Phase-2SubmissionTemplate

Student Name: [JAGADEESHWARI J]

RegisterNumber:[422723104043]

Institution: [V.R.S. College of engineering and technology]

Department:[Computer Science and Engineering]

Date of Submission: [10.05.2025]

Topic : [Revolutionizing customer support with an intelligent chatbot for automated assistance]

GithubRepositoryLink:https://github.com/Jagadeeshwari279/Jagadeeshwari.git

---

1. Problem Statement

This project aims to develop an AI-powered chatbot capable of understanding customer queries and responding with relevant, accurate information. The problem type is primarily a combination of:

👉Text classification (intent detection)

👉Named Entity Recognition (NER)

👉Sequence generation or retrieval (response generation)

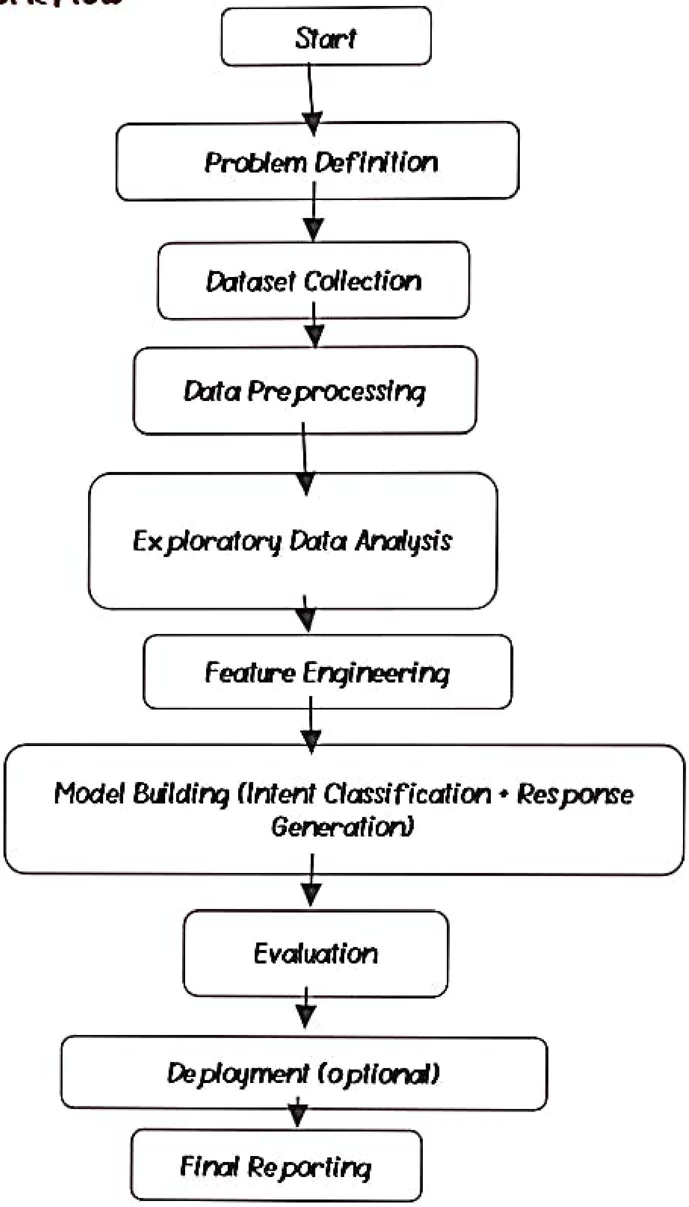
2. Project Objective

👉 **Technical Goal:** Build a machine learning/NLP-based chatbot that can classify intents, extract entities, and provide automated, accurate replies.

👉 **Model Objective:** High classification accuracy for intent detection and response relevance for output generation.

👉 **Post-EDA Refinement:** Identified the need for better handling of ambiguous queries and added fallback handling.

3. Project Workflow



4. Data Description

👉**Dataset Name :** niraliivaghani chatbot dataset

👉**Link :** https://www.kaggle.com/datasets/niraliivaghani/chatbot-dataset

👉**Type of Data :** unstructured

👉**Number of Records and Features:** ~100,000 records with fields like user query, intent label, timestamp, and response

👉**Nature of Data:** Static

👉**Target Variable:** Intent category (e.g., billing issue, login failure, general inquiry)

5. Data Preprocessing

Link:https://colab.research.google.com/drive/1fmbgnwcCdFTD80C4UagdaWLNaGgJ2_E3?usp=sharing

5.1 Handling values :

Handling Missing Values : Removed incomplete conversations

5.2 Removing or Justifying Duplicate Records

Duplicates: Dropped repeated entries

5.3 Text Cleaning

👉 Lowercasing, punctuation removal, stopword removal, lemmatization labels

👉Encoding: Label encoding for intent labels

5.4 Encoding:

Label encoding for intent labels

5.5 Vectorization

TF-IDF / word embeddings (e.g., GloVe or BERT)

5.6 Normalization

Not required for text, but token length truncation applied.

6. Exploratory Data Analysis (EDA)

👉**Univariate:** Word clouds and frequency plots for top words in each intent

👉**Multivariate:** Count plots by intent class analysis of average query length by intent

👉**Insights:** Majority of queries fall into 5-6 main categories Certain intents have very distinctive keywords (e.g., "reset," "password" → login issue)

7. Feature Engineering

👉**Custom Features:** Message length, number of keywords matched

👉**NLP features :** TF-IDF scores, POS tags

👉**Dimensionality Reduction:** Used PCA or UMAP for visualizing intent clusters

👉**Justification:** Enhanced model's ability to separate intents using semantic clues

8. Model Building

Models Used:

👉Logistic Regression ((baseline classifier))

👉BERT-based fine-tuned Transformer for intent classification

👉

👉**Justification:** BERT captures contextual semantics better than traditional models

👉**Split:** 80/20 train-test split using train_test_split.

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

9. Visualization of Results & Model Insights

👉Confusion matrix for intent classification

👉 ROC Curve (if applicable)
👉t-SNE plot for visualizing intent separability
👉Top influential words/features per intent class

Findings

👉BERT achieved F1-score > 90%

Errors occurred mainly in overlapping or ambiguous intents

10. Tools and Technologies Used

👉Language: Python
👉IDE: Jupyter Notebook / VS Code
👉Libraries: pandas, numpy, scikit-learn, NLTK, spaCy, Transformers
👉Visualization: seaborn, matplotlib, Plotly

11. Team Members and Roles

JAGADEESHWARI J - Documentation and visualization

JAMUNA RANI V- Feature engineering

ISHWARIYA A - Model development

JAYABHARATHI J - Data cleaning, EDA