LAWRENCE TECHNOLOGICAL UNIVERSITY
MATH AND COMPUTER SCIENCE
COLLABORATIVE RESEARCH PROJECT 1

SOCIAL MEDIA SENTIMENT ANALYSIS SYSTEM

SULEMAN ALI SAIF MIR
SUMAYYA SANA SYEDA
July 19, 2024

ABSTRACT

The Social Media Sentiment Analysis System is designed to analyze public sentiment on various topics such as products, events, and political issues by leveraging data from social media platforms. This system aims to provide real-time sentiment analysis, visualize sentiment trends, and offer insights for businesses, researchers, and policymakers.

INTRODUCTION

Project Overview

The increasing prevalence of social media has generated vast amounts of data that reflect public opinion on a myriad of topics. This project aims to harness this data to analyze sentiment, thereby providing valuable insights for decision-making processes in various domains.

Objectives

- To develop a system capable of gathering and analyzing social media data.
- To provide real-time sentiment analysis and visualization of sentiment trends.
- To create an intuitive user interface for searching and viewing sentiment results.

Scope

- Platforms: Initially focused on Twitter, with potential expansion to other social media platforms.
- Analysis: Utilizing natural language processing (NLP) techniques for sentiment analysis.
- Visualization: Interactive dashboards and comprehensive reports.
- Deployment: A web application accessible via cloud services.

General Description

Product Perspective

The system integrates with social media platforms via APIs and employs web scraping tools for data collection. It preprocesses, analyzes, and visualizes the data to provide actionable insights into public sentiment.

Product Functions

- Data Collection: Using APIs (e.g., Tweepy for Twitter) and web scraping tools (e.g., BeautifulSoup).
- Data Cleaning: Removing noise such as special characters, URLs, and stop words, followed by tokenization and lemmatization.
- Sentiment Analysis: Utilizing tools like VADER and TextBlob for sentiment classification and training machine learning models (e.g., logistic regression, SVM, LSTM).
- Visualization: Creating bar charts, word clouds, and sentiment timelines using tools like Plotly and Tableau.
- User Interface: Developing a web application using frameworks such as Flask or Django, deployed on cloud platforms like Heroku or AWS.

User Characteristics

- End-Users: Individuals or organizations interested in public sentiment.
- Researchers: Analysts conducting sentiment studies.
- Businesses: Companies analyzing market trends and customer feedback.
- Policymakers: Government officials monitoring public opinion.

Specific Requirements

Functional Requirements

1. Data Collection:
   - Use Tweepy to gather data from Twitter.
   - Implement web scraping with BeautifulSoup for additional platforms.

2. Data Cleaning:
   - Remove special characters, URLs, and stop words.
   - Tokenize and lemmatize/stem text data.

3. Sentiment Analysis:
   - Use VADER and TextBlob for initial sentiment classification.
   - Train machine learning models such as logistic regression, SVM, and LSTM.

4. Visualization:
   - Create bar charts, word clouds, and sentiment timelines.
   - Develop interactive dashboards with Tableau.

5. User Interface:
   - Design a web application using Flask or Django.
   - Deploy the application on Heroku or AWS.

Non-Functional Requirements

- Performance: The system should provide sentiment analysis results within 5 seconds and handle up to 10,000 requests per hour.
- Usability: The user interface should be intuitive and accessible to users with disabilities.
- Reliability: Ensure 99.9% uptime and implement failover mechanisms for continuity.
- Security: Secure user data and implement authentication and authorization.
- Maintainability: Follow best practices for code organization and documentation, ensuring ease of updates.

Data Analysis and Modeling

Data Cleaning

Data cleaning involves converting text to lowercase, removing text within square brackets, URLs, HTML tags, punctuation, newline characters, and words containing numbers. This ensures the data is normalized and ready for analysis.

Jaccard Similarity

The Jaccard Similarity index is used to measure the similarity between two sets of tokens, calculated as the size of the intersection divided by the size of the union of the sets. This metric is particularly useful for comparing the full text and selected text to understand their overlap.

python

```
def jaccard(str1, str2):
    a = set(str1.lower().split())
    b = set(str2.lower().split())
    c = a.intersection(b)
    return float(len(c)) / (len(a) + len(b) - len(c))
```

Data Visualization

The visualization component includes generating bar charts, word clouds, and sentiment timelines to present the data in an intuitive manner. These visualizations help in identifying trends and patterns in sentiment data.

Conclusion

The Social Media Sentiment Analysis System provides a comprehensive solution for analyzing and visualizing public sentiment on social media platforms. By leveraging advanced NLP techniques and interactive visualizations, the system offers valuable insights for businesses, researchers, and policymakers. Future enhancements could include emotion detection, aspect-based sentiment analysis, and support for multiple languages.