# Phase-2 Submission Template

Student Name: [D Vijayalakshmi]

Register Number: [421623104169]

Institution: [Mailam Engineering College]

Department: [CSE]

Date of Submission: [Insert Date]

Github Repository Link: [Update the project source code to your Github Repository]

---

## 1. Problem Statement

*[It is extracting accurate emotional insights from this unstructured text data is a complex challenge.The nuances of human languag,sarcasm,slang,abbreviations,and multiple content make it difficult to intergret sentiments.*

- ***Dataset:*** *Text are often short and informal.Emojis,hashtags,and slang require special handling.Label may reflect multiple emotions.*
- *This is a multiclass classification problem the model aims to classify text into one of several emotional categories (eg: joy,anger,sadness,fear,surprise)*
- ***Social impact:****Helps intracking public mood and mental health patterns during crises (eg:COVID-19).****Automation****:Reduces the need for manual monitoring of social media.* ***Researxh value:****Supports studies in psychology,sociologyand marketing by providing large-scale emotion data.]*

## 2. Project Objectives

- ***1)****To collect and preprocess real world social media data(eg:tweets,comments) for effective emotions.****2)****To develop a multiclass classification model that can accurately categorize user conversation such as joy,anger,sadness.etc.****3)****To apply and compare different ML and DL algm.*

- *Initially,the focus was only on basic sentiment(positive/negative/neutral) but after exploring the dataset. The goal has evoled from simple sentiment analysis to emotion.*
- *Build a reliable and scalable system that helps businesses,researchers,and social organizations decode public emotions.]*

## 3. Flowchart of the Project Workflow

*[**Data Collection->Data Cleaning->EDA->Feature Engineering->Model Building->Model Evaluation->Visualization & Interpretation->Deployment/Demo Interface**]*

## 4. Data Description

- *Dataset: Twitter Sentiment/Emotion Dataset.*
- *Type of data: Unstructured Data-includes tweets,comments,or short social media texts.*
- *Number of Records and Features: Eg(if using Kaggle dataset) Rows:~20000 to 50000 tex entries column/Features:Text-the actual social media.Emotion/Sentiment-the target label.*
- *Static dataset:pre-collected and labeled data.*
- *Target variable (if supervised learning):Sentiment Label:Positive,Negative,Neutral.]*

## 5. Data Preprocessing

- *Handle missing values:It is text and emotion sentiment columns.Dropped records with missing text or labels using dropna()since incomplete data can't be used for training.*
- *Duplicate records:Duplicate text entries were identified to avoid model bias.*

- *Detect and treat outliers:Outliers in text data were reviewed based on unusually short texts(eg:1-2 characters)*
- *Text Cleaning:Convrted text to lowercase removed.URLs,mentions(@user),hashtags(#hashtags).Emojis and special characters.*
- *Encode Target Labels:label bonding was applied to convert textual labels (like joy,anger)into numeric format for model training.*
- *Vectorization:Used TF-IDF for vectorizer to convert cleaned text into*
- *numerical vectors.*

*Train-Test Split:Splitting dataset into training and testing sets for model validation.*

## 6. Exploratory Data Analysis (EDA)

*[Perform detailed statistical and visual exploration of the data.*

- *Univariate Analysis:*
  - *Target Variable(sentiment/emotion)distribution used countplot to check the balance of emotion classes.*
- *Bivariate/Multivariate Analysis:*
  - *Some emotions may tend to be expressed in longer messages(eg:sadness)compared to others.*
  - *Correlation between text length and encoded label.*
- *Insights Summary:*
  - *TF-IDF or word embeddings capture important keywords or patterns tied to emotions.*

  *Text length can be used as an auxiliary feature.*

  *Presence of emojis, exclamation marks can also be useful.*

## 7. Feature Engineering

●*As the core data is text, converting it to numerical features is essential:*

*- Justification: Captures the importance of words relative to all documents, reducing the effect of common words like "the".*

●*Created new features based on the length of the message:*

*Justification: Emotions like anger or joy might correlate with longer or shorter messages.*

## 8. Model Building

● *1. Logistic Regression: A popular choice for binary classification problems, logistic regression is suitable for sentiment analysis.*

*2. Random Forest: An ensemble method that combines multiple decision trees, random forest is robust and handles high-dimensional data well.*

● *1. Logistic Regression: Selected for its simplicity, interpretability, and effectiveness in binary classification problems.*

*2. Random Forest: Chosen for its ability to handle complex data, reduce overfitting, and provide feature importance scores.*

● *1. Data Split: Split the data into training (80%) and testing sets (20%) with stratification to maintain class balance.*

*2. Text Preprocessing: Apply techniques like tokenization, stopword removal, and vectorization (e.g., TF-IDF).*

●*Matrices: Use accuracy, precision, recall, and F1-score to evaluate the performance of both models.*

*2. Comparison: Compare the performance of logistic regression and random forest models.*

## 9. Visualization of Results & Model Insights

● *Confusion matrix: This is a fundamental tool for evaluating the performance of a classification model. It shows the counts of:True Positives (TP): The model correctly predicted a positive sentiment.True Negatives (TN): The model correctly predicted a negative sentiment.False Positives (FP): The model incorrectly predicted a positive sentiment when it wasactually negative (Type I error)*

●*A higher AUC generally means better model performance. By examining the shape of the curve, you can understand the trade-off between sensitivityand specificity at different classification thresholds. For example, if you need to minimizefalse negatives.*

## 10. Tools and Technologies Used

● *Python:Python*: The primary programming language used for this project.*

● *Google Colab:Google Colab*: A cloud-based notebook environment used for development, testing, and deployment. Jupyter Notebook: An interactive development environment used for exploratory data analysis and visualization.*

● *pandas*: A library for data manipulation and analysis.*

*- numpy: A library for numerical computing.*

*- seaborn: A library for data visualization.*

*- matplotlib: A library for creating static, animated, and interactive visualizations.*

## 11. Team Members and Contributions

*[Team Members:*

*1. S Supriya*

*2. A Sunmathi*

*3. S Vaishnavi*

*4. D Vijayalakshmi*

○ ***S Supriya:****Data cleaning and EDA: Responsible for cleaning and preprocessing the social media conversation data, as well as performing exploratory data analysis (EDA) to understand the data distribution and patterns.*

- ***A Sunmathi:****Feature engineering: Responsible for extracting relevant features from the text data, such as sentiment scores, word frequencies, and topic modeling.*
- ***S Vaishnavi:****Model development: Responsible for developing and training machine learning models for sentiment analysis, including logistic regression, random forest, and other algorithms.*
- ***D Vijayalakshmi:****Documentation and reporting:Responsible for documenting the project methodology, results, and insights, as well as creating reports and presentations to stakeholders*