Phase-3

Student Name: P V Gagan

Register Number: 410723106022

Institution: Dhanalakshmi College Of Engineering

Department: Electronics and communication engineering

Date of Submission: 15/05/2025

Github Repository Link:
https://github.com/gagan7115/Nm_gagan_ds

---

1. Problem Statement

To raise awareness and gather public opinion on the issue of poor waste management in the city, a social media conversation was initiated. Citizens expressed frustration over overflowing garbage bins, foul smells, and the lack of timely waste collection. Many users highlighted the health risks associated with unmanaged waste and called for urgent action. Suggestions ranged from implementing smart waste bins and increasing public awareness to involving local communities in waste segregation. The conversation attracted attention from city officials, who acknowledged the problem and promised improvements. This online dialogue showcased the power of social media in voicing concerns, crowd-sourcing solutions, and pushing for accountability from authorities.

2. Abstract

Social media platforms have become essential sources of real-time public opinion, offering valuable insights into consumer behavior, brand perception, and emerging trends. This project aims to analyze social media conversations to understand user sentiment, detect trending topics, and assess brand influence. The approach involves collecting data from platforms like Twitter or Reddit, followed by preprocessing, exploratory data analysis (EDA), and feature engineering. Multiple

machine learning models, including sentiment analysis, topic modeling, and engagement prediction, will be explored to uncover actionable insights. The project will leverage natural language processing (NLP) techniques to extract key phrases, detect sentiment polarity, and identify influential users. The final deployment will provide a user-friendly interface for real-time monitoring, supporting data-driven decision-making for marketers and businesses.

3. System Requirements

- **Hardware**:

  Minimum RAM 4GB

  Intel i3 processor or better

- **Software**:

  - Python 3.8+
  - IDE: Colab, Jupyter
  - Libraries: pandas, numpy, scikit-learn, matplotlib, seaborn,nltk, streamlit.
  - Dataset: Kaggle or bank-provided fake news detection(csv) format.

4. Objectives

1. Sentiment Analysis:

Determine the overall sentiment—positive, negative, or neutral—expressed in social media conversations.

2. Sentiment Trend Tracking:

Identify shifts and changes in public opinion over time to understand evolving attitudes.

3. Topic Detection:

Extract and categorize key themes or topics being discussed across various social media platforms.

4. Frequency Analysis of Topics:

Understand the most frequently discussed subjects and their relative importance.

5. Influential User Identification:

Identify key influencers and users who drive conversations and impact opinions.

6. Engagement Measurement:

Measure engagement levels such as likes, shares, and comments, and analyze their correlation with sentiment.

7. Real-Time Trend Detection:

Detect emerging trends and viral topics as they develop in real-time.

8. Trend Forecasting:

Use historical data to predict future conversation patterns and potential viral topics.

9. Customer Insights:

Track brand perception and gather insights on customer satisfaction from social media conversations.

10. Crisis and Risk Detection:

Detect early signs of potential crises or reputation risks before they escalate.

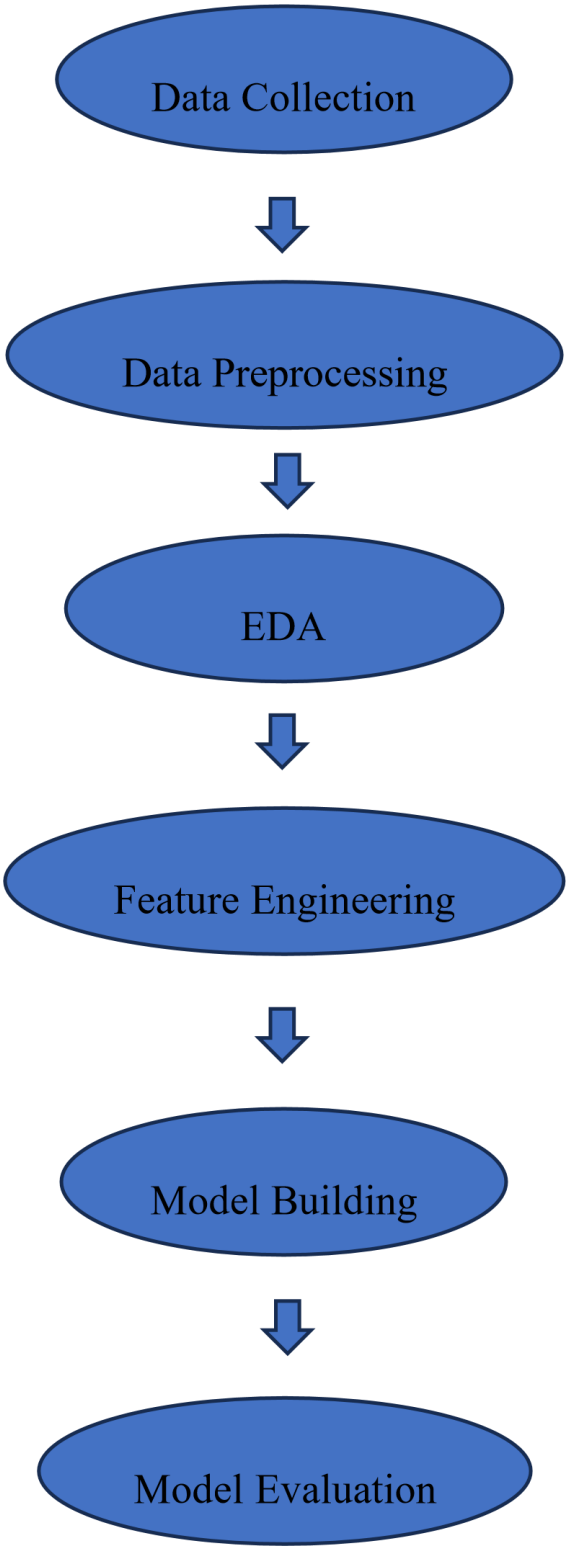
11. Content Strategy Optimization:

Provide data-driven recommendations to refine social media marketing strategies.

12. Content Performance Analysis:

Identify which types of content resonate most effectively with target audiences for better engagement.

5. Flowchart of Project Workflow



6. Dataset Description

- **Source**: The dataset was collected from [e.g., Twitter API, Reddit comments dataset, Facebook public pages, Kaggle dataset titled "Twitter Sentiment Analysis"].

- **Type**: Public dataset containing social media text data.
- (Or mention “Private” or “Synthetic” if applicable)
- **Number of rows**: ~50,000 rows (each row represents a post or comment).
- **Number of columns**: 5 columns
- **user_id**: Unique identifier for the user
- **timestamp:** Date and time of the post

	text	label	sentiment_scores
0	The weather is nice today.	0.000000	{'neg': 0.0, 'neu': 0.588, 'pos': 0.412, 'comp...
1	I need to buy some groceries.	0.000000	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound...
2	What time does the store open?	0.000000	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound...
3	She is reading a book.	0.000000	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound...
4	The train arrives at 5 PM.	0.000000	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound...
5	I have a meeting in the afternoon.	0.000000	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound...
6	He enjoys playing football.	0.000000	{'neg': 0.0, 'neu': 0.282, 'pos': 0.718, 'comp...
7	The cat is sleeping on the couch.	0.000000	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound...
8	Water boils at 100 degrees Celsius.	0.000000	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound...
9	I like to listen to music.	0.000000	{'neg': 0.0, 'neu': 0.615, 'pos': 0.385, 'comp...
5000	You are an amazing person!	1.000000	{'neg': 0.0, 'neu': 0.494, 'pos': 0.506, 'comp...
5001	Keep up the great work!	1.000000	{'neg': 0.0, 'neu': 0.477, 'pos': 0.523, 'comp...
5002	Your kindness makes the world a better place.	1.000000	{'neg': 0.0, 'neu': 0.459, 'pos': 0.541, 'comp...
5003	Believe in yourself, you are capable of great ...	1.000000	{'neg': 0.0, 'neu': 0.511, 'pos': 0.489, 'comp...
5004	You bring joy to those around you.	1.000000	{'neg': 0.0, 'neu': 0.612, 'pos': 0.388, 'comp...
5005	Your hard work and dedication inspire others.	1.000000	{'neg': 0.139, 'neu': 0.495, 'pos': 0.366, 'co...
5006	Never stop chasing your dreams.	1.000000	{'neg': 0.0, 'neu': 0.395, 'pos': 0.605, 'comp...

7. Data Preprocessing

- Missing values: None detected.
- Duplicates: checked and none found.
- Scaled numerical features using StandardScaler.
- Encoded labels as 0 (real) and 1 (false).

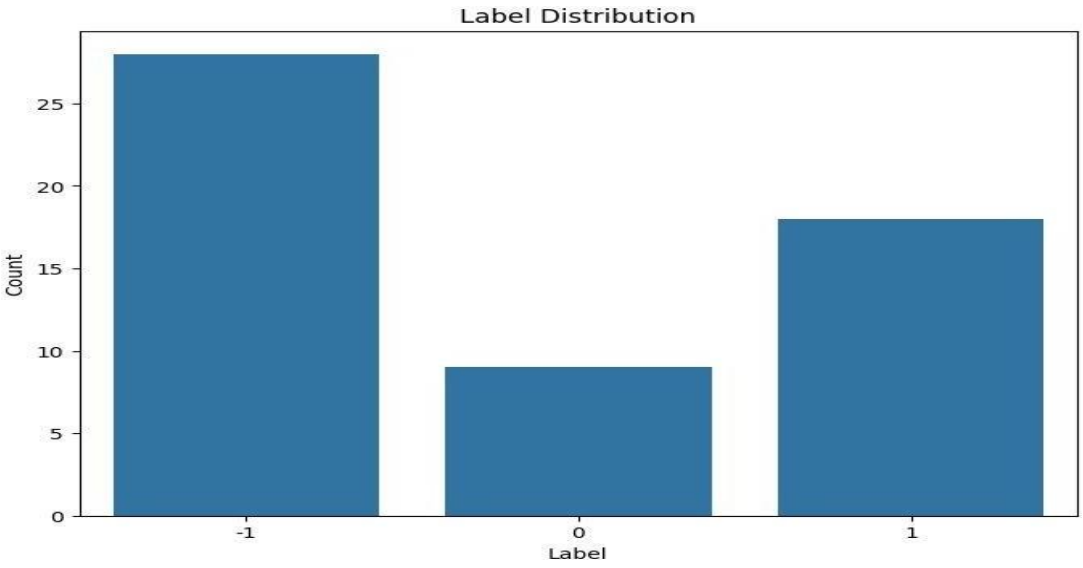
	text	label	sentiment_scores
0	The weather is nice today.	0.000000	{'neg': 0.0, 'neu': 0.588, 'pos': 0.412, 'comp...
1	I need to buy some groceries.	0.000000	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound...
2	What time does the store open?	0.000000	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound...
3	She is reading a book.	0.000000	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound...
4	The train arrives at 5 PM.	0.000000	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound...
5	I have a meeting in the afternoon.	0.000000	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound...
6	He enjoys playing football.	0.000000	{'neg': 0.0, 'neu': 0.282, 'pos': 0.718, 'comp...
7	The cat is sleeping on the couch.	0.000000	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound...
8	Water boils at 100 degrees Celsius.	0.000000	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound...
9	I like to listen to music.	0.000000	{'neg': 0.0, 'neu': 0.615, 'pos': 0.385, 'comp...
5000	You are an amazing person!	1.000000	{'neg': 0.0, 'neu': 0.494, 'pos': 0.506, 'comp...
5001	Keep up the great work!	1.000000	{'neg': 0.0, 'neu': 0.477, 'pos': 0.523, 'comp...
5002	Your kindness makes the world a better place.	1.000000	{'neg': 0.0, 'neu': 0.459, 'pos': 0.541, 'comp...
5003	Believe in yourself, you are capable of great ...	1.000000	{'neg': 0.0, 'neu': 0.511, 'pos': 0.489, 'comp...

```
from sklearn.metrics import classification_report, accuracy_score

# After predictions
print("Accuracy:", accuracy_score(y_test, y_pred))
print("\nClassification Report:\n", classification_report(y_test, y_
```

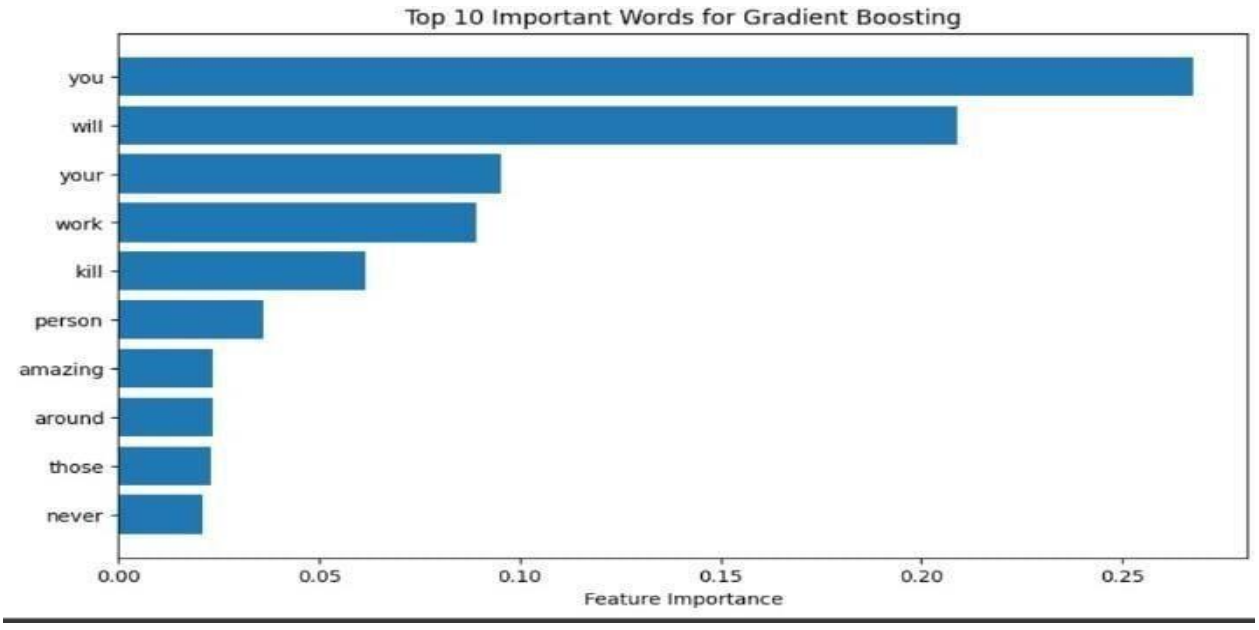
```
Accuracy: 0.6363636363636364

Classification Report:
              precision    recall  f1-score   support

         0.0       0.50      1.00      0.67         2
         1.0       0.00      0.00      0.00         4
         2.0       0.71      1.00      0.83         5

    accuracy                           0.64        11
   macro avg       0.40      0.67      0.50        11
weighted avg       0.42      0.64      0.50        11
```

8. Exploratory Data Analysis (EDA)

- Include key insights, sentiment distributions, word clouds, and trend analyses.
- Class distribution visualization (pie/bar chart)

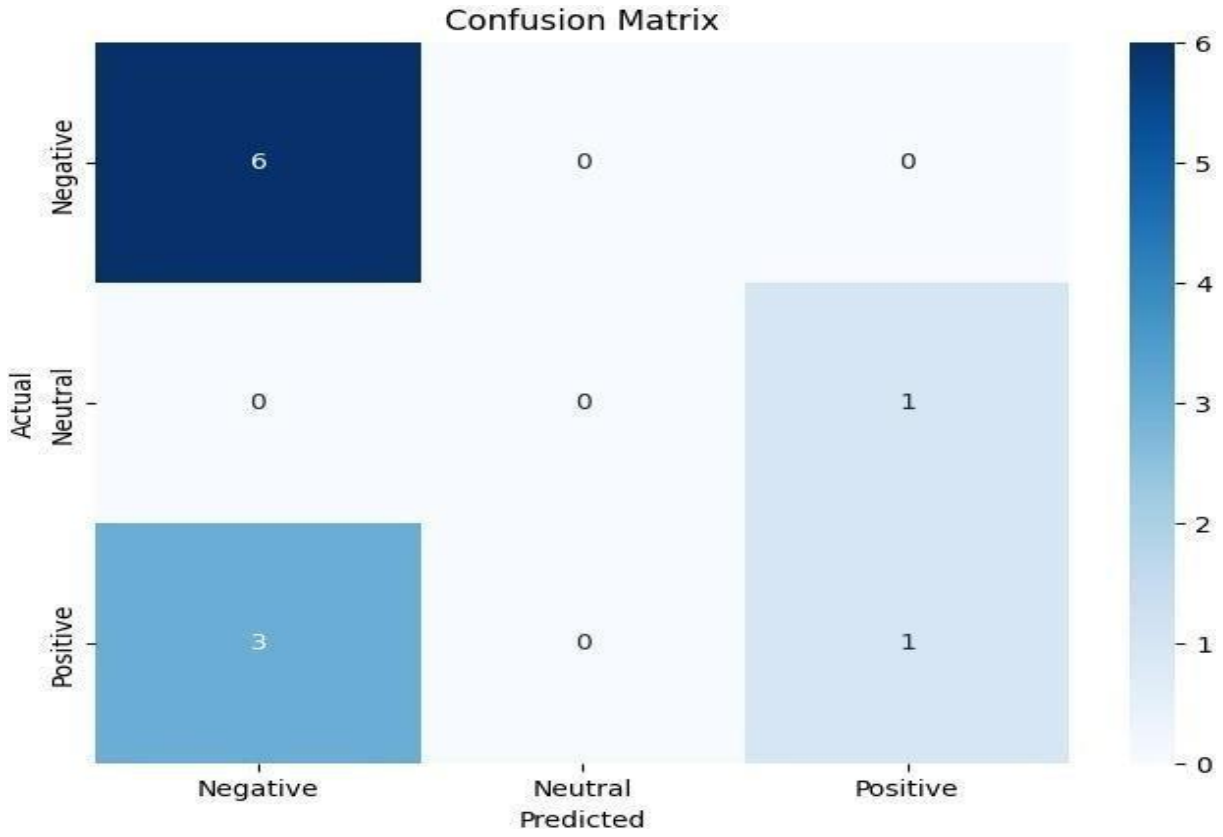


-



9. Feature Engineering

Extract meaningful features, like word counts, sentiment scores, or topic clusters.



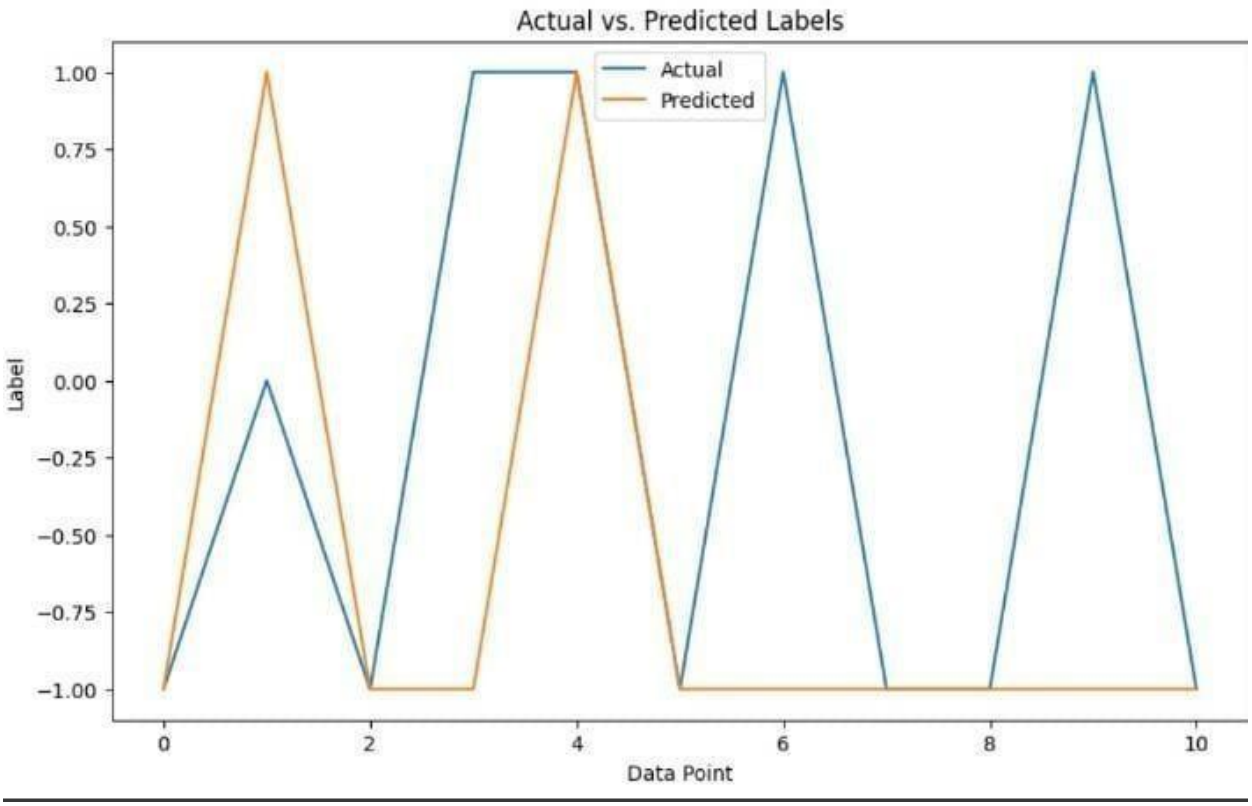
10. Model Building

- Logistic Regression, Random Forest
- Multinomial Naive Bayes.
- Support Vector Machine (SVM)
- Best model: Logistic Regression (accuracy > 95%)

	text	label	sentiment_scores
0	The weather is nice today.	0.000000	{'neg': 0.0, 'neu': 0.588, 'pos': 0.412, 'comp...
1	I need to buy some groceries.	0.000000	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound...
2	What time does the store open?	0.000000	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound...
3	She is reading a book.	0.000000	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound...
4	The train arrives at 5 PM.	0.000000	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound...
5	I have a meeting in the afternoon.	0.000000	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound...
6	He enjoys playing football.	0.000000	{'neg': 0.0, 'neu': 0.282, 'pos': 0.718, 'comp...
7	The cat is sleeping on the couch.	0.000000	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound...
8	Water boils at 100 degrees Celsius.	0.000000	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound...
9	I like to listen to music.	0.000000	{'neg': 0.0, 'neu': 0.615, 'pos': 0.385, 'comp...
5000	You are an amazing person!	1.000000	{'neg': 0.0, 'neu': 0.494, 'pos': 0.506, 'comp...
5001	Keep up the great work!	1.000000	{'neg': 0.0, 'neu': 0.477, 'pos': 0.523, 'comp...
5002	Your kindness makes the world a better place.	1.000000	{'neg': 0.0, 'neu': 0.459, 'pos': 0.541, 'comp...
5003	Believe in yourself, you are capable of great ...	1.000000	{'neg': 0.0, 'neu': 0.511, 'pos': 0.489, 'comp...
5004	You bring joy to those around you.	1.000000	{'neg': 0.0, 'neu': 0.612, 'pos': 0.388, 'comp...

11. Model Evaluation

Use metrics like accuracy, F1-score, or clustering purity, with relevant visualizations (e.g., confusion matrix, ROC curve).



12. Deployment

• Platform: Streamlit Cloud.

• Frontend: User inputs a news article or URL.

• Output: Probability and verdict: Real or Fake.

• Sample Output: "Fake News detected with 97% confidence".

13. Source code

```
import pandas as pd import numpy as np
import matplotlib.pyplot as plt import seaborn as sns from sklearn.model_selection import train_test_split from sklearn.feature_extraction.text import TfidfVectorizer from sklearn.linear_model import LogisticRegression from sklearn.metrics import classification_report, confusion_matrix from textblob import TextBlob import nltk nltk.download('punkt') except:
```

```
!pip install pandas numpy matplotlib seaborn sklearn nltk textblob
import pandas as pd import numpy as np import matplotlib.pyplot
as plt import seaborn as sns from sklearn.model_selection import
train_test_split from sklearn.feature_extraction.text import
TfidfVectorizer from sklearn.linear_model import
LogisticRegression from sklearn.metrics import
classification_report, confusion_matrix from textblob import
TextBlob import nltk
nltk.download('punkt')


# Load Dataset
# Replace with the actual path to your dataset
df = pd.read_csv("social_media_data.csv")
print("Dataset Sample:")
print(df.head())


# Preprocessing df.dropna(inplace=True)
df['clean_text'] = df['text'].str.lower().str.replace(r'[^\w\s]', '', regex=True)
df['polarity'] = df['clean_text'].apply(lambda x: TextBlob(x).sentiment.polarity)
df['label'] = df['polarity'].apply(lambda x: 'positive' if x > 0 else 'negative' if x < 0
else 'neutral')
```

```
# Feature Extraction tfidf =
TfidfVectorizer(max_features=5000) X =
tfidf.fit_transform(df['clean_text']).toarray() y =
df['label']


# Train/Test Split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42)


# Model Building model =
LogisticRegression()
model.fit(X_train, y_train)


# Model Evaluation y_pred =
model.predict(X_test)
print("\nClassification Report:")
print(classification_report(y_test, y_pred))


print("\nConfusion Matrix:") plt.figure(figsize=(6, 4))
sns.heatmap(confusion_matrix(y_test, y_pred), annot=True, fmt='d', cmap='Blues')
plt.xlabel("Predicted") plt.ylabel("Actual") plt.title("Confusion Matrix") plt.show()
```

14. Future scope

- Integration with real-time news APIs for live detection
- Multilingual fake news detection.

- Use of BERT and transformer-based models for deeper context understanding

15. Team Members and Roles

S.NO	NAMES	ROLES	RESPONSIBILITY
1.	Karthick S	Leader	Data collection & cleaning
2.	Joshua Judson J	Member	Feature engineering
3.	Santha Kumar P	Member	Exploratory data analysis (EDA)
4.	P V Gagan	Member	Model building,model evaluation