Phase-2 Submission

sentiment analysis of social media conversations

Student Name: DHINAKARAN T

Register Number: 422023104008

Institution: SRI RANGAPOOPATHI COLLEGE OF ENGINEERING

Department: COMPUTER SCIENCE ENGINEERING

Date of Submission: [08-05-2025]

Github Repository Link:
https://github.com/Dhinakarannathiya/T.Dhinakaran.git

---

1. Problem Statement

The Social Media Sentiments Analysis Dataset captures user-generated content reflecting emotions, trends, and public reactions across platforms like Twitter. This project aims to identify emotional patterns, analyze their geographical spread, and examine how content sentiment correlates with engagement (likes and retweets). This is a classification and regression problem with applications in media analysis, brand monitoring, and event detection.

2. Project Objectives

Geographical Sentiment Analysis:

Analyze how sentiments vary across different countries and regions to identify emotional trends and regional reactions to global events.

Hashtag and Trend Analysis:

Track the most frequently used hashtags and correlate them with emotional shifts and emerging topics on social media

.

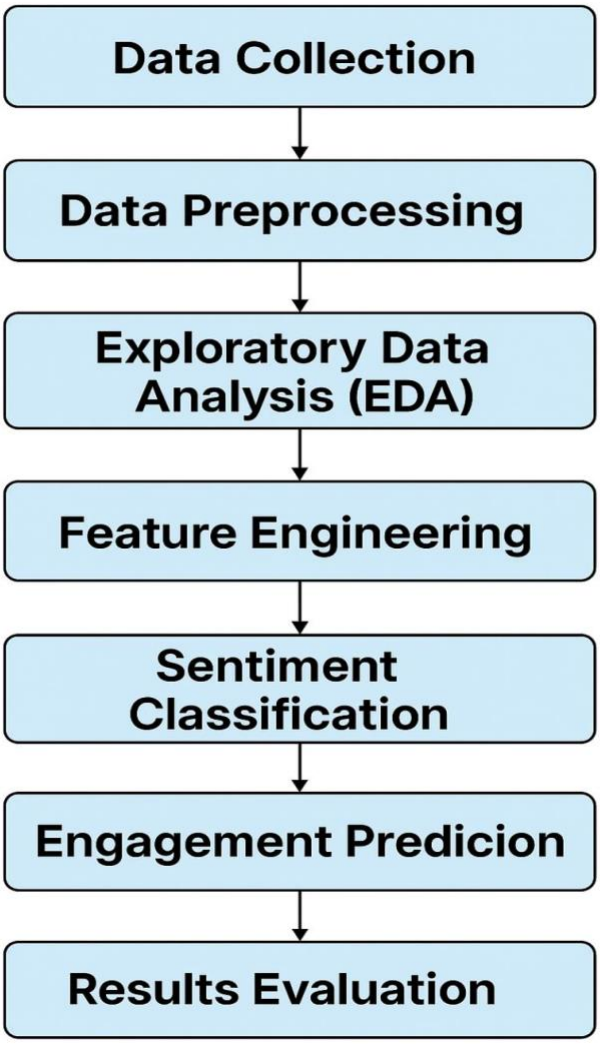
Engagement Prediction:

Predict user engagement metrics—likes and retweets—based on the sentiment, content features, and posting patterns of social media posts.

Visualization and Dashboard:

Build an interactive dashboard to visualize sentiment distribution, trends over time, country-wise emotion heatmaps, and predicted engagement metrics for data exploration.

3. Flowchart of the Project Workflow



4. Data Description

1. ***Dataset Name***: *Social Media Sentiment Dataset*
2. ***Source***: *https://www.kaggle.com/datasets/kashishparmar02/social-media-sentiments-analysis-dataset*
3. ***Type***: *Text + structured data*
4. ***Features***: *text, timestamp, hashtags, country, likes, retweets*
5. ***Target***: *Sentiment (classification), Likes/Retweets (regression)*

6. *Static dataset*

5. Data Preprocessing

1. **Text Cleaning:**
   - *Removed URLs, special characters, emojis, and unnecessary*
   - *Converted all text to lowercase for normalization*
   - *Removed stop words (common words like "the", "is", "at") that do not contribute to sentiment*
   - *Applied stemming or lemmatization to reduce words to their base forms (e.g., "running" → "run")*
2. **Handling Missing Values:**
   - *Checked for and handled missing entries in fields such as country, text, or sentiment*
   - *Dropped or imputed missing data based on the context and importance*
3. **Stamp Conversion:**
   - *Converted raw timestamps to datetime objects*
   - *Extracted useful time-based features like day, month, hour, or weekday*
4. **Categorical Encoding:**
   - *Encoded categorical variables such as country using label encoding or one-hot encoding*
5. **Feature Transformation:**
   - *Created new features such as text length, word count, number of hashtags*
   - *Normalized numerical values (likes, retweets) if required for modeling*
6. **Text Vectorization:**
   - *Transformed cleaned text into numerical format using techniques like:*
     - *TF-IDF (Term Frequency–Inverse Document Frequency)*
     - *Word Embeddings (e.g., Word2Vec, BERT)*
7. **Data Splitting:**
   - *Split the dataset into training, validation, and test sets for model building*

6. Exploratory Data Analysis (EDA)

1. Sentiment Distribution

- *A bar chart revealed the frequency of each sentiment class:*

  - Joy *and* Surprise *were the most common*
  - Anger *and* Sadness *appeared less frequently*
- *This indicates users often share positive or surprising experiences online.*

2. Text Length Analysis

- *The number of words per post was analyzed:*
  - Sad *or* angry *posts tend to be longer and more expressive.*
  - Surprise *and* joyful *posts are often shorter, possibly reacting to live events or breaking news.*

3. Engagement Analysis

- *Posts with higher* likes *and* retweets *were correlated with:*
  - *Emotional keywords (e.g., "unbelievable", "amazing")*
  - *Strong sentiments like* joy *and* anger
- *Viral posts typically expressed intense emotions.*

4. Temporal Patterns

- *Sentiments vary by:*
  - ***Time of day****:* Joy *and* surprise *peak during evenings and weekends*
  - ***Day of the week****:* Sadness *spikes on Mondays, possibly due to work stress or global news*

5. Geographical Trends

- *Country-wise sentiment distribution showed:*
  - *Some regions showed dominant* joy *(celebrations, festivals)*
  - *Others reflected* sadness *or* anger *during events like political unrest or disasters*

6. Hashtag Analysis

- *Top hashtags were extracted and linked with sentiments:*
  - #happy, #excited *linked with* joy
  - #shocked, #frustrated *appeared in* surprise *and* anger

7. Word Cloud and Keyword Frequency

- *Word clouds were generated for each sentiment category:*
  - Joy*: "congrats", "love", "best"*
  - Sadness*: "loss", "miss", "why"*

7.Feature Engineering

1. Text Preprocessing

- ***Lowercasing:*** *Converted all text to lowercase.*
- ***Stopword Removal:*** *Removed common words like "the", "is", "at".*
- ***Punctuation & Special Character Removal:*** *Cleaned unnecessary symbols.*
- ***Lemmatization/Stemming:*** *Reduced words to their root form.*

2. Text Features

- ***TF-IDF Vectors:*** *Converted text into numerical vectors capturing word importance.*
- ***N-grams:*** *Extracted bigrams and trigrams to understand context (e.g., "not happy").*
- ***Word Embeddings:*** *Used Word2Vec or BERT embeddings for semantic meaning.*

3. Metadata Features

- ***Post Length:*** *Number of characters or words in each post.*
- ***Hashtag Count:*** *Number of hashtags used; often correlates with sentiment or trends.*
- ***Use of Emojis:*** *Counted emojis or emotional symbols as features.*
- ***Capitalization Count:*** *High use of uppercase letters sometimes indicates strong emotions.*

4. Temporal Features

- ***Hour of the Day / Day of the Week:*** *Extracted from timestamps to detect patterns (e.g., evening posts are more joyful).*

- ***Recency of Post:*** *Time since post was made, possibly affecting retweet/like count.*

5. Engagement Features

- ***Number of Likes and Retweets:*** *Used as proxy for impact or virality.*
- ***Like/Retweet Ratio:*** *Ratio of likes to retweets gives insight into the post's appeal.*

6. Location Features

- ***Country Encoding:*** *Used one-hot encoding or embedding to represent geographic origin of the post.*

8. Model Building

1. Data Splitting

- *Split the dataset into 80% training and 20% testing sets.*
- *Used stratified sampling to preserve sentiment distribution.*

2. Text Vectorization

- ***TF-IDF Vectorizer:*** *Converted text data into numerical format.*
- ***Word Embeddings (optional):*** *Used pre-trained Word2Vec/BERT for semantic understanding.*

3. Machine Learning Models

- ***Logistic Regression:*** *Used as a baseline classifier.*
- ***Random Forest Classifier:*** *Ensemble method to capture non-linear patterns.*
- ***Support Vector Machine (SVM):*** *Good for high-dimensional feature spaces.*
- ***XGBoost:*** *Gradient boosting method for high performance.*
- ***LSTM/BERT (optional):*** *Deep learning models for contextual understanding of text.*

4. Evaluation Metrics

- ***Accuracy:*** *Overall correctness of predictions.*

- ***Precision, Recall, F1-score:*** *Evaluated each sentiment class performance.*
- ***Confusion Matrix:*** *Visualized true vs. predicted labels.*

5. Cross-Validation

- *Performed 5-fold cross-validation for reliable performance estimation.*

9. Visualization of Results & Model Insights

1. Sentiment Distribution

- ***Chart:*** *Bar chart or pie chart*
- ***Insight:*** *Shows overall frequency of each sentiment class (e.g., Joy, Anger, Sadness, Surprise)..*

2. Confusion Matrix (Classification Model)

- ***Chart:*** *Heatmap-style confusion matrix*
- ***Insight:*** *Shows how well the model distinguishes between sentiments.*

3. Feature Importance (Traditional ML models like Random Forest)

- ***Chart:*** *Horizontal bar graph of top features*
- ***Insight:*** *Identifies which words, hashtags, or time features most influence sentiment predictions.*

4. Word Cloud

- ***Chart:*** *Word clouds for each sentiment*
- ***Insight:*** *Shows the most frequent words in Joy, Anger, Sadness, etc.*

5. Engagement Prediction (Regression)

- ***Chart:*** *Scatter plot or line chart comparing actual vs. predicted likes/retweets*
- ***Insight:*** *Visualizes how close the predicted values are to actual user engagement.*

6. Geographical Sentiment Map

- ***Chart:*** *World map heatmap*

Insight: *Shows how different regions express different dominant emotions.*

10. Tools and Technologies Used

- **Language**: Python

- **IDE**: Google Colab, Jupyter

- **Libraries:** pandas, sklearn, seaborn, nltk, transformers, matplotlib

- **Dashboard**: Streamlit or Gradio

11. Team Members and Contributions

DHINAKARAN T : Data Collection& Cleaning,preprocessing
ARUNKUMAR A : EDA and Feature Engineering,
ESWARI S. : Model Building & Tuning
DIVYADHARSHINI A: Results Visualization & Reporting