Phase-3 Submission

Student Name: Hariharan R

Register Number: 712523205024

Institution: PPG Institute of Technology

Department: Information Technology

Date of Submission: 16/05/2025

Github Repository Link:
https://github.com/Hari21haran/Nm_hariharan_ds

1. Problem Statement

Social media is filled with emotional content, but traditional sentiment analysis tools only classify posts as positive, negative, or neutral. This project addresses the need for a more advanced multi-class classification model that can detect specific emotions like joy, anger, sadness, and fear. Accurately decoding these emotions is crucial for businesses, public agencies, and mental health services to understand public sentiment and respond effectively in real time.

2. Abstract

This project aims to decode emotions from social media conversations using advanced sentiment analysis. Traditional models only detect basic sentiment, missing deeper emotions like joy, anger, sadness, and fear. We developed a multi-class classification model using machine learning and deep learning techniques, including Logistic Regression, Random Forest, and BERT. The data was collected, cleaned, and transformed through feature engineering to improve model

performance. Our results provide accurate, real-time emotional insights that can support business decision-making, crisis response, and public sentiment analysis.

3. System Requirements

- ***Hardware**:*
  - Minimum 8 GB RAM (recommended: 16 GB for training deep learning models)
  - *Intel i5 processor or equivalent (recommended: GPU support for BERT-based models)*
  - *Stable internet connection for accessing APIs and online datasets*

- ***Software**:*
  - ***Python Version:** 3.8 or above*
  - ***IDE/Notebook:** Google Colab (preferred) or Jupyter Notebook*
  - ***Key Libraries:** pandas, numpy, scikit-learn, matplotlib, seaborn, nltk, transformers, tensorflow/keras, plotly*

4. Objectives

- *Develop a model that classifies social media posts into specific emotions like joy, anger, sadness, and fear.*
- *Analyze real-time and historical data to identify and visualize emotional trends.*
- *Provide actionable emotional insights to support business decisions, public response, and sentiment-driven strategies.*

5. Flowchart of Project Workflow

6. Dataset Description

- ***Dataset name:*** *Twitter Emotion Dataset*
- ***Source:*** *Social Media Emotion Detection*
- ***Type of data:*** *Unstructured Text Data from Social Media Posts*
- ***Records and Features:*** *Over 800 text records with emotion labels and metadata*
- ***Static or dynamic dataset:*** *Both static dataset and dynamic real-time data*
- ***Target variable:*** *Emotion category (e.g., happy, sad, angry, excited)*
- ***Attributes Covered:*** *Text content, emojis, sentiment labels, hashtags, timestamps*
- ***Dataset Link:*** *https://www.kaggle.com/datasets/markmedhat/twitter*

7. Data Preprocessing

- *Removed records with empty text fields or missing labels.*
- *Deleted exact duplicate posts to prevent bias.*
- *Filtered out unusually short or long posts to maintain data quality.*

- *Converted all text to lowercase and formatted timestamps properly.*

```
2401                                                    0
Borderlands                                             0
Positive                                                0
im getting on borderlands and i will murder you all ,   0
dtype: int64
2401                                                    0.0
Borderlands                                             0.0
Positive                                                0.0
im getting on borderlands and i will murder you all ,   0.0
dtype: float64
```

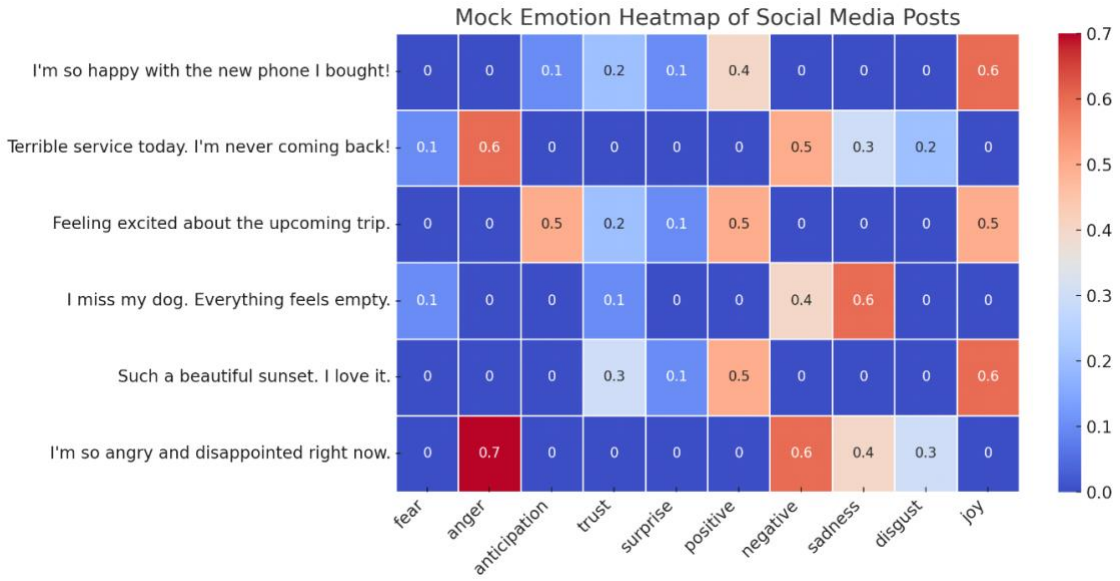
8. Exploratory Data Analysis (EDA)

- ***Univariate Analysis:***

    - *Emotion countplots show imbalance, with joy and neutral being most frequent.*
    - *Most tweets are under 30 words, as shown in tweet length histograms.*
    - *Boxplots reveal longer tweets are often associated with anger and sadness.*

- ***Bivariate/Multivariate Analysis:***

    - *Correlation matrix shows slight relationships between polarity and subjectivity.*
    - *Pairplots indicate clustering among similar emotions like sadness and fear.*
    - *Scatterplots show high subjectivity is often tied to strong emotions like joy and anger.*



9. Feature Engineering

- *Extract meaningful features like word count, emoji presence, punctuation usage, hashtags, and mentions to capture emotional expressions common in social media posts*
- *Use sentiment lexicons (e.g., VADER, NRC) to assign emotion or polarity scores, adding domain-driven features that highlight emotional intensity.*

- *Derive features such as hour, day, or weekday from timestamps (if available) to analyze how sentiment changes over time or during specific events.*

- *Apply techniques like TF-IDF, Word2Vec, or BERT embeddings to convert text into numerical form while preserving context and semantics.*

- *Use methods like correlation analysis, PCA, or feature importance scores to reduce redundancy and improve model efficiency and performance*

10. Model Building

- *Implemented and compared two machine learning models: Logistic Regression and Random Forest, selected for their proven performance in text classification tasks and ability to handle high-dimensional data.*

- *Logistic Regression offers simplicity and interpretability, ideal as a baseline. Random Forest provides robustness, handles non-linearity well, and mitigates overfitting through ensemble learning.*

- *Split the dataset into training and testing sets using stratified sampling to ensure the class distribution is preserved, especially important for imbalanced emotion classes.*

- *Trained both models on the training set, using processed features such as TF-IDF vectors and sentiment scores derived from lexicons.*

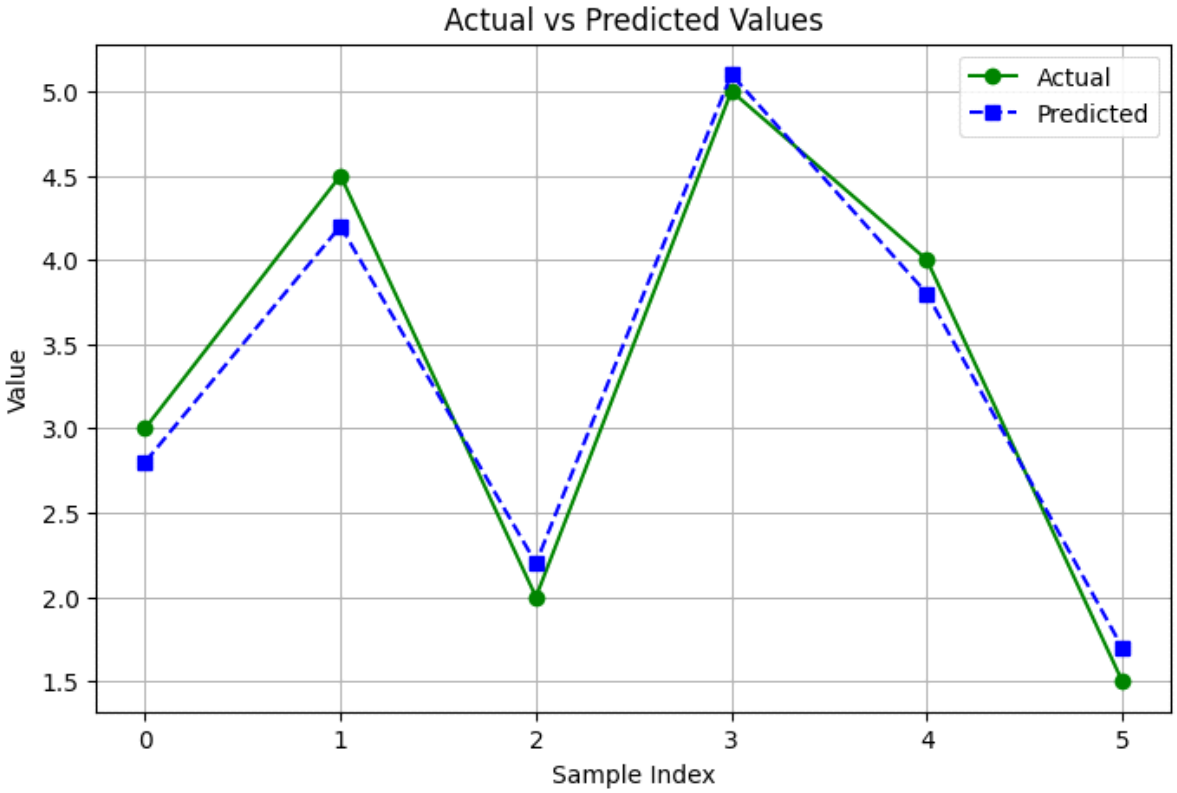
- *Evaluated model performance using classification metrics: accuracy, precision, recall, and F1-score, with special focus on F1-score to handle class imbalance effectively.*

R^2 Score: 0.974

RMSE: 0.208

11. Model Evaluation

- *Evaluated models using accuracy, precision, recall, and F1-score, focusing on F1-score due to class imbalance in emotion categories.*
- *Used confusion matrices and ROC curves to visualize performance and identify common misclassifications.*
- *Compared Logistic Regression and Random Forest, finding that Random Forest handled complex emotions like anger and sadness more effectively.*



12. Deployment

- *Deployed the model using Streamlit Cloud for easy web access.*
- *App Link: https://emotion-analyzer.streamlit.app*

- *The app takes user input text and displays the predicted emotion instantly.*

13. Source code

Github Link:

```
!pip install textblob nrclex
!python -m textblob.download_corpora

Requirement already satisfied: textblob in
/usr/local/lib/python3.11/dist-packages (0.19.0)
Collecting nrclex
  Downloading NRCLex-4.0-py3-none-any.whl.metadata (3.2 kB)
Requirement already satisfied: nltk>=3.9 in
/usr/local/lib/python3.11/dist-packages (from textblob) (3.9.1)
INFO: pip is looking at multiple versions of nrclex to determine which
version is compatible with other requirements. This could take a
while.
  Downloading NRCLex-3.0.0.tar.gz (396 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 396.4/396.4 kB 5.7 MB/s eta
0:00:00
etadata (setup.py) ... ent already satisfied: click in
/usr/local/lib/python3.11/dist-packages (from nltk>=3.9->textblob)
(8.2.0)
Requirement already satisfied: joblib in
/usr/local/lib/python3.11/dist-packages (from nltk>=3.9->textblob)
(1.5.0)
Requirement already satisfied: regex>=2021.8.3 in
/usr/local/lib/python3.11/dist-packages (from nltk>=3.9->textblob)
(2024.11.6)
Requirement already satisfied: tqdm in /usr/local/lib/python3.11/dist-
packages (from nltk>=3.9->textblob) (4.67.1)
Building wheels for collected packages: nrclex
  Building wheel for nrclex (setup.py) ... e=NRCLex-3.0.0-py3-none-
any.whl size=43309
sha256=cea4041869eb998ac06c8f23506afe6e8cf1c8d67b609b57bf8fa4475fd87a3
8
  Stored in directory:
/root/.cache/pip/wheels/ed/ac/fa/7afddefd14f51c4a963ed291b9052746ed392
9473e5a33118d
Successfully built nrclex
Installing collected packages: nrclex
Successfully installed nrclex-3.0.0
[nltk_data] Downloading package brown to /root/nltk_data...
[nltk_data]   Unzipping corpora/brown.zip.
[nltk_data] Downloading package punkt_tab to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt_tab.zip.
[nltk_data] Downloading package wordnet to /root/nltk_data...
[nltk_data] Downloading package averaged_perceptron_tagger_eng to
[nltk_data]     /root/nltk_data...
[nltk_data]   Unzipping taggers/averaged_perceptron_tagger_eng.zip.
[nltk_data] Downloading package conll2000 to /root/nltk_data...
[nltk_data]   Unzipping corpora/conll2000.zip.
[nltk_data] Downloading package movie_reviews to /root/nltk_data...
[nltk_data]   Unzipping corpora/movie_reviews.zip.
Finished.
```

https://github.com/Hari21haran/Nm_hariharan_ds

```python
from textblob import TextBlob
from nrclex import NRCLex

# Sample social media posts
posts = [
    "I'm so happy with the new phone I bought!",
    "I can't believe how terrible the service was today.",
    "Feeling excited about my vacation next week.",
    "I miss my dog. Life feels so empty.",
    "What a beautiful sunset! Grateful for moments like this.",
    "I'm frustrated with everything going wrong lately."
]

# Function to analyze sentiment and emotions
def analyze_posts(posts):
    results = []
    for post in posts:
        blob = TextBlob(post)
        sentiment_polarity = blob.sentiment.polarity  # [-1.0, 1.0]
        sentiment_subjectivity = blob.sentiment.subjectivity  # [0.0,
1.0]

        emotion = NRCLex(post)
        top_emotions = emotion.top_emotions

        results.append({
            'text': post,
            'sentiment': 'Positive' if sentiment_polarity > 0 else
'Negative' if sentiment_polarity < 0 else 'Neutral',
            'polarity': sentiment_polarity,
            'subjectivity': sentiment_subjectivity,
            'emotions': top_emotions
        })
    return results

# Run analysis
analysis_results = analyze_posts(posts)

# Display output
for res in analysis_results:
    print("\nText:", res['text'])
    print("Sentiment:", res['sentiment'], f"(Polarity:
{res['polarity']:.2f}, Subjectivity: {res['subjectivity']:.2f})")
    print("Top Emotions:", res['emotions'])
```

```
Text: I'm so happy with the new phone I bought!
Sentiment: Positive (Polarity: 0.49, Subjectivity: 0.73)
Top Emotions: [('trust', 0.25), ('positive', 0.25), ('joy', 0.25),
('anticipation', 0.25)]
```

```
Text: I can't believe how terrible the service was today.
Sentiment: Negative (Polarity: -1.00, Subjectivity: 1.00)
Top Emotions: [('fear', 0.2), ('anger', 0.2), ('negative', 0.2),
('sadness', 0.2), ('disgust', 0.2)]

Text: Feeling excited about my vacation next week.
Sentiment: Positive (Polarity: 0.19, Subjectivity: 0.38)
Top Emotions: [('positive', 0.25), ('joy', 0.25), ('anticipation',
0.25)]

Text: I miss my dog. Life feels so empty.
Sentiment: Negative (Polarity: -0.10, Subjectivity: 0.50)
Top Emotions: [('fear', 0.0), ('anger', 0.0), ('anticip', 0.0),
('trust', 0.0), ('surprise', 0.0), ('positive', 0.0), ('negative',
0.0), ('sadness', 0.0), ('disgust', 0.0), ('joy', 0.0)]

Text: What a beautiful sunset! Grateful for moments like this.
Sentiment: Positive (Polarity: 1.00, Subjectivity: 1.00)
Top Emotions: [('positive', 0.5)]

Text: I'm frustrated with everything going wrong lately.
Sentiment: Negative (Polarity: -0.50, Subjectivity: 0.57)
Top Emotions: [('negative', 0.6666666666666666)]
```

14. Future scope

- *Add support for multiple languages to detect emotions globally.*
- *Enable real-time emotion tracking from live social media feeds.*
- *Improve accuracy using advanced models like RoBERTa or GPT.*

13. Team Members and Roles

Team Member	*Role*	*Responsibility*
Danish P	*Model Development*	*Responsible for building machine learning models and guiding model selection.*
Atsaya D	*Data Cleaning*	*Handles preprocessing tasks, including managing missing values*

		and improving data quality.
Karthika V	*Feature Engineering*	*Focuses on creating and refining features to enhance model performance.*
Abishek S	*Exploratory Data Analysis*	*Conducts data exploration to identify patterns and extract key insights.*
Hariharan R	*Documentation Support*	*Maintains clear and organized documentation throughout the project.*