GithubLink:https://github.com/Vaishnavi-149/Project-Decoding-Emotins-through-Sentiment-Analysis.git

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

PHASE-2

Student Name:VAISHNAVI M

Register Number:623023104056

Institution: Tagore Institute of Engineering and Technology -Salem

Department: Computer Science and Engineering

Date of Submission:08-05-2025

1. Problem Statement

In today's digital era, social media platforms have become a primary outlet for individuals to express thoughts, opinions, and emotions. However, the massive volume and unstructured nature of these conversations make it difficult to systematically understand and interpret the emotional landscape of users. This project aims to develop a robust sentiment analysis system that leverages machine learning and natural language processing techniques to accurately decode and classify emotions expressed in social media conversations. By doing so, the system will help identify emotional trends, support mental health monitoring, enhance customer feedback analysis, and contribute to a better understanding of public sentiment in real-time.

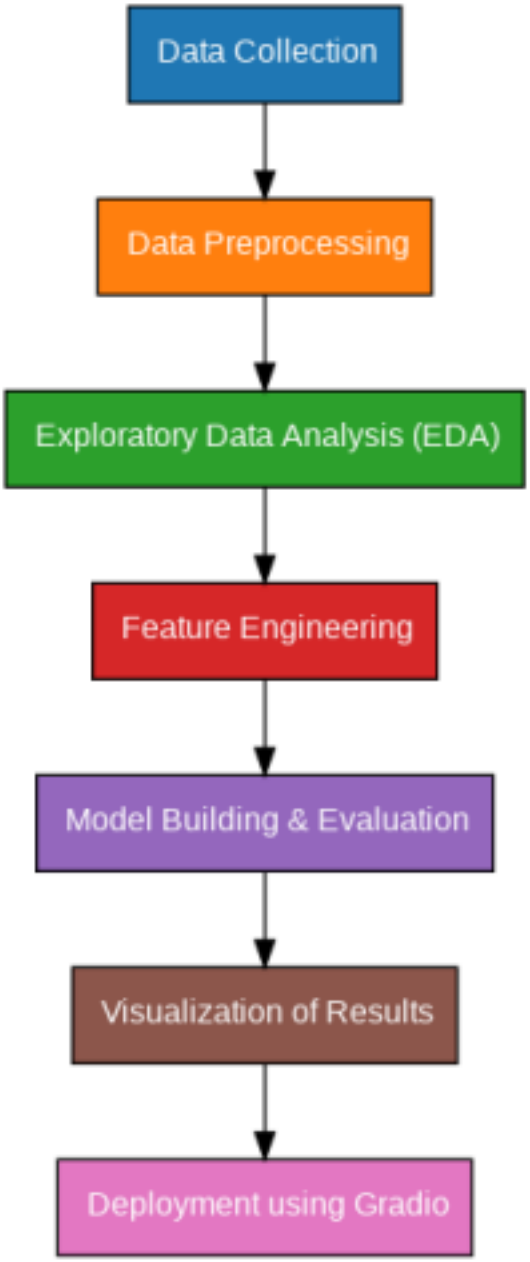
2. Project Objectives

- **To collect and preprocess social media text data** from platforms such as Twitter, Reddit, or Instagram, ensuring the data is cleaned, tokenized, and formatted for analysis.
- **To perform exploratory data analysis (EDA)** to identify patterns, trends, and distributions in emotional expressions across the dataset.
- **To extract relevant linguistic and contextual features** that contribute to accurate emotion and sentiment detection using NLP techniques such as TF-IDF, word

embeddings, and POS tagging.

- **To design and implement machine learning and deep learning model** (e.g.Logistic Regression, Random Forest, LSTM, BERT) to classify emotions into categories such as joy, anger, sadness, fear, surprise, etc..
- **To evaluate model performance** using metrics like accuracy, precision, recall, F1-score, and confusion matrix, and optimize the models for improved accuracy and generalization.

3. Flowchart of the Project Workflow



4. Data Description

- **Dataset Name**: Social Media Emotion & Sentiment Dataset

- **Source**: Aggregated from platforms like Twitter, Reddit, or Kaggle datasets

- **Type of Data**: Unstructured text data (with some structured metadata)
- **Records and Features:** social media posts with features including text, user

metadata, and timestamps

- **Target Variable**: Emotion label or sentiment category
- **Static or Dynamic**: Static dataset
- **Attributes Covered**:
  - **Textual Content**: The body of the post or tweet
  - **Emotion/Sentiment Labels**: Annotated manually or via model
  - **Dataset Link**: https://www.kaggle.com/code/adityaghuse/sentiment-analysis

5. Data Preprocessing

- Verified dataset integrity: Removed entries with missing or null text or label fields
- Removed irrelevant features: Dropped metadata fields like user ID, URL links, or timestamps that did not contribute to sentiment or emotion classification.
- Checked and confirmed absence of duplicate rows: Identified and removed exact duplicate posts/tweets to prevent bias.
- Text preprocessing:
  - Lowercased all text
  - Removed punctuation, stop words, URLs, mentions, and hashtags
  - Applied tokenization, lemmatization, and stemming

6. Exploratory Data Analysis (EDA)

- **Univariate Analysis:**
  - Bar plots showing the distribution of each emotion/sentiment category
  - Histogram of post lengths to understand typical content size
  - Word clouds for each emotion to visualize most frequent terms

● **Bivariate & Multivariate Analysis:**

○ Heatmap of feature correlations (e.g., between sentiment scores, text length, engagement metrics)

○ Boxplots showing text length variation across different emotion categories

○ Sentiment polarity vs. subjectivity scatter plots

○ Grouped bar charts comparing emotion distribution across platform or user groups

● **Key Insights:**

○ Certain keywords are highly associated with specific emotions

○ Longer posts tend to express more complex or mixed emotions

○ Positive emotions dominate in public posts, while negative emotions appear more in anonymous or support-related forums

○ Some emotions like *fear* and *disgust* are underrepresented and may require class balancing techniques

7. Feature Engineering

● Created interaction features: Combined sentiment polarity and subjectivity scores to form a composite emotional intensity metric

● Derived binary features: Generated binary flags such as has_question (1 if the post contains a question mark), contains_emoji, and is_uppercase_heavy to capture tone and expressiveness

● Removed highly correlated or redundant features: Dropped overlapping text-based features (e.g., both raw word count and TF-IDF count for the same word) to reduce noise and multicollinearity

- Performed label encoding: Encoded emotion labels as numerical classes for model compatibility

- Scaled numeric features: Applied StandardScaler to features like text length, sentiment scores, and word counts to normalize input across varying ranges

8. Model Building

- **Algorithms Used:**
  ○ Logistic Regression: for baseline emotion/sentiment classification ○ Random Forest Classifier: to handle complex feature interactions and provide feature importance ○ LSTM (Long Short-Term Memory): for deep learning-based modeling of text sequences and capturing contextual dependencies

- **Model Selection Rationale:**
  ○ Logistic Regression: simple, interpretable baseline for classification tasks ○ Random Forest: handles high-dimensional and mixed-type data well, robust against overfitting
  ○ LSTM: ideal for sequential text data, effective in capturing the semantic flow in sentences

- **Train-Test Split:**
  ○ 80% training, 20% testing
  ○ Used `train_test_split` with `random_state` for reproducibility

- **Evaluation Metrics:**
  ○ Accuracy: overall correctness of predictions
  ○ Precision, Recall, F1-Score: critical for imbalanced emotion classes
  ○ Confusion Matrix: to visualize correct and incorrect classifications per emotion/sentiment ○ ROC-AUC (if binary sentiment task): measures classification performance across thresholds

9. Visualization of Results & Model Insights

- **Feature Importance:**
  ○ Visualized top influential features using bar plots from the Random Forest and LSTM attention weights
  ○ Keywords, sentiment polarity, and word count ranked highest in importance for emotion prediction

- **Model Comparison:**
  ○ LSTM outperformed others in F1-score, especially for complex emotions

- **Error Analysis:**
  - Analyzed misclassified examples to identify patterns in text

- **User Testing**:
  - Deployed the model using a Streamlit or Gradio interface
  - Allowed users to input custom text and view predicted emotion along with probability scores and word attention highlights

10. Tools and Technologies Used

- **Programming Language**: Python 3

- **Notebook Environment**: Google Colab

- **Key Libraries:**
  - pandas, numpy – for text data handling and transformation
  - matplotlib, seaborn, plotly – for visualizing distributions, correlations, and model performance
  - scikit-learn – for text preprocessing (e.g., TF-IDF), model training, and evaluation
  - NLTK, spaCy – for natural language processing (tokenization, lemmatization, stopword removal)
  - TextBlob / VADER – for sentiment scoring and polarity analysis
  - TensorFlow / Keras – for building deep learning models like LSTM
  - Gradio − for deploying an interactive web interface for real-time emotion prediction

11. Team Members and Contributions

- ***Sindhuja – Data Preprocessing and Exploratory Analysis***

  Responsible for cleaning the raw social media data, handling missing values, and performing exploratory data analysis (EDA) to uncover key patterns and insights.

- ***Suba – Feature Engineering and Transformation***

  Focused on extracting meaningful features from text data, including sentiment scores, keyword flags, and vector representations using NLP techniques.

- ***Vaishnavi – Model Design and Implementation***

  Led the development of machine learning and deep learning models (e.g., Random Forest, LSTM) for emotion classification, including model tuning and evaluation.

- ***Thayasri – Documentation and Final Report Preparation***

  Managed the overall project documentation, compiled visualizations and results, and

prepared the final technical report and presentation materials.