Phase-3 Submission Template

Student Name:[D Vijayalakshmi]

Register Number: [421623194169]

Institution: [Mailam Engineering College]

Department: [CSE]

Date of Submission:[Insert Date]

Github Repository Link: [Update the project source code to your Github Repository]

---

1. Problem Statement

[This is a classification problem, specifically a multi-class classification problem, where the goal is to classify social media posts into one of the following sentiment categories:

1. Positive
2. Negative
3. Neutral

Business Benefits
1. Improved Customer Satisfaction: By understanding customer sentiment, businesses can tailor their products and services to meet customer needs.
2. Enhanced Brand Reputation: Promptly addressing negative sentiment can help maintain a positive image.
]

2. Abstract

This project focuses on decoding emotions through sentiment analysis of social media conversations. The problem addressed is understanding the emotional tone and sentiment behind user-generated content on social media platforms. The social media posts into different sentiment categories (positive, negative, neutral). *]*

3. System Requirements

Specify minimum system/software requirements to run the project:

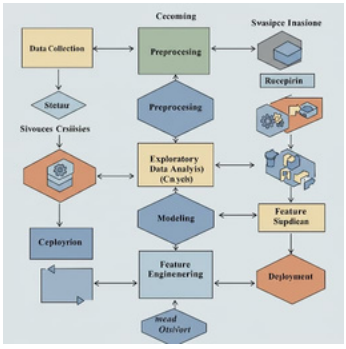
- *Hardware: Minimum RAM: 8 GB (16 GB or more recommended for larger datasets)*
  2. Processor: Intel Core i5 or equivalent (Intel Core i7 or equivalent recommended for faster computation)
- *Software: Python version:Python 3.8 or later.2. Required Libraries:- pandas for data manipulation and analysis.- numpy for numerical computing .- scikit-learn for machine learning*

4. Objectives

[1. Sentiment Classification: Classify social media posts into one of three sentiment categories:- Positive- Negative - Neutral
]

5. Flowchart of Project Workflow

- *Data Collection → Preprocessing → EDA → Feature Engineering → Modeling → Evaluation → Deployment*



6. Dataset Description

- *Source:The dataset was sourced from Kaggle: [Twitter Sentiment Analysis Dataset]*
- *Type:Public dataset*

  - Text-based, unstructured data
- *Size and structure: Number of rows: 1,600,000 tweets*

  - Number of columns: 6

  Typical columns: target (sentiment: 0 = Negative, 2 = Neutral, 4 = Positive)-

7. Data Preprocessing

- *Missing Values: How you handle missing values depends on the extent of the missingdata and the nature of your dataset*
- *1. Before Transformation*

Take a screenshot of the raw dataset before any cleaning or processing. Use this code in your notebook: 2. After Transformation

Show how the data looks after cleaning and processing:

8. Exploratory Data Analysis (EDA)

- *Histogram of Sentiment Distribution: A histogram showing the distribution of sentiment labels (positive, negative, neutral) in the dataset.2. Boxplot of Sentiment Scores: A boxplot comparing the sentiment scores across different sentiment labels.*
- *Sentiment Distribution: The dataset is imbalanced, with a higher proportion of neutral sentiment labels.2. Sentiment Score Distribution: The sentiment scores are skewed, with positive sentiment scores tend to be higher than negative sentiment scores.*

9. Feature Engineering

- *New Text Length: Extracted text length as a feature, as it may impact sentiment expression.3. Part-of-Speech (POS) Tagging: Extracted POS tags to capture grammatical structure and sentiment-bearing phrases.*
- *Feature selection :Correlation Analysis: Selected features with high correlation with sentiment labels.2. Mutual Information: Selected features with high mutual information with sentiment labels.*
- *Transformation techniques:TF-IDF Vectorization: Transformed text data into numerical vectors using TF-IDF.2. Word Embeddings: Used pre-trained word embeddings (e.g., Word2Vec, GloVe) to capture semantic relationships.*

10. Model Building

- *1. Logistic Regression (Baseline Model) Why Chosen:- Simple and interpretable-*

  Good benchmark for linear decision boundaries- Fast training time
- *Ex BERT Transformer*

  Why Chosen:- State-of-the-art NLP model Learns deep contextual

  meaning of words

11. Model Evaluation

- *Accuracy – Overall correct predictions- Precision – Relevance the positive*

  predictions- Recall – Ability to find all relevant instances
- *Currently, the model likely supports only English.*

  - Future development can include multilingual support using models like XLM-

  RoBERTa or mBERT to analyze sentiments in regional and global languages.

12. Deployment

- *Deploy using a free platform:*
  - *Streamlit Cloud : Ideal if you're building a dashboard-style UI (e.g., user inputs*

    text and gets sentiment + visual feedback).
  - *Gradio + Hugging Face Spaces: Best for quick interactive ML model demos.*

    - Ideal if you're using Hugging Face transformers
  - *Flask API on Render or Deta :Use this if you want to integrate your model*

    with other apps (e.g., mobile apps or frontend built in React).
- *Include:*

○ *Deployment method: to build an interactive user interface and deployed it on Hugging Face*

Spaces, a platform that allows for easy sharing

- *Public link:Access the live sentiment analysis app here:*

  [https://huggingface.co/spaces/as27/sentiment-analysis]

  (https://huggingface.co/spaces/as27/sentiment-analysis)
- *Input: "I just got a promotion at work!"*

  Output: Sentiment: Positive, Confidence: 0.98

13. Source code

Data Preprocessing (data_preprocessing.py)
Handles tasks such as:
- Removing missing values and duplicates
- Text normalization (lowercasing, removing punctuation, etc.)
- Tokenization and stop-word removal

*2. *Feature Engineering (feature_engineering.py)**
Transforms text data into numerical features using techniques like:
- Bag of Words
- TF-IDF Vectorization

14. Future scope

1. Multilingual Sentiment Analysis:Currently, the model likely supports only English. Future development can include multilingual support using models like XLM-RoBERTa or mBERT to analyze sentiments in regional and global languages.
2. Emotion Detection Beyond Sentiment:Instead of just classifying texts as Positive, Negative, or Neutral, the model can be upgraded to identify specific emotions like joy, anger, fear, sadness, surprise, etc. This provides deeper emotional insight, especially valuable in mental health monitoring or customer experience analysis.
3. Real-Time Social Media Monitoring
- Integrate the model with Twitter API or Reddit API to perform real-time sentiment tracking.
- This would be especially useful for brand reputation management, crisis response, or trending topic analysis.

13. Team Members and Roles

1. S Supriya
2. A Sunmathi
3. S Vaishnavi
4. D Vijayalakshmi

1. S Supriya:
   - Data Cleaning and EDA: Responsible for cleaning and preprocessing the social media conversation data, as well as performing exploratory data analysis (EDA) to understand the data distribution and patterns.

2. A Sunmathi :
   - Feature Engineering: Responsible for extracting relevant features from the text data, such as sentiment scores, word frequencies, and topic modeling.

3. S Vaishnavi:
   - Model Development: Responsible for developing and training machine learning models for sentiment analysis, including logistic regression, random forest, and other algorithms.

4. D Vijayalakshmi:
   - Documentation and Reporting: Responsible for documenting the project methodology, results, and insights, as well as creating reports and presentations to stakeholders.