**Phase-2 Submission**

**Student Name:** PRIYAN P

**Register Number:** 712523205045

**Institution:** PPG INSTITUTE OF TECHNOLOGY

**Department:** B.TECH INFORMATION TECHNOLOGY

**Date of Submission:** [16/05/2025]

**Github Repository Link:** [**https://github.com/P-priyan-05/NM\_PRIYAN\_DS**](https://github.com/P-priyan-05/NM_PRIYAN_DS)

### **1. Problem Statement**

### *This project aims to build a model to* ***decode emotions from social media text data*** *using* ***sentiment analysis techniques****. This is a* ***multi-class classification*** *problem, where each text sample is categorized into one of several emotional states such as happy, sad, angry, fearful, surprised, or neutral.*

### **2. Project Objectives**

### *To identify and classify the emotions expressed in user-generated content from social media platforms.*

### *To apply* ***natural language processing (NLP)*** *techniques for data preprocessing and feature extraction.*

### *To implement and compare multiple* ***machine learning and deep learning models*** *for emotion classification.*

### *To optimize model performance using suitable* ***evaluation metrics*** *and* ***hyperparameter tuning****.*

### 

### **3. Flowchart of the Project Workflow**

### *[Visually represent the entire workflow from start to finish.]*

### 

### **4. Data Description**

***Dataset Name:*** *Emotion Detection from Text (e.g., Twitter Emotion Dataset or Kaggle Emotion Dataset)*

***Source:*** *Kaggle / Twitter API / Public NLP repositories*

***Data Type:*** *Unstructured (textual data)*

***No. of Records:*** *~20,000–50,000 posts/comments*

***Features:*** *Text, Emotion Label*

***Dynamic or Static:*** *Static dataset*

***Target Variable:*** *Emotion class (happy, sad, angry, fear, etc.)*

### **5. Data Preprocessing**

* *Handle missing values (removal, imputation, etc.).*
* *Remove or justify duplicate records.*
* *Detect and treat outliers.*
* *Convert data types and ensure consistency.*
* *Encode categorical variables (label encoding, one-hot encoding).*
* *Normalize or standardize features where required.*
* *Document and explain each transformation step clearly in code and markdown.]*

### 

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

*Removed HTML tags, special characters, and emojis.*

*Converted text to lowercase.*

*Tokenization and lemmatization applied.*

*Stop words removed using NLTK.*

*Encoded target labels using LabelEncoder.*

*Vectorized text using:*

### 

### **7. Feature Engineering**

Removed HTML tags, special characters, and emojis.

Converted text to lowercase.

Tokenization and lemmatization applied.

Stop words removed using NLTK.

Encoded target labels using LabelEncoder.

Vectorized text using:

### 

### **8. Model Building**

### *Logistic Regression*

### *Random Forest*

### *Multinomial Naive Bayes*

### *Support Vector Machine (SVM)*

### *LSTM (Long Short-Term Memory Neural Network)*

### **9. Visualization of Results & Model Insights**

*Confusion Matrix: To evaluate class-wise performance*

*ROC Curve: Multi-class ROC-AUC visualization*

*Feature Importance: For Random Forest and Logistic Regression*

*Word Clouds: Top words for each emotion*

L**10. Tools and Technologies Used**

***Programming Language:*** *Python*

***IDE/Notebook:*** *Jupyter Notebook / Google Colab*

***Libraries:***

### **11. Team Members and Contributions**

***[****List names and responsibilities.*

* *Clearly mention who worked on:*
  + *Data cleaning*
  + *EDA*
  + *Feature engineering*
  + *Model development*
  + *Documentation and reporting]*