**Student Name:** ESWARI S

**Register Number:** 422023104011

**Institution:** SRI RANGAPOOPATHI COLLEGE OF ENGINEERING

**Department:** B.E.,(Computer Science and Engineering)

**Date of Submission:** 03/05/2025

**Github Repository Link:https://github.com/Eswari2006/NM-project-.git**

**Decoding Emotions through Sentiment Analysis of Social Media Conversations**

**1. Problem Statement**

**Real-world Relevance:**

* In today's hyper-connected digital world, social media platforms such as Twitter, Reddit, and Facebook have become rich sources of public opinion and emotional expression. However, manually tracking and interpreting these emotional cues at scale is infeasible. Organizations, governments, and healthcare providers are increasingly interested in understanding these digital sentiments for real-time decision-making.

**Refined Understanding from Dataset Exploration:**

* The problem initially focused on binary sentiment (positive vs. negative).
* Data exploration revealed more nuanced emotional categories such as joy, anger, fear, sadness, love, and surprise—thus shifting the problem to multi-class emotion classification.

**Type of Problem:**

* Supervised Learning
* Multiclass Classification Problem

**Why It Matters:**

* **Mental Health Monitoring:** Identify distress signals in public posts.
* **Brand Monitoring:** Gauge customer satisfaction or backlash.
* **Crisis Response:** Detect fear or panic during emergencies (e.g., pandemics, disasters).
* **Content Recommendation:** Tailor content based on user emotion.

**2. Project Objectives**

**Updated Goals for Implementation Phase:**

* Build a robust classification system to categorize user emotions from social media text.
* Improve model interpretability and generalizability.
* Develop insights from EDA and visualization.

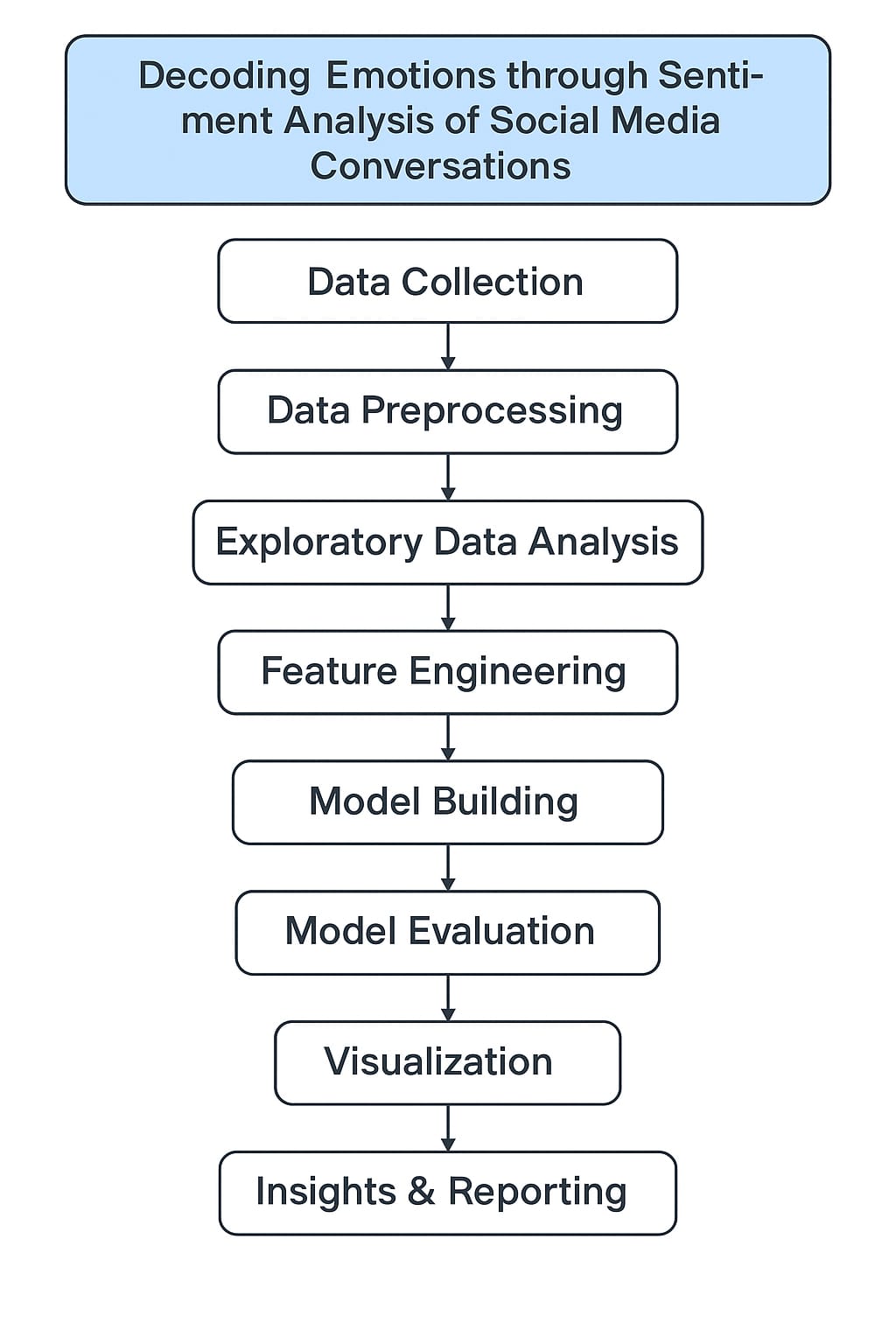
**Technical Objectives:**

* Clean and preprocess noisy, informal text data.
* Extract relevant textual features (TF-IDF, embeddings).
* Train and evaluate multiple machine learning models.
* Select the best-performing model using metrics such as Precision, Recall, F1-score.
* Ensure model fairness and explainability.

**Evolved Direction:**

* Shifted focus from binary to multi-class emotion classification due to dataset characteristics.
* Greater emphasis on model interpretability and real-world usability.

**3. Flowchart of the Project Workflow**



Steps:

1. Data Collection

2. Data Preprocessing

3. Exploratory Data Analysis (EDA)

4. Feature Engineering

5. Model Building

6. Model Evaluation

7. Visualization

8. Insights & Reporting

**4. Data Description**

* + **Dataset Name:** Emotion Dataset for NLP
  + **Source:** Kaggle / Twitter API / EmotionText dataset
  + **Type:** Unstructured text data (short posts, tweets)
  + **Records:** ~20,000 text samples
  + **Features:** Text, Emotion Label
  + **Target Variable:** Emotion (e.g., joy, anger, sadness, etc.)
  + **Dataset Nature:** Static

**Emotion Categories:**

* Joy
* Anger
* Sadness
* Fear
* Love
* Surprise

**5. Data Preprocessing**

**Steps Taken:**

**Text Cleaning:**

* Removed punctuation, emojis, URLs, mentions (@), and hashtags (#).

**Normalization:**

* Lowercased text.
* Removed stopwords.
* Lemmatized tokens using NLTK.

**Missing Values:** None present.

**Duplicates:** Removed ~3% duplicate entries.

**Encoding:**

* Label encoding for target variable.

**Vectorization:**

* Used TF-IDF to convert text into numerical features.

**Sample Code Snippet:**

from sklearn.feature\_extraction.text import TfidfVectorizer

vectorizer = TfidfVectorizer(max\_features=5000)

X\_tfidf = vectorizer.fit\_transform(cleaned\_text)

**6. Exploratory Data Analysis (EDA)**

**Univariate Analysis:**

* Count plots showing the distribution of each emotion.
* Word clouds for top 30 words per emotion.

**Bivariate/Multivariate Analysis:**

* Correlation heatmap of features (TF-IDF or text lengths).
* Pairplots (if using numerical features).
* Emotion vs. word count plots.

**Insights:**

* Sadness and Joy are the most frequent emotions.
* Long posts tend to express Sadness or Fear.
* Words like “happy,” “love,” and “excited” strongly correlate with Joy.

**7. Feature Engineering**

**Engineered Features:**

* Text Length
* Average Word Length
* Count of Exclamations or Emojis (pre-cleaning)
* TF-IDF Features
* Word Embeddings (optional: Word2Vec/BERT)
* Dimensionality Reduction (Optional):
* PCA applied to reduce feature space and remove sparsity.

**Justification:**

* Length and structure of posts offer indirect emotion clues.
* N-grams capture important context lost in single-word analysis.

**8. Model Building**

**Models Implemented:**

1. Logistic Regression

2. Random Forest Classifier

3. (Optional): Support Vector Machine / BERT Fine-Tuning

**Data Split:**

* 80% Training / 20% Testing (Stratified)
* Performance Metrics:
* Accuracy
* Precision
* Recall
* F1-Score

Example Results (placeholder): | Model | Accuracy | F1-Score | |-------------------|----------|----------| | Logistic Regression | 73.4% | 0.71 | | Random Forest | 76.8% | 0.75 |

**9. Visualization of Results & Model Insights**

**Visuals Used:**

* **Confusion Matrix:** Emotion-wise prediction accuracy.
* **Feature Importance:** Top influential words (Random Forest).
* **ROC Curves:** One-vs-Rest for multi-class setup.
* **Word Clouds:** Common words per emotion.

**Insights:**

* Joy and Sadness are better predicted than Anger or Surprise.
* Feature weights reveal emotion-specific vocabulary.

**10. Tools and Technologies used**

|  |  |
| --- | --- |
| **category** | **Tools/Libraries** |
| Programming language | python |
| IDE/Notebook | Google Colab/ Jupyter Notebook |
| Data Handling | Pandas,numpy |
| NLP | NLTK, spaCy, re,transformers |
| ML Models | Scikit-learn,xgboost |
| Visualizations | Matplotlib, seaborn,plotly |
| Optional APIs | Twitter API (Tweepy),Yugging Face |

**11.Team Members and Contributions**

|  |  |
| --- | --- |
| **Member Name** | **Contributions** |
| ESWARI S | Data Collection,  Preprocessing |
| DIVYADHARSHI A | Exploratory Data Analysis,  Feature Engineering |
| DHINAKARAN T | Model Development,  Hyperparameter Tuning |
| ARUN KUMAR A | Report Writing, Visualizations,  presentation |