CS306(Machine Learning): Project Report

Project Title: Squid Game Sentiment Analysis Dashboard

Objective:

The objective of this project is to develop an interactive and user-friendly sentiment analysis dashboard that allows users to:

- Analyze the sentiment of individual text inputs.
- Process and analyze sentiment from bulk data provided in CSV format.
- Generate insightful visualizations and metrics summarizing sentiment data.

Problem Statement:

Sentiment analysis is essential for understanding public opinion, user feedback, and emotional tones within text data. However, existing sentiment analysis tools often lack simplicity or visualization capabilities. This project addresses the need for a straightforward yet robust dashboard where users can input text, upload datasets, and receive actionable insights.

Methodology:

1. Requirement Analysis

- **Objective**: Identify the key features and functionalities required for the dashboard.
    - Single-text sentiment analysis.
    - Bulk sentiment analysis via CSV upload.
    - Visualizations for data insights.
    - Results download for offline use.

---

2. System Design

2.1 User Interface (Frontend)

- **Streamlit Framework**:
  - Used to create an intuitive and responsive interface.
  - Sidebar for navigation and main area for analysis and visualizations.
  - Widgets for text input, file upload, and download options.

2.2 Data Processing (Backend)

- **Text Cleaning**:
  - Leveraged `cleantext` for preprocessing input text by removing unnecessary elements (e.g., stopwords, punctuation).
  - Ensured standardized and clean text for accurate analysis.
- **Sentiment Analysis**:
  - Used `TextBlob` to compute:
    - **Polarity**: A score ranging from -1 (negative sentiment) to 1 (positive sentiment).
    - **Subjectivity**: A score ranging from 0 (objective text) to 1 (highly subjective text).
  - Applied a custom function using the **tanh** function to scale and classify polarity scores into categories (Positive, Neutral, Negative).

2.3 Data Aggregation and Visualization

- **Data Aggregation**:
  - Used `Pandas` to group, filter, and compute summary metrics on the processed data.
- **Visualizations**:
  - Created bar charts for sentiment distributions using `Matplotlib` and `Seaborn`.
  - Plotted boxplots to display polarity score distributions.
  - Used time-series line plots for datasets containing timestamp columns.

---

3. Implementation

3.1 Text Input Analysis

- The user inputs text through a text box.
- The input text is processed and analyzed with `TextBlob` for polarity and subjectivity.
- Results are displayed in a tabular format using Streamlit's `metric` widget.

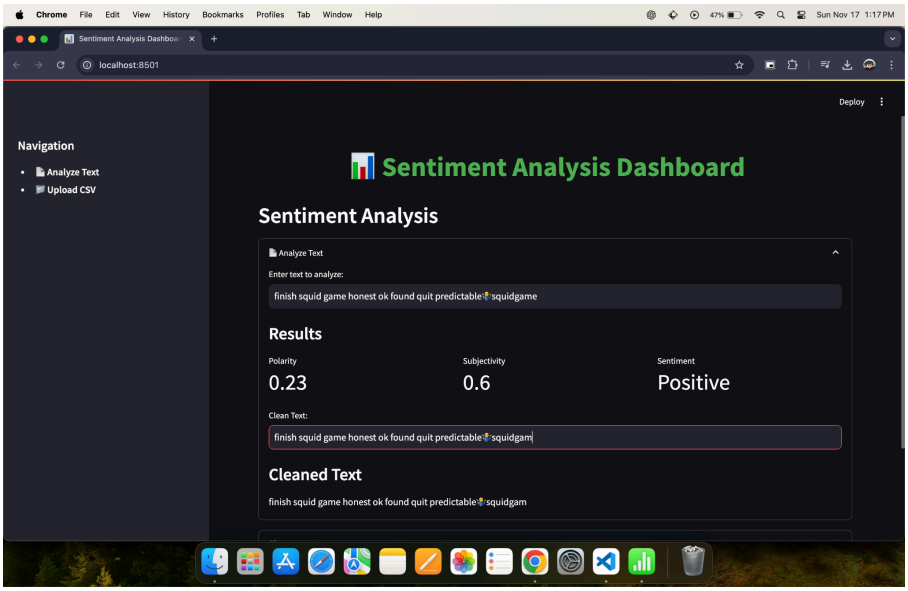
3.2 Bulk CSV Analysis

1. **File Upload**:

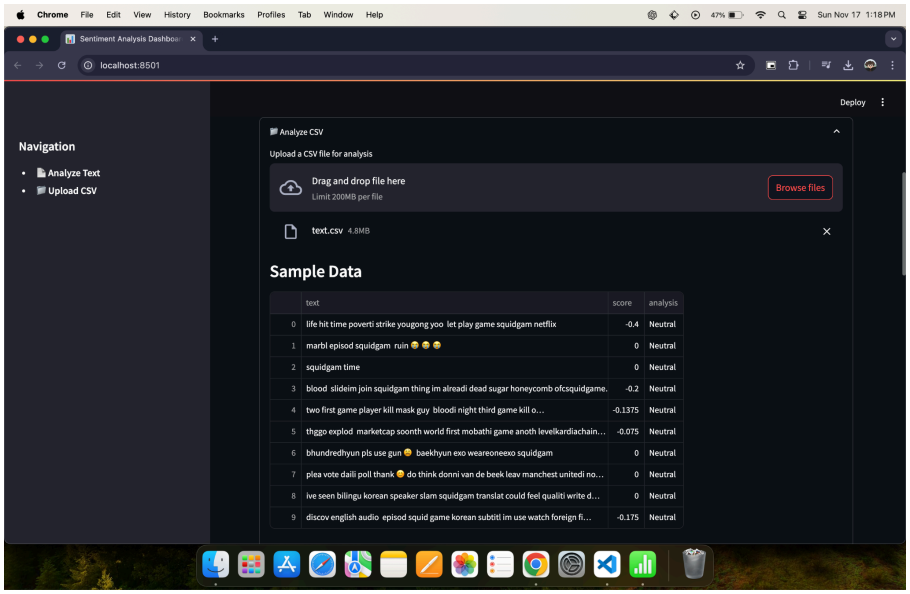
    - Users upload a CSV file with a mandatory `text` column.
2. **Text Cleaning**:
    - Clean each text entry using `cleantext` for preprocessing.
3. **Sentiment Analysis**:
    - Apply `TextBlob` to calculate polarity and derive sentiment (Positive, Neutral, or Negative) using the tanh-based scaling.
4. **Data Visualization**:
    - Generate bar charts and boxplots for sentiment distributions.
    - Plot time-series data trends if the dataset includes a `timestamp` column.
5. **Results Export**:
    - Processed data, including scores and sentiment classifications, can be downloaded as a CSV file.

---

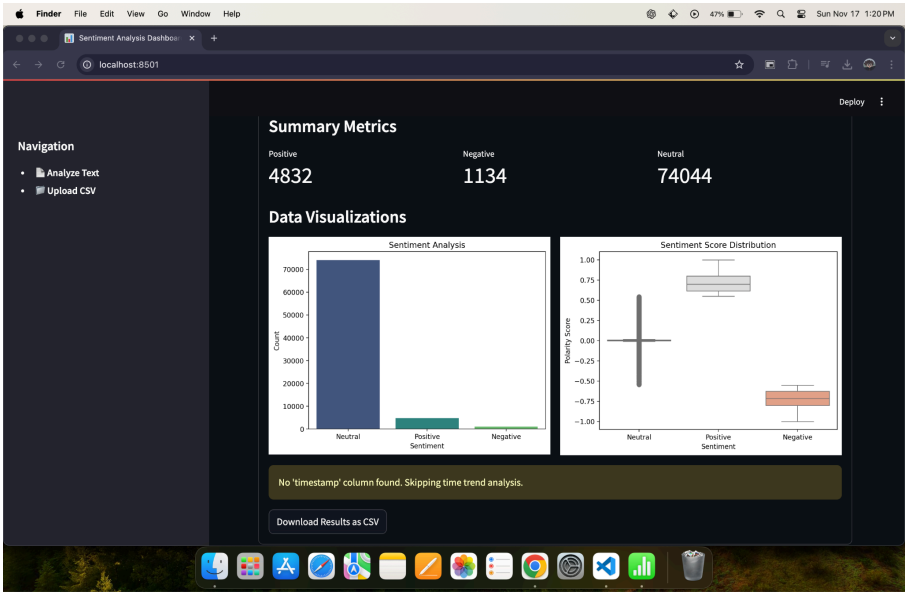
4. Pictures Demonstrating the Application Interface



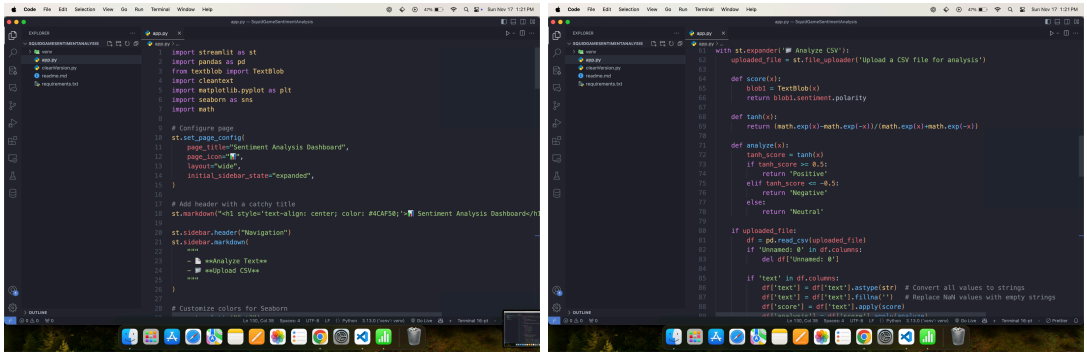
(Fig 4.1: Analyze Text page)



(Fig 4.2: Upload CSV page)



(Fig 4.3: Report Analysis of the output)



(Fig 4.4: Snapshot of the code)

---

5. Evaluation and Testing

- The application was tested with various datasets, including short texts, large CSV files, and time-sensitive data.
- Performance benchmarks were conducted to ensure fast processing and smooth visualization rendering.

Technologies and Libraries Used:

1. Streamlit

- Provides a web-based interface for real-time analysis.
- Displays metrics, visualizations, and input widgets.

2. Pandas

- Processes and aggregates large datasets efficiently.
- Handles missing or malformed data entries.

3. TextBlob

- Performs sentiment scoring with polarity and subjectivity metrics.
- Provides a simple, intuitive interface for analyzing text.

4. Cleantext

- Cleans raw text by removing unwanted characters, spaces, and stopwords.
- Enhances the quality of input data for better sentiment analysis.

5. Matplotlib and Seaborn

- Creates visually engaging plots to showcase results.
- Summarizes sentiment trends, distributions, and patterns effectively.

6. Math Library

- Implements the **tanh** function to normalize sentiment scores and classify them into qualitative categories.

7. Streamlit Caching

- Improves performance by caching repeated computations, such as CSV file conversions.

Key Features Delivered:

1. **Single Text Analysis**:
    - Real-time metrics for polarity, subjectivity, and sentiment classification.
    - Text cleaning functionality for better preprocessing.
2. **Bulk Text Analysis**:
    - Automated sentiment scoring for datasets.
    - Sentiment classification into Positive, Neutral, or Negative categories.
    - Summary metrics and downloadable results.
3. **Visualizations**:
    - Bar charts and boxplots for distribution analysis.
    - Time-series sentiment trends for timestamped data.
4. **Export Functionality**:
    - Download processed results as a CSV file.

Conclusion

This project successfully leverages multiple Python libraries to build a robust and interactive sentiment analysis dashboard. The integration of text cleaning, real-time processing, sentiment scoring, and visualization ensures a comprehensive and user-friendly solution for analyzing textual data.

Benefits

- Streamlined sentiment analysis for individual users and businesses.
- Scalable for various dataset sizes and use cases.
- Easy-to-understand insights through visualizations and metrics.

Future Enhancements

- Incorporate support for multilingual text analysis.
- Provide advanced visualization options like word clouds for better context.

Date: 17th Nov. 2024

Prepared by:
Sudipto Ray (Roll No.: 2201206)
B.Tech (3rd Year)
Indian Institute of Information Technology, Guwahati
Department of Computer Science and Engineering (CSE)