ISSN 2706-6827 Vol. 5, Issue 13: 26-30, DOI: 10.25236/IJFS.2023.051305

Respond to Weibo Comments While Incorporating the Sentiment Evolution Trend Depicted by a Sankey Chart

Shu Wu[1], Xiaobei Wang[1], Hongwei Yu[1], Xucong Zhang[1], Weiyang Xia[1], Dingyi Cheng[1], Alia Erbolat[2], Beibei Tang[3], Kangrui Sun[4,*]

[1]*Sino-German College, University of Shanghai for Science and Technology, Shanghai, China*
[2]*School of Journalism and Communication, Shanghai University, Shanghai, China*
[3]*Golden Concord Holdings, Suzhou, China*
[4]*Yizhu Intelligent Technology, Hangzhou, China*
[*]*Corresponding author*

Abstract: *In the context of media public opinion becoming the determination of people's views and behaviors, the analysis and control of public opinion has become an important part of understanding the needs and demands of the public and maintaining the stability of the public mood. [1] This project aims to use the program algorithm to obtain the remarks related to the specified event in online social media (microblog), and analyze it through the algorithm, so as to provide a complete map of the development trend and tendency of public opinion events, and provide reference for the development and coping strategies under the current time node. In response to the comments on Weibo (Sichuan Middle School Incident), the program uses algorithms such as text detection, exception filtering, sorting, and self-designed and trained neural networks to automatically obtain information, establish subjects and time nodes, Analyze the proportion of different public opinion development tendencies and output them in the form of maps, reduce the expenditure of human resources, effectively reduce the emotional bias in the process of obtaining public opinion information, effectively and rationally analyze the tendency and evolution direction of public opinion, show the dynamic evolution process, and provide theoretical reference for response and governance.*

Keywords: *public opinion, reference for the development, Circulation of Industrial Data, Shanghai*

1. Introduction

Weibo comments often reflect the trend of a public opinion after public discussion and fermentation. However, the cleaning of this data and the establishment of models are often difficult because of the redundant content.

For this factor, we use an optimized algorithm to minimize the impact of comment errors (e.g., @, rub popularity), and display the trend of public opinion through Sangi charts.

Here are the high-level steps of our study.

1.1. Data Extraction

This section focuses on the problems and optimization methods in the process of obtaining and processing microblog data. The project we chose is the Scrapy project "weibosearch" in github, which obtains relatively comprehensive Weibo content and large amount of information, but there are some problems in practical applications, such as slow acquisition speed, inability to be accurate to hours or minutes, and a large amount of duplicate data when obtaining a large amount of data. Therefore, we have carried out a series of optimizations to improve the efficiency and accuracy of data acquisition, and also processed the acquired data to provide effective data for subsequent topic classification, correlation, and visualization.

1.2. Summary of data

This section focuses on the necessary filtering of the comments obtained in the extract section.

ISSN 2706-6827 Vol. 5, Issue 13: 26-30, DOI: 10.25236/IJFS.2023.051305

Identifiers made through Python will contain @Content filtering removes identifiers to prevent interference with topic definition. Then, after comparing the singlepass and LDA models, the LDA model with better results was selected, and the comments were divided by the development time of time. Finally, the optimal number of titles is defined by the elbow method, and the themes of each period are extracted.

1.3. Confirm the association relationship between child events

In the tasks in this section, we aim to perform text association analysis, filter out highly similar text, and explore the association between texts. Therefore, we propose a correlation analysis method based on text similarity, and design the corresponding experimental process and evaluation indicators

1.4. Subsequent evolution of events

Regarding the drawing of the Sankey diagram, on the basis of using the LDA model to treat comments similarly, we integrate the obtained data, classify the sub-events shown in each comment, obtain the number of comments of each sub-event, and divide the number of sub-events by relative relationship to obtain the transfer rate between sub-events.

The main contributions of this paper include, but are not limited to:

• An association analysis method based on text similarity is proposed, which can accurately measure the similarity between texts.

• We designed and implemented an automated text processing and analysis tool to enhance the efficiency and accuracy of text relevance analysis.

• We have provided a scalable solution capable of handling large-scale text datasets and generating highly relevant text groups.2.Related research

LDA was proposed by Blei, David M., Ng, Andrew Y., and Jordan in 2003 to speculate on the subject distribution of documents. It can give the topic of each document in the document set in the form of a probability distribution, so that after analyzing some documents to extract their topic distribution, it can be subject clustered or text classified according to the topic distribution [2-3]., LDA uses a bag-of-words model. The so-called bag-of-word model is a document, we only consider whether a word appears, not the order in which it appears. In the bag-of-word model, "I like you" and "You like me" are equivalent. One model that is the opposite of the bag-of-word model is the n-gram, which takes into account the order in which words appear. [4] In the past, scholars at home and abroad often directly cited the LDA model to the comments of some websites, and due to the limitations of extranet websites, the topic differentiation of the LDA model would be very good.

With the popularity of social media platforms, Weibo, as one of the largest social media in the Chinese world, has become an important channel for obtaining public opinion and exploring social hot spots. The crawling, analysis and processing of microblog content has become the basis of many studies. For example, all kinds of microblog-based public opinion analysis, social network analysis, sentiment analysis and other research need to obtain a large amount of microblog data as the basis for research. At the same time, due to the large and diverse data of microblogs, how to effectively obtain and process these data has also become an important research topic.

For the acquisition of Weibo data, it is common practice to use the API (Application Programming Interface) provided by Weibo to crawl data. API is a data interface provided by the Weibo platform for developers, allowing developers to obtain public data on Weibo through a predetermined method. However, due to Weibo's limitations on API calls (such as daily call limit, search result page limit, etc.), it is often inefficient to directly use APIs for large-scale data crawling.

In the absence of sufficient data through APIs, many studies have turned to the method of directly crawling microblogging pages. At present, research on microblog data acquisition mainly focuses on scrapy-based crawler projects or simulated browser access. Scrapy is a powerful crawling framework that efficiently crawls and processes web page data. However, due to Weibo's anti-crawling strategy and search restrictions, how to use Scrapy to bypass the anti-crawler mechanism and parse the web page structure to crawl Weibo data has become a challenge. [5]

Sankey diagrams require specific data visualization tools or programming languages, such as Matplotlib in Python, D3 in R or D3 in JavaScript, or Excel as .js. Creating a Sankey chart using Excel is a straightforward process, and the general steps are as follows:

ISSN 2706-6827 Vol. 5, Issue 13: 26-30, DOI: 10.25236/IJFS.2023.051305

1) Data Preparation: Begin by organizing the data you intend to present. Ensure that the data includes essential information such as the source node, destination node, and the traffic or degree of association between these nodes.

2) Node List Creation: In Excel, generate a node list featuring the source and destination nodes within a single column. Make certain that there are no duplicate nodes in this list.

3) Flow Data Table Generation: Create a flow data table in Excel, with rows and columns representing combinations of source and destination nodes. Populate the table with numeric values corresponding to the degree of association between nodes.

4) Rectangular Block Chart Design: Utilize a rectangular block chart in Excel to mimic the nodes of a Sankey chart. Generate a rectangle for each node and adjust its size based on the node's traffic or degree of association.

5) Connecting Rectangular Boxes: Employ tools like lines, arrows, or curves in Excel to connect the rectangular boxes representing source and destination nodes. These connections illustrate the flow or association between nodes.

6) Adding Tags and Values: Enhance the chart's clarity by attaching tags to each node and connector. This enables readers to grasp the significance of the nodes and the numerical values of the flow. You can employ Excel's text box and label features to incorporate these labels. However, making a Sankey chart in Excel is a compromise approach, just static and does not provide the full Sankey chart functionality like professional data visualization tools.

2. Algorithm design

2.1. Weibo data acquisition optimization

Weibo data acquisition mainly uses the Scrapy crawler framework. During the optimization process, we first solved the problem of a large amount of duplicate data. In the settings.py ITEM_PIPELINES, we used Duplicates Pipeline for filtering duplicate data. To prevent data duplication, you can achieve this by reading the existing CSV file and appending new data to it, allowing the old data to remain intact.

When processing a large amount of data, the crawler of the original project will not be able to obtain all the data due to the Weibo search page limit (50 page limit). When the crawler of the original project is faced with data larger than 46 pages, it will subdivide the search conditions, accurate to the hour, and then accurate to the province, city, and administrative district, so as to break through the 50-page limit. However, the filtering function by region in the Weibo search service is not accurate, and can only be obtained if the user checks and fills in the publishing address, and in fact, only 1/20 of users will fill in this information when posting Weibo, and most of the content is only "such and such a restaurant", etc., which is meaningless to limit the search information. That's why we've eliminated regional filtering and instead searched by time to improve the efficiency of data acquisition. After optimization, the crawler acquisition speed has increased by nearly 500 times.

At the same time, the search function of Weibo itself is also flawed. If the search is limited by a large time span, all relevant Weibo blogs cannot be obtained. For example, a search is limited to one month, but only nearly ten Weibo posts near the end of the time can be obtained. The search time is limited to the start time to the next day, and when multiple searches are conducted, the number of Weibo posts can be obtained in each search far exceeding 10. Therefore, the crawler has been improved, and in the case of a large search time span, it will force a split, split the search time range to one day, and if the search results exceed a certain number of pages, then subdivide the search into each hour. Although this will reduce the efficiency of obtaining microblogs, it can effectively obtain far more data than the original and increase the accuracy of analysis results.

After obtaining the Weibo data, we need to sort and clean the data. We used the pandas library to process csv files. Since the Scrapy crawler processes data requests asynchronously, it needs to sort the obtained chronological microblogs. After sorting, we use the resample function of the pandas library to calculate the total number of microblogs every 4 hours and per day (the time interval is used as an adjustable variable), and then use matplotlib for visualization. After visualization, it is fairly intuitive to see at what point in time there is a lot of discussion. At the same time, the official certified media Weibo (a small number) after the classification of the LDA topic model is used as the basis for the time period division of a large number of Weibo topic classifications in the future, so as to show the direction of

ISSN 2706-6827 Vol. 5, Issue 13: 26-30, DOI: 10.25236/IJFS.2023.051305

public opinion themes in different time periods and at different stages of event development.

2.2. Weibo data cleaning and LDA theme model

In the process of data cleaning, it is inevitable to encounter various problems. Some problems such as malicious screen brushing, rubbing the heat to hang labels, etc., need to be deleted in advance in the part where the data is picked. Some of the gaps in the comments can be eliminated in the data cleansing section. At the same time, for the @ keyword with frequent Weibo comments, the LDA model program is likely to include the @ name into a theme, so we made a part of the elimination program to delete the fields containing @: to ensure that the model will not be disturbed when running.

At the same time, due to the current prevalence of Internet memes, the LDA model cannot correctly stop these meme statements. Therefore, we add the function of a special thesaurus for general LDA models.

When we enter the corresponding stop text in this txt, the LDA model will treat this word as a priority stop structure, thus greatly protecting the correct distinction of topics.

2.3. Weibo comment topic association

In this section, we use advanced text similarity calculation methods, such as those based on the bag-of-word model and TF-IDF, to calculate the degree of similarity between texts.

TF-IDF (Term Frequency-Inverse Document Frequency): TF-IDF is a commonly used feature representation method used to measure the importance of words in text. It combines word frequency (TF) and inverse document frequency (IDF) to determine the weight of words by calculating their frequency in text and their rarity throughout the document collection

• Word weight calculation formula: TF-IDF (w, d) = TF(w, d) * IDF(w).

• Inverse document frequency calculation formula in word weight: IDF(w) = log(N/DF(w)) where N represents the total number of documents in the document collection, and DF(w) represents the number of documents containing the word w.

•Set similarity thresholds, filter out pairs of text with similarity above the threshold, and group them into related text groups.

• Generate new Excel tables or other forms of output based on the results of text groups to demonstrate highly relevant text groups.

2.4. Comment on Topic Evolution

From the general trend of the incident, we can know: on June 7, 2023, "Uncle was suspected of being exposed by a woman for secretly photographing" appeared on the Weibo hot search list and triggered continuous attention and discussion, and the incident was exposed at 11:39 on June 7, 2023, and gradually subsided on July 5, 2023. Combined with Weibo historical data and news, it can be seen that the murdered woman quickly fermented after exposing the incident through Weibo, which triggered questions from the public and the media. The uncle was suspected of secretly photographing and was exposed by a woman, and the incident appeared on the Weibo hot search list. At 16:42 on June 9, 2023, Sichuan University responded that it was investigating. At 17:31 on June 9, 2023, the woman's identity was exposed, which once again caused heated discussions. At 19:55 on June 11, 2023, the woman issued an apology. At 10:15 on June 20, 2023, Sichuan University responded that after verification, the student was given the sanction of staying in the party for probation, and the incident gradually subsided in the time after that.

We have counted the number of comments and sub-events in the form of the same or similar that are important in the development of the event.

From these sub-events mentioned above, we can roughly divide the entire incident into five stages, the first stage is the beginning of public opinion (the woman's post is secretly photographed); the second stage contains the first comment of the media (reporting this incident), the first official response (need to be investigated) and the first comment of netizens (whether it should be expelled, whether it can be forgiven, etc.); the third stage contains the second comment of the media ("unwarranted") style rights protection is not desirable), the official second response (the online rumor is not an official voice, it attaches great importance to it and will be dealt with according to law) and the second comment of

ISSN 2706-6827 Vol. 5, Issue 13: 26-30, DOI: 10.25236/IJFS.2023.051305

netizens (Sichuan University pretending to die, online violence, etc.): the fourth stage is the official third response (giving the sanction of probation in the party), and the fifth stage is the third comment of netizens after seeing the punishment result (whether the result is reasonable, whether the result is serious, etc.). Finally, a dynamic Sankey diagram is drawn.

3. Conclusion

By optimizing the crawler code of the Scrapy project "weibosearch", we achieved the goal of improving the amount and efficiency of data acquisition under the hot topics of large amounts of data. After obtaining the data, we processed and analyzed the data, successfully obtained hot topics and public opinion tendencies, and provided effective observation and analysis methods for public opinion.

By associating sections, our method is able to effectively identify highly similar texts and group them into groups of related text. And under the given similarity threshold, we successfully controlled the number of results and generated text groups that met the relevance requirements. Finally, by analyzing the text groups in the results, we find that text association analysis is of great significance for discovering the association relationship and common theme between texts.

It can be noted from the Sangi diagram and statistics that the development of public opinion as a whole conforms to the life cycle theory, both germinating, maturing, and fading, but it is worth noting that netizens have repeatedly appeared in the evaluation of Sichuan University's handling of the incident, such as untimeliness and inaction, in addition, some companies even issued statements rejecting graduates of Sichuan University. It can be seen that Sichuan University's handling of this incident is still flawed, and the data collected from the statistical table shows that the first and second official responses did not have a certain calming effect on the development of public opinion.

Acknowledgement

This research was funded by China Youth and Children Research Association (Grant No. 2023B18), the Fund of University of Shanghai for Science and Technology (Nos. 22SLCX-ZD-005, SH2023253 SH2023254, XJ2023515); the Fund of University of Shanghai for Science and Technology (2023QYKTLX3-18); China Foundation for Youth Entrepreneurship and Employment (2023A03-01); the Fund of University of Shanghai for Science and Technology, Research on the factors influencing the flexible employment choice of female college students in the context of common wealth and counter measures.

References

[1] Chen J, Gong Z, Liu W. A. Dirichlet process biterm-based mixture model for short text stream clustering [J]. Applied Intelligence, 2020, 50 (5): 1609-1619.
[2] Nimala K, Jebakumar R. A. Robust User Sentiment Biterm Topic Mixture Model Based on User Aggregation Strategy to Avoid Data Sparsity for Short Text [J]. Journal of Medical Systems, 2019, 43 (4).
[3] Zhu L, Xu H, Xu Y, et al. A joint model of extended LDA and IBTM over streaming Chinese short texts [J]. Intelligent Data Analysis, 2019, 23 (3): 681-699.
[4] Murshed B A H, Abawajy J, Mallappa S, et al. Enhancing Big Social Media Data Quality for Use in Short-Text Topic Modeling[J]. Ieee Access, 2022, 10: 105328-105351.
[5] Deng Xiaolu, Yao Song. Research on Sina Weibo data crawler based on Scrapy [J]. Modern Information Technology, 2023, 7(03):44-47. DOI:10. 19850/j. cnki. 2096-4706. 2023. 03. 010