**Phase-1**

**DECODING EMOTIONS THROUGH SENTIMENT ANALYSIS OF SOCIAL MEDIA CONVERSATIONS**

**Student Name:** Dhanasekaran.T

**Register Number:** 511723106301

**Institution:** Pallavan college of engineering

**Department:** Electronics and communication

**Date of Submission:** 25/04/2025

# *1.PROBLEM STATEMENT*

In the era of digital communication, social media platforms have become key spaces where individuals express opinions, emotions, and reactions in real time. However, the vast volume and unstructured nature of social media content make it difficult to manually analyze emotional trends or public sentiment.

This project addresses the need to automatically identify and decode human emotions embedded in social media conversations using sentiment analysis and natural language processing (NLP) techniques. By analyzing textual data from platforms like Twitter or Reddit, the system aims to classify sentiments (e.g., positive, negative, neutral) and deeper emotional tones (e.g., joy, anger, sadness, fear).

Understanding these emotional patterns is valuable across various domains including mental health monitoring, brand perception analysis, and public opinion tracking, making it a relevant and powerful application of AI.

# *2.OBJECTIVES OF THE PROJECT*

The primary objective of this project is to develop an NLP-driven system capable of identifying and classifying sentiments and emotions expressed in social media conversations. The model will analyze textual data to understand public mood, emotional trends, and context-driven reactions.

Specific objectives include:

• To collect and preprocess social media text data (e.g., from Twitter or Reddit) for sentiment analysis.

• To perform text cleaning and transformation, including tokenization, stopword removal, stemming, and lemmatization.

• To conduct exploratory data analysis (EDA) on language patterns, frequent words, and sentiment distributions.

• To build and train NLP models (e.g., Logistic Regression, Naive Bayes, LSTM, or Transformers like BERT) for sentiment and emotion classification.

• To evaluate model performance using classification metrics such as precision, recall, F1-score, and confusion matrix.

• To visualize emotional trends over time or by topic using plots or dashboards.

• To deploy the model as a user-friendly web app or API that allows input of text and returns sentiment/emotion predictions.

These objectives aim to bridge human emotion with machine understanding, offering valuable insights from unstructured social media data

# *3.SCOPE OF THE PROJECT*

This project focuses on building an end-to-end sentiment and emotion analysis system for social media conversations using natural language processing (NLP) and machine learning techniques. It includes all major stages of a typical NLP pipeline, from data acquisition to model deployment.

Features to be included:

• Collection of textual data from social media platforms (e.g., Twitter API or Reddit datasets).

• Preprocessing of unstructured text including cleaning, normalization, and tokenization.

• Training sentiment classification models to detect general sentiment (positive, negative, neutral).

• Extension to multi-class emotion detection (e.g., joy, anger, sadness, fear).

• Visualization of emotional trends, word clouds, and sentiment timelines.

• Interactive deployment of the final model as a web application or API.

Limitations and Constraints:

• The project is limited to English language social media data.

• Due to API restrictions or dataset size, data volume may be capped.

• The system is for research and educational purposes only and not intended for clinical or commercial use.

• Real-time monitoring or multilingual analysis is outside the current scope but could be explored in future work.

• The model’s performance may be affected by slang, sarcasm, or abbreviations, which are prevalent in social media language.

By defining these boundaries, the project stays focused and feasible while still delivering meaningful insights from large-scale social interactions.

# *4.DATA SOURCES*

To train and evaluate the sentiment and emotion classification models, this project will use publicly available datasets containing social media conversations, as well as potentially real-time data via APIs.

Primary Data Sources:

• Twitter API (Tweepy / snscrape)

o Tweets collected using relevant hashtags or keywords.

o Requires developer access and API credentials.

o Useful for gathering real-time or domain-specific emotional content.

• Reddit Datasets (via Pushshift or Kaggle)

o Subreddit posts and comments on various topics.

o Rich in emotional expression and topic diversity.

• Kaggle Datasets

o Pre-labeled sentiment and emotion datasets such as:

Sentiment140 Dataset: Over 1 million tweets labeled with positive/negative/neutral sentiments.

Emotion Dataset from Twitter: Contains tweets labeled with emotions like joy, anger, fear, sadness, etc.

• GoEmotions Dataset by Google

o A large dataset of Reddit comments labeled across 27 emotion categories plus neutral.

o Suitable for fine-grained emotion classification.

Data Nature:

• Public: All selected datasets are publicly accessible and suitable for academic use.

• Static & Dynamic: Kaggle datasets are static, while API-based sources (Twitter/Reddit) allow real-time dynamic data collection.

These sources provide a robust foundation for training emotion-aware NLP models that can generalize across different types of social media conversations.

# *5.HIGH-LEVEL METHODOLOGY*

* **Data Collection**

• Use Twitter API (via Tweepy or snscrape) or Pushshift Reddit API to gather posts/comments containing emotional content.

• Alternatively, utilize pre-labeled datasets from platforms like Kaggle or Google Research (GoEmotions).

* **Data Cleaning**

• Remove URLs, mentions (@username), hashtags, emojis, special characters, and HTML tags.

• Convert text to lowercase and correct spelling variations or slang.

• Remove stopwords and unnecessary punctuation.

* **Exploratory Data Analysis (EDA)**

• Visualize frequent words using word clouds and bar plots.

• Analyze sentiment/emotion distribution and text length.

• Use n-gram analysis (bigrams, trigrams) to identify common phrases.

* **Feature Engineering**

• Text vectorization using methods such as:

o TF-IDF

o Word2Vec / GloVe

o Transformer-based embeddings (e.g., BERT embeddings via HuggingFace)

* **Model Building**

• Train and compare different classifiers:

o Traditional ML: Logistic Regression, Naive Bayes, SVM

o Deep Learning: LSTM, Bi-LSTM

o Transformers: BERT, RoBERTa, fine-tuned on emotion/sentiment tasks

* **Model Evaluation**
* Use classification metrics such as:

Accuracy, Precision, Recall, F1-Score

Confusion Matrix

ROC-AUC (for binary/multiclass sentiment).

* **Visualization & Interpretation**

• Visualize predicted vs. actual emotions using heatmaps and pie charts.

• Track sentiment/emotion trends over time (e.g., daily/weekly spikes).

• Use attention visualizations (if using transformers) to interpret text influence.

* **Deployment**
* Streamlit (for fast deployment)

Flask (for API-based deployment)

• Enable users to input social media text and receive real-time sentiment/emotion predictions.

# *6.TOOLS AND TECHNOLOGIES*

* **Programming Language**

Python: Chosen for its readability, versatility, and extensive support in the NLP and ML community.

* **Notebook/IDE**

• Google Colab: For cloud-based development and training with GPU support.

• Jupyter Notebook: For local experimentation, visualization, and iterative testing.

• VS Code: For developing scripts, API deployment, and full-stack integration.

* **Libraries**

• Data Handling & EDA:

pandas, numpy, matplotlib, seaborn, wordcloud

• Text Preprocessing:

re (regex), nltk, spaCy, emoji, beautifulsoup4

• Vectorization & Embeddings:

scikit-learn (TF-IDF), gensim (Word2Vec), transformers (BERT, RoBERTa via HuggingFace)

• Modeling & Evaluation:

scikit-learn, keras, tensorflow, pytorch, transformers

• Visualization & Interpretation:

plotly, streamlit, LIME, SHAP (for model explainability)

* **Optional Tools for Deployment**

• Streamlit: For building interactive and user-friendly web apps for sentiment/emotion input and output.

• Flask: For creating an API to serve predictions to other services.

• Gradio (optional): For quickly wrapping ML models in a web interface with minimal code.

• Docker (optional): For containerization and scalable deployment in production environments

# *7.TEAM MEMBERS AND ROLES*

This project is collaboratively executed by a team of five, each contributing their expertise across different stages of the sentiment analysis pipeline.

|  |  |
| --- | --- |
| **Team members** | **Roles and responsibilities** |
| **Lokesh S.** | Project Coordinator & Model Architect – Oversees project workflow, selects appropriate models, and fine-tunes sentiment/emotion classifiers. |
| **Vannamathi K.** | Data Acquisition & Preprocessing Lead – Handles data scraping from social media platforms, API integration, and all text preprocessing steps. |
| **Dhanasekaran T** | NLP Engineer – Focuses on feature engineering, embedding techniques, and deep learning models like LSTM or BERT. |
| **Venkades.T** | Visualization & Analytics Specialist – Responsible for EDA, sentiment trend visualization, and interpretation of model outputs using SHAP or LIME. |
| **Vijayalaskhmi.S** | Deployment Engineer – Manages deployment of the model using tools like Streamlit or Flask and ensures a functional, user-friendly web interface. |