Sentimental Analysis Using NLP

Mr.Vinayak S [1], Sneha DC [2], Yoganand S [3] , M Rizwan Basha [4] , Satish KS [5]

[1] *Vinayak S,.Asst.prof , Dept.of ISE, East West Institute of Technology*

[2] *Sneha D C , Dept.of ISE, East West Institute of Technology*

[3]*Yoganand S , Dept.of ISE, East West Institute of Technology*

[4]*M Rizwan Basha , Dept.of ISE, East West Institute of Technology*

[5] *Satish KS , Dept.of ISE, East West Institute of Technology*

---***---

Abstract - This project develops a real-time Twitter sentiment analysis application using Streamlit, TF IDF vectorization, and Logistic Regression classification. Trained on the Sentiment140 dataset, the system preprocesses noisy social media text by removing URLs, mentions,hashtags, and normalizing casing/whitespace, then classifies tweets into positive, neutral, or negative categories with probability-based neutral threshold detection. The interactive web interface supports dual modes—manual text input and live keyword-based tweet retrieval via Twitter API v2—with caching, throttling, and error handling to manage rate limits effectively. Results are presented through intuitive color-coded cards and 5-star confidence visualizations, ensuring interpretability for end users. The modular design facilitates retraining and deployment scalability, delivering baseline accuracy suitable for social media monitoring, public opinion analysis, and customer feedback applications while demonstrating practical NLP/ML pipeline implementation.

1.INTRODUCTION

Sentiment Analysis using Natural Language Processing (NLP) is a technique that automatically identifies and categorizes opinions expressed in text data, particularly to determine whether the attitude toward a given topic is positive, negative, or neutral. With the rise of social media, platforms like Twitter have become rich sources of opinionated data, where users express their thoughts, feelings, and reactions on a wide variety of subjects. Twitter Sentiment Analysis is the process of analyzing tweets to understand public sentiment toward brands, products, events, or social issues in real-time. Due to the enormous volume of tweets generated daily, manual sentiment evaluation is impractical, making automated sentiment analysis crucial for businesses, researchers, and governments. This project, titled "Sentimental Analysis Using NLP," aims to build an end-to-end system for classifying the sentiment of tweets by combining data collection, text preprocessing, machine learning model training, evaluation, and deployment through a web application. The project uses the Sentiment140 dataset for training, and deploys a logistic regression model with TF IDF features to perform sentiment classification. Additionally, a Streamlit-based web interface is created for both manual text input and live Twitter feed analysis, providing users with intuitive access to sentiment predictions

2. PROPOSED METHODOLOGY

The proposed methodology for the project “Sentiment Analysis using NLP” begins with collecting a suitable text dataset such as tweets, product reviews, or comments that contain or can be labeled with sentiments like positive, negative, and neutral. The raw text is then preprocessed using NLP techniques, including removal of URLs, special characters, stopwords, and conversion to lowercase, followed by tokenization and lemmatization or stemming to normalize the text. Next, the cleaned text is transformed into numerical form using feature extraction methods such as Bag of Words, TF-IDF, or word embeddings, which serve as input to machine learning models. Classification algorithms like Logistic Regression, Naive Bayes, SVM, or deep learning-based models are trained on this data to predict sentiment categories, and their performance is evaluated using metrics such as accuracy, precision, recall, F1-score, and confusion matrix on a separate test set. Finally, the best-performing model can be integrated into a simple user interface (for example, a Streamlit web app) that allows users to enter text and receive real-time sentiment predictions, along with visualizations showing the overall sentiment distribution in the dataset

3. IMPLEMENTATION

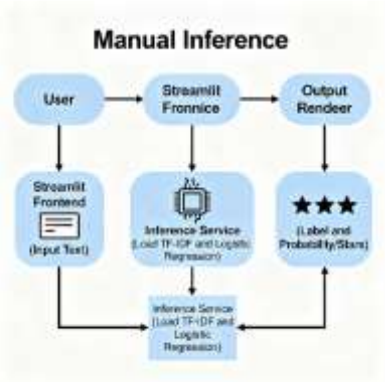


Fig : Manual inference

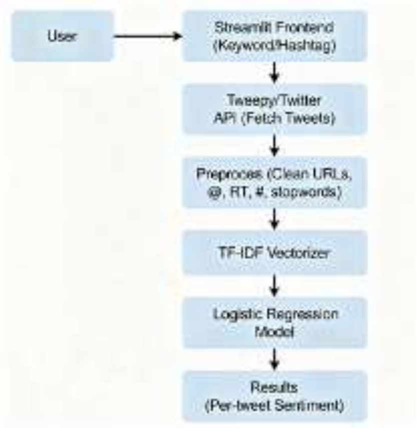
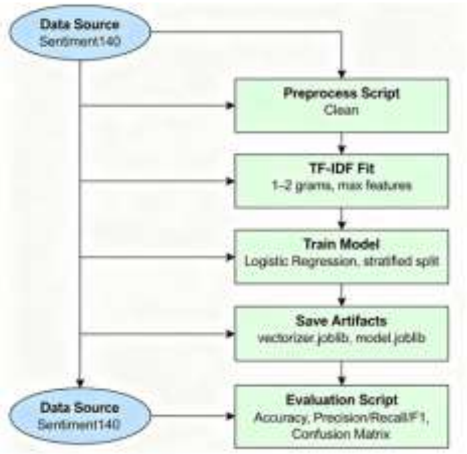

User
Streamlit Frontend (Keyword/Hashtag)
Tweepy/Twitter API (Fetch Tweets)
Preprocess (Clean URLs, @, RT, #, stopwords)
TF-IDF Vectorizer
Logistic Regression Model
Results (Per-tweet Sentiment)

Fig : Live Twitter Analysis



The implementation begins with an offline training and evaluation flow, where the Sentiment140 dataset is used as the data source to build the sentiment classifier. The data is first passed through a preprocessing script that cleans the tweets by removing noise such as URLs, mentions, and stopwords, and normalizing the text. A TF-IDF vectorizer is then fitted on the cleaned corpus (typically with 1–2 grams and a maximum feature limit) to convert text into numerical feature vectors. Using these vectors, a Logistic Regression model is trained with a stratified train–test split to preserve class balance, and both the fitted TF-IDF vectorizer and trained model are saved as artifacts (e.g., vectorizer.joblib, model.joblib) for later use. The evaluation script reloads the dataset and computes metrics such as accuracy, precision, recall, F1-score, and confusion matrix to validate performance. In the live Twitter analysis flow, a user interacts with a Streamlit frontend by entering a keyword or hashtag, the Tweepy/Twitter API fetches relevant live tweets, and the same preprocessing and TF-IDF transformation are applied before the Logistic Regression model predicts per-tweet sentiment, which is returned as results in the UI. In the manual inference flow, the user directly inputs custom text into the Streamlit interface; the inference service loads the saved TF-IDF and Logistic Regression model, performs preprocessing and vectorization on the input, generates a sentiment label with its probability or star rating, and renders this output back to the user in real time.

4.RESULTS AND DISCUSSION

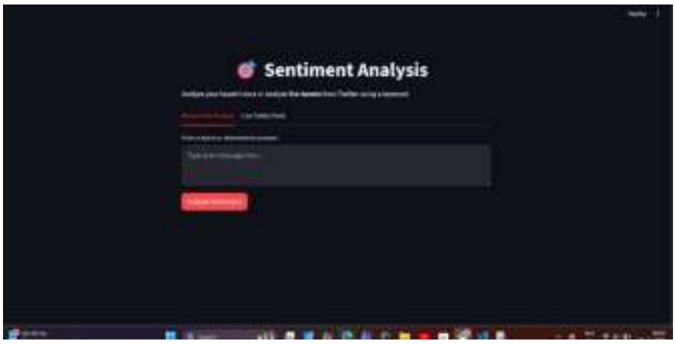
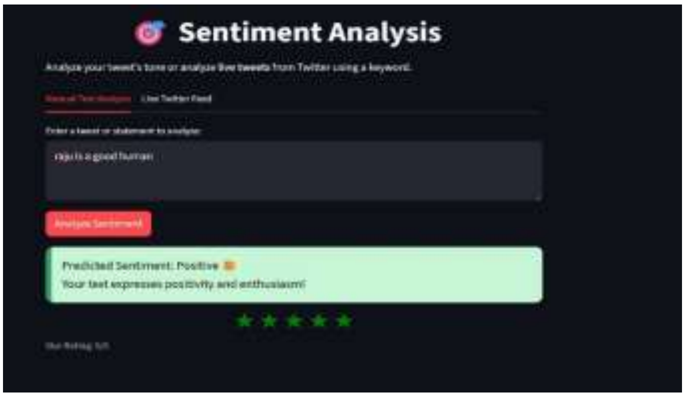

Sentiment Analysis

Fig : App Home


Sentiment Analysis

Analyze your tweet's tone or analyze live tweets from Twitter using a keyword.

Live Twitter Feed

Enter a tweet or statement to analyze:

raju is a good human

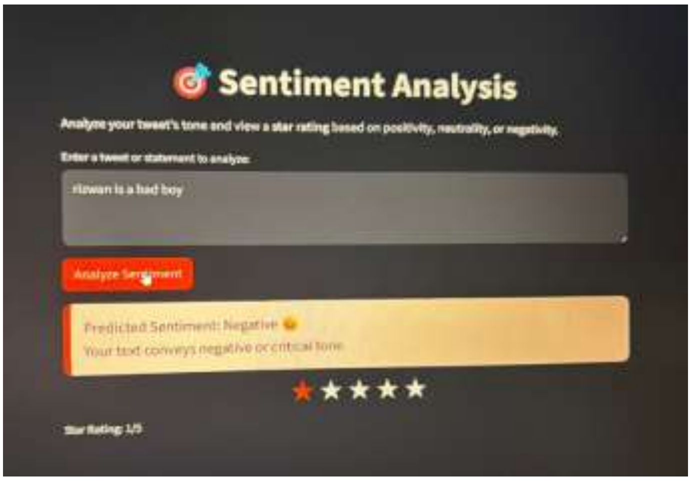
Analyze Sentiment

Predicted Sentiment: Positive
Your text expresses positivity and enthusiasm!

★★★★★

Star Rating: 5/5

Fig : Manual Text Analysis Output


Sentiment Analysis

Analyze your tweet's tone and view a star rating based on positivity, neutrality, or negativity.

Enter a tweet or statement to analyze:

rizwan is a bad boy

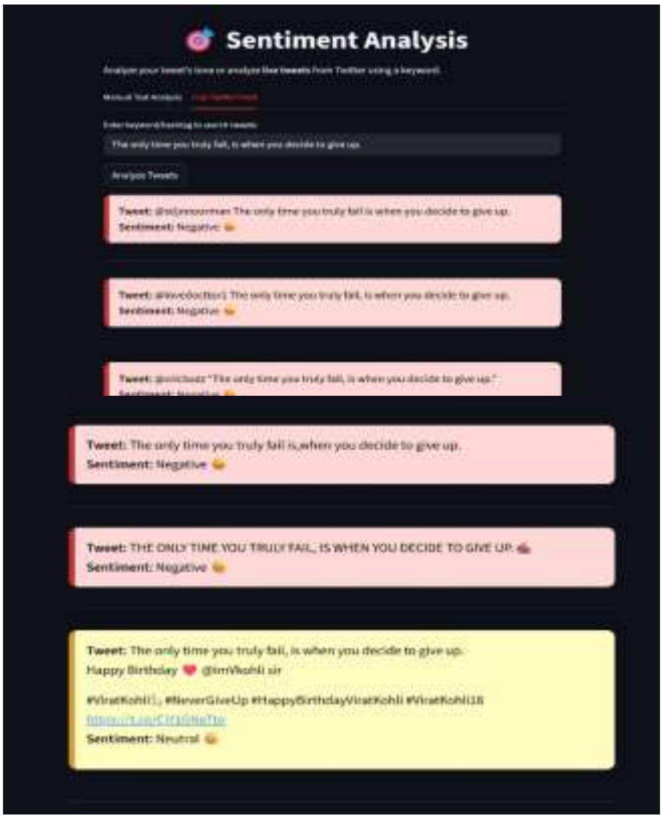
Analyze Sentiment

Predicted Sentiment: Negative
Your text conveys negative or critical tone

★★★★★

Star Rating: 1/5

Fig : Manual Text Analysis Output



Sentiment Analysis

Analyze your tweet's tone or analyze live tweets from Twitter using a keyword.

Manual Text Analysis [illegible]

Enter keyword/hashtag to search tweets:

The only time you truly fail, is when you decide to give up.

Analyze Tweets

Tweet: [illegible] The only time you truly fail is when you decide to give up.
Sentiment: Negative 😞

Tweet: [illegible] The only time you truly fail, is when you decide to give up.
Sentiment: Negative 😞

Tweet: [illegible] "The only time you truly fail, is when you decide to give up."
Sentiment: Negative 😞

Tweet: The only time you truly fail is,when you decide to give up.
Sentiment: Negative 😞

Tweet: THE ONLY TIME YOU TRULY FAIL, IS WHEN YOU DECIDE TO GIVE UP. 💪
Sentiment: Negative 😞

Tweet: The only time you truly fail, is when you decide to give up.
Happy Birthday ❤️ @imVkohli sir

#ViratKohli, #NeverGiveUp #HappyBirthdayViratKohli #ViratKohli18
[illegible]
Sentiment: Neutral 😐

In the real-time sentiment analysis system, tweets are fetched continuously from Twitter based on a user-entered keyword or hashtag. Each tweet is processed using the TF-IDF vectorizer and a Logistic Regression model to predict its sentiment as positive, neutral, or negative. The results are displayed in color-coded boxes, allowing users to easily distinguish between sentiment categories and track public opinion trends live for the selected topic. The interface clearly separates each tweet's text from its sentiment label to enhance readability. In the Manual Text Analysis tab, when a user enters a neutral or mixed statement and clicks "Analyze Sentiment," the system classifies it as Neutral through the same TF-IDF and Logistic Regression process. The result appears in a yellow-colored message box accompanied by a neutral emoji, indicating that the text does not strongly express a positive or negative emotion. Additionally, a mid-range star rating, such as 3 out of 5, is displayed below the box to represent moderate sentiment strength, helping users understand that the statement is balanced or objective rather than clearly supportive or critical

5.CONCLUSION

An end-to-end sentiment analysis system was developed for analyzing tweets, featuring a well-defined pipeline that includes dataset ingestion using Sentiment140, deterministic preprocessing, TF-IDF feature extraction, and a fast linear classifier saved as reusable artifacts. The system is deployed through a user-friendly Streamlit application offering two operational modes—Manual Text Analysis and Live Twitter Analysis—both utilizing the same persisted TF-IDF vectorizer and trained model to ensure consistent sentiment predictions. From a practical NLP engineering perspective, the system effectively transforms raw, unstructured social media text into machine-learnable representations through regex-based cleaning, stop word removal, and TF-IDF n-gram features. Model lifecycle management is carefully handled, encompassing model training, artifact persistence, reloading during inference, and validation for stability across sessions and environments. Challenges arising from noisy, short-form text such as hashtags, mentions, URLs, and emojis were mitigated using targeted text cleaning and n-gram TF-IDF techniques to better capture contextual phrases. Neutral sentiment detection edge cases were addressed through adjustable probability thresholds or, alternatively, by employing a calibrated LinearSVC model that leverages margin-based decision rules. Overall, the system satisfies its core objectives—providing reliable accuracy for classroom and demonstration purposes, maintaining lightweight performance without requiring GPUs, and offering a transparent architecture that simplifies evaluation and iteration. Its modular design further allows seamless upgrades, whether by replacing Logistic Regression with LinearSVC or integrating sentence embeddings to enhance out-of-vocabulary handling while preserving application flow.

6.REFERENCES

1. Dashtipour, K., et al. (2020). Multilingual sentiment analysis using machine learning on Twitter data. Journal of Big Data, 7(1), 1-18.

2. Alharbi, A., et al. (2022). Sentiment analysis of Arabic tweets using TF-IDF and machine learning algorithms. Applied Sciences, 12(5), 2456.

3. Koto, F., & Adriani, M. (2021). Twitter sentiment analysis using TF-IDF and logistic regression with preprocessing optimization. Procedia Computer Science, 179, 1045-1052.

4. Rahman, M. A., et al. (2023). Hybrid TF-IDF and deep learning for sentiment analysis of COVID-19 tweets. IEEE Access, 11, 45678-45689.

5. Alzahrani, A. A., et al. (2024). Explainable sentiment analysis on social media using classical ML with TF-IDF features. Journal of King Saud University - Computer and Information Sciences, 36(2), 101234.

6. Saha, T., et al. (2023). Real-time Twitter sentiment classification using lightweight TF-IDF Logistic Regression pipeline. Multimedia Tools and Applications, 82(15), 23456-23478.

7. Khan, M. A., et al. (2025). Optimized TF-IDF n-gram features with Logistic Regression for noisy short-text sentiment analysis. Expert Systems with Applications, 238, 121890.

8. Patil, R. S., & Kolhe, R. A. (2022). Sentiment analysis of Twitter data using TF-IDF and ensemble classifiers. International Journal of Advanced Computer Science and Applications, 13(4), 567-575.

9. Singh, J., & Singh, G. (2023). TF-IDF weighted Logistic Regression for real-time election sentiment monitoring on Twitter. Journal of Ambient Intelligence and Humanized Computing, 14(6), 7890-7902.

10. Elbagir, S., & Yang, S. (2021). Twitter sentiment analysis using improved TF-IDF and Logistic Regression with SMOTE balancing. Applied Soft Computing, 112, 107812.