Project Title: Decoding Emotions Through Sentiment Analysis of Social Media Conversations

PHASE-1

1. Problem Statement

In the digital age, social media platforms have become key outlets for public expression. Analyzing these vast conversational datasets offers unique opportunities to understand societal emotions in real-time. This project focuses on decoding emotional states from user-generated content using sentiment analysis techniques. By identifying emotional patterns and trends in online conversations, we aim to provide actionable insights for businesses, mental health monitoring, and public sentiment tracking.

2. Objectives of the Project

- To develop a sentiment analysis model capable of classifying emotional tones in social media conversations.

- To identify prevalent emotional states (e.g., joy, anger, sadness) expressed in public discourse.

- To provide visual insights into emotional trends and shifts over time or in response to specific events.

- To create an interactive dashboard or notebook that showcases sentiment distribution and emotional dynamics.

3. Scope of the Project

Features to Analyze:

- Textual data from social media (e.g., tweets, Facebook posts, Reddit comments)

- Metadata (e.g., timestamps, hashtags, user engagement)

Constraints and Limitations:

- Only publicly available datasets will be used.

- Text-based sentiment; non-verbal cues (e.g., emojis, videos) may not be deeply integrated.

- Focused on English language content.

- Classification models may be limited to predefined emotional categories.

4. Data Sources

- Dataset: Social media posts with labeled sentiment/emotion (e.g., Sentiment140, GoEmotions)

- Source: Public repositories like Kaggle, Google Research

- Type: Public

- Nature: Static dataset downloaded and used throughout the project

- Example Link: https://github.com/google-research/google-research/tree/master/goemotions

5. High-Level Methodology

Data Collection

- Acquire datasets from open repositories

- Load into Google Colab using pandas

Data Cleaning

- Remove special characters, links, and stopwords

- Tokenization and lemmatization

- Handle imbalanced datasets using techniques like SMOTE

Exploratory Data Analysis (EDA)

- Visualize sentiment distribution

- Word clouds and n-gram analysis

- Analyze trends across time and topics

Feature Engineering

- Convert text to numerical features using TF-IDF or word embeddings

- Include sentiment lexicon scores (e.g., VADER, TextBlob)

Model Building

- Use models like Logistic Regression, Naive Bayes, LSTM, BERT

- Compare performance and select the most accurate model

Model Evaluation

- Metrics: Accuracy, Precision, Recall, F1-Score

- Use k-fold cross-validation for robustness

Visualization & Interpretation

- Emotion distribution charts and timelines

- Highlight key words contributing to emotion detection

Deployment

- Create a Streamlit or Gradio interface for real-time sentiment analysis

6. Tools and Technologies

- Programming Language: Python

- Notebook/IDE: Google Colab

- Libraries:

  - Data Handling: pandas, numpy

  - NLP & Sentiment: nltk, spacy, transformers, TextBlob

  - Visualization: matplotlib, seaborn, plotly

  - Modeling: scikit-learn, tensorflow, huggingface

- Tools for Deployment:

  - Streamlit or Gradio