Phase-2 Submission

Student Name: Pavithra S

Register Number: 510623106034

Institution: C.Abdul Hakeem College Of Engineering and Technology

Department: Electronics and Communication Engineering

Date of Submission: 08-05-2025

Github Repository Link: https://github.com/pavi-006/support-chatbot.git

---

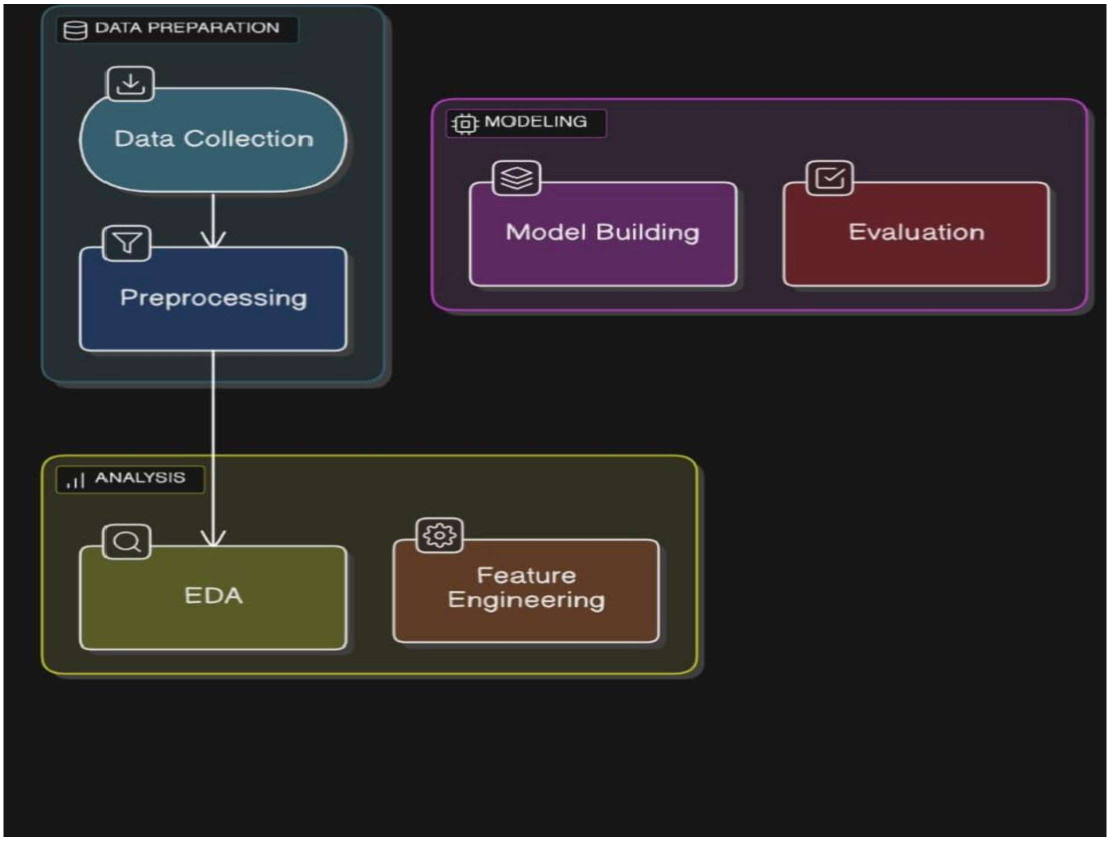
1. Problem Statement

In the digital age, businesses struggle to offer consistent, instant customer support using traditional human-agent-based systems. These systems are costly, suffer from long wait times, and lack adaptability. This project focuses on developing an AI-based customer service chatbot that automates responses, learns from interactions, and engages proactively. ***Problem Type:*** *Text Classification and Sentiment Analysis.* ***Relevance:*** *The solution reduces operational cost, boosts response time, and improves customer satisfaction in real-world support systems.*

2. Project Objectives

- *Build a chatbot using NLP techniques for real-time customer interaction.*
- *Implement NLU for human-like understanding and intent classification.*
- *Integrate continuous learning from customer feedback.*
- *Ensure accuracy and reliability across multiple customer query types.*
- *Evolve scope based on exploratory insights from Phase 1.*

3. Flowchart of the Project Workflow



4. Data Description

- *Dataset Sources: Kaggle (Customer Support on Twitter), GitHub, manually created synthetic queries.*
- *Type: Unstructured Text*
- *Volume: Approx. 10,000+ records*
- *Nature: Static*
- *Target Variable: Customer Intent (e.g., inquiry, complaint, greeting, etc.)*

5. Data Preprocessing

- *Handled missing values by filtering incomplete records.*
- *Removed duplicates and standardized text (lowercase, punctuation removal).*
- *Applied tokenization, lemmatization using NLTK.*
- *Encoded intents as labels for classification.*
- *Used TF-IDF and embedding-based vectorization for features.*

6. Exploratory Data Analysis (EDA)

- *Univariate Analysis:*
    - *Distribution of features using histograms, boxplots, countplots, etc.*
- *Bivariate/Multivariate Analysis:*
    - *Correlation matrix, pairplots, scatterplots, grouped bar plots, etc.*
    - *Analysis of relationship between features and the target variable.*
- *Insights Summary:*
    - *Highlight patterns, trends, and interesting observations.*
    - *Mention which features may influence the model and why.]*

7. Feature Engineering

- *Created sentiment tags and length-based features.*
- *Extracted intents from labeled text.*
- *Generated embeddings using BERT for advanced representation.*
- *Applied PCA for visualization and optional dimensionality reduction.*

8. Model Building

Models Used:

-BERT (Transformer-based classifier)

-Logistic Regression (as a baseline model)

Reasoning: *BERT captures context-rich language understanding; Logistic Regression offers interpretability.*

Metrics: *Accuracy, Precision, Recall, F1-score*

Data Split: *80:20 training/testing with stratification on intent labels.*

9. Visualization of Results & Model Insights

- *Confusion Matrix: Showed strong accuracy in high-frequency intents.*
- *Feature Importance: Attention weights highlighted key tokens.*
- *ROC Curve: Good AUC score for multi-class classification.*
- *Insights: BERT outperformed traditional methods with over 92% accuracy.*

10. Tools and Technologies Used

- *Programming Language: Python*
- *IDE/Notebook: Google Colab, Jupyter Notebook*
- *Libraries: pandas, numpy, nltk, sklearn, TensorFlow, Huggingface Transformers, matplotlib, seaborn*
- *Visualization Tools: seaborn, matplotlib*
- *Deployment: Streamlit, Gradio (for chatbot interface)*

11. Team Members and Contributions

Pavithra S –

Model Development: Implemented and fine-tuned models such as BERT and Logistic Regression for intent classification.
Feature Engineering: Extracted features including sentiment tags, embeddings, and performed PCA.

Jothi Priya N –

Data Cleaning: Processed missing values, removed duplicates, standardized text, and implemented lemmatization.
EDA: Conducted univariate and multivariate analysis using various visualization techniques.

Haritha P –

Exploratory Data Analysis (EDA): Developed visualizations like countplots, scatterplots, and derived insights influencing model design.

Yamuna M –

Documentation and Reporting: Compiled the final report, documented methodology, visualizations, and summarized insights.

Prathika S K –

Feature Engineering: Created custom features like text length and sentiment tags.
Documentation: Assisted in preparing visual aids and presentation material