Phase-1 Submission Template

Student Name: Ajithkumar E Register

Number: 620123106002

Institution: AVS ENGINEERING COLLEGE

Department: ECE

Date of Submission: 30.04.2025

1. Problem Statement

In today's digital age, people frequently express their thoughts and emotions on social media platforms such as Twitter, Instagram, and Reddit. Understanding these emotions at scale can provide valuable insights for businesses, governments, and researchers. However, decoding these sentiments is challenging due to the informal and context-driven nature of social media language. This project aims to analyze and classify emotions from social media conversations using sentiment analysis techniques to better understand user behavior and public mood.

2. Objectives of the Project - To collect and preprocess real-time or historical social media data. - To classify text data into sentiment categories such as positive, negative, neutral, or emotional tones like joy, anger, sadness, etc. - To identify and visualize trends in public sentiment around specific topics or events. - To evaluate the effectiveness of different machine learning and NLP models in emotion

3. Scope of the Project - Analysis of tweets or posts related to specific hashtags, keywords, or events. - Implementation of multiple NLP models including traditional (Naive Bayes, SVM) and deep learning (LSTM, BERT). - Limitations: - Data collection may be limited to a specific timeframe or API quota. - Sentiment classification may face challenges with sarcasm, slang, and multilingual content. - Real-time deployment is optional and dependent on access to APIs.

4. Data Sources - Dataset Sources: - Twitter API (for real-time tweet collection) - Kaggle or UCI datasets for historical sentiment-labeled tweets - Type: Public (or via API access) - Nature: Combination of static (pre-collected) and dynamic (real-time via Twitter API)

5. High-Level Methodology Data Collection: - Collect tweets using Twitter API with Tweepy or snscrape. - Supplement with public datasets (e.g., Sentiment140, Kaggle tweet sentiment datasets).

Data Cleaning:

- Remove URLs, mentions, hashtags, emojis, stopwords.
- Normalize text using lowercase conversion, lemmatization, and punctuation removal.

6. Tools and Technologies

Exploratory Data Analysis (EDA):

- Use word clouds, frequency plots, and sentiment distribution graphs to explore data.
- Track sentiment over time or per topic.

Feature Engineering:

- Convert text into numerical form using TF-IDF, Word2Vec, or BERT embeddings.
- Create features like post length, hashtag count, or emoji presence.

Model Building:

- Start with Logistic Regression, Naive Bayes, and SVM for baseline.
- Experiment with LSTM and Transformer-based models like BERT for better accuracy.

Model Evaluation:

- Use Accuracy, Precision, Recall, F1-Score, and Confusion Matrix.
- Evaluate using stratified cross-validation.

Visualization & Interpretation:

- Generate interactive dashboards showing emotion trends.
- Use bar plots, pie charts, line graphs for sentiment over time and top keywords.

Deployment (Optional):

- Build a Streamlit web app or Jupyter dashboard to visualize and interact with sentiment predictions.

Programming Language: Python

Notebook/IDE: Google Colab, Jupyter Notebook

Libraries:

- Data Processing: pandas, numpy, re

- NLP: NLTK, spaCy, transformers (HuggingFace)

- Visualization: matplotlib, seaborn, plotly, wordcloud

- Modeling: scikit-learn, TensorFlow, Keras, HuggingFace Transformers

- API Access: Tweepy, snscrape

|

Optional Tools for Deployment:

- Streamlit, Gradio, or Flask for web-based interface

|

7. Team Members and

|

Roles

```
|---------------|-----------------------------------------------|
```

| Ajithkumar E | Data Collection, Preprocessing, Documentation |

| Meganadhan M | Model Building and Evaluation

| Ayyanar S | Exploratory Data Analysis and Visualization

| Gowtham V | Feature Engineering and Deployment (Streamlit) |