Phase-2

Student Name: Yashwanth KS

Register Number: 410723104098

Institution: Dhanalakshmi College of Engineering

Department: computer science engineering

Date of Submission: 05-05-2025

Github Repository Link: https://github.com/KsYashwanth1109/Nm-yashwanth

---

1. Problem Statement

With the exponential growth of social media platforms like Twitter, Facebook, and Instagram, people now express their thoughts, opinions, and emotions online more than ever. These platforms have become a rich repository of user-generated content, which provides unique opportunities to understand public sentiment across various domains such as politics, mental health, marketing, and customer service.

2. Project Objectives

1. To collect a diverse dataset of social media text data such as tweets or Facebook comments.

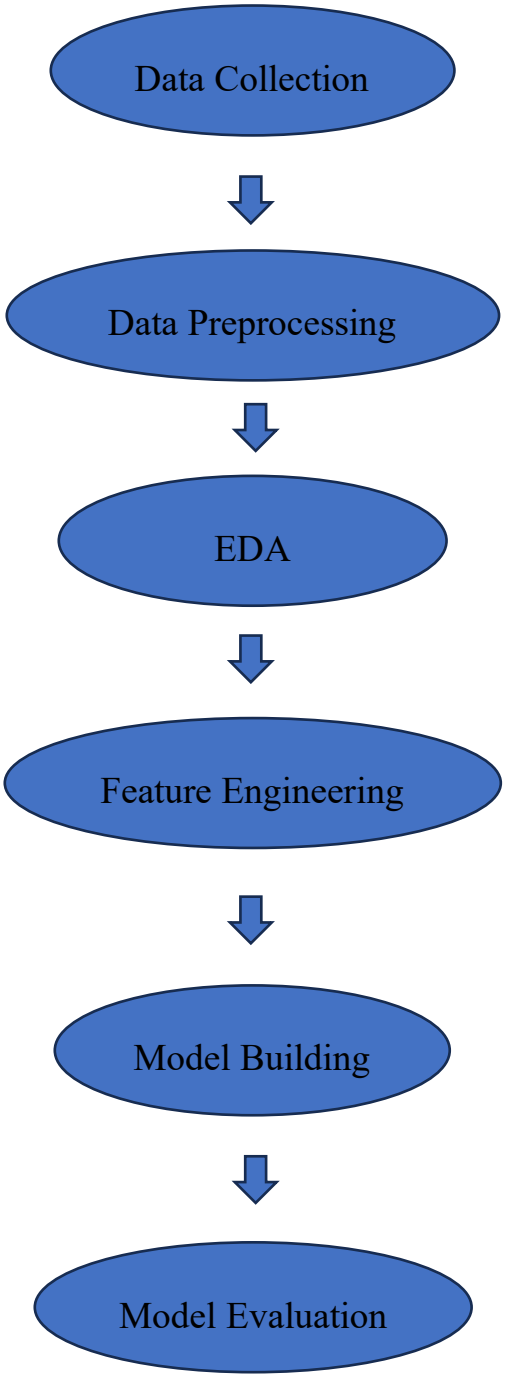
2. To preprocess the textual data by cleaning, normalizing, and tokenizing the content.

3. To build models capable of performing sentiment classification with high accuracy and robustness.

4. To visualize emotional trends across time, location, or topics to better understand public sentiment.

5. To address challenges such as sarcasm detection, handling class imbalance, and dealing with multilingual content.

3. Flowchart of the Project Workflow



4. Data Description

Dataset Source: Data is collected from the Twitter API, Reddit threads, or publicly available sentiment datasets on platforms like Kaggle.

Data Type: Unstructured text data containing user-generated content.

Features: Tweet/comment text, timestamps, user metadata (optional), sentiment labels.

Target Variable: Sentiment classification—commonly Positive, Negative, and Neutral.

Nature of Data: Can be static (archived datasets) or dynamic (real-time data from social APIs).

5. Data preprocessing

Text Cleaning: Remove punctuation, numbers, stop words, URLs, hashtags, and emojis.

Text Normalization: Lowercasing, correcting spelling errors, and expanding contractions.

Tokenization and Lemmatization: Breaking text into individual words and reducing them to their root form.

Handling Imbalanced Classes: Apply oversampling (e.g., SMOTE) or undersampling techniques if sentiment categories are not evenly distributed.

6. Exploratory data analysis(EDA)

Analyze most frequently occurring words across sentiment classes.

Generate word clouds to visualize dominant words in positive, negative, and neutral texts.

Assess the distribution of sentiments to understand biases in the data.

Study trends over time or across hashtags to identify sentiment shifts.

7. Feature engineering

Text Vectorization: Convert text into numerical format using TF-IDF, Bag of Words, or advanced embeddings like Word2Vec or GloVe.

Contextual Embeddings: Use pre-trained language models like BERT to capture contextual relationships in sentences.

Linguistic Features: Add parts of speech tags, sentiment lexicons, or n-gram frequency counts.

8. Model building

Choose appropriate models such as:

Logistic Regression and SVM for baseline performance

Random Forest for robustness and interpretability

LSTM and BERT for deep learning-based performance

Fine-tune models using cross-validation and hyperparameter tuning.

Evaluate performance using metrics:

Accuracy

Precision, Recall, and F1-score

Confusion Matrix and ROC/AUC

9. Visualization of results and data insigths

Confusion Matrix: Visual representation of prediction performance.

ROC/AUC Curve: To analyze model performance at various classification thresholds.

Sentiment Over Time: Line plots to visualize how sentiment changes over time.

Keyword Contribution: Identify key terms that contribute most to each sentiment class.

10. Tools and technologies used

Programming Language: Python

NLP Libraries: NLTK, spaCy, TextBlob, HuggingFace Transformers

Data Handling: pandas, numpy

Modeling: scikit-learn, TensorFlow, Keras

Visualization: matplotlib, seaborn, plotly

IDEs: Google Colab, Jupyter Notebook

Version Control: GitHub for tracking project development

11. Team Members and Contributions

NAME	ROLE	RESPONSIBLE

Ravi kumaar	Member	Data Collection, Data Cleaning
umesh	Member	Visualization & Interpretation
Yashwanth	Member	Exploratory Data Analysis (EDA), Feature Engineering
Sri kanth	Leader	Model Building, Model Evaluation