B.M.S COLLEGE OF ENGINEERING BENGALURU
Autonomous Institute, Affiliated to VTU

FDS AAT
Report on

Twitter Sentiment Analysis

Submitted in partial fulfillment of the requirements for AAT

Bachelor of Engineering
in
Computer Science and Engineering(Data Science)

Submitted by:

Manoj R
[1BM22CD037]

Mayur Hegde
[1BM22CD040]

Department of Computer Science and Engineering
B.M.S College of Engineering
Bull Temple Road, Basavanagudi, Bangalore 560 019
2023-2024

B.M.S COLLEGE OF ENGINEERING

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING (DATA SCIENCE)

DECLARATION

We, Manoj R [1BM22CD037], Mayur Hegde [1BM22CD040] students of 3rd Semester, B.E, Department of Computer Science and Engineering (Data Science), BMS College of Engineering, Bangalore, hereby declare that, this AAT Project entitled "Twitter Sentiment Analysis" has been carried out in Department of CSE(DS), BMS College of Engineering, Bangalore during the academic semester October 2023 - March 2024. We also declare that to the best of our knowledge and belief, the AAT Project report is not from part of any other report by any other students.

Signature of the Candidates

Manoj R [1BM22CD037] Mayur Hegde [1BM22CD040]

BMS COLLEGE OF ENGINEERING

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING (DATA SCIENCE)

CERTIFICATE

This is to certify that the AAT Project titled "**Twitter Sentiment Analysis"** has been carried out by Manoj R [1BM22CD037] and Mayur Hegde [1BM22CD040] during the academic year 2023-2024.

Signature of the Faculty in Charge

Table of Contents

1. Introduction

1.1 Research Objective

Social media platforms like Twitter offer a treasure trove of data reflecting public opinion, emerging trends, and user behavior in real-time. This report explores the potential of Twitter data for social research by analyzing a collection of tweets. The primary objective of this research is to uncover hidden patterns within the data, understand user sentiment towards various topics, and gain insights into the issues capturing user attention.

1.2 Data Source Selection and Justification

The chosen dataset for this analysis consists of a collection of tweets, potentially retrieved from Twitter's public API or obtained from a reputable third-party source. Publicly available Twitter data offers a rich and valuable resource for researchers due to several compelling advantages:

- **Real-time Nature:** Twitter data captures opinions and discussions as they unfold, providing a snapshot of current events and trending topics.
- **Breadth of Coverage:** The platform encompasses a diverse range of users and viewpoints, offering insights into public sentiment across various demographics and social groups.
- **Data Volume:** The sheer volume of tweets allows researchers to analyze large datasets and identify statistically significant patterns.

2. Hardware and Software Requirements

2.1 Hardware

The hardware requirements for this project will depend on the chosen dataset size and the complexity of the analysis. A computer with sufficient processing power and memory to handle the chosen dataset size is essential. Depending on the chosen approach, this could range from a standard laptop to a machine with high processing capabilities.

2.2 Software

- **Python (Programming Language):** Python is the primary programming language chosen for this project due to its readability, extensive libraries, and large developer community.
- **Libraries:**
  - **pandas (Data Manipulation and Analysis):** Pandas offers powerful data structures and tools for cleaning, transforming, and analyzing Twitter data.
  - **nltk (Natural Language Processing):** NLTK provides functionalities

for various NLP tasks like tokenization, stop word removal, and stemming/lemmatization, essential for text analysis.

- **matplotlib/seaborn (Data Visualization):** These libraries allow researchers to create informative and visually appealing charts and graphs to present their findings.
- **TextBlob (Sentiment Analysis):** TextBlob simplifies sentiment analysis by providing pre-trained models to categorize tweet sentiment as positive, negative, or neutral.

3. Types of Analysis

The analysis will delve into several key areas to extract meaningful information from the tweet data.

- **Engagement Analysis:** User engagement metrics like retweets and likes provide insights into content popularity and user interaction. Analyzing these metrics can help identify tweets that resonate with the audience and spark discussions.
- **Sentiment Analysis:** Identifying the overall sentiment expressed within the tweets is crucial for understanding public opinion on various topics. Sentiment analysis can reveal positive or negative trends surrounding specific events, products, or social issues.
- **Text Analysis:** By analyzing the content of tweets, researchers can uncover frequently used words, phrases, and potential topics of discussion. Techniques like N-grams (sequences of n words) or topic modeling can be used to identify emerging themes and areas of user interest.
- **Optional Analysis:** Depending on the research goals and the dataset characteristics, the analysis can be extended to include:
  - Network analysis: Exploring user interactions and relationships within the Twitter network can reveal communities of interest and how information flows between users.
  - Time series analysis: Examining trends over time can identify how sentiment or topic popularity evolves, potentially uncovering seasonal patterns or reactions to current events.

4. Implementation

4.1 Source Code Considerations

The specific source code will vary depending on the chosen libraries, analysis techniques, and the structure of the Twitter data (CSV, JSON, etc.). However, a general outline might involve the following steps:

1. **Data Loading and Cleaning:**
   - Load the Twitter data using pandas.
   - Perform data cleaning tasks like removing irrelevant information (e.g., user IDs, URLs), handling missing values, and potentially identifying and removing duplicate tweets.

2. **Text Preprocessing:**
   - Preprocess the tweet text using NLTK or similar libraries. This may involve:
     - Converting text to lowercase
     - Removing special characters and URLs
     - Tokenizing words (splitting text into individual words)
     - Removing stop words (common words like "the", "a", "an") that don't contribute much meaning
     - Optionally, stemming/lemmatization (reducing words to their base form)

3. **Engagement Analysis:**
   - Calculate engagement metrics like average retweets and likes per tweet to understand user interaction levels.

2. **Sentiment Analysis :**

    - Use TextBlob or other sentiment analysis libraries to determine the sentiment polarity (positive, negative, or neutral) of each tweet.

3. **Text Analysis:**

    - Analyze the preprocessed text data to identify:
        - Word frequency distribution: Identify the most frequently used words to understand the general themes discussed within the tweets.
        - N-grams (sequences of n words): Analyze frequently occurring sequences of words (bigrams, trigrams) to uncover potential topics and phrases used by the community.
        - Optionally, topic modeling techniques can be applied to discover latent topics within the data.

4. **Visualization:**

    - Utilize libraries like matplotlib and seaborn to create informative visualizations that effectively communicate the analysis findings. This could include:
        - Bar charts or histograms to represent the distribution of retweets, likes, and sentiment scores.
        - Word clouds to visualize the most frequent words and highlight prominent topics.
        - Scatter plots to explore relationships between variables (e.g., retweets vs. sentiment).

4.2 Python Data Structures and Libraries Used

- **pandas:** DataFrames and Series for data manipulation, cleaning, and analysis.

- **NLTK:** Libraries for tokenization, stop word removal, stemming/lemmatization (text processing).
- **matplotlib/seaborn:** Libraries for creating various charts and visualizations.
- **TextBlob:** Sentiment analysis library for classifying tweet sentiment.

4.3 Experimental Analysis and Results

Note: Due to the absence of the actual code and data, this section cannot provide specific results. However, we can outline the general approach to presenting the findings.

- **Engagement Analysis:** Report the average number of retweets and likes per tweet, potentially comparing these metrics across different categories (e.g., positive vs. negative sentiment).
- **Sentiment Analysis:** Describe the overall sentiment distribution within the data (percentage of positive, negative, and neutral tweets). Analyze if sentiment varies based on specific topics or keywords.
- **Text Analysis:** Present the most frequently used words and bigrams/trigrams. Discuss the potential topics these words and phrases might indicate.
- **Visualization:** Include charts and graphs that effectively represent the analysis findings. Ensure the visualizations are clear, well-labeled, and support the conclusions drawn from the data.

5. Conclusion

The Twitter data analysis, using Python and its rich ecosystem of libraries, has provided valuable insights into user behavior, sentiment, and potential trending topics. The analysis of engagement metrics, sentiment polarity, and text content has helped reveal patterns within the data. This information can be used for various purposes, such as understanding public opinion on current events, identifying areas of concern, or gauging the effectiveness of social media campaigns.

6. References

- List any resources used for the research, including libraries, tutorials, or online articles relevant to the analysis techniques employed.

Note: This report serves as a general framework. The specific content of each section will depend on the chosen analysis techniques, the Twitter data characteristics, and the research goals.