Phase-1 Submission

Student Name: Sharon Cynthiya J

Register Number: 712523104058

Institution: PPG Institute of Technology

Department: BE. Computer Science and Engineering

Date of Submission: 25.04.2025

1.Problem Statement

In today's digital age, people often share their emotions on platforms like Twitter, Facebook, and Reddit. Understanding these emotions is valuable for businesses, mental health experts, and policymakers. Since manual analysis of this vast, unstructured data isn't practical, an automated sentiment analysis system can provide timely insights to support better decision-making.

2.Objectives of the Project

- To collect and analyze social media conversations for sentiment and emotional patterns.
- To classify posts into emotion categories such as joy, anger, sadness, or surprise.
- To develop machine learning and deep learning models that can accurately predict sentiments.
- To visualize emotional trends across different topics and over time.

3.Scope of the Project

What We'll Analyze:

- The sentiment of each post (positive, negative, or neutral)
- Emotional categories using models like NRC Emotion Lexicon or BERT
- Popular hashtags and keywords related to emotional content

Project Limitations:

- Focus is limited to English-language posts
- Only publicly available data will be used
- The output will be shown via dashboards or notebooks—not full web apps

4.Data Sources

Source:

- Twitter API : https://developer.twitter.com/en/docs/twitter-api
- Reddit API : https://www.reddit.com/dev/api/
- Kaggle Datasets (e.g., Sentiment140, Emotion Dataset): https://www.kaggle.com/datasets

Nature: Public data

Type: Dynamic (updated in real time if API is used) or static if using downloaded datasets

5.High-Level Methodology

- **Data Collection:** Using Twitter API, Reddit API, or public datasets.
- **Data Cleaning:** Removing noise like stop words, hashtags, links, emojis, and duplicates.
- **Exploratory Data Analysis (EDA):** Creating word clouds, sentiment graphs, and trend plots.
- **Feature Engineering:** Representing text using TF-IDF, Word2Vec, GloVe, or BERT encodings.
- **Model Building:** Using models like Logistic Regression, Naive Bayes, Random Forest, LSTM, and BERT to classify **sentiment/emotions.**
- **Model Evaluation:** Measuring accuracy, precision, recall, F1-score, and confusion matrix.
- **Visualization:** Presenting insights using Seaborn, Plotly, and Matplotlib.
- **Deployment:** Creating an interactive dashboard with Streamlit or Gradio, or providing a well-documented notebook.

6.Tools and Technologies

Programming Language: Python

Development Environment: Google Colab or Jupyter Notebook

Key Libraries:

- *Data Processing*: pandas, numpy, re
- *Visualization*: matplotlib, seaborn, plotly, wordcloud

- *Modeling & NLP*: scikit-learn, tensorflow, transformers, nltk, textblob

Deployment Tools: Stream lit, Gradio.

7.Team Members and Roles

S.NO	NAME	ROLE
1.	Sandhiya M	Data Collection, Preprocessing, and Exploratory Analysis
2.	Devadharshini V	Feature Engineering and Model Development
3.	Sharon Cynthiya J	Model Evaluation and Performance Tuning
4	Badmasri V	Data Visualization and Insight Presentation
5.	Keerthika D	Dashboard Deployment and Final Report Compilation