Data in Brief 53 (2024) 110232

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Invited Data Manuscript

A structured sentiment analysis dataset based on public comments from various domains

Zhongliang Wei*, Shunxiang Zhang

School of Computer Science and Engineering, Anhui University of Science & Technology, Huainan, China

ARTICLE INFO

Article history:
Received 11 December 2023
Revised 12 February 2024
Accepted 16 February 2024
Available online 22 February 2024

Dataset link: A Structured Sentiment Analysis Dataset (Original data)

Keywords:
Sentiment classification
Text mining
Triple classification
Natural language processing

ABSTRACT

A structured sentiment analysis dataset, derived from social media comments, is introduced in this paper. The dataset spans 22 diverse domains and comprises over 200,000 reviews, providing a rich resource for sentiment analysis tasks in the Chinese language context. Each comment within the dataset has been manually annotated with a sentiment label, either positive, negative, or neutral, and grouped by topic. This meticulous annotation process ensures the dataset's reliability for training, validating, and testing sentiment analysis models. The construction of the dataset involved a three-step process. Initially, data was collected from the topics that garnered high attention and discussion rates, thereby reflecting the authentic opinions of users. Following data collection, preprocessing was undertaken to remove extraneous elements, while preserving emoticons that are crucial for sentiment analysis. The final step involved manual annotation by researchers, who assigned sentiment labels to each comment based on various factors. The dataset stands as a valuable contribution to the field of natural language processing, particularly for sentiment analysis tasks in the Chinese language context.

* Corresponding author.
E-mail address: zhlwei@aust.edu.cn (Z. Wei).

https://doi.org/10.1016/j.dib.2024.110232

Specifications Table

Subject	Computer Science, Natural Language Processing.
Specific subject area	Chinese language, Structured dataset, Sentiment analysis, machine learning, deep learning
Data format	Analyzed
Type of data	Text, Table
Data collection	The data were acquired by manually scraping relevant data from Chinese social media platform: Weibo. These data were categorized into 22 domains based on the topics and comments. The data collection was conducted through web browsers and the data were stored in Microsoft EXCEL files.
Data source location	The structured sentiment analysis dataset was constructed at Anhui University of Science and Technology, Huainan, China.
Data accessibility	Repository name: Zenodo DOI:10.5281/zenodo.10488076 Direct URL to Data: https://zenodo.org/records/10488077

1. Value of the Data

- This dataset is a Chinese text sentiment analysis dataset, covering 22 domains and nearly 200,000 pieces of data information, and it can be used to analyze the sentiment of Chinese text. It is suitable for researchers who are interested in exploring the sentiment dynamics and variations across different topics and domains in the Chinese social media context.
- To the best of our knowledge, this dataset is a rare Chinese text dataset with structured collection for topics and reviews. It can facilitate the development and evaluation of sentiment analysis models that can leverage the structured information.
- Researchers can use this dataset for sentiment analysis by dictionary-based methods, machine learning-based methods or deep learning-based methods. It can also serve as a benchmark dataset for comparing the performance of different methods and models on Chinese sentiment analysis tasks.
- This dataset provides a different perspective on Chinese media information, and researchers can incorporate this structured information into their proposed sentiment computing models. For example, researchers can use this dataset to investigate how the sentiment of a topic is influenced by the sentiment of its related reviews.

2. Background

Sentiment analysis, alternatively recognized as opinion mining or emotion AI, is a cognitive process that involves the systematic examination, processing, summarization, and inference of subjective text imbued with emotional connotations.

Integral to the effectiveness of sentiment analysis is the requirement for a substantial volume of text data to train models. Weibo is the largest Chinese text data generation platform compared to other Chinese social media. According to Weibo Reports Third Quarter 2023 Unaudited Financial Results [1], in September 2023, Weibo had 605 million monthly active users and 260 million daily active users, so it is necessary to construct a dataset on Weibo for Chinese sentiment analysis.

In the course of our investigation into sentiment analysis [2–4], we have deliberately directed our attention towards the pivotal role that datasets play in shaping the outcomes of sentiment analysis models. As our research unfolded, we observed a discernible pattern in the structure of topics and reviews prevalent in social media contexts [5]. This observation prompted the conceptualization and implementation of a dataset aligning with similar structures in 2022. Significantly, recent developments indicate a noteworthy interest from other researchers who are actively pursuing avenues stemming from this line of research [6].

Fig. 1. The overall structure of the dataset.

3. Data Description

The dataset denoted as "ch_22d_org" is characterized by its composition of content spanning across 22 distinct domains. Each domain is systematically assigned a numerical order for reference, delineated as follows: Emotion, Celebrities, Finance, Law, Sports, Cinema, Shows, Campus, Tourism, TV Dramas, Technology, Health, Games, Military, Digital, Constellations, Fitness, Comedy, Animation, International, Covid-19, and Government. Notably, the dataset's organizational structure entails the segregation of content from each domain into individual Excel files, facilitating a clear demarcation of thematic categories. The comprehensive arrangement of domains and their respective content is visually represented in Fig. 1, elucidating the overarching framework of the dataset.

The uniform structure of each Excel file within the "ch_22d_org" dataset is characterized by a consistent arrangement of fields. These fields, encapsulated within the title field, include numerical identifiers, category classification, topical information, reviews, and corresponding labels. The correlation between the "topic" and "review" fields within each data table adheres to a 1-to-10 relationship. This structural characteristic, as expounded in Section VALUE OF THE DATA, underscores the organized structure of the dataset. This structure is a distinctive feature that sets it apart from other Chinese datasets. This unique 1-to-10 relationship is illustratively exemplified in Table 1, providing a tangible representation of the structured framework inherent in the dataset.

As shown in Table 1, a noteworthy attribute of the dataset lies in the consistent alignment of topics, each corresponding to 10 distinct reviews. The label field within the table is intentionally annotated manually, featuring classifications into positive, negative, and neutral categories. This meticulous labeling process is integral to the structured nature of the dataset, a quality that permeates consistently across its entirety. This structural coherence is evident not only in the thematic content but also in ancillary fields, such as the serial number field.

Examining the serial number field as an illustrative example, the structured comment groups are organized systematically. The serial numbers, ranging from 17-000011-0 to 17-000011-9, delineate a cohesive unit of comments. The leftmost numerical identifier '17' signifies the 17th domain, corroborated by the category label 'fitness.' The intermediary numerical sequence '000011' signifies the initial group of structured comment information, a numerical range that extends, in some instances, up to 001200 within specific Excel files. Finally, the numeric range '0 to 9' corresponds sequentially to different comments under the same topic within this structured group.

To further elucidate the dataset's compositional characteristics, Table 2 presents a comprehensive overview of comment counts and label statistics for each domain. This tabulation serves to provide quantitative insights into the distribution of comments and associated sentiment labels, contributing to a more nuanced understanding of the dataset's content and structure. emotion, celebrities, finance, law, sports, cinema, variety shows, campus, tourism, TV dramas, tech-

Table 1
The organizational structure of data.

No.	Domain	Topic	Review	Label
17-000011-0	Fitness	Trying to lose fat and build muscle this year	I love it. It's my dream body.	Positive
17-000011-1	Fitness	Trying to lose fat and build muscle this year	I wouldn't believe you if you said you didn't work out at all.	Negative
17-000011-2	Fitness	Trying to lose fat and build muscle this year	That's a great body.	Positive
17-000011-3	Fitness	Trying to lose fat and build muscle this year	Do you have a link to that top?	Neutral
17-000011-4	Fitness	Trying to lose fat and build muscle this year	It's a great body. It grows where it's supposed to.	Positive
17-000011-5	Fitness	Trying to lose fat and build muscle this year	I want to build muscle.	Neutral
17-000011-6	Fitness	Trying to lose fat and build muscle this year	I'm envious.	Positive
17-000011-7	Fitness	Trying to lose fat and build muscle this year	What a perfect body!	Positive
17-000011-8	Fitness	Trying to lose fat and build muscle this year	It's a shame to lose weight.	Negative
17-000011-9	Fitness	Trying to lose fat and build muscle this year	Don't be like this.	Negative

Table 2
The number of comments and label statistics for each domain.

Domain	File Name	Number of Topics	Number of Reviews	positive	negative	neutral
Emotion	01-情感.xlsx	1100	11,000	1566	715	8719
Celebrities	02-明星.xlsx	1100	11,000	7135	1684	2181
Finance	03-财经.xlsx	1100	11,000	2935	3287	4778
Law	04-法律.xlsx	1100	11,000	5744	2465	2791
Sports	05-体育.xlsx	688	6880	3239	1655	1986
Cinema	06-放映厅.xlsx	1100	11,000	5454	3855	1691
Shows	07-综艺.xlsx	1100	11,000	5424	1621	3955
Campus	08-校园.xlsx	1100	11,000	2206	5897	2897
Tourism	09-旅游.xlsx	1200	12,000	6734	309	4957
TV Dramas	10-电视剧.xlsx	1082	10,820	6451	2735	1634
Technology	11-科技.xlsx	403	4030	1575	942	1513
Health	12-养生.xlsx	650	6500	4773	1587	140
Games	13-游戏.xlsx	1100	11,000	4812	2655	3533
Military	14-军事.xlsx	1100	11,000	4013	3828	3159
Digital	15-数码.xlsx	1100	11,000	5886	3306	1808
Constellations	16-星座.xlsx	1110	11,100	5716	3356	2028
Fitness	17-健生.xlsx	860	8600	4663	2032	1905
Comedy	18-搞笑.xlsx	1100	11,000	4401	4266	2333
Animation	19-动漫.xlsx	1100	11,000	4238	2060	4702
International	20-国际.xlsx	800	8000	5077	1820	1103
Covid-19	21-疫情.xlsx	70	700	302	208	190
Government	22-政务.xlsx	571	5710	2606	924	2180
Total		20,624	206,240	94,950	51,207	60,183

nology, health, games, military, digital, constellations, fitness, comedy, animation, international, epidemic, and government.

Illustrated in Fig. 2 is a graphical representation delineating the proportions of positive, negative, and neutral labels within the dataset. This visualization serves to intuitively convey the distribution of label types both collectively across all data and individually for each of the 22 domains. The figure offers a comprehensive overview, enabling a nuanced understanding of the sentiment label distribution within the dataset.

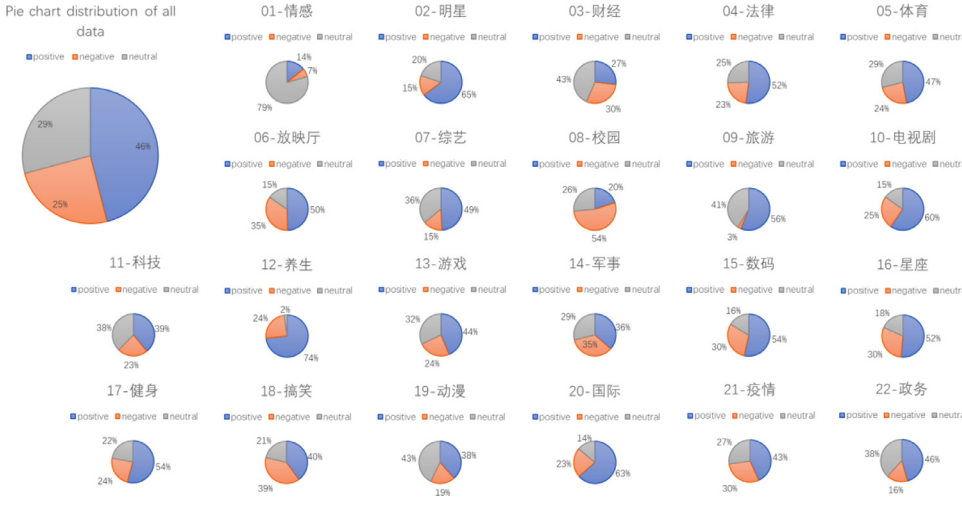

Fig. 2. The percentage of three labeled information in pie charts.

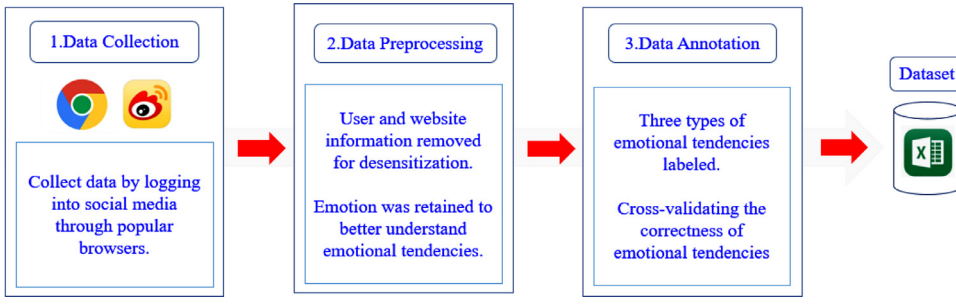

Fig. 3. Dataset construction process.

By presenting the proportional breakdown of sentiment labels, Fig. 2 facilitates a visual exploration of how positive, negative, and neutral sentiments are distributed across the entire dataset and within specific domains. This graphical representation contributes to the interpretability of the dataset's sentiment dynamics, allowing researchers to discern patterns and variations in sentiment expressions across diverse domains.

4. Experimental Design, Materials and Methods

The construction of this dataset was built in the following three steps, as shown in Fig. 3.

1. Data collection for our study involved a substantial workload, particularly across 22 distinct domains. To streamline the process, we allocated the task among 11 students, each responsible for gathering data from two specific domains. The data was acquired by accessing designated Weibo Topics through a web browser. Starting from February 8, 2022, and ending on May 31, 2022, a period of nearly 100 days, the data were obtained by accessing specified Weibo topics through the web browser. These topics were selected based on their daily popularity, which is characterized by a high level of attention and discussion, as well as the diversity and authenticity of Weibo users' opinions. For each topic, we selectively collected 10 reviews based on the popularity of the reviews. We consider that

the more popular reviews are more compatible with the current topic. The resulting data was meticulously organized, with different types of information classified and stored in separate Excel files. Each Excel file corresponds to a specific domain and is named using the "No.-domain.xlsx" convention.

2. Following data collection, we undertook a data preprocessing step to enhance the quality and uniformity of the collected information. The process involved simple cleaning procedures, such as the removal of website links starting with "http/https" and sensitive user information initiated with "@". Using the Python re library, we implemented regular expressions to identify and replace these unwanted elements. Emoticons, such as "[开心] " and "[悲伤] ", were retained during this stage, as they contribute to expressing emotions and enriching the text's expressiveness. The preprocessed data was then overwritten onto the original Excel files, preserving the initial format and structure.
3. Upon completing the preprocessing phase, the dataset underwent manual annotation, wherein each Weibo comment was systematically assigned a sentiment label to signify its emotional orientation. The sentiment labels employed comprised three categories: positive, negative, and neutral. The assessment of the emotional stance of Weibo comments was based on various factors such as tone, vocabulary, and the presence of emoticons. Specifically, comments expressing positive emotions like agreement, support, or satisfaction were designated as positive, while those conveying negative sentiments such as opposition, criticism, or dissatisfaction were labeled as negative. Instances where comments did not overtly express emotions or exhibited unclear emotional ambiguity were categorized as neutral. The labeling procedure was initially completed by the researcher responsible for gathering information about the current domain, followed by meticulous cross-checking and validation by another researcher to ensure accuracy and consistency of the sentiment labels. These researchers were all second-year or third-year students pursuing their master's degrees, and all of their research areas were also in natural language processing, that ensured the professionalism and accuracy of the data collection and data labeling. Because of the complexity of the Chinese language, about 10% of the data was ambiguous during the labeling process, which required active communication and collaboration between the two researchers to resolve.

Concluding these sequential steps, the finalized dataset emerged, poised for application in sentiment analysis tasks within the domain of natural language processing. This curated dataset serves as a systematically organized and structured emotional data resource, encapsulating Weibo users' opinions and sentiments. The dataset, now publicly accessible on an open platform, facilitates researchers in selectively utilizing the entirety or specific subsets of the data tailored to their research content for sentiment analysis endeavors.

Limitations

The data collection methodology employed in constructing this dataset was limited to categorizing sentiments into three broad classes: positive, negative, and neutral. Notably, it did not extend to encompassing multi-class classification, which would involve discerning specific emotions such as joy, anger, grief, and happiness within the comments. This apparent limitation suggests an avenue for future research to delve deeper into a more nuanced sentiment analysis framework.

Furthermore, the dataset lacks explicit sentiment orientation labels for each individual Topic. Researchers engaging in the complex task of sentiment analysis with this dataset are required to independently assess and determine the sentiment orientation of each Topic. This omission poses an additional layer of complexity, as the absence of predefined sentiment orientation adds an element of subjectivity to the interpretation of results. Future efforts may consider incorporating this aspect into dataset augmentation, providing a more comprehensive resource for sentiment analysis tasks.

Ethics Statement

The collected data has been fully anonymous, and the data redistribution policies of social media platforms have been complied with [7].

Data Availability

A Structured Sentiment Analysis Dataset (Original data) (zenodo).

CRediT Author Statement

Zhongliang Wei: Conceptualization, Methodology, Data curation, Visualization, Writing – original draft, Writing – review & editing; **Shunxiang Zhang:** Validation, Supervision, Writing – review & editing.

Acknowledgements

Funding: This work was supported by the Natural Science Research Project of Anhui Educational Committee (grant number: KJ2021A0449), and The University Synergy Innovation Program of Anhui Province (grant number: GXXT-2021-008). In addition, the authors would like to express their gratitude to the graduate students who participated in the collection, preprocessing and labelling of the data.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Sina Weibo, Weibo Reports Third Quarter 2023 Unaudited Financial Results. https://www.prnewswire.com/news-releases/weibo-reports-third-quarter-2023-unaudited-financial-results-301982934.html. (Accessed 12 February 2024).
[2] S.X. Zhang, Z.L. Wei, Y. Wang, T. Liao, Sentiment analysis of Chinese micro-blog text based on extended sentiment dictionary, Fut. Gener. Comput. Syst. 81 (2018) 395–403, doi:10.1016/j.future.2017.09.048.
[3] S.X. Zhang, Z.Y. Hu, G.L. Zhu, et al., Sentiment classification model for Chinese micro-blog comments based on key sentences extraction, Soft Comput. 25 (2021) 463–476, doi:10.1007/s00500-020-05160-8.
[4] S.X. Zhang, H.B. Yu, G.L. Zhu, An emotional classification method of Chinese short comment text based on ELECTRA, Conn. Sci. 34 (1) (2022) 254–273, doi:10.1080/09540091.2021.1985968.
[5] Z.L. Wei, W.J. Liu, G.L. Zhu, S.X. Zhang, M.-Y. Hsieh, Sentiment classification of Chinese Weibo based on extended sentiment dictionary and organisational structure of comments, Conn. Sci. 34 (1) (2022) 409–428, doi:10.1080/09540091.2021.2006146.
[6] K. Liu, M. Hai, Rumor detection of Covid-19 related microblogs on Sina Weibo, Procedia Comput. Sci. 221 (2023) 386–393, doi:10.1016/j.procs.2023.07.052.
[7] Sina Weibo, Weibo Online Service Agreement. https://open.weibo.com/wiki/. (Accessed 10 December 2023).