Phase-1SubmissionTemplate

Student Name: Hemachandran G

RegisterNumber:410723104024

Institution: Dhanalakhmi College of Engineering

Department: Computer Science

Date of Submission: 28.04.2025

Decoding emotion through sentimental analysis of social media conversations

1. ProblemStatement

- With millions of people expressing their thoughts daily on social media, understanding public emotions has become both valuable and challenging. Traditional sentiment analysis often misses the complexity of human emotions.
- This project aims to decode nuanced emotions—like joy, anger, and sadness—from social media conversations using advanced sentiment analysis. Accurate emotion detection can support better decision-making in areas like mental health, marketing, and public policy.

2. Objectives of the Project

- To analyze and classify emotions from social media conversations using sentiment analysis.
  The project aims to detect emotions like joy, anger, sadness, and fear through NLP techniques.
  It helps in understanding public sentiment for applications in marketing, health, and policy-making.

3. Scope of the Project

- The project aims to decode human emotions by applying sentiment analysis

techniques to social media conversations.
It focuses on extracting, classifying, and interpreting emotional cues from user-generated content.
Major platforms like Twitter, Facebook, and Instagram will be analyzed for public posts and comments.

- Natural Language Processing (NLP) and machine learning models will be employed for accurate sentiment detection.
  Emotions will be categorized into primary types such as joy, anger, sadness, fear, and surprise.
  The analysis will help identify emotional trends, public opinion, and potential societal impact.
  This research can support applications in marketing, public policy, and mental health monitoring.

4. Data Sources

- This project will use publicly available social media data from platforms like Twitter, Reddit, and Facebook.
  Twitter's API can provide real-time and historical tweets for emotion analysis.
  Reddit comments and posts offer rich conversational data with diverse emotional expressions. Datasets like Kaggle sentiment datasets may support model training .

5. High-Level Methodology

- **DataCollection**- Data will be obtained via **API access** to platforms like Twitter (X) or Reddit using official developer APIs. Where necessary, **web scraping** techniques will be used for publicly available data, adhering to platform policies.
- **DataCleaning**–
  Gather a balanced dataset with positive, negative, and neutral examples
  Include diverse social media conversations (tweets, comments, posts)
- **Text Cleaning**:
  Remove special characters, URLs, mentions
  Handle emojis (convert to text or remove)
  Expand contractions (I'm → I am)
  Correct common misspellings

Text Normalization:

Convert to lowercase
Remove stopwords
Perform lemmatization/stemming.

- **Exploratory Data Analysis (EDA)** –
- we would examine:
  Class distribution (positive/negative/neutral)
  Word frequency distributions
  Word clouds for each sentiment class
  Text length distributions
  Correlation between text features and labels

- **FeatureEngineering**–we will create new features such as sentiment scores, emoji sentiment indicators, and hashtag frequency to capture emotional nuances. Existing text data will be transformed using techniques like TF-IDF, word embeddings (e.g., Word2Vec), and lemmatization to enhance model understanding. These transformations aim to improve the accuracy and depth of emotion classification.

- **Model Building** – we plan to experiment with models like Logistic Regression and Random Forest for baseline classification due to their simplicity and interpretability. These models can effectively learn

emotional patterns and complex language structures present in social media conversations.

- **ModelEvaluation**–

Key metrics to track:

Accuracy
Precision, Recall, F1-score (especially for minority classes)
Confusion matrix
ROC-AUC for binary classification

- **Visualization&Interpretation**–

Important visualizations:
Confusion matrix heatmap
ROC curves
Precision-Recall curves
Feature importance (for interpretable models)
Word clouds by sentiment class.

- **Deployment** – We plan to deploy the project as an interactive web app or dashboard using tools like Streamlit or Dash for real-time sentiment visualization. The backend will be supported by Python and integrated with preprocessed data and trained models for smooth interaction.

6. ToolsandTechnologies

We plan to use **Python** as the primary programming language due to its rich ecosystem for data science. Key libraries include **Pandas** and **NumPy** for data manipulation, **NLTK**, **spaCy**, and **TextBlob** for text preprocessing and sentiment analysis, and **scikit-learn**, **TensorFlow**, and **PyTorch** for building and training models. Visualization will be handled with **Matplotlib**, **Seaborn**, and **Plotly**, while deployment tools like **Streamlit** or **Dash** will be used to create interactive dashboards.

- **ProgrammingLanguage**–The primary programming language for *Decoding Emotions through Sentiment Analysis of Social Media Conversations* will be **Python or SAS**. It is widely used in data science for its extensive libraries and frameworks that support natural language processing, machine learning, data visualization and data analytics.
- **Notebook/IDE**–**Jupyter Notebook** will be used for local development and visualization of results.
  VS Code may be utilized for writing and organizing modular Python scripts.
  (These platforms support integration with popular libraries like TensorFlow, PyTorch, and scikit-learn for sentiment analysis.)

- **Libraries** –Several key libraries will be used to handle data collection, processing, modeling, and visualization. For sentiment and emotion analysis, libraries like **TextBlob** and **Transformers** from Hugging Face will enable the use of both simple sentiment tools and advanced pre-trained models like BERT. Additionally, **Pandas** and **NumPy** will handle data manipulation, while **Matplotlib** and **Seaborn** will be used for visualizing insights and model performance.).

7. **Optional Tools for Deployment**– Tools and frameworks like **Streamlit** and **Gradio** will be considered for creating interactive and user-friendly web interfaces.

   To build a proper sentiment analysis system:
   Collect a balanced dataset with real social media conversations
   Include multiple sentiment classes (at least positive/negative/neutral)
   Ensure diversity in text length, topics, and writing styles
   Consider multi-label classification for complex emotions

8. TeamMembersandRoles

Team member names	Roles	Responsibility
K G Deepaprakesar	Leader	Data collecting
Lohit A S	Member	Data cleaning

Karthick V	**Member**	**Data preparation**
Hemachandran G	**Member**	**Data visualization**