Phase-1 Submission

Student Name: *Subash S*

Register Number: *712523106018*

Institution: *ppg institute of technology*

Department: *B.E. Electronics and Communication Engineering*

Date of Submission: 30/04/2025

1.Problem Statement

Social media captures vast public emotions and opinions but is often unstructured. Understanding these emotions is crucial for businesses, policymakers, and mental health experts. This project applies sentiment analysis to decode emotions from social media data. It enables better decision-making and real-time insight into public sentiment.

2.Objectives of the Project

- *Build a model to detect and classify emotions from social media text.*
- *Collect and preprocess data from platforms like Twitter and Reddit.*
- *Analyse emotional trends across time, topics, or events.*
- *Visualize insights through dashboards for easy interpretation.*
- *Deliver actionable insights to support informed decision-making.*

3.Scope of the Project

Features to be Analysed / Built

- *Collection of real-time or historical social media data (e.g., tweets, Reddit posts).*
- *Text preprocessing (tokenization, stop-word removal, lemmatization).*
- *Emotion and sentiment classification (e.g., happy, sad, angry, neutral).*
- *Visualization of emotion trends over time or in response to events.*

Limitations and Constraints

- *Analysis limited to English-language text (due to model/language constraints).*
- *Dependence on publicly available datasets and APIs (e.g., Twitter API).*
- *Sentiment detection may not fully capture sarcasm, slang, or cultural nuances.*
- *Deployment limited to offline/demo environments (no full-scale production).*
- *Use of predefined NLP models (e.g., BERT, ROBERTA) due to time/resource constraints.*

4.Data Sources

The project will utilize both static and dynamic datasets to analyse emotional sentiment in social media conversations.

Social media APIS (Dynamic Data):

Sentiment140, Emotion Dataset from Kaggle, or GoEmotions (Google) will be used for training and validation.

5.High-Level Methodology

Data Collection *–* ***Sources:*** *Data will be obtained from public APIs like Twitter API and Reddit API for real-time social media posts.*

Supplementary Datasets: *Pre-labelled datasets such as Sentiment140, GoEmotions, and Kaggle emotion datasets will be downloaded for training and validation.*

Tools: *Python libraries like TWEEPY or PRAW will be used for data extraction from APIs.*

Data Cleaning

Issues Addressed: *Handling missing values, removing duplicates, cleaning special characters, links, hashtags, emojis, and stop words.*

Approach: *Text normalization (lowercasing, lemmatization), tokenization using NLTK or spacy, and removal of irrelevant content (e.g., advertisements or spam).*

Exploratory Data Analysis (EDA)

Techniques: *Frequency analysis, word clouds, sentiment distribution charts, and time-based trend graphs.*

Tools: *matplotlib, seaborn, portly, and pandas for visualizing emotion/sentiment patterns across different topics or time periods.*

Feature Engineering

Text Features: *TF-IDF vectors, word embeddings (Word2Vec, Glove), or transformer-based embeddings (BERT).*

Additional Features: *Post length, use of punctuation, and emojis may be considered to enhance emotional context.*

Model Building

Traditional ML*: Logistic Regression, Naive Bayes, SVM*

***Deep Learning**: LSTM, Bi-LSTM*

***Transformer Models:** BERT, Roberta for emotion classification*

***Justification:** Transformer models like BERT are well-suited for understanding context and emotion in short, informal text like tweets or comments.*

Model Evaluation

***Metrics:** Accuracy, Precision, Recall, F1-Score, Confusion Matrix*

***Validation Strategy:** Train-test split or k-fold cross-validation to ensure robustness and avoid overfitting.*

Visualization & Interpretation

***Output Presentation:** Interactive dashboards, time-series plots, and emotion heatmaps to display trends and model outputs.*

***Tools:** Stream lit or Dash for dashboards; seaborn, matplotlib, and plotly for visual analysis.*

Deployment

***Mode:** A simple interactive web app or notebook-based demo will be developed.*

***Tools:** stream lit, Flask, or a JUPYTER Notebook with embedded visualizations to allow users to explore results and test live inputs.*

6.Tools and Technologies

Programming Language: *Python, due to its extensive support for data analysis, natural language processing (NLP), and machine learning (ML) tasks.*

Notebook/IDE: *GOOGLE COLAB and JUPYTER Notebook*

Libraries:

- ***Data Processing:*** *Pandas, NumPy, re*
- ***Natural Language Processing:*** *NLTK and spacy, Text Blob and VADER, Transformers*
- ***Data Visualization:*** *Matplotlib and seaborn, Word cloud*
- ***Modelling and Machine Learning:*** *scikit-learn, TensorFlow/Py torch*

7. Team Members and Roles

TEAM MEMBER	***ROLE***	***RESPONSIBILITIES***
SUBASH. S	*Team leader, reviewer & tester*	*Plans and coordinates tasks, main Conduct testing, report bugs, provide suggestions, and contribute to final report and presentation.*
RAJESHKUMAR.N	*coordinator*	*Assists analyses, report and reviews code* and documents the Process
KAMALESH.S	*Frontend & Documentation Support*	*assists in building the UI with Stream lit maintains ticket logs.*
PRASANNA.S. D	*Research Engineering*	Analyses data research and data collection.
SANJAY. E	*Data & Research* Specialist	*Collect datasets, support Data preprocessing and Documentation.*