

Comparing User Activity on X and Mastodon

Shiori Hironaka
Kyoto University
Kyoto, Japan

hironaka@media.kyoto-u.ac.jp
0000-0001-7994-2858

Mitsuo Yoshida
University of Tsukuba
Tokyo, Japan

mitsuo@gssm.otsuka.tsukuba.ac.jp
0000-0002-0735-1116

Kazuyuki Shudo
Kyoto University
Kyoto, Japan

0000-0002-3939-9800

Abstract—The “Fediverse”, a federation of decentralized social media servers, has emerged after a decade in which centralized platforms like X (formerly Twitter) have dominated the landscape. The structure of a federation should affect user activity, as a user selects a server to access the Fediverse and posts are distributed along the structure. This paper reports on the differences in user activity between Twitter and Mastodon, a prominent example of decentralized social media. The target of the analysis is Japanese posts because both Twitter and Mastodon are actively used especially in Japan. Our findings include a larger number of replies on Twitter, more consistent user engagement on mstdn.jp, and different topic preferences on each server.

Index Terms—Fediverse, Social media analysis, User activity, Topic analysis

I. INTRODUCTION

Social media has become an integral part of people’s daily lives, serving as a platform for communication, information gathering, and content sharing. These platforms generate vast amount of data that capture various aspects of human activity and interaction. Researchers have utilized social media data to analyze social trends [1], [2] and gain insights into public opinions and reactions to social issues [3]–[5].

Traditionally, centralized social media platforms like X¹ (formerly Twitter) have dominated the landscape, with a single company acting as the service provider. However, recent years have seen the emergence and growth of decentralized social media platforms. These decentralized networks, often referred to as the “Fediverse,” consist of multiple servers operated by different service providers. They form a large, interconnected social network using a unified protocol implemented across server clusters. Users can choose and connect to their preferred servers to access these services.

One prominent example of decentralized social media is Mastodon², which implements the ActivityPub protocol. While the user base of decentralized platforms has been growing, it remains unclear whether their user activities mirror those of established centralized platforms like Twitter.

Japan presents an interesting case study in this context. As Twitter’s second-largest market after the United States, it has been widely adopted for various purposes, including accessing

real-time news and updates, communicating with friends, and sharing opinions on current events and hobbies. Concurrently, Japan has experienced significant adoption of decentralized social media platforms.

This research aims to investigate the differences in user activities between Twitter and Mastodon from a user and topic perspective. By examining these platforms, we seek to understand how user activities differ between centralised and decentralised social media ecosystems. Our findings include a larger number of replies on Twitter, more consistent user engagement on mstdn.jp, and different topic preferences on each server.

II. DATA

A. Collection

We collect data using streaming API of Twitter and Mastodon. Mastodon is a decentralized social media platform where services are independent of each other in units called instances (servers). Users are free to create and operate instances. Even if users do not open an instance, users can use social media by joining other people’s instances. When collecting data on Mastodon, it is necessary to collect data for each instance separately. In this paper, we used data from mstdn.jp, the largest Mastodon instance in Japan, and pawoo.net an instance that focuses on topics related to illustration.

From Twitter, posts are collected using the 1% sampled streaming API, limited to Japanese. Local timelines were collected from two instances of Mastodon, mstdn.jp and pawoo.net. An instance’s local timeline contains posts from users within that instance whose posts are set to public. Although Mastodon does not limit the language of posts to Japanese.

As a result, we collected 125,703,990 posts from Twitter, 999,392 posts from mstdn.jp, and 425,138 posts from pawoo.net, between May 3 and June 23, 2023³.

B. Aggregation

The data we collected consists of streams of posts. Each post is accompanied by the user’s profile information at the time of posting. Therefore, we aggregate these posts by user and extract the most recent profile information for each user.

There are three types of posts:

- Original posts: Posts that are neither replies nor shares.

³After June 23, 2023, Twitter’s API is no longer available.

The research was supported by ROIS NII Open Collaborative Research 2024-24FS02.

¹<https://x.com/> (accessed 2024-10-15)

²<https://joinmastodon.org/> (accessed 2024-10-15)

- Reply posts: Posts with a reply-to post ID in their metadata.
- Share posts: Posts referred to as “boosts” on Mastodon and “reposts” or “retweets” on Twitter.

For each user, we calculate the following user attributes:

- Total posts during the period: The number of posts made by the user that appeared in the stream during the collection period.
- Proportion of reply posts: The ratio of reply posts to the total number of collected posts for the user.
- Active days: The number of days between the user’s first and last posts within the collection period.

We conduct analyses for each user attributes. In addition to these attributes, we conduct analyses using the number of followees and followers, and follower–followee ratio.

C. Limitation

The data used in this analysis was collected through streaming APIs. There are several limitations to consider:

- For Twitter, we used a 1% sampled data stream, resulting in a lower proportion of collected posts compared to Mastodon.
- Mastodon data was collected from local timelines. Consequently, reblogged posts are not included in our dataset.
- While the Twitter data consists entirely of Japanese posts due to language filtering during collection, the Mastodon data was collected without language restrictions. As a result, the Mastodon dataset may include posts in languages other than Japanese.

These limitations should be taken into account when interpreting the results of our comparative analysis between Twitter and Mastodon.

III. ANALYSIS OF USER ATTRIBUTES

A. Total Posts During the Period

Platforms with a high number of posts are considered actively used platforms. First, Figure 1 shows the probability density histograms of the number of posts for each post type. Share posts do not exist on Mastodon, so they are not used in this study. The number of posts refers to the total number of posts across all post types.

From Figure 1, it can be observed that Twitter tends to have many users with a low number of posts. Only Twitter data is 1% sampled, while Mastodon data is 100% complete. As a result, many Twitter users have no observed posts, and even when posts are observed, only 1/100 of the actual posts are captured. This is likely the cause of the observed distribution.

Since Twitter posts may actually exist at 100 times the observed rate, Figure 1 also includes a plot where the observed number of Twitter posts is multiplied by 100. Considering any type of post, it can be inferred that Twitter has the highest number of posts. Comparing Mastodon instances, it was found that mstdn.jp tends to have more posts than pawoo.net.

B. Proportion of Reply Posts

Replies constitute a form of direct communication in social media, facilitating user-to-user conversations. A high proportion of replies indicates that a platform is used for direct interactions and dialogic communication.

We calculated the ratio of reply posts to total posts for each user in our collected data. Since only a small fraction of users utilize replies, and displaying data for all users would make the results difficult to interpret, we focused on users who actively use replies. We plotted data only for users with more than 5 reply posts, as shown in Figure 2. Among users who posted at least once during the collection period, the percentage of users with more than 5 reply posts was 4.1% for Twitter, 5.0% for mstdn.jp, and 1.8% for pawoo.net. These users are represented in the plot.

Figure 2 reveals that Twitter tends to have a higher proportion of reply posts compared to Mastodon. This result may be attributed to either more frequent use of replies on Twitter or fewer replies appearing on Mastodon’s public timeline. Mastodon offers more flexible privacy settings for posts compared to Twitter. Consequently, Mastodon users may choose not to display their replies on the local timeline, potentially resulting in fewer replies in our aggregated data. Twitter lacks such granular settings; thus, if an account is public, replies are displayed on the public timeline by default.

C. Active Days

To examine whether platforms are used continuously, we calculated the active days for each user. We define active days as the number of days between a user’s first and last observed posts in our collected data. Users with fewer than two posts are assigned zero active days. Given our total data collection period of 52 days, the maximum possible active days is 52.

Figure 3 presents the probability density histogram of active days. The results indicate that mstdn.jp exhibits a higher proportion of users with longer active days, suggesting more consistent and enthusiastic user engagement compared to other servers. Conversely, pawoo.net demonstrates a higher proportion of users with active days of just a few days. Twitter falls between these two extremes; however, direct comparison is not feasible due to the 1% sampling of Twitter data.

D. Degree Distribution

We extracted the number of followees and followers for each user from their profile information. Due to the difference in social network sizes between Twitter and the Fediverse, the scale of followees and followers for each user also differs. To compare the distribution of followees and followers across platforms, we plotted probability density histograms.

The results are shown in Figure 4. Our analysis revealed that mstdn.jp exhibited the highest proportion of users with zero followees, while pawoo.net demonstrated the highest proportion of users with zero followers. Furthermore, mstdn.jp tended to have fewer users with a high number of followers. Moreover, both Mastodon instances (mstdn.jp and pawoo.net) showed a tendency towards having more users with fewer

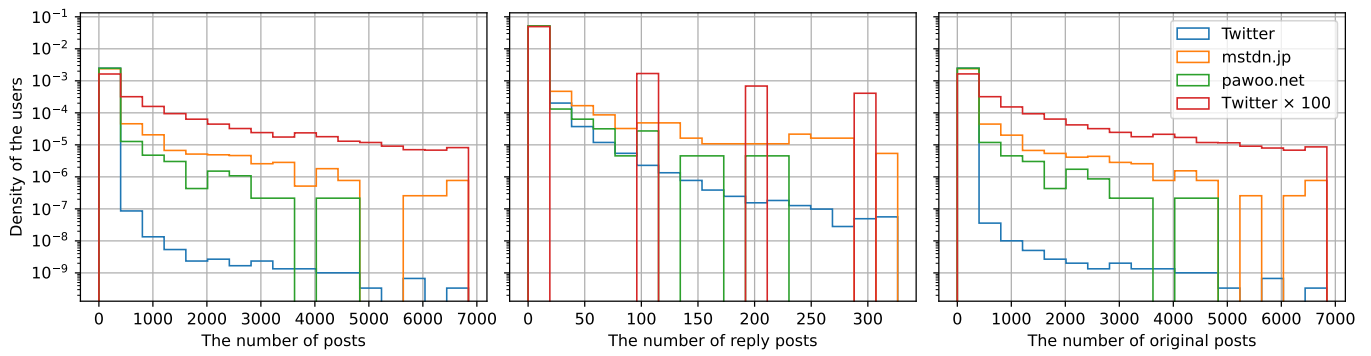


Fig. 1. Density histogram of the number of posts, reply posts, and original posts.

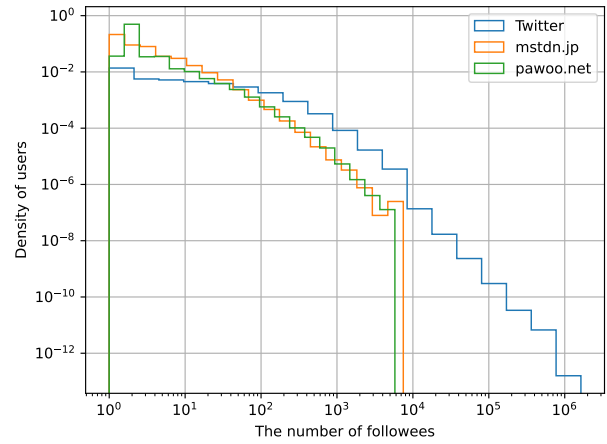
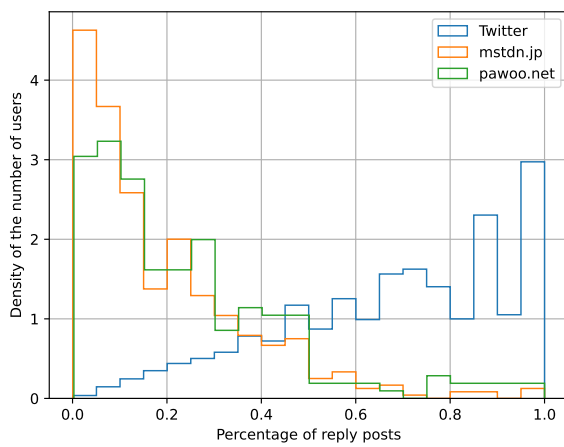
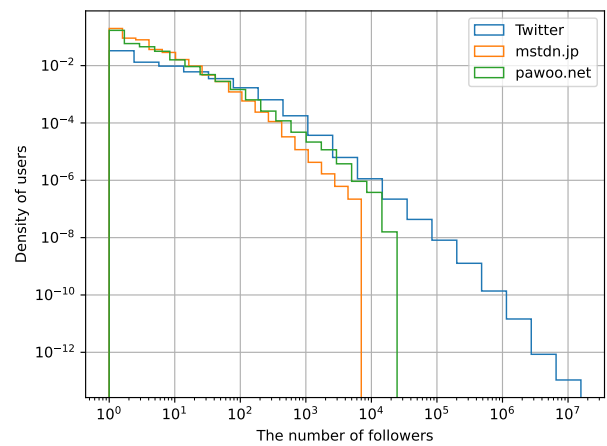
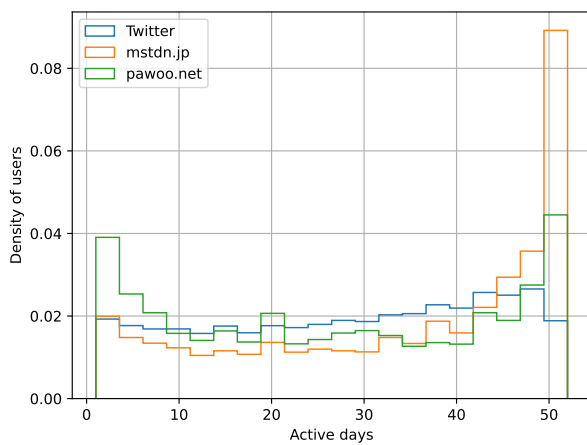


Fig. 2. Density histogram of the percentage of reply posts across platforms. Twitter tends to have a higher proportion of reply posts compared to Mastodon.

(a) Number of followers.



(b) Number of followers.

Fig. 3. Density histogram of active days across platforms. Users on mstdn.jp exhibit a tendency towards more regular platform engagement.

Fig. 4. Density histogram of number of followers and followers across platforms.

followers and followers compared to Twitter. This observation may indicate that the Fediverse is still an emerging social network that has not yet reached its full potential in terms of user connections and network density.

E. Follower–Followee Ratio

The number of followees and followers are key characteristics of a user’s established social network, reflecting how users utilize social media platforms. However, these values are influenced by the total number of users on each platform, making direct comparisons between Twitter and the Fediverse challenging.

Previous studies have classified users based on their follower–followee ratio [6], [7]. Therefore, we calculate the ratio of followees to followers. The follower–followee ratio is calculated using the following formula:

$$\text{Ratio} = \log_{10} \left(\frac{\text{Number of Followees} + 1}{\text{Number of Followers} + 1} \right) \quad (1)$$

We apply a logarithmic transformation to the ratio. Consequently, users with positive values have more followees than followers, while users with negative values have more followers than followees.

Figure 5 shows the probability density histogram of the calculated follower–followee ratios for each platform. To exclude inactive users, we only plotted data for users with a combined total of followees and followers greater than 20. Our analysis reveals several key observations. First, mstdn.jp exhibits a high proportion of users with follower–followee ratios close to 0, indicating a good balance between followees and followers. This suggests that these users primarily use social media for interaction purposes. Second, Twitter has more users with positive values compared to other platforms, indicating a higher proportion of users with subscribing purposes. Lastly, pawoo.net has more users with negative values, suggesting the presence of some influential users on this platform. These findings highlight the different user behaviors and network structures across the studied platforms, reflecting diverse user activities and social dynamics. The variations in follower–followee ratios provide insights into how users on each platform engage with content and build their social networks.

IV. ANALYSIS OF TOPICS

To compare across platforms, we first train topic models. Then, using the estimated topics, we investigate the characteristics of topics in posted content for each platform.

A. Data Preprocess

To prepare the collected posts for analysis, the following preprocessing steps were performed:

- 1) Extract post content and normalize strings.
- 2) Determine vocabulary.
- 3) Create a corpus for topic analysis.

First, reposts and boosts (both features similar to retweets) were excluded from the collected posts. Since Mastodon posts

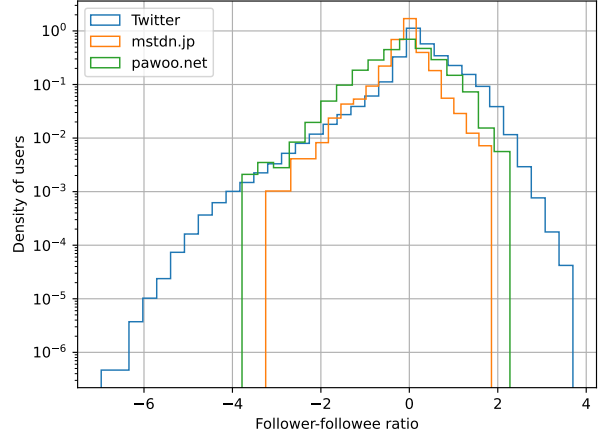


Fig. 5. Density histogram of follower–followee ratios across platforms.

contain HTML tags, these tags were removed. Additionally, URLs and user IDs were removed from both Twitter and Mastodon posts, and string normalization was performed using neologdn⁴.

Next, Japanese morphological analysis was then performed using Vibrato⁵ and mecab-ipadic v2.7.0. Words whose parts of speech were identified as nouns other than non-independent, verbs other than non-independent or suffixes, adjectives, adverbs, adnominals, interjections, or symbols were kept in the documents, while others were removed. Verbs were lemmatized to their base forms. Single-character hiragana or alphabet words and words composed solely of numbers were also removed. Words appearing in more than 10% of all posts were excluded, and the top 50,000 words with the highest document frequency were selected as the vocabulary.

A corpus of documents was created that contained only words from the vocabulary present in each post. Posts that did not contain words from the vocabulary were excluded.

Finally, we got 17,432,493 posts from Twitter, 767,746 posts from mstdn.jp, and 268,614 posts from pawoo.net. The number of posts from Twitter is more than ten times higher than that from Mastodon. The average number of words per document after preprocessing was 8.01 for Twitter, 8.81 for mstdn.jp, and 10.55 for pawoo.net.

B. Preparation of Topic Model

The Biterm Topic Model (BTM) [8] is employed for our analysis. BTM has been reported to be particularly effective for short texts [9]. Given that this study involves topic analysis of over 18 million posts, it is crucial to utilize an efficient method to handle such a large dataset. We experimented with various numbers of topics (K) for the BTM. Ultimately, we determined that $K = 30$ was sufficient to capture the essential

⁴<https://github.com/ikegami-yukino/neologdn>

⁵<https://github.com/daac-tools/vibrato>

characteristics of the platform. Therefore, this number of topics was adopted for our analysis.

Suppose that given corpus D ($d \in D$) and the number of topics K , θ_k , the probability of topic k , and ϕ_{kv} , the probability of word v choosing topic k , can be learned. Then, from these parameters, the topic distribution $P(z = k | d)$ for a document can be estimated. In this study, we used the bitermplus⁶ package for BTM computations, with hyperparameters set to $\alpha = 6.25$ and $\beta = 0.01$.

To examine the characteristics of a set of documents, we define the topic distribution for the document set. The topic distribution for document set D_F is calculated using the following equation:

$$\frac{1}{N_F} \sum_{d \in D_F} P(z = k | d) \quad (2)$$

where N_F is the number of posts in D_F ($|D_F| = N_F$).

C. Result

We investigate the characteristics of the posted content for each platform using traintd model. The topic distributions for each platform, computed from the BTM with $K = 30$, are shown in Figure 6. Overall, the broad trends of posting topics for each platform have been captured.

According to Figure 6, the most common topics in Twitter’s posts are Topics 0, 13, 14, 15, and 26. Topic 0 consists mainly of emojis and words expressing happiness and support for content creators or favorites. Topic 13 is about Gacha games and events, Topic 14 is about exchanging goods, Topic 15 is about Seven Eleven-related advocacy campaigns, and Topic 26 is about Lawson-related advocacy campaigns. It is noteworthy that campaigns to encourage user contributions as part of corporate marketing strategies are prevalent on Twitter, but not observed on Mastodon.

According to Figure 6, the prevalent topics in mstdn.jp posts are Topic 1, Topic 10, and Topic 11. Topic 1 consists of everyday posts, Topic 10 contains words related to life and family, and Topic 11 contains words related to society and politics, with mature expressions.

According to Figure 6, the most common topics in pawoo.net posts are Topic 17 and Topic 28, with Topic 16 being as common as Twitter. Topic 17 contains words related to anime, AI, illustration, Topic 28 contains mostly English words, and Topic 16 contains words related to the illustration community. This reflects the preference of the Mastodon instance for creative activities centered around illustration.

V. RELATED WORK

The study of user activities on social media platforms has been a long-standing area of research in the field of social computing [10]–[15]. These patterns provide valuable insights into user behavior, information dissemination, and community dynamics. However, as Trifiro and Gerson [16] point out, many existing methodologies are constrained to single social media

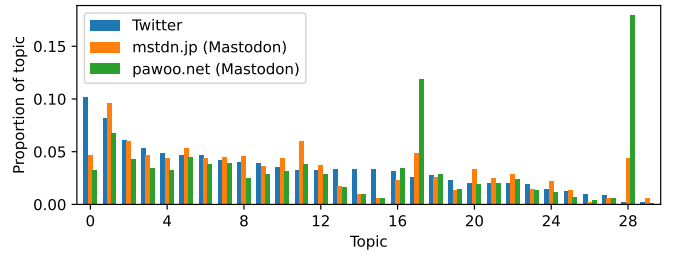


Fig. 6. Topic distributions for each platform ($K = 30$).

platforms, highlighting a significant gap in approaches that can capture cross-platform or generalized usage patterns.

In recent years, the landscape of social media has evolved with the emergence and growth of decentralized platforms. These platforms, exemplified by Mastodon and others in the Fediverse, introduce unique characteristics that set them apart from traditional centralized platforms like Twitter [17]. Several studies have begun to explore these decentralized ecosystems. Zignani et al. [18] provided one of the first comprehensive analyses of the Mastodon network structure, while La Cava et al. [19], [20] explored the specific features and user behaviors within decentralized social media environments. Khateeb et al. [21] also analyzed trends within a single Mastodon instance. To the best of our knowledge, no research has yet undertaken a comprehensive comparison of user activities between Twitter and Mastodon, nor attempted to compare topics across multiple Mastodon instances. Our study aims to address this gap by providing the first direct comparison of user activities between Twitter and Mastodon.

VI. CONCLUSION

This study aimed to investigate the differences in user activities between centralized (Twitter) and decentralized (Mastodon) social media platforms, focusing on user attributes and topic distributions. Our analysis revealed several key findings:

- 1) Post frequency: Twitter users generally posted more frequently than Mastodon users, with mstdn.jp users being more active than pawoo.net users.
- 2) Reply behavior: Twitter showed a higher proportion of reply posts compared to Mastodon instances, suggesting more direct user-to-user interactions.
- 3) User engagement: mstdn.jp demonstrated more consistent user engagement over time compared to pawoo.net and Twitter.
- 4) Network structure: Mastodon instances showed a tendency towards smaller number of followers and followers compared to Twitter, possibly reflecting the nascent stage of the Fediverse.
- 5) Topic distribution: Each platform exhibited distinct topic preferences, with Twitter featuring more marketing campaigns, mstdn.jp focusing on everyday life and societal issues, and pawoo.net centering around creative activities.

⁶<https://github.com/maximtrp/bitermplus>

These findings highlight the unique characteristics of centralized and decentralized social media platforms, reflecting differences in user behavior, community dynamics, and content focus. Our research contributes to the understanding of how decentralized social media ecosystems differ from traditional centralized platforms, providing insights for future development and research in this evolving landscape. Furthermore, this study provides insights into key aspects for future research to focus on when comparing different social media platforms.

REFERENCES

- [1] J. Benhardus and J. Kalita, "Streaming Trend Detection in Twitter," *International Journal of Web Based Communities*, vol. 9, no. 1, pp. 122–139, 2013.
- [2] A. Bogdanowicz and C. Guan, "Dynamic topic modeling of twitter data during the COVID-19 pandemic," *PLOS ONE*, vol. 17, no. 5, p. e0268669, 2022.
- [3] M. J. Kushin, M. Yamamoto, and F. Dalisay, "Societal Majority, Facebook, and the Spiral of Silence in the 2016 US Presidential Election," *Social Media + Society*, vol. 5, no. 2, p. 2056305119855139, 2019.
- [4] S. Chou Jen, O. Masanao, S. Takeshi, N. Ken, S. Kanji, M. Junichiro, and S. Ichiro, "Constructive Approach for Early Extraction of Viral Spreading Social Issues from Twitter," in *Proceedings of the 12th ACM Conference on Web Science*, 2020, pp. 96–105.
- [5] T. Hu, M. H. Ribeiro, R. West, and A. Spitz, "Quotatives Indicate Decline in Objectivity in U.S. Political News," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 17, pp. 363–374, 2023.
- [6] W. Yan, Y. Zhang, and W. Bromfield, "Analyzing the follower–followee ratio to determine user characteristics and institutional participation differences among research universities on ResearchGate," *Scientometrics*, vol. 115, no. 1, pp. 299–316, 2018.
- [7] H. Oshimo, S. Hironaka, M. Yoshida, and K. Umemura, "Follower–Followee Ratio Category and User Vector for Analyzing Following Behavior," in *Proceedings of the 9th International Conference on Advanced Informatics: Concepts, Theory and Applications*, 2022.
- [8] X. Cheng, X. Yan, Y. Lan, and J. Guo, "BTM: Topic Modeling over Short Texts," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 12, pp. 2928–2941, 2014.
- [9] J. Qiang, Z. Qian, Y. Li, Y. Yuan, and X. Wu, "Short Text Topic Modeling Techniques, Applications, and Performance: A Survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 3, pp. 1427–1445, 2022.
- [10] B. Poblete, R. Garcia, M. Mendoza, and A. Jaimes, "Do All Birds Tweet the Same? Characterizing Twitter Around the World," in *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, 2011, pp. 1025–1030.
- [11] P. A. Longley, M. Adnan, and G. Lansley, "The Geotemporal Demographics of Twitter Usage," *Environment and Planning A: Economy and Space*, vol. 47, no. 2, pp. 465–484, 2015.
- [12] K. Das and S. K. Sinha, "A Survey on User Behaviour Analysis in Social Networks," *International Journal of Computer Science and Information Security*, vol. 14, no. 11, pp. 895–908, 2016.
- [13] X. Luo, C. Jiang, W. Wang, Y. Xu, J.-H. Wang, and W. Zhao, "User behavior prediction in social networks using weighted extreme learning machine with distribution optimization," *Future Generation Computer Systems*, vol. 93, pp. 1023–1035, 2019.
- [14] P. M. Valkenburg, I. I. van Driel, and I. Beyens, "The associations of active and passive social media use with well-being: A critical scoping review," *New Media & Society*, vol. 24, no. 2, pp. 530–549, 2022.
- [15] Z. Xue, Q. Li, and X. Zeng, "Social media user behavior analysis applied to the fashion and apparel industry in the big data era," *Journal of Retailing and Consumer Services*, vol. 72, p. 103299, 2023.
- [16] B. M. Trifiro and J. Gerson, "Social Media Usage Patterns: Research Note Regarding the Lack of Universal Validated Measures for Active and Passive Use," *Social Media + Society*, vol. 5, no. 2, p. 2056305119848743, 2019.
- [17] A. Raman, S. Joglekar, E. D. Cristofaro, N. Sastry, and G. Tyson, "Challenges in the Decentralised Web: The Mastodon Case," in *Proceedings of the Internet Measurement Conference*, 2019, pp. 217–229.
- [18] M. Zignani, C. Quadri, S. Gaito, H. Cherifi, and G. P. Rossi, "The Footprints of a "Mastodon": How a Decentralized Architecture Influences Online Social Relationships," in *IEEE Conference on Computer Communications Workshops*, 2019, pp. 472–477.
- [19] L. La Cava, S. Greco, and A. Tagarelli, "Understanding the growth of the Fediverse through the lens of Mastodon," *Applied Network Science*, vol. 6, no. 1, pp. 1–35, 2021.
- [20] —, "Information consumption and boundary spanning in Decentralized Online Social Networks: The case of Mastodon users," *Online Social Networks and Media*, vol. 30, p. 100220, 2022.
- [21] S. Al-khateeb, "Dapping into the Fediverse: Analyzing What's Trending on Mastodon Social," in *Social, Cultural, and Behavioral Modeling*, 2022, pp. 101–110.