**Phase-2 Submission Template**

**Student Name:** VETHAVALLI J

**Register Number:** 613023243060

**Institution:** VIVEKANANDHA COLLEGE OF TECHNOLOGY FOR WOMEN

**Department:** B.TECH- ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

**Date of Submission:** 03.05.2025

**Github Repository Link:**

**[https://github.com/Vetha262/project1.git](https://github.com/anusreemanusreem08/project )**

### **Problem Statement**

*In the digital age, social media platforms serve as a primary outlet for individuals to express their thoughts, opinions, and emotions. However, the massive volume and unstructured nature of social media conversations make it challenging to accurately understand public sentiment and emotional trends. Traditional sentiment analysis methods often fall short in capturing nuanced emotions such as sarcasm, mixed feelings, or context-specific sentiments. This project aims to develop a robust sentiment analysis model that can decode a wide range of human emotions from social media conversations, enabling deeper insights into public opinion, mental health trends, and social behaviours.*

**2.Project Objectives**

*1. To collect and preprocess social media data from platforms like Twitter, Facebook, or Reddit for emotion and sentiment analysis.*

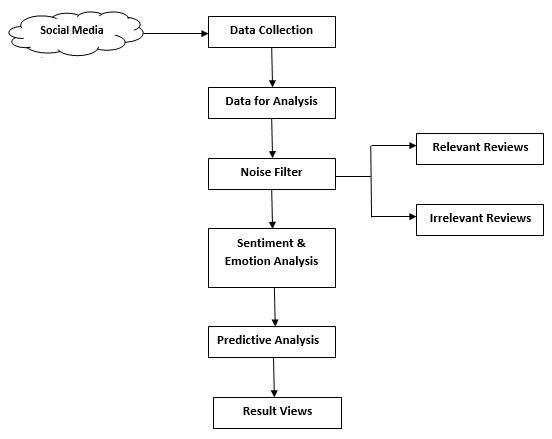
*2. To develop a sentiment analysis model capable of identifying a wide range of emotions (e.g., joy, anger, sadness, surprise, fear, etc.) beyond basic positive/negative/neutral classification.*

*3. To apply Natural Language Processing (NLP) techniques such as tokenization, lemmatization, and part-of-speech tagging to understand text structure and meaning.*

*4. To implement and compare machine learning and deep learning algorithms (e.g., SVM, Random Forest, LSTM, BERT) for effective emotion classification.*

*5. To visualize emotional trends and patterns across different topics, timeframes, or user groups using tools like word clouds, sentiment timelines, or heatmaps.*

### **Flowchart of the Project Workflow**

****

### **4. Data Description**

* ***Dataset Name****: social media emotions & sentiment dataset*
* ***Source****: Twitter API/google Goemotions dataset /Kaggle public repositories*
* ***Type of Data****: unstructured text data*
* ***Records and Features****: 58000 posts/comments with attributes including texts, emotion labels, and metadata.*
* ***Target Variable****: emotion label (eg: joy, anger, sadness, fear, surprise, etc.)*
* ***Static or Dynamic****: Static dataset (can also be extend with live data using APIs)*
* ***Attributes Covered****: textual content: main body of the post or comment.*

*Dataset Link: GoEmotions- Google Dataset or Kaggle Sentiment Analysis Datasets*

### **Data Preprocessing**

*The data preprocessing phase is crucial to prepare raw social media text for emotion and sentiment analysis. The following steps were performed:*

*Text Cleaning*

1. *Removed unwanted elements such as:*
2. *URLs (https://...)*
3. *Mentions (@username)*
4. *Hashtags (#topic)*
5. *Emojis and special symbols Converted text to lowercase to maintain consistency.*
6. *Tokenization: Split each sentence or post into individual tokens (words), enabling easier analysis and vectorization.*
7. *Stop-word Removal: Removed common stop-words (e.g., is, the, a, of) which do not contribute significant meaning to the emotional context.*
8. *Lemmatization/Stemming: Reduced words to their root form (e.g., “running” → “run”) to normalize different word variants.*
9. *Noise Filtering: Eliminated non-English characters and repetitive letters (e.g., “so happy” → “so happy”) using regex*..

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

***Univariate Analysis:***

*Emotion Distribution: Bar plots created to understand the frequency distribution of each emotion category (e.g., joy, sadness, anger, fear, surprise, etc.).*

*Text Length Distribution:*

*Histograms plotted to examine the number of words or characters per post/comment.*

***Bivariate & Multivariate Analysis:***

*Emotion vs Text Length:*

*Boxplots used to analyze how the length of posts varies across different emotions.*

*Co-occurrence Matrix:*

*Heatmaps created to show common word pairs and their frequency.*

***Key Insights:***

*Some emotions (e.g., joy or neutral) occur more frequently than others.*

*Posts expressing anger or sadness tend to be longer and more descriptive.*

*Emotion-specific keywords (e.g., "happy", "hate", "excited", "cry") significantly influence classification.*

### **7. Feature Engineering**

* *Text Length Features: Word count, character count, and average word length.*
* *Punctuation & Emoji Counts: Exclamation marks, question marks, emojis as emotion indicators.*
* *Sentiment Scores: Polarity and subjectivity extracted using Text Blob or VADER.*
* *TF-IDF & Embeddings: Text converted using TF-IDF, Word2Vec, or BERT for semantic understanding*

### **8. Model Building**

* *Model Selection: Tried multiple classifiers—Logistic Regression, Naive Bayes, SVM, Random Forest, and deep learning models (LSTM, BERT).*
* *Vectorization: Used TF-IDF for traditional models and BERT embeddings for deep learning.*
* *Training: Models trained on pre- processed and labeled text data with an 80:20 train-test split.*
* *Evaluation Metrics: Accuracy, Precision, Recall, and F1-score used to assess performance*

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

### *Confusion Matrix: Displayed true vs. predicted emotions, highlighting model accuracy across classes.*

### *Bar Charts: Compared model performance (accuracy, precision, recall) across different algorithms.*

### *Emotion Prediction Samples: Visual snapshots of real social media posts with predicted emotion labels.*

### **10. Tools and Technologies Used**

* ***Programming Language****: Python3*
* ***Notebook Environment****: Google Collab*
* ***Key Libraries****:*

*○ pandas, numpy for data handling*

*○ matplotlib, seaborn, plotly for visualizations*

*○ scikit-learn for preprocessing and modelling*

*○ Gradio for interface deployment*

### **11. Team Members and Contributions**

* ***Data collection and text pre-processing:*** *Anusree M*
* ***Feature engineering:*** *Arunaethi M S*
* ***Model development:*** *Arunika E*
* ***Result visualization:*** *Chella*
* ***Final report writing****: Vethavalli J*