Visualization of Topic Trending and Sentiment Analysis with Sina Weibo Data

Yifan Chen
yc6340
yc6340@nyu.edu

Abstract

Social media platforms have profoundly affected every aspect of people's lives. With the emergence of social media, researchers have gained a powerful tool for understanding public behavior and attitudes. This project aims to develop a visual analysis of topic trending and sentiments on Sina Weibo's hot search list, inspired by Pustilnik and Besio's research on social media text stream visualization. The system comprises three parts: visualization of trending topics, visualization of sentiment analysis, and identification of keywords. This approach will provide valuable insights into Chinese society's changing concerns and emotions.

1. Introduction

In recent decades, social networking platforms have revolutionized how people communicate, interact, and share information. With the ability to share opinions, experiences, and emotions across borders, social media has become a powerful tool for understanding people's behavior and attitudes [1]. Among these platforms, Sina Weibo has become one of the largest social platforms in China. According to Weibo's financial report at the end of Q4 2022, it reached 586 million monthly active users [2]. Weibo's rich and diverse textual data provides a robust database for analyzing and tracking Chinese socio-cultural behaviors, preferences, and sentiments [3]. Therefore, studying Weibo is of significant value.

Although extensive research has been conducted on sentiment analysis and trend analysis of social media platforms such as Twitter, Facebook, and LinkedIn, only some studies have investigated Chinese social media [4] [5]. Moreover, previous research on Weibo has not provided an intuitive way to present their analysis results [6]. To address this gap, this project proposes a system for visualizing trending topics and analyzing the sentiment of Weibo's hot searches. This system draws inspiration from Pustilnik and Besio's research on social media text stream visualization [7], which proposes a visualization method that can be applied to this project to visualize Weibo data in an intelligible presentation. The system is composed of three parts, namely: (1) visualization of the hotness of trending topics over time, (2) visualization of sentiment changes among Weibo users, and (3) identification of keywords associated with trending topics. This approach will provide valuable insights into the Chinese population's perception of various topics, including health, sports, entertainment, and more. By utilizing Weibo's rich and diverse textual data, researchers, social media experts, and policymakers will better understand social changes in Chinese society.

The rest of this paper is structured as follows. The demonstration section will illustrate the proposed system's composition and capabilities. Then, the implementation details section will explain how the system was built. The future work section will discuss potential directions for further development. Finally, the conclusion summarizes the essential findings and contributions of this project.

2. Demonstration

This section will demonstrate the visualization of Weibo Topic Trending and Sentiment Analysis. There are three types of visualization to present the analysis results: bar chart, line chart, and word cloud.

2.1. Total Hotness of Trending Topics

The number of clicks on hot searches reflects people's attention to the topic during a specific period. Therefore, to investigate at which time people pay more attention to popular topics, a time series analysis of Weibo hot searches is necessary. Figure 1 is a bar chart that displays the total hot trend over time. Each bar's color represents the overall hotness, with darker colors indicating more clicks. Additionally, a red line is included in the chart, representing the average level of hotness across all time. This visualization provides a clear overview of the evolution of trending topics' hotness in different periods and can reveal notable changes in topic popularity.

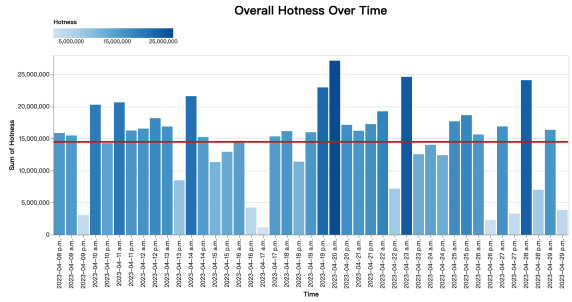

Figure 1. Overall Hotness Over Time

2.2. Sentiment Analysis

Sentiment analysis is crucial for understanding people's emotions, opinions, and attitudes toward various topics. It helps to gain valuable insights into people's perceptions of different issues. Figure 2 is a line chart that visualizes the sentiment trends of various hot search topics on Weibo over time. Each line in the chart represents a different topic type, such as ceremony, drama, and movie. The color of each line corresponds to a specific topic type, and the chart's background color reflects the sentiment at each time point. The sentiment is color-coded by a diverging color map to enhance the visualization, with green indicating positive emotions, yellow representing neutral, and red indicating negative. This visualization can help track the popularity of different topic types and sentiment changes over time, revealing people's attitudes toward social and cultural phenomena.

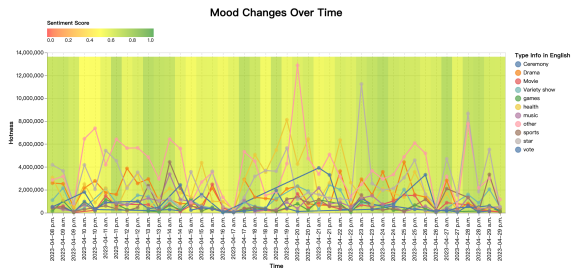

Figure 2. Mood Changes Over Time

2.3. Keyword Analysis

Word Cloud is a powerful tool that can vividly visualize text data. It can highlight and colorize each text corpus's most frequent words and topics. In this project, Figure 3 is implemented as a Word Cloud to display the most frequent keywords associated with each hot search topic. The size of each keyword in the Word Cloud represents its frequency of occurrence. Moreover, users can observe how the keywords have changed over time by sliding a slider on the top of the figure. This visualization method lets users identify the most relevant keywords associated with each hot search topic and track their changes over time. Overall, this keyword analysis can help understand the key themes and topics people discuss on Weibo and how they change over time.

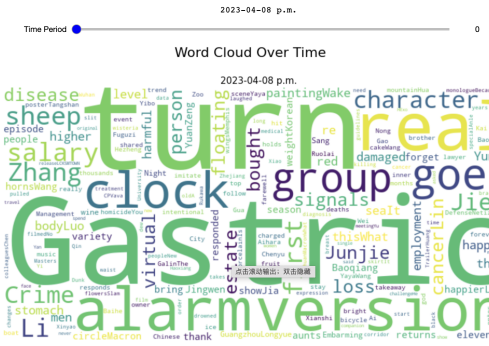

Figure 3. Word Cloud Over Time

3. Implementation Details

3.1. Data Acquisition and Processing

3.1.1 Data Collection

The data collection process involved web scraping the top hot search keywords on Weibo. The Python programming language was used for the web scraping task, with the libraries requests, BeautifulSoup, lxml, and re being employed to handle the HTTP requests, HTML parsing, and regular expression matching. To avoid being detected as a crawler and potentially getting blocked by the server, headers were added to disguise the requests as coming from a regular browser, with a delay of 1 second between each request to reduce the server load. The collected data includes the ranking, hot word in both Chinese and English, type of the hot word in both Chinese and English, hotness score, and the hyperlink to the corresponding search results page. The hot word and type information were translated from Chinese to English using the Googletrans library.

Data collection was conducted periodically to capture real-time changes in hot keywords. To fully account for changes in hot topics over a relatively large time frame, I scraped the data twice daily, once in the morning and once in the afternoon, from April 8, 2023, to February 29, 2023. Finally, 2174 hot topics information was acquired and saved to a CSV file.

3.1.2 Data Processing

The data preprocessing stage began with selecting relevant columns for sentiment analysis and cleaning the dataset. Specifically, a subset of columns was chosen for sentiment

analysis, including the hot word in Chinese, its corresponding English translation, type information in English, hotness score, rank, and time. Any rows with missing data were removed, resulting in a cleaned dataset suitable for further analysis. Furthermore, The 'Time' column was converted to a pandas timestamp datatype to facilitate time-based analysis. Also, the 'Hotness' column was converted to a numeric datatype to prepare the data for quantitative analysis. Any non-numeric values in the 'Hotness' column were replaced with NaN values, ensuring a uniform data type throughout the dataset.

3.2. Sentiment Analysis

In the sentiment analysis stage, the SnowNLP library was utilized to compute sentiment scores. To accomplish this, a function that takes Chinese hot search text as input and returns a sentiment score reflecting the probability of positive emotions was defined. The sentiment scores were limited to a range of 0 to 1, where 0 indicates negative sentiment, 0.5 indicates neutral sentiment, and 1 means positive sentiment. Following the sentiment analysis, any rows containing missing values were removed, resulting in a clean dataset suitable for further analysis.

3.3. Data Visualization

3.3.1 Visualization of Total Hotness Trend

The total hotness trend was visualized using the Altair and Seaborn Python libraries. The initial step involved grouping the data by morning and evening time categories. Subsequently, an interactive bar chart was generated using Altair. The x-axis represents time, and the y-axis represents the overall value of the hotness score for each time group. The hotness quantity was color-encoded using a sequential color map. A vertical line was overlaid on the chart to highlight the mean value of the hotness score across all time groups. Overall, this visualization approach enables a comprehensive view of the hotness trend over time, facilitating straightforward interpretation and analysis of the underlying patterns.

3.3.2 Visualization of Sentiment Analysis

This project utilized the Altair Python library to visualize sentiment analysis. The data were grouped by the time of day and the type of information in English. A line chart was created, with the x-axis representing time and the y-axis representing the hotness for each hot search type. The hotness score is mapped to the color of the bars using a sequential color map. Additionally, the color of the background was mapped to the average sentiment score using a color scale ranging from red (negative) to yellow (neutral) to green (positive). The resulting visualization provides a comprehensive view of the changes in mood over time and the relationship between hotness and sentiment for different types of information.

3.3.3 Visualization of Word Cloud

To visualize the most frequently used keywords over time, this project utilized the WordCloud library and implemented a slider widget using ipywidgets. The slider widget enables users to view the word cloud at different periods. The data used to generate the word cloud was obtained from the sum of the hot words in English for each time. When a user changes the slider's value, the on_value_change function is called. This function extracts the relevant text data for the selected period and passes it to the generate_wordcloud function. The generate_wordcloud function creates a word cloud from the hot search text data to display the most frequent keywords. The size of each keyword in the word cloud is determined by its frequency of occurrence in the text data. The higher the frequency, the larger the keyword appears in the word cloud. This graph aims to highlight the keywords that get the most attention visually. Users can easily compare the differences in the most clicked terms between different periods using a slider widget.

4. Future Work

The current study analyzed public sentiment and interest trends based on data collected within one month. However, a more extended period could provide a more comprehensive understanding of the changes and trends in public opinion. Therefore, future research could conduct a longitudinal analysis to capture the dynamics of public sentiment over a more extended period, such as one year or multiple years. In addition to conducting a longitudinal analysis, future research could explore more advanced visualization techniques to provide more insights and engage users more effectively. While the current study used a bar chart, line chart, and word cloud, advanced visualization techniques such as network analysis, geographic mapping, and interactive visualizations could provide a more dynamic representation of the data.

5. Conclusion

In summary, this project developed a system for visualizing trending topics and analyzing the sentiment of Sina Weibo's hot searches. The system comprises three parts: total hotness visualization, sentiment analysis visualization, and keyword identification. The system provides valuable insights into Chinese society's changing concerns and emotions. It can be helpful for researchers, social media experts, and policymakers to understand social changes in Chinese society. Future work could include expanding the system's capabilities to longer-time data and developing more advanced visualization techniques.

References

[1] Andrei Sechelea, Tien Do Huu, Evangelos Zimos, and Nikos Deligiannis. Twitter data clustering and visualization. In *2016 23rd International Conference on Telecommunications (ICT)*, pages 1–5, 2016. 1

[2] Y Wu. Weibo's revenue drops 20% last quarter as membership business declines for the first time in the second half of last year, 2023. Accessed: May 5, 2023, Retrieved from: https://m.thepaper.cn/newsDetail_forward_2212577. 1

[3] Hu S. Weibo – how is china's second largest social media platform being used for social research, 2022. Accessed: April 11, 2023, Retrieved from: https://blogs.lse.ac.uk/impactofsocialsciences/2020/03/26/weibo-how-is-chinas-second-largest-social-media-platform-being-used-for-social-research/. 1

[4] Andrei Sechelea, Tien Do Huu, Evangelos Zimos, and Nikos Deligiannis. Twitter data clustering and visualization. In *2016 23rd international conference on telecommunications (ICT)*, pages 1–5. IEEE, 2016. 1

[5] Huang M. Huang V. Wang Y. Chu, C. Sentiment analysis and topic trending analysis with weibo data. sfu professional computer science, 2020. Accessed: April 11, 2023, Retrieved from: https://medium.com/sfu-cspmp/sentiment-analysis-and-topic-trending-analysis-with-weibo-data-7ff75e178037. 1

[6] Yunzhi Ye and Xiao Tan. Visualization of sina weibo propagation and sentiment analysis. In *Proceedings of the 2019 International Conference on Computer, Network, Communication and Information Systems (CNCI 2019)*, pages 46–53. Atlantis Press, 2019/05. 1

[7] Martin Pustilnik and Mariano Besio. Social media text streaming visualization. In *2019 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 144–145, 2019. 1