.

**Phase-2 Submission – Machine Learning Project**

**Student Details**  
**Name:** Ragul.S  
**Register Number:** 422223243045  
**Institution:** Surya Group of Institutions  
**Department:** Artificial intelligence and data science   
**Date of Submission:** 29/04/2025  
**GitHub Repository:**[**https://github.com/RAGULMR/Ragul-NM-Phase2**](https://github.com/RAGULMR/Ragul-NM-Phase2)

**1. Problem Statement**

This project aims to automate and enhance customer support services by developing a Generative AI chatbot. The chatbot is trained on real-world customer conversation data sourced from Twitter, focusing on understanding and responding to customer queries effectively.

**Problem Type:** Classification (Intent Classification, Sentiment Analysis, Response Prediction)  
**Relevance:** Automating customer service boosts satisfaction, minimizes response time, reduces operational costs, and ensures 24/7 support availability.

**2. Project Objectives**

**Technical Goals:**

* Preprocess and clean Twitter-based customer support data.
* Build models to classify customer **intent** and **sentiment**.
* Predict suitable chatbot responses.
* Evaluate accuracy and real-world applicability.

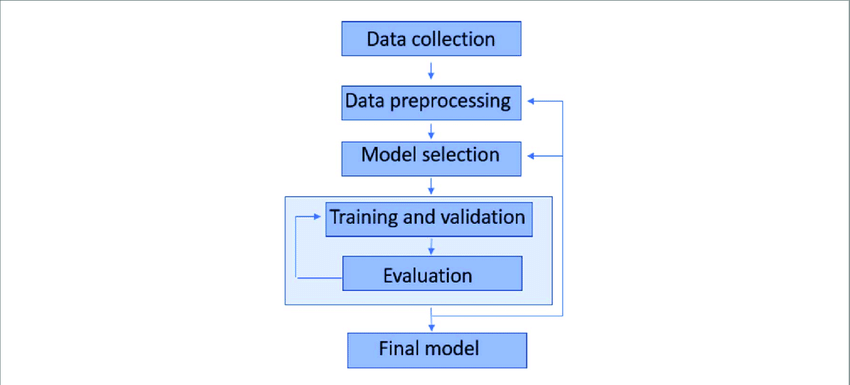
**Model Aims:**

* Ensure high classification accuracy for understanding customer queries.
* Generate meaningful, context-aware chatbot responses applicable in real-world support systems.

**Data Insights & Adjustments:**

* Sentiment classification is emphasized if emotional content is significant.
* Intent classification remains the focus for service-based queries.
* Additional features or dataset balancing may be introduced as needed.

**3. Project Workflow (Flowchart)**

*(*

**4. Data Description**

* **Dataset:** Twitter Customer Support Dataset
* **Source:** Kaggle
* **Format:** Text (Tweets & Replies)
* **Size:** ~3,000 conversations
* **Features:** tweet text, response, author\_id, inbound
* **Target:** Intent label or Sentiment score
* **Nature:** Static data

**5. Data Preprocessing**

* **Missing Values:** Dropped rows with missing queries or responses.
* **Duplicates:** Removed identical query-response rows.
* **Outliers:**
  + Filtered overly short (<5 words) and extremely long queries.
* **Text Cleaning:**
  + Lowercased text, removed emojis, hashtags, links, and stopwords.
* **Encoding:** Label encoding applied to target classes.
* **Text Vectorization:** TF-IDF used to convert text into numerical format.
* **Train-Test Split:** 80% training, 20% testing.

**6. Exploratory Data Analysis (EDA)**

* **Univariate Analysis:** Frequency plots for intents and sentiments.
* **Multivariate Analysis:**
  + Relationship between tweet length and sentiment.
  + Co-occurrence of keywords and intent.
* **Insights:**
  + Common intents: Refund Request, Order Status Inquiry
  + Longer queries often signal complaints

**7. Feature Engineering**

* **Engineered Features:**
  + Message length (word/character count)
  + Presence of service-related keywords (e.g., “refund”, “cancel”)
* **Techniques:**
  + TF-IDF
  + (Optional) Word2Vec for deeper embedding

**8. Model Building**

* **Models Used:**
  + Logistic Regression (baseline)
  + Random Forest Classifier (for non-linear patterns)
* **Split:** 80%-20% training/testing
* **Evaluation Metrics:** Accuracy, Precision, Recall, F1-Score

**9. Results Visualization & Insights**

1. **Confusion Matrix:**  
   Shows misclassification patterns (e.g., “Order Status” vs. “Delivery Delay”).
2. **ROC Curve:**  
   Visualizes classifier performance across thresholds (AUC indicates strength).
3. **Feature Importance Plot:**  
   Highlights most impactful features like keyword presence or message length.
4. **Model Comparison:**  
   Visual evaluation of model metrics for performance selection.

**10. Tools and Technologies**

* **Language:** Python
* **IDE/Notebooks:** Google Colab, Jupyter
* **Libraries:** pandas, numpy, seaborn, matplotlib, scikit-learn, nltk, transformers
* **Visualization:** matplotlib, seaborn, Plotly

**11. Team Members and Contribution**

Ragul..SManaged planning, preprocessing

Kabilan.M: Evaluation., Conducted EDA

Vasanthakumar.J**:** , created visualizations, assisted feature engineering.

Gopinath.D: , chatbot logic, and testing framework