நான் முதல்வன்

ORACLE

AdroIT Technologies®
Innovative Solutions Pvt LTD

Phase-2

Student Name: Harshan M

Register Number: 714223106010

Institution: Tamilnadu college of Engineering

Department: Electronics and communication engineering

Date of Submission: 09/05/2025

Github Repository Link: https://github.com/harshanm06/Sentiment-analysis.git

---

1. Problem Statement

Social media platforms have become a major outlet for public expression. However, the vast and unstructured nature of these conversations makes it difficult to understand users' emotional states in real time. This project aims to perform sentiment analysis to decode user emotions such as happiness, sadness, anger, etc., from social media conversations using natural language processing (NLP) and machine learning techniques.

2. Project Objectives

Extract and preprocess social media conversation data.

Develop models to classify text into emotional categories (e.g., joy, sadness, anger, fear, surprise).

Optimize model performance for real-time or batch deployment.

Ensure interpretability and explainability of model decisions.

3. Flowchart of the Project Workflow

Data Collection → Data Preprocessing → EDA → Feature Engineering → Model Training → Evaluation → Insights & Visualization → Deployment.

4. Data Description

Source: [e.g., Kaggle’s Emotion Dataset or Twitter API]

Type: Text (unstructured)

No. of Records: ~25,000 text entries

Features: Text, Emotion Label

Target Variable: Emotion (joy, sadness, anger, etc.)

Dataset Type: Static

5. Data Preprocessing

Removed duplicates and null values

Lowercased all text

Removed punctuation, numbers, stopwords

Tokenized and lemmatized words

Encoded emotion labels

Used TF-IDF or word embeddings for feature extraction

6. Exploratory Data Analysis (EDA)

Univariate: Countplots for emotion distribution

Bivariate: Word clouds per emotion, most frequent words

Multivariate: Sentiment score distributions across emotions

Insights: Imbalance in label distribution; anger and sadness texts tend to use stronger negative words

7. Feature Engineering

Created sentiment scores (using TextBlob/VADER)

Extracted length-based features (word count, character count)

Used TF-IDF vectorization and Word2Vec embeddings

Dimensionality reduction using PCA (optional)

8. Model Building

Models Used: Logistic Regression, Random Forest, and BERT (optional advanced)

Train/Test Split: 80/20

Evaluation Metrics: Accuracy, Precision, Recall, F1-Score

BERT showed highest accuracy (~85%), Random Forest showed ~75%

9. Visualization of Results & Model Insights

Confusion matrix for each model

ROC curves for multi-class classification

Feature importance for tree-based models

LIME/SHAP used for model interpretability (optional)

10. Tools and Technologies Used

Language: Python

IDE: Jupyter Notebook, Google Colab

Libraries: pandas, numpy, sklearn, seaborn, matplotlib, nltk, transformers, TextBlob

Visualization: seaborn, matplotlib, Plotly

11. Team Members and Contributions

Harshan M – Data Collection, Cleaning, Model Training

Akaash P - EDA, Feature Engineering, Model Evaluation

Suriya P –Visualization, Documentation,Reporting.