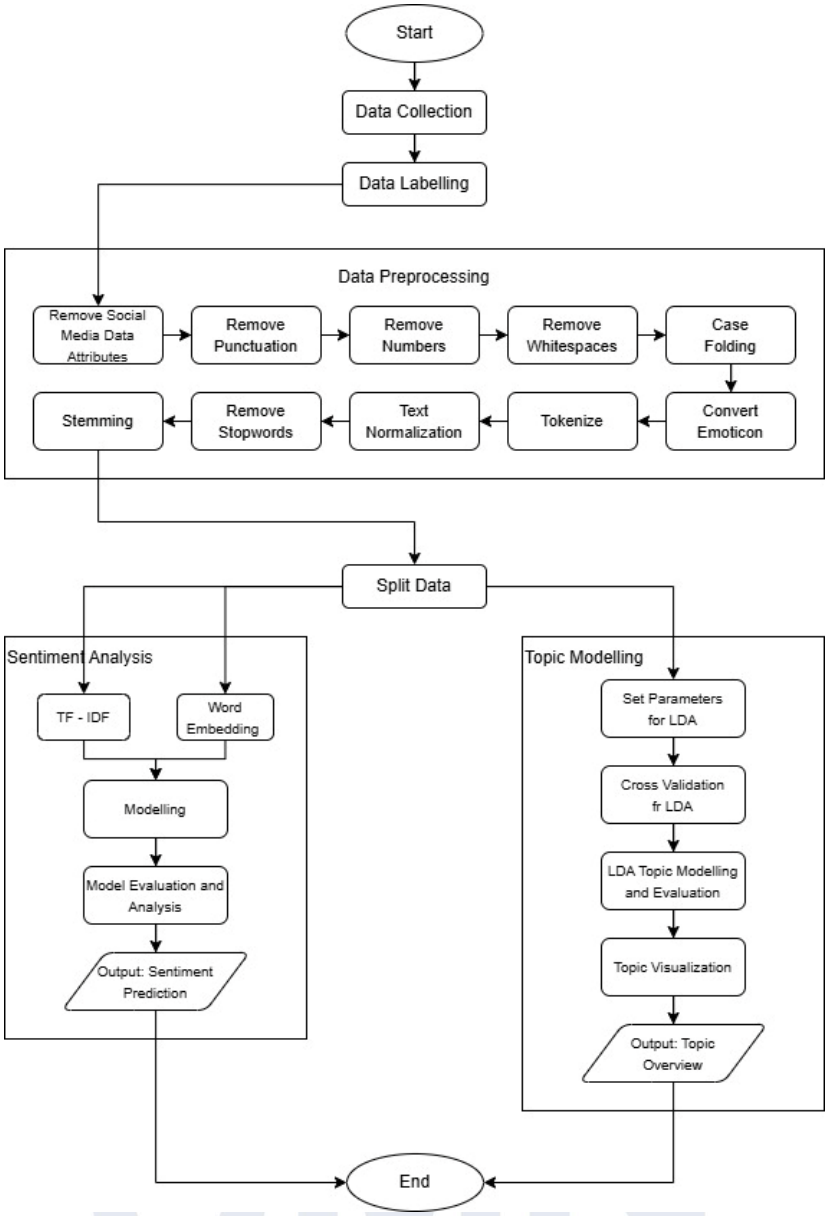
3.1 Gambaran Umum Objek Penelitian

Pada penelitian ini, objeknya adalah data *post* dari pengguna X yang berkaitan dengan Shopee PayLater serta komentar Instagram resmi Shopee PayLater. Data tersebut berisi opini masyarakat mengenai objek yang diteliti yaitu layanan Shopee PayLater yang merupakan layanan *paylater* yang paling banyak digunakan di kalangan Gen Z dan milenial. Data *post* yang dikumpulkan akan diberi label dengan membagi sentimen pengguna ke dalam 2 klasifikasi kategori yaitu sentimen positif dan sentimen negatif. Setelah pemberian label, data akan diproses sebelum masuk ke tahap pemodelan dengan algoritma *Long Short-Term Memory* dan *Support Vector Machine* untuk analisis sentimen, serta *Latent Dirichlet Allocation* untuk pemodelan topik. Data *post* pada X yang diambil pada penelitian ini berada dalam rentang waktu 1 Januari 2024 hingga 1 Januari 2025 yaitu dalam jangka 1 tahun. *Post* yang dikumpulkan mencakup berbagai aspek seperti pengalaman pengguna, keluhan, kepuasan, serta diskusi umum tentang layanan Shopee PayLater.

3.1 Metode Penelitian

3.1.1 Alur Penelitian

Alur penelitian untuk analisis sentimen dan pemodelan topik mengenai platform *paylater* di Indonesia akan digambarkan sebagai berikut:



Gambar 3.1 Alur Penelitian

Gambar 3.1 merupakan *flowchart* yang menjelaskan alur penelitian apa saja yang dilakukan. Alur penelitian didapat dari penelitian terdahulu yang meneliti objek Bank Digital dengan menggunakan analisis sentimen dan pemodelan topik [19]. Meskipun demikian, alur penelitian pada studi ini telah dimodifikasi untuk memenuhi kebutuhan spesifik, yaitu pengaplikasian algoritma LSTM dan SVM, di mana pada algoritma LSTM tidak menggunakan TF-IDF (*Term Frequency-Inverse Document Frequency*), melainkan *Word Embedding* sehingga dilakukan penyesuaian pada bagian analisis sentimen di alur penelitian.

Metode TF-IDF hanya digunakan untuk representasi teks pada algoritma SVM. TF-IDF efektif dalam menangkap bobot kepentingan suatu kata dalam dokumen relatif terhadap korpus, yang cocok untuk model berbasis *feature-engineering* seperti SVM. Sementara itu, untuk algoritma LSTM, TF-IDF tidak digunakan. Sebagai alternatif, metode *Word Embedding* diterapkan untuk merepresentasikan teks. *Word embedding* mampu merepresentasikan kata-kata dalam *dense vector* yang menangkap makna semantik dan hubungan kontekstual antar kata. Pendekatan ini lebih sesuai untuk model *deep learning* seperti LSTM, karena memungkinkan model untuk belajar pola kompleks dari data sekuensial secara lebih efektif tanpa memerlukan *feature engineering* manual. Oleh karena itu, penyesuaian telah dilakukan pada bagian analisis sentimen di alur penelitian untuk mendukung perbedaan metode representasi teks LSTM dan SVM.

Penjelasan setiap langkah pada alur penelitian sebagai berikut:

1. *Data Collection*

   Pada tahap *data collection*, data dikumpulkan melalui proses crawling yang dilakukan terhadap X selama tahap pengumpulan informasi dengan mengambil data berupa tweet pengguna X menggunakan tweet-harvest berdasarkan kriteria *keywords* platform *paylater*. Pada media sosial Instagram, data yang didapatkan berupa komentar dari post Instagram akun resmi platform *paylater* dengan batasan komentar per post sebesar 50.

2. *Data Labeling*

   Pada tahapan ini, dilakukan pemberian label pada setiap baris data tweet dan komentar yang dilabeli dengan sentimen positif dan sentimen negatif. Algoritma yang digunakan untuk analisis sentimen menggunakan *supervised learning*, yang cara kerjanya membutuhkan label pada setiap datanya agar model dapat melakukan klasifikasi. Untuk menjamin kualitas dan objektivitas pelabelan, serta meminimalkan bias subjektif dari satu individu, proses ini melibatkan tiga orang *annotator* dengan latar belakang yang relevan dengan objek penelitian. Ketiga annotator tersebut sebagai berikut:

1. Peneliti Utama: Memiliki pemahaman mendalam mengenai tujuan penelitian, definisi operasional sentimen positif dan negatif dalam konteks ShopeePayLater, serta konteks umum layanan keuangan digital dan e-commerce.

2. Mahasiswa Tingkat Akhir Jurusan Sastra Indonesia**:** Dipilih karena memiliki kepekaan terhadap nuansa bahasa, gaya bahasa informal, slang, dan ekspresi sentimen yang beragam di media sosial seperti X dan Instagram. Keahlian ini membantu dalam menginterpretasi makna teks secara lebih akurat.

3. Pengguna Aktif X dan dan Layanan Shopee PayLater**:** Dipilih untuk memberikan perspektif pengguna nyata yang familiar dengan platform media sosial yang dianalisis dan layanan Shopee PayLater. Pengalaman ini membantu dalam memahami konteks percakapan dan sentimen dari sudut pandang pengguna.

Apabila terdapat perbedaan dalam pemberian label untuk data tertentu (misalnya, satu *annotator* memberikan label 'positif' sementara dua lainnya memberikan label 'negatif'), maka akan dilakukan diskusi bersama antar ketiga *annotator*. Diskusi ini bertujuan untuk menyamakan persepsi berdasarkan pedoman pelabelan yang telah ditetapkan. Label akhir untuk label pada data tersebut akan diambil berdasarkan *majority vote* atau suara terbanyak karena jumlah *annotator* yang ganjil. Proses ini juga merupakan bentuk implementasi praktis dari prinsip validasi *inter-rater agreement*, yang penting dalam penelitian berbasis *supervised learning*.

3. *Remove Social Media Data Attributes*

   Pada tahapan ini, dilakukan penghapusan atribut sosial media yang berupa tagar ‘#’, link, mention ‘@’. Hal ini dilakukan karena atribut tersebut tidak diperlukan dalam memprediksi klasifikasi sentimen.

4. *Remove Punctuation*

   Pada tahapan ini, dilakukan penghapusan tanda baca. Pada python, terdapat library yang bernama string yang berisi kumpulan tanda baca sebagai berikut: !"#$%&'()*+,- ./:;<=>?@[\]^_`{|}~.

5. *Remove Numbers*

Pada tahapan ini, dilakukan penghapusan angka pada setiap baris data karena angka tidak memiliki makna dalam prediksi klasifikasi sentimen. Angka yang dihapus berupa 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, dst.

6. *Case Folding*

   Pada tahapan ini, dilakukan perubahan semua huruf dari UPPERCASE, Capital Each Word, tOGGLE cASE, menjadi lowercase.

7. *Remove Whitespaces*

   Pada tahapan ini, whitespace yang berlebih akan dihapus karena tidak memiliki makna pada data teks.

8. *Convert Emoji*

   Pada data media sosial, terdapat emoji yang dapat menggambarkan emosi, untuk itu dilakukan convert emoji dengan menggunakan library emot yang dapat mengubah emoji menjadi kata “sedih” dan “senang”.

9. *Tokenization*

   Pada tahapan ini, dilakukan pemecahan kalimat menjadi satuan kata.

10. *Text Normalization*

    Pada tahapan ini dilakukan normalisasi teks untuk mengubah kata atau singkatan yang tidak formal menjadi bentuk kata formal atau baku yang sesuai dengan standar KBBI.

11. *Remove Stop Words*

    Pada tahapan ini, dilakukan penghapusan stop words. Stop words adalah kata-kata yang tidak memberikan informasi penting. Contoh kata stop words dalam bahasa Indonesia yaitu “yang”, “di”, “ke”, dst. Pada python, kata stop words ada pada library nltk.corpus.

12. *Stemming*

    Pada tahapan ini, kata yang sudah dipisah dari bentuk kalimat, dipisah lagi menjadi hanya kata dasar saja. *Library* python untuk *stemming* dengan Bahasa Indonesia yaitu sastrawi.

13. *Split Data*

    Pada tahapan ini, dilakukan pemisahan data menjadi *training set*, *validation*, dan *testing set*. Pemisahan data akan dibagi kedalam pro. Pada pemisahan

data untuk pemodelan topik, dilakukan per kategori label sentimen yaitu positif dan negatif.

14. *Create* TF-IDF

    Pada tahapan ini, sebelum melakukan analisis sentimen dibuat TF-IDF (*Term Frequency-Inverse Document Frequency*) untuk pembobotan kata dari tiap dataset yang digunakan pada penelitian ini. Proses TF-IDF terdiri dari dua tahap, yaitu perhitungan frekuensi kata dalam suatu dokumen dibagi dengan jumlah kata dalam dokumen tersebut atau disebut *Term Frequency* (TF). Tahap *Inverse Document Frequency* (IDF) di mana kata diberi bobot berdasarkan perhitungan logaritma dari banyaknya dokumen dibagi dengan banyaknya dokumen yang mengandung kata tersebut. Tujuan dari TF-IDF adalah untuk memberikan bobot pada kata-kata berdasarkan seberapa sering mereka muncul dalam dokumen tertentu serta seberapa umum kata tersebut di seluruh koleksi dokumen (korpus) [39]. TF-IDF dapat dihitung dengan rumus berikut:

$$TF - IDF(t,d) = TF(t,d) \times IDF\ (t)$$

Rumus 3.1 Rumus TF-IDF

Dengan rumus IDF sebagai berikut:

$$IDF(t) = \log(\frac{N}{DF(t)})$$

Rumus 3. 2 Rumus IDF

15. *Word Embedding*

    Word embedding adalah teknik representasi teks dalam bentuk vektor numerik berdimensi rendah yang menangkap hubungan semantik antar kata. Dibandingkan dengan metode tradisional seperti TF-IDF yang hanya melihat frekuensi kata, *word embedding* dapat memahami makna dan hubungan antar kata berdasarkan konteksnya dalam teks. *Word embedding* dapat menangkap hubungan semantik dan sintaksis antara kata-kata, memungkinkan penggunaan dalam berbagai aplikasi seperti klasifikasi teks, analisis sentimen, dan penerjemahan bahasa [40].

16. *Long Short Term Memory* (LSTM)

Pada tahapan ini, data yang sudah melalui tahap labeling dan preprocessing akan dilakukan modeling dengan menggunakan algoritma LSTM.

17. *Support Vector Machine Modeling*

    Setelah pemodelan dengan algoritma LSTM akan dilakukan pemodelan dengan SVM untuk melakukan klasifikasi sentimen.

18. *Model Comparison Analysis*

    Pada tahapan ini, model LSTM, dan SVM akan dibandingkan dengan menggunakan *confusion matrix*.

19. *Set LDA Parameters*

    Pada tahapan ini, ditentukan parameter untuk algoritma LDA. Parameter yang ditentukan adalah jumlah topik optimal yang akan dihasilkan dari model.

20. *Cross Validation for LDA*

    Tahapan cross-validation dilakukan untuk memastikan pemodelan topik dengan LDA sudah menghasilkan jumlah topik yang optimal. Validasi ini dinilai dengan nilai coherence, atau nilai kemiripan diantara banyak kata yang tergabung dalam satu topik.

21. *LDA Topic Modeling and Evaluation*

    Pada tahapan ini, pemodelan topik menggunakan algoritma LDA dilakukan dengan jumlah topik (K) optimal yang telah ditentukan melalui proses *cross*-validation. *Output* dari model LDA berupa distribusi probabilitas kata-kata untuk setiap topik, serta distribusi probabilitas topik untuk setiap data opini. Setelah model LDA menghasilkan topik-topik, langkah selanjutnya adalah evaluasi kualitatif, interpretasi, dan penamaan topik. Setiap topik diberi label tema yang singkat, deskriptif, dan secara akurat menangkap maksud dari tema tersebut. Misalnya, kumpulan kata kunci seperti “promo”, “diskon”, “cashback”, “gratisongkir” dapat diinterpretasikan dan dinamai sebagai tema “Promosi dan Keuntungan”. Validasi interpretasi dilakukan melalui *peer debriefing* berupa diskusi dengan dosen pembimbing mengenai kesesuaian tema dengan kata kunci serta penyajian transparan, di mana pada sub-bab Hasil dan Diskusi terdapat

daftar kata kunci utama dan kalimat representatif yang disajikan setiap topik. Cara ini memungkinkan pembaca untuk menilai validitas interpretasi.

22. *Topic Visualization*

    Pada tahapan ini, setelah topik-topik diinterpretasikan, dinamai, dan dikaitkan dengan sentimen, hasil pemodelan topik kemudian divisualisasikan. Visualisasi dapat berupa *topic overview* yang menyajikan nama topik, kata kunci utama per topik, proporsi topik dalam keseluruhan korpus, serta kaitan sentimennya.

3.1.2 Metode Data Mining

Metode *Data Mining* yang digunakan dalam penelitian ini menggunakan *framework* CRISP-DM (*Cross-Industry Standard Process for Data Mining*). Metode *data mining* ini memiliki 6 tahapan yang diterapkan dalam penelitian ini sebagai berikut:

1. *Business Understanding*

   Pada tahapan *business understanding*, dilakukan pemahaman mengenai mengapa analisis sentimen Shopee PayLater dilakukan, dan apa tujuan serta manfaat dari penelitian ini. Penelitian ini bertujuan untuk menganalisis sentimen publik pada media sosial X dan Instagram dengan objek Shopee PayLater. Penelitian ini juga menerapkan pemodelan topik untuk mendapatkan pola topik sehingga dapat menambah wawasan mengenai Shopee PayLater dan dapat dijadikan sebagai bahan pertimbangan untuk pihak pengembang aplikasi agar meningkatkan layanan mereka.

2. *Data Understanding*

   Pada tahapan *data understanding*, dilakukan pemahaman mengenai data yang ada. Tahapan ini meliputi keseluruhan proses dari pengumpulan data, eksplorasi data, pengecekan pada kualitas data, dan membuat deskripsi singkat pada data. Proses pengumpulan data dalam periode X bulan dari X hingga X. Data diperoleh dengan menggunakan tools Google Colaboratory dan bahasa pemrograman Python.

3. *Data Preparation*

Pada tahapan *data preparation*, dilakukan persiapan data untuk tahap berikutnya yaitu pemodelan. Hal yang dilakukan pada tahapan ini yaitu *data labeling* dan *data preprocessing* pada alur penelitian. Setelah data sudah bersih dan siap, akan dilanjutkan ke tahap pemodelan.

4. *Modeling*

   Pada tahapan *modeling*, dilakukan pemodelan yang diawali dengan splitting data menjadi training set dan testing data. Bagian analisis sentimen, data akan dilakukan tahapan yang terdapat pada alur penelitian, dimulai dari TF-IDF, kemudian data akan dianalisa menggunakan dua model yaitu LSTM dan SVM.

5. *Evaluation*

   Pada tahapan *evaluation*, model dengan hasil akurasi paling baik akan dipilih untuk hasil klasifikasi sentimen, dan output dari pemodelan topik akan berupa *topic overview*.

6. *Deployment*

   Pada tahapan *deployment*, model yang terpilih akan di *deploy* dalam sebuah *website* sederhana di mana penggunanya dapat memasukkan sebuah dataset yang berisi komentar X dan/atau komentar di Instagram, dan program dapat memprediksi sentimen tersebut ke dalam kategori positif atau negatif, serta menghasilkan topik dari hasil pemodelan topik.

3.2 Teknik Pengumpulan Data

Pada penelitian ini, jenis data yang digunakan adalah data primer yang dikumpulkan langsung dari sumber, dan spesifik untuk tujuan penelitian. Data media sosial X didapatkan dengan melakukan crawling data X dengan tools Google Colaboratory dan bahasa pemrograman python. Data media sosial Instagram didapatkan dengan menggunakan tools Apify yaitu dengan mengambil bagian komentar pada post dari akun resmi Instagram @shopeepay_id dan @spaylater_id.

3.2.1 Populasi dan Sampel

Populasi pada penelitian ini adalah opini masyarakat dari sosial media X dan Instagram mengenai Shopee PayLater. Sampel pada penelitian ini adalah data media sosial berisi opini masyarakat mengenai Shopee PayLater yang

diperoleh dalam kurun waktu 1 tahun. Data ini diperoleh dari hasil *crawling* data X dan scraping data komentar Instagram. Sampel diambil dengan teknik *purposive sampling*, yaitu teknik pengambilan sampel di mana peneliti memilih data berdasarkan kriteria tertentu yang telah ditetapkan sesuai dengan tujuan penelitian. Dalam penelitian ini, kriteria yang digunakan mencakup relevansi opini terhadap Shopee PayLater, sesuai dengan kata kunci dan unggahan khusus mengenai topik yang ingin diteliti.

3.2.2 Periode Pengambilan Data

Pada penelitian ini, periode pengambilan data adalah selama 1 tahun dimulai dari 1 Januari 2024 hingga 1 Januari 2025. Rentang waktu ini dipilih untuk memastikan bahwa data yang dikumpulkan mencerminkan tren dan perubahan opini masyarakat terhadap Shopee PayLater dalam kurun waktu yang cukup panjang. Dengan periode satu tahun, penelitian ini dapat menangkap berbagai perubahan opini yang mungkin dipengaruhi oleh perubahan kebijakan Shopee PayLater, atau peristiwa ekonomi yang berhubungan dengan layanan tersebut. Data yang diambil dalam kurun waktu yang lebih panjang memberikan peluang untuk mengamati beragam opini yang mungkin terjadi akibat perubahan regulasi, atau pengalaman pengguna yang berkembang seiring waktu. Hasil penelitian diharapkan lebih akurat dalam menggambarkan persepsi masyarakat terhadap Shopee PayLater selama periode yang ditentukan.

3.3 Variabel Penelitian

3.3.1 Variabel Independen

Variabel independen yang digunakan dalam penelitian ini adalah opini masyarakat mengenai objek penelitian yaitu Shopee PayLater yang terdapat pada media sosial X dan Instagram. Data pada media sosial X berupa opini publik dalam bentuk *post* dengan kata kunci dan tagar (*hashtag*) yang sudah ditentukan yaitu mengenai Shopee PayLater. Data pada media sosial Instagram berupa opini publik dalam bentuk komentar pada post Instagram akun resmi @shopeepay_id dan @spaylater_id.

3.3.2 Variabel Dependen untuk Analisis Sentimen dan Pendekatan Pemodelan Topik

Dalam penelitian ini, terdapat dua jenis analisis utama dengan karakteristik variabel yang berbeda:

1. Analisis Sentimen

   Variabel dependen adalah variabel yang nilainya dipengaruhi atau diprediksi oleh variabel independen. Dalam konteks analisis sentimen yang bersifat *supervised learning*, variabel dependen yang digunakan adalah label sentimen pada setiap baris data data opini pengguna layanan. Label sentimen ini terdiri dari dua kategori, yaitu sentimen positif dan sentimen negatif. Model klasifikasi sentimen akan dilatih untuk memprediksi label sentimen ini berdasarkan fitur-fitur yang diekstrak dari teks opini (variabel independen).

2. Pemodelan Topik

   Latent Dirichlet Allocation (LDA) merupakan metode *unsupervised learning*. Berbeda dengan metode *supervised* seperti klasifikasi sentimen, LDA tidak bekerja dengan konsep variabel dependen yang diprediksi. Sebaliknya, LDA bertujuan untuk menemukan struktur topik laten yang tersembunyi dalam kumpulan dokumen yang berupa data opini.

3.4 Teknik Analisis Data

Pada penelitian ini, data yang ada bersifat kualitatif karena berbentuk opini masyarakat mengenai Shopee PayLater. Analisis ini dilakukan dengan menggunakan bahasa pemrograman python dengan *tools* Google Colaboratory. Data akan didapatkan dengan menggunakan *tools* Tweet Harvest untuk *crawling* data pada platform X dan data komentar pada Instagram akan didapatkan menggunakan tools Apify.