Github Link:

Project Title:Decoding emotions through sentiment analysis of social media conversations

PHASE-2

1.Problem Statement:

In today's digital age, social media platforms serve as a significant outlet for people to express their thoughts, feelings, and opinions.

However, the massive and unstructured nature of social media data makes it challenging to accurately interpret the emotional states conveyed in these conversations.

Traditional sentiment analysis methods often focus only on basic polarity (positive, negative, neutral), failing to capture the nuanced spectrum of human emotions.

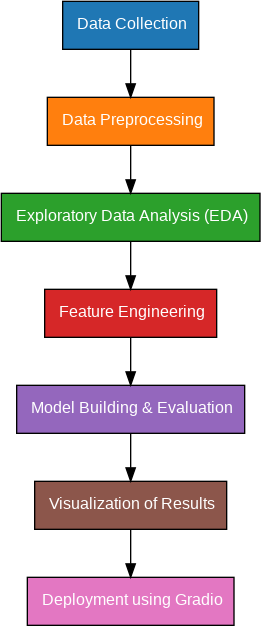
2.Project Objectives:

To collect and preprocess social media data from platforms such as Twitter, Reddit, or Facebook using APIs and web scraping tools.

To implement sentiment and emotion classification models that go beyond polarity detection by identifying specific emotions such as joy, anger, sadness, fear, surprise, and disgust.

To compare and evaluate various NLP and machine learning techniques (e.g., Naive Bayes, LSTM, BERT) for their effectiveness in emotion recognition.

3.Flowchart of the Project Workflow:



4.Data Description

The dataset for this study consists of user-generated text collected from multiple social media platforms, including Twitter, Reddit, and Facebook (public posts only), selected for their high user engagement and diverse content. The data includes posts, comments, replies, and hashtags, which reflect spontaneous emotional expression.

1. Source Platforms:

Twitter

Reddit

Facebook

2. Data Features

3. Data Size

4. Emotion Annotation

5. Preprocessing Steps

5.Data Preprocessing:

Effective preprocessing is crucial for preparing raw social media text for sentiment and emotion analysis. Social media data often contains noise such as slang, emojis, URLs, and grammatical inconsistencies. The following steps are applied to clean and standardize the data for accurate model training and evaluation.

1. Data Cleaning

2. Text Normalization

Spelling Correction.

Contractions Expansion

Lemmatization/Stemming

3. Tokenization

4. Stopword Removal

5. Emoji and Emoticon Handling

6. Slang and Abbreviation Expansion

7. Language Detection and Filtering

8. Handling Imbalanced Data

6.Exploratory Data Analysis:

EDA is essential for understanding the structure, patterns, and quality of the collected social media data before applying sentiment and emotion classification models. It helps in identifying trends, anomalies, and relationships in the dataset.

1. Data Overview

Dataset Size

Platform

Time Range

Text Length

2. Class Distribution

Emotion Labels

Imbalance Check

Visualization

3. Word Frequency Analysis

Unigrams and Bigrams

Emotion-Specific Keywords

4. Temporal Trends

5. Correlation Analysis

Engagement Metrics vs Emotion

Platform vs Emotion Distribution

6. Text Characteristics

7.Feature Eng

Feature engineering transforms raw text data into meaningful input representations that improve model performance for emotion classification. For sentiment analysis of social media conversations, both text-based and metadata features can be extracted.

1. Text-Based Features

2. Metadata Features (Optional)

3. Feature Selection

8.Model Building:

The goal of this stage is to develop machine learning and deep learning models capable of accurately classifying emotions in social media text. Multiple models are tested to compare their performance on emotion detection tasks.

1. Problem Formulation

2. Model Selection

3. Model Training

4. Model Evaluation

5. Model Deployment (Optional

9.Visualization of Results & Model Insights:

1. Confusion Matrix

2. Precision, Recall, and F1-Score Bar Plot (Per Class)

3. Class Distribution (Emotion Label Proportions)

4. ROC Curve

5. Model Training and Validation Curves

6. Feature Importance (For Tree-based Models)

7. Word Cloud (For Emotion-Related Keywords)

8. Distribution of Sentiment Scores

10.Tools and Technologies used:

1. Programming Language

Python

2. Libraries and Frameworks

a. Data Preprocessing and Analysis

Pandas

NumPy

NLTK (Natural Language Toolkit

spaCy

b. Text Vectorization

c. Machine Learning Algorithm

11.Team Members and Contributions:

* + *Clearly mention who worked on:*
    - *Data cleaning :M.Dhiliban*
    - *EDA :M.Dhiliban*
    - *Featureengineering :C.Durga devi*
    - *Model development :C.Durga devi*
    - *Documentation and reporting :A.Gogul*