# Social Media Sentiment Analysis Project

**Mudavath Mounika,**
**Indian Institute of Technology Kharagpur,**
**Kharagpur, Kolkata, West Bengal,**
**mounikamudavath333@gmail.com**

## 🔹 Introduction

This project focuses on analyzing Twitter data to classify the sentiment of tweets as **positive** or **negative**. Using **Natural Language Processing (NLP)** and **Machine Learning** techniques, the system processes raw tweets, cleans the text, removes unnecessary characters, and applies feature extraction methods to prepare the data for classification.

## 🔹 Dataset

- **Source**: Sentiment140 dataset
- **Size**: 1.6 million tweets
- **Features**:
  - `sentiment` → Label (0 = Negative, 4 = Positive)
  - `text` → Raw tweet text

The dataset is large and managed using **Git LFS** on GitHub.

## 🔹 Methodology

1. **Data Preprocessing**
   - Removed URLs, mentions, numbers, and special characters
   - Converted text to lowercase
   - Reduced repeated characters
   - Removed stopwords
   - Tokenization and text normalization
2. **Feature Extraction**
   - Used **TF-IDF Vectorizer** to transform text into numerical features
   - Limited vocabulary size to 5,000 most frequent words
3. **Model Training**
   - Applied **Logistic Regression** as the classifier
   - Trained the model on 80% of the data and tested on 20%
4. **Evaluation Metrics**
   - Accuracy
   - Precision, Recall, and F1-score

## 🔹 Results

- The Logistic Regression model provided **good accuracy** on sentiment classification.
- TF-IDF with Logistic Regression proved efficient for large-scale text classification.

## 🔹 Conclusion

This project demonstrates how **Machine Learning and NLP** can be used to classify sentiment from social media text. The pipeline is efficient, scalable, and can be extended with deep learning models (e.g., LSTMs, BERT) for further improvement.

## 🔹 Future Work

- Use advanced word embeddings (Word2Vec, GloVe, BERT)
- Apply deep learning models for higher accuracy
- Deploy the model as a web app for real-time sentiment analysis