A Multi-theme Rich-content Social Media Data Collection

Xinyi Li
Jiayin Wang
Tsinghua University
Beijing, China
lixinyi22@mails.tsinghua.edu.cn
jiayinwangthu@gmail.com

Qinglang Guo
China Academic of Electronics and Information Technology
Beijing, China
gql1993@mail.ustc.edu.cn

Zhiqiang Guo
Min Zhang*
Tsinghua University
Beijing, China
georgeguo.gzq.cn@gmail.com
z-m@tsinghua.edu.cn

Abstract

Social computing is a critical research domain, with a focus on analyzing both user behavior and information diffusion. However, existing social media datasets are typically constructed within specific time frames or centered around specific user groups, leading to incomplete or biased representations of themes and content. Consequently, there is a gap in datasets designed to support information-centered research. To address this, we introduce a **M**ulti-**T**heme social media dataset with rich content, Weibo-MT, that encompasses four diverse themes: travel, movie, psychology, and car. The Weibo-MT dataset includes posts, comments, and user profile information for both poster and commentor. Statistics of the distributions of posts and users show significant diversity across different themes, indicating a high degree of theme heterogeneity. This dataset holds substantial potential for advancing social simulations, information diffusion, user behavior research across diverse interest areas, and can be used as a valuable resource for cross-platform/cross-domain services such as recommendation. The dataset is publicly available at https://github.com/lixinyi22/Weibo-MT.

CCS Concepts

• **Human-centered computing** → **Collaborative and social computing**; • **Information systems** → **Web mining**.

ACM Reference Format:
Xinyi Li, Jiayin Wang, Qinglang Guo, Zhiqiang Guo, and Min Zhang. 2024. A Multi-theme Rich-content Social Media Data Collection. In *Proceedings of The Web Conference (WebConf '24).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

As social networks become vital platforms for communication and information dissemination, social media data has emerged as an essential foundation for research. Diverse themes and populations converge on social media platforms, interacting and generating new content. Up to now, This complex process of interpersonal interaction produces data that is still difficult to replicate through other simulations or models.

Data from mainstream social media platforms such as Twitter, Facebook, and Sina Weibo has been applied in fields like social network analysis and information diffusion for a long time. Existing studies have thoroughly analyzed these platforms' user and platform characteristics [4][12] and leveraged such data to train models across various domains or evaluate model performance. However, most existing datasets are custom-built to meet the specific needs of individual studies. This results in missing data fields or inconsistent organizational structures when the datasets are applied to other research contexts, making them less suitable for cross-domain studies. Additionally, many existing datasets are constructed by either focusing on a core group of users or randomly sampling all available information within a specific time frame. Subsequent keyword or event extraction from posts often limits the datasets' utility for research centered on news or themes, leading to incomplete data for such studies.

In this paper, we introduce a novel social media dataset collected from Sina Weibo, organized with four themes: travel, psychology, movie, and car. The dataset consists of 7,705 posts with rich content, 202,238 comments, and 54,401 user profiles, encompassing both post publishers and commenters. The major contributions of the new data collection are:

- We provide a novel multiple-theme based social media collection that enables comparisons or integration with datasets from other sources, such as recommendation collections.
- While preserving user privacy, the dataset offers rich natural text in posts, comments sentiment-label, and anonymous user attributes. It supports in-depth content understanding such as LLM-based analyses and user preference research.
- The collected posts span an average of 2,771 days for each theme. By considering the relative timing of posts and comments, the dataset allows for reconstructing information diffusion processes over time.

2 Related Work

The analysis of modern social media data (e.g., Weibo, Twitter) has provided abundant insight for multiple research domains including information diffusion [11] and social networks [9]. Kwak et al. [5] studied the topological characteristics of Twitter through crawled Twitter datasets. Hodas et al. [3] conducted a social contagion rule research by collecting all tweets and the friend and follower information for all tweeting users. Weng et al. [10] collected a more recent Twitter dataset that includes multiple relations: reciprocal follower/followee, retweeting, and mentioning interactions between users. Cao et al. [1] published a Weibo dataset that

*Corresponding author.

WebConf '24, 28 April - 2 May, 2025, Sydney, Australia

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

contains the retweet cascade of each message by identifying the retweeting path from the text of its retweets. Zhang et al. [13] collected a richer Weibo data with all the users' profiles, such as name, gender, verification status, to model the social influence locality of retweeting behaviors. Weibo21 [7] is a multi-domain fake news dataset with news content and comments.

Despite their contributions to advancing research in the social media domain, these collected datasets are often fixed in their construction, lack comprehensive content information, and are not universally applicable across multiple research fields. Furthermore, these datasets predominantly emphasize the topological relationships between users while overlooking the textual content of posts. Lastly, existing datasets are primarily centered on mainstream English-language social media platforms such as Twitter and Facebook, with limited representation from platforms with different cultural tones and user dynamics.

3 Dataset Construction

3.1 Data Collection

Firstly, we focus on four main themes: travel, movie, and car, psychology. We collected 774,413 posts from Sina Weibo platform using the official API by searching related hashtags and keywords. Specifically, we selected the '20 Minute Park Theory' as the focal theme within the domain of psychology. Then, based on these original posts, we extracted 14,914,660 comments, along with the user profiles of both the post creators and the commentators.

We conducted a detailed hashtag analysis within the corpus of posts, quantifying their frequency distribution. Subsequently, we identified hashtags with a high degree of relevance to the subject and filtered these posts, thereby guaranteeing a more concentrated theme relevance.

3.2 Data Cleaning

To maintain a dataset devoid of discriminatory content, we employed the OpenAI GPT-4o-mini API for the textual screening of each post. The OpenAI GPT series API includes a content moderation system that provides severity warnings for prompts containing elements of hate, sexual content, violence, or self-harm[1]. When harmful content is detected in the input, the API either returns an error or tags a content_filter label in the finish_reason field of the response. By leveraging this feature, we filtered the text of all posts. The results confirmed that no discriminatory content exists in the dataset, ensuring its safety for use in various machine-learning tasks without introducing potential biases.

Additionally, an analysis of post lengths revealed that all posts contain a minimum of five characters, negating the need for further filtering. A summary of the basic statistics for the final dataset is presented in Table 1. Note that before quantifying the character count in the posts, we removed all hyperlinks in the text and substituted emoji names that were enclosed in square brackets ("[]") with a corresponding single character, to ensure that the character count won't miscounted.

[1]https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/content-filter?tabs=definitions%2Cuser-prompt%2Cpython-new

Table 1: Basic Information about our Weibo-MT dataset.

	Travel	Car	Movie	Psychology	Total
#post	3,304	2,756	1,828	1,095	7,705
Avg. post length	223.06	151.13	103.54	59.33	182.74
#comment	34,962	39,639	132,805	4,650	202,238
#user	9,142	13,502	29,395	4,462	54,401
#user w/post	141	45	59	1,077	1,280
#user w/comment	9,020	13,466	29,343	3,433	53,208
Time Range (Days)	4,382	4,783	1,740	180	4,782

3.3 Data Anonymization

To ensure user privacy, we performed anonymization checks and processing on post, comment, and user data.

First, we replaced all post IDs, comment IDs, and user IDs with anonymized identifiers. This change ensures that these IDs cannot be traced back to records in the social media platform's backend. Second, we converted all timestamps into relative times based on a randomly assigned reference time. This approach preserves the temporal relationships between posts and comments while obscuring the actual timestamps, making it impossible to retrieve original posts within specific time frames. Lastly, all non-numerical attributes in user profiles were transformed into categorical representations, which eliminates the possibility of inferring user identities based on specific information.

Since original posts cannot be accurately retrieved on Sina Weibo based solely on the natural text content, the natural text included in our dataset does not pose risk of exposing user privacy.

3.4 Sentiment Analysis

In this study, we utilized the DistilBERT-based Multilingual Sentiment Classification Model[2] accessible through the Hugging Face platform to conduct sentiment analysis on the natural language text contained within user comments. The comments were systematically classified into five sentiment categories: *Very Negative*, *Negative*, *Neutral*, *Positive*, and *Very Positive*. This methodological approach not only retains a substantial degree of semantic information but also anonymizes the textual content, thereby safeguarding user privacy. Consequently, this anonymization facilitates the dataset's applicability to a broader array of downstream tasks.

4 Dataset Statistics

We conducted a comprehensive analysis of the dataset distribution from two distinct perspectives: post and user, particularly focusing on various theme. Our findings indicate substantial disparities in the distribution patterns of posts and user interactions across themes.

4.1 Post-center Analysis

Figure 1 reports the distribution of comment counts for posts under various themes, alongside a cumulative curves. We can identify notable disparities across themes. The cumulative percentage curves for the travel and car themes exhibit a relatively smooth trajectory. In contrast, the movie theme's comment count distribution

[2]https://huggingface.co/tabularisai/multilingual-sentiment-analysis

Table 2: Dataset Attribute Overview

	Attributes
Post	post_id, user_id, publish_time, comment_count, like_count, post_content, post_hashtags.
Comment	comment_id, post_id, user_id, publish_time, like_count, comment_count, sentiment.
User	user_id, gender, followee_count, followers_count, location, birthday, registry_time, certification, credit, constellation, comment_count, repost_count, like_count, post_count.

is characterized by a more uniform spread, as evidenced by the cumulative curve's nearly linear progression. This phenomenon is due to limited users within the movie theme, where a select few users contribute the majority of posts. The popularity of most posts from these users tends to be similar, resulting in analogous levels of engagement. Conversely, the comment count distribution towards psychology theme reveals a significant long-tail effect.

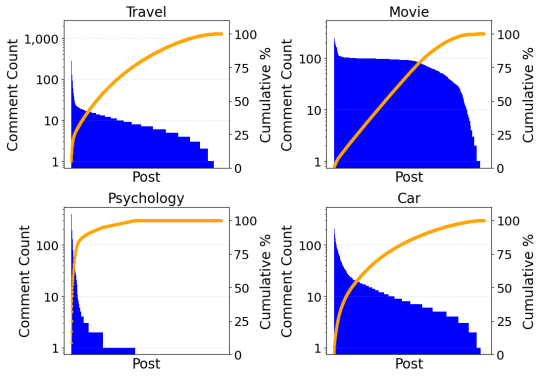

Figure 1: Comment Count and Cumulative % of four themes.

Furthermore, in Figure 2, we present the temporal distribution of comment counts for posts of different themes. To facilitate the observation of the overall distribution, outliers have been removed from boxplot except psychology. Travel, movie and car themes exhibit a trend where the number of post comments initially increases over time, followed by a gradual decrease in the growth rate as time progresses. This pattern aligns with a general understanding of social information diffusion. Although the psychology theme shows less pronounced trends in the boxplot due to a lower total number of comments, a similar pattern can still be observed from the outliers. This also suggests a Matthew effect, where increased visibility leads to a further acceleration in engagement as the number of viewers grows.

We further analyze the sentiment tendency of comments under various themes, shown in Figure 3. The results reveal that, except for the psychology theme, most sentiments are predominantly *Neutral*, followed by *Very Positive* and *Very Negative*. This suggests a balanced sentiment distribution, with users expressing distinct attitudes rather than ambivalence. In contrast, the psychology theme exhibits a strong positive bias, likely influenced by the "20 Minute Park Theory," highlighting the mental health benefits of spending time in nature. Notably, the presence of *Very Negative* comments of this theme may offer additional insights for sentiment analysis.

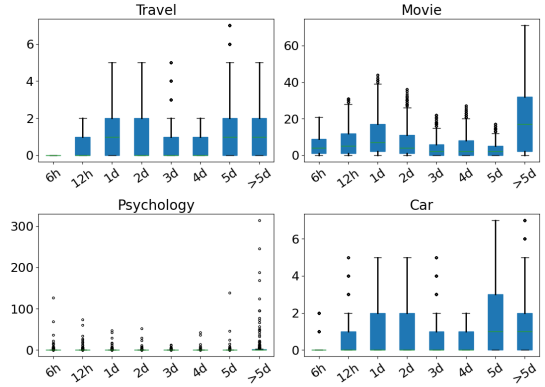

Figure 2: Comment Time Distribution

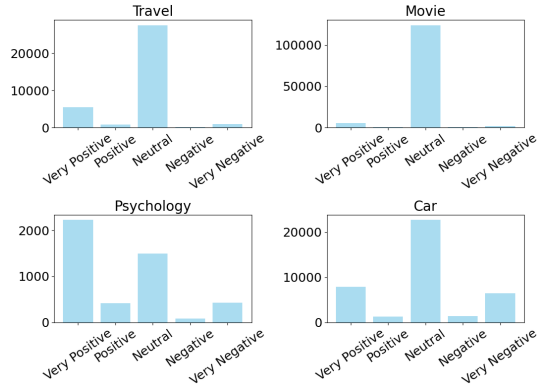

Figure 3: Comments' Sentiment Distribution

4.2 User-center Analysis

The user profiles include 14 attributes covering three aspects: Basic Information, Social Interaction, and Credit and Label, shown in Table 2. Figure 4 shows the distribution of user post count and cumulative curves of four themes. The travel and cars themes exhibit a long-tail distribution, while the movie theme displays a smoother exponential curve. The pyschology theme approaches uniformity, with a linear growth pattern evident in its cumulative curve.

Additionally, user comment count distributions reveal that the commenting user set differs from that of postings in Figure 5. Engagement is dominated by a small number of active users across travel, movie, and car themes, whereas the pyschology theme shows a more equitable distribution of posts among users.

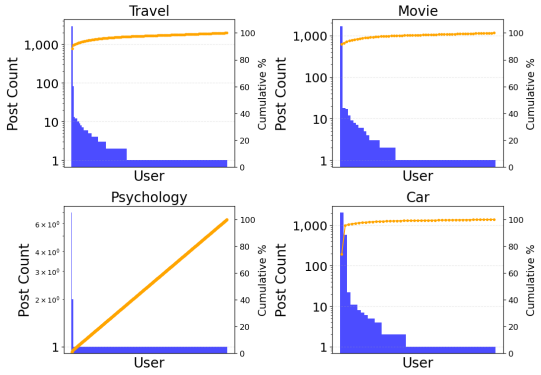

Figure 4: Post Count and Cumulative % for theme

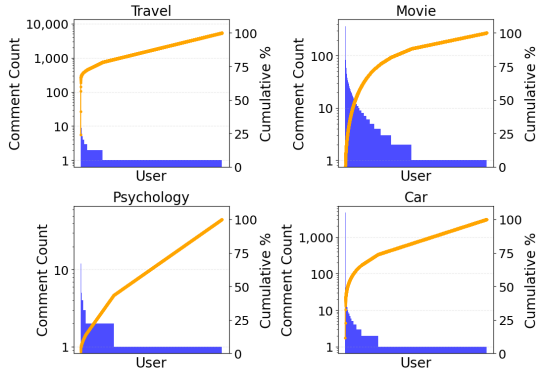

Figure 5: Comment Count and Cumulative % for theme

5 Potential Application of Dataset

5.1 Social Simulation and Information Diffusion

Research on social simulation and information diffusion aims to predict future states of information spread, including user involvement, diffusion paths, and the effects of various information types, such as rumors and public opinion [2]. Key tasks include cascade size prediction, popularity prediction, and next-user prediction and so on. Our dataset comprises post timestamps and related comments, along with user involvement, facilitating the construction of information diffusion paths. It serves as a fundamental resource for various studies in Information Diffusion.

5.2 Cross-platform Recommendation

This dataset can be integrated with related recommendation datasets under the same theme from other platforms to support the development of cross-platform recommendation systems. For example, in the domain of movie, datasets such as MovieLens, Douban Movie [14], Movie Review Dataset from Rotten Tomatoes [8], and IMDB Movie Reviews [6] provide rich item and user information. Similarly, in the travel domain, datasets like Yelp and other commercial or industry-specific datasets offer extensive item and user data. By comparing these with post content and comment ratings in the present dataset, additional dimensions of information can be utilized for training recommendation system models.

5.3 User Profiling

This dataset provides raw post content while ensuring user privacy protection, offering valuable natural language text for analysis and simulation. Additionally, it includes rich user profile attributes and historical behavior records, which can serve as crucial data and test samples for user modeling and simulations.

6 Conclusion and future work

In this paper, we introduce a multi-theme social media dataset, Weibo-MT, that includes realistic posts and comments across four themes: travel, movie, car, and psychology. Prioritizing user privacy, this dataset offers rich, multi-dimensional insights into user interactions. It serves as a valuable resource for social simulation, information diffusion, user modeling and cross-platform recommendations, enabling comparative analyses across diverse themes. Future research could leverage this dataset to perform fine-grained modeling of user behavior and information dissemination patterns, contributing to improved understanding and strategies in social computing.

References

[1] Qi Cao, Huawei Shen, Keting Cen, Wentao Ouyang, and Xueqi Cheng. 2017. Deephawkes: Bridging the gap between prediction and understanding of information cascades. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 1149–1158.

[2] Fuxia Guo, Xiaowen Wang, Yanwei Xie, Zehao Wang, Jingqiu Li, and Lanjun Wang. 2024. A Survey of Datasets for Information Diffusion Tasks. *arXiv preprint arXiv:2407.05161* (2024).

[3] Nathan O Hodas and Kristina Lerman. 2014. The simple rules of social contagion. *Scientific reports* 4, 1 (2014), 4343.

[4] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What is Twitter, a social network or a news media?. In *Proceedings of the 19th international conference on World wide web (WWW '10)*. Association for Computing Machinery, New York, NY, USA, 591–600. https://doi.org/10.1145/1772690.1772751

[5] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What is Twitter, a social network or a news media?. In *Proceedings of the 19th international conference on World wide web*. 591–600.

[6] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning Word Vectors for Sentiment Analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Portland, Oregon, USA, 142–150. http://www.aclweb.org/anthology/P11-1015

[7] Qiong Nan, Juan Cao, Yongchun Zhu, Yanyan Wang, and Jintao Li. 2021. MDFEND: Multi-domain fake news detection. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 3343–3347.

[8] Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the ACL*.

[9] Jiezhong Qiu, Jian Tang, Hao Ma, Yuxiao Dong, Kuansan Wang, and Jie Tang. 2018. Deepinf: Social influence prediction with deep learning. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 2110–2119.

[10] Lilian Weng, Filippo Menczer, and Yong-Yeol Ahn. 2013. Virality prediction and community structure in social networks. *Scientific reports* 3, 1 (2013), 1–6.

[11] Cheng Yang, Hao Wang, Jian Tang, Chuan Shi, Maosong Sun, Ganqu Cui, and Zhiyuan Liu. 2021. Full-scale information diffusion prediction with reinforced recurrent networks. *IEEE Transactions on Neural Networks and Learning Systems* 34, 5 (2021), 2271–2283.

[12] Louis Yu, Sitaram Asur, and Bernardo A. Huberman. 2011. What Trends in Chinese Social Media. https://doi.org/10.48550/arXiv.1107.3522 arXiv:1107.3522 [physics].

[13] Jing Zhang, Biao Liu, Jie Tang, Ting Chen, and Juanzi Li. 2013. Social influence locality for modeling retweeting behaviors. In *Twenty-third international joint conference on artificial intelligence*. Citeseer.

[14] Feng Zhu, Yan Wang, Chaochao Chen, Guanfeng Liu, and Xiaolin Zheng. 2020. A Graphical and Attentional Framework for Dual-Target Cross-Domain Recommendation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*. 3001–3008.