FediData: A Comprehensive Multi-Modal Fediverse Dataset from Mastodon

Min Gao
Shanghai Key Lab of Intelligent Information Processing, College of Computer Science and Artificial Intelligence, Fudan University
Shanghai, China
mgao21@m.fudan.edu.com

Haoran Du
Shanghai Key Lab of Intelligent Information Processing, College of Computer Science and Artificial Intelligence, Fudan University
Shanghai, China
hrdu24@m.fudan.edu.cn

Wen Wen
Shanghai Key Lab of Intelligent Information Processing, College of Computer Science and Artificial Intelligence, Fudan University
Shanghai, China
wwen24@m.fudan.edu.cn

Qiang Duan
Department of Information Sciences and Technology, The Pennsylvania State University
Abington, Pennsylvania, United States
qduan@psu.edu

Xin Wang*
Shanghai Key Lab of Intelligent Information Processing, College of Computer Science and Artificial Intelligence, Fudan University
Shanghai, China
xinw@fudan.edu.cn

Yang Chen*
Shanghai Key Lab of Intelligent Information Processing, College of Computer Science and Artificial Intelligence, Fudan University
Shanghai, China
chenyang@fudan.edu.cn

Abstract

Recently, decentralized online social networks (DOSNs) such as Mastodon have emerged quickly, bringing new opportunities for studies in user behavior modeling and multi-modal learning. However, their decentralized architecture presents two key challenges: 1) Distributed data and inconsistent access strategies across several individual instances make a unified collection difficult; 2) user-generated content (UGC) contains multiple modalities while lacking standard organization and high-quality annotation. To address these issues, we constructed FediData, a comprehensive multi-modal dataset from Mastodon. Our dataset integrates user profiles, text, images, and social interactions. To validate FediData's usefulness, we designed and analyzed several tasks and systematically evaluated the performance of existing state-of-the-art methods. Our analysis reveals the unique challenges of DOSNs and highlights the value of FediData in DOSN-related studies. We believe FediData could serve as a foundational dataset for advancing user behavior analytics, multi-modal learning, and future decentralized web research. All data and documentation are available in a Zenodo repository at https://zenodo.org/records/15621243 (DOI: 10.5281/zenodo.15621243).

CCS Concepts

• **Information systems → Social networks**.

*Xin Wang and Yang Chen are Corresponding Authors.

CIKM '25, Seoul, Republic of Korea

ACM ISBN 979-8-4007-2040-6/2025/11
https://doi.org/10.1145/3746252.3761634

Keywords

Decentralized Online Social Networks; Fediverse; Mastodon; Multi-Modal Dataset

ACM Reference Format:
Min Gao, Haoran Du, Wen Wen, Qiang Duan, Xin Wang, and Yang Chen. 2025. FediData: A Comprehensive Multi-Modal Fediverse Dataset from Mastodon. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM '25), November 10–14, 2025, Seoul, Republic of Korea.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3746252.3761634

1 Introduction

Recently, decentralized online social networks (DOSNs) have developed and gained public attention. Unlike traditional social networks, DOSNs provide users with more freedom to store their data and communicate in a distributed mechanism [1, 15]. Among them, Mastodon is one of the most representative DOSNs, consisting of multiple servers operated by different organizations [2, 5, 17, 21]. Mastodon's distributed inter-connectivity relies on the ActivityPub protocol[1]. Supported by this architecture, users generate and share content in a wide range of modalities, such as text, images, and social interactions. Such multi-modal data reflects users' real social behaviors and expressions. This user-generated content (UGC) data provides key support for research tasks such as multi-modal sentiment computing [16], user behavior modeling [15], and toxicity content moderation [4]. Additionally, it lays an important foundation for exploring cross-modal semantic alignment and fusion mechanisms.

Although DOSNs contain rich multi-modal UGC data, publicly available multi-modal datasets are still extremely scarce. Among most existing studies, only a few have released open-source datasets

[1]An open standard developed by the World Wide Web Consortium (W3C) to enable federated communication between different social platforms. This protocol is an open standard developed by the World Wide Web Consortium (W3C) to enable federated communication between different social platforms, allowing content to be synchronized across nodes while retaining each node's control over data locally. Please refer to https://www.w3.org/TR/activitypub/ for more information.

[13, 18, 28], while only covering one or two modalities. For example, FederatedSharing [14] contains user metadata and social relationships, while Fedivertex[2] contains social graphs from seven DOSNs. Moreover, several efforts have been made to provide tools for collecting datasets from DOSNs. For example, FediLive [20] and Mastodoner [27] are two good cases. The former was designed to collect real-time snapshots of users from Mastodon, while the latter was designed to collect both user data and instance activity data. While existing open-source datasets and tools offer some support for dataset crawling, they still require additional time, effort, and resources to collect satisfactory datasets. Consequently, the current datasets and collection tools face challenges in advancing research domains, such as multi-modal learning and computational social science, as well as in supporting critical studies, like sentiment analysis and social bot detection using multi-modal data of DOSN users.

Constructing a multi-modal dataset from decentralized platforms presents several unique challenges (CHs). ***CH1:*** During data collection, Mastodon's distributed architecture introduces challenges such as varied data formats and rate limits, making a unified collection difficult. ***CH2:*** UGC data could exhibit high heterogeneity, including images, texts, user metadata, and social links, while lacking standard organization and high-quality annotation. Therefore, these challenges hinder researchers from accessing and analyzing public Mastodon data effectively.

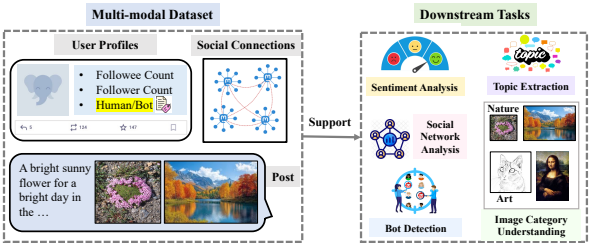


Figure 1: Illustration of FediData Dataset Framework.

To tackle the above challenges, we propose FediData, the first open multi-modal dataset collected from Mastodon, which is dedicated to providing realistic and reliable data support for social behavior modeling, multi-modal learning, and studies on user interaction patterns. To collect data from Mastodon, we design and implement a distributed data collection and processing framework tailored for decentralized scenarios. We collect various multi-modal user data, including text, images, user metadata, and social relationships. As shown in Figure 1, the constructed dataset has the following key features: (1) Fusion of multi-modal heterogeneous data: FediData covers tweets, images, user profiles, and social relationships, and supports a variety of downstream tasks, such as sentiment analysis [19], topic extraction [19], and social bot detection [11, 19, 23]. (2) Alignment and annotation of multi-modal data: During the data collection process, the modalities are aligned with precise timestamps, preserving the complex social behaviors in the user interaction process. Moreover, high-quality annotation of social bots on this dataset supports bot-related studies. To further explore the potential of the dataset, we design a series of benchmarking tasks and systematically evaluate the performance of existing methods on this dataset to validate its practical value. Our contributions are threefold:

- We collected and open-sourced FediData, a large-scale, real-world dataset sourced from a federated social platform called Mastodon. FediData incorporates diverse content types, including textual posts, images, user profiles, and social links, which fills the gap of social media datasets in decentralized platforms.
- FediData is equipped with high-quality, multi-modal fine-aligned data structures and manually annotated user labels. This offers reliable training and evaluation benchmarks essential for studies in multi-modal learning, emotion analysis, and social bot detection.
- To validate the practical value and research potential of the FediData dataset, we design several important benchmarking tasks and a targeted evaluation of the performance of existing representative methods and tools. These tasks not only reveal the value of our dataset but also provide a research foundation for the development of related fields.

Ethical Considerations. We adhere to explicit protocols to ensure ethical data handling throughout data collection and preprocessing. We comply with the rate limits, the Robots Exclusion Protocol, and DOSNs' content moderation policies to avoid unauthorized access to private data. This study was reviewed and approved by the Institute of Science and Technology at Fudan University. We release the dataset and documentation to support reproducibility and promote future research in multi-modal learning within DOSNs.

2 Dataset Construction

Data Collection. To construct FediData, we designed a simple and effective collection algorithm tailored for DOSNs. We collected publicly available user data from multiple instances of Mastodon using its official REST API[3]. We identified active instances and then used standardized APIs and scraping protocols to collect posts, images, user profiles, and follow relationships. We performed a BFS-based traversal through the following relationships, starting from many active seed users. Because each instance imposes rate limits (300 requests per 5 minutes), we incorporated a dynamic rate adjustment strategy to pause queries to an instance once its limit is reached, and then automatically resume collection when permissible. Following FediLive [20], we mapped unique identifiers by combining usernames and instance names (e.g., username@instance_name) and similarly annotating post IDs. This strategy prevents ID collisions, ensuring accurate identification of users and posts. In this study, we crawled the data in two phases. The first phase crawled users' posts, profiles, and social relationships during October 25-30, 2024, using seven Linux servers with Intel Xeon Silver 4114 (40 cores) and 187 GB memory, ensuring efficient and reliable data collection. The second phase crawled user avatars and images from user posts, conducted on a Linux server equipped with an AMD EPYC 7402P processor (24 physical cores/48 threads), 125 GB memory, and about 1.5 TB of local disk space, providing sufficient and stable computing and storage resources for multi-threaded parallel image crawling.
Data Pre-processing. We performed several pre-processing operations on the raw collected data to ensure data availability and privacy compliance. First, we uniformly mapped all user IDs and

[2]https://www.kaggle.com/datasets/marcdamie/fediverse-graph-dataset

[3]https://docs.joinmastodon.org/client/intro/

used a hash function to convert the original IDs into irreversible anonymous identifiers to prevent leakage of user-identifiable information. Meanwhile, to further protect user privacy, we anonymized usernames with hashing techniques [7]. Finally, we standardized the timestamp format, converted the local time in different servers to UTC, and filtered out invalid samples with abnormal format, missing modals, or empty content. After the above processing, the final dataset is cleaner and more unified, facilitating the training and evaluation of downstream tasks. Finally, we obtained 64,345 users from 493 instances containing 3,044,116 social links, with 725,282 posts and 765,019 images.

Data Annotation. To support downstream tasks such as social bot detection and user behavior modeling, we annotated user identity attributes (whether they are social bots or not). We first sampled users based on the social following graph with seed users and employed the Metropolis-Hastings Random Walk (MHRW) algorithm [25] to ensure an unbiased sampling. Consequently, we obtained a subgraph with 12,548 nodes and 1,048,576 edges. We manually annotated whether a user is a social bot by referring to discriminative criteria in existing studies [9], including features such as highly automated posting behavior, frequent retweeting of duplicate content, posting of tweets containing suspicious links, and lack of real personal information. Three researchers with relevant backgrounds independently labeled the data double-blindly to enhance annotation quality. Samples with annotation disagreements were discussed to achieve a consensus. Finally, 1,109 suspected social bot accounts were labeled, and annotation consistency was evaluated with a Cohen's kappa [8] value of 0.876.

3 Data Analysis in Practical Scenarios

In this section, we perform a comprehensive data analysis utilizing FediData for three practical scenarios.

3.1 Topic Extraction & Sentiment Analysis

Based on our dataset, we perform topic extraction and sentiment analysis using natural language processing (NLP). Specifically, we use `gpt-4.1-mini` to extract the topic and identify the sentiment tendency of each post. The sentiment of each tweet is classified according to Plutchik's three sentiment categories [19], including positive, neutral, and negative. We follow the topic list in [6, 19] to extract topics for all posts. Figure 2 shows an example of prompts and output to extract emotions and topics using ChatGPT. To explore content differences across different instances on a distributed social platform, we select two representative instances (*mastodon.social* and *linuxrocks.online*) and analyze their topic distributions. As shown in Figure 3, the *mastodon.social* instance exhibits a significant topic diversity, reflecting the wide range of interests and expressions among its largest user base. In contrast, more than 50% of the content in *linuxrocks.online* is related to technology, highlighting its strong professional orientation. This significant variation reveals the unique architectural advantages of Mastodon, with each instance showcasing distinct content and community cultures that provide users with diverse social spaces.

To further explore word frequency statistics and sentiment analysis across different instances, we conduct additional analysis based on the two selected instances, focusing on two topics: technology and gaming. In subfigures 4(a)-(b), the inconsistency in the prominent words within the word clouds of the two instances suggests variations in user focus and language style across different instances. Additionally, we observe significant differences in user text in different instances according to subfigures 4(c)-(d). For the topics of "technology" and "gaming", *mastodon.social* and *linuxrocks.online* exhibit relatively similar sentiment distributions, with a higher proportion of positive sentiment and lower proportions of neutral and negative sentiment. Notably, *linuxrocks.online* shows more significant differences in sentiment distribution, while those in *mastodon.social* are relatively smaller. This further indicates that the user base on *mastodon.social* is more diverse, accommodating a wider range of perspectives and emotions, while the *linuxrocks.online* instance might place greater emphasis on rational discussion and technical sharing, resulting in larger differences in the emotional distribution of technical topics.

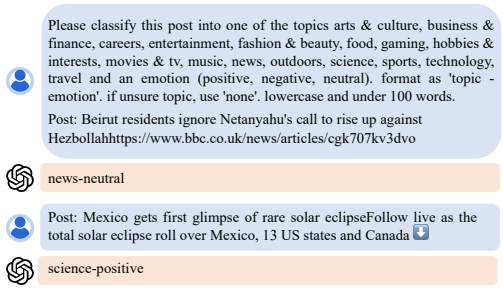


Figure 2: Two examples of our prompt design for extracting topics and emotions.

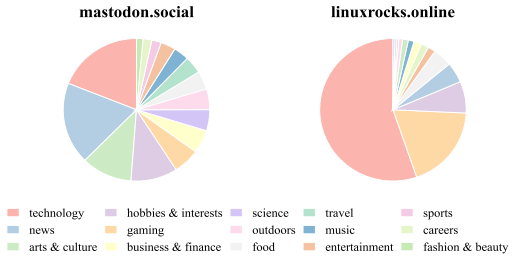


Figure 3: Topic distribution of two representative instances.

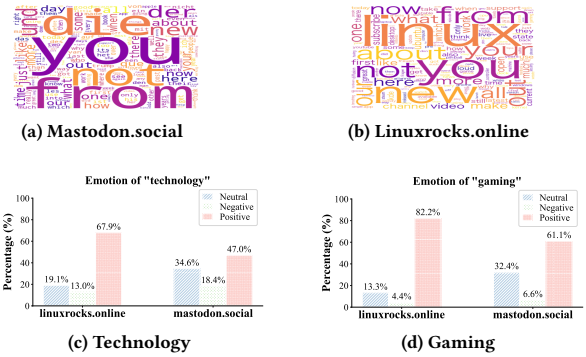


Figure 4: Word cloud visualization and emotion distribution.

3.2 Social Bot Detection

We construct a social bot detection task based on labeled user identity information to evaluate the practical application value of the dataset in a distributed environment. Specifically, to validate the

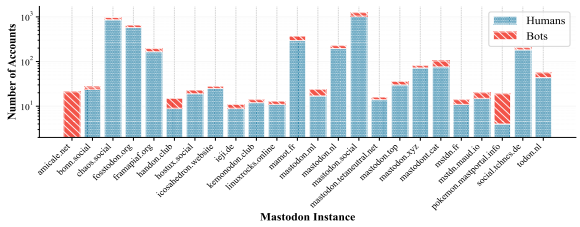


Figure 5: Distributions of the number of accounts (humans and bots) across different instances.

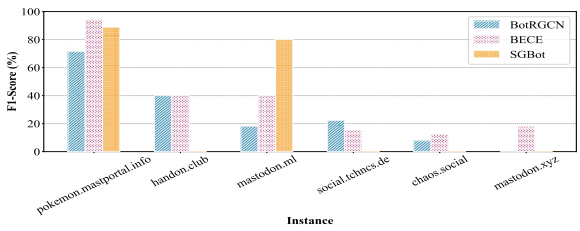


Figure 6: Performance (F1-Score) of representative methods for social bot detection on FediData.

effectiveness of existing representative social bot methods in our dataset, we select and implement several representative methods, such as BECE [23], SGBot [26], and BotRGCN [10]. We evaluate each method running in each instance, and then synthesize the results in these instances. Please note that existing methods do not consider the image modality of accounts. In this work, we additionally use the CLIP model [24] to extract embeddings from the image data posted by accounts and fuse them with other modalities by mapping them into the same embedding space for a fair comparison. For effective training, we select instances with more than 10 accounts (including humans and bots), and at least 10 social connections. The distributions of the total accounts and bots within each instance are shown in Figure 5. Then, we split the account data on each eligible instance into training/validation/testing parts in the ratio of 2:1:1 and finally evaluate the results on the test set on each instance as the basis for each representative detection method.

Based on the above setup, we test the performance of these models on selected instances. For clarity, we choose the top three instances (*pokemon.mastportal.info, handon.club*, and *mastodon.ml*) and the bottom three instances (*social.tchncs.de, chaos.social*, and *mastodon.xyz*) based on the number of bots for each instance. Then, we report the detection results of these representative methods in Figure 6. Surprisingly, these methods show low F1-Score values for several instances, and the performance varies across different instances, indicating that existing detection methods designed for centralized platforms do not always perform well for decentralized scenarios. This might be due to the heterogeneity of account data across different instances. This observation inspires future bot detection work to consider the characteristics of users on different instances for constructing effective detection methods.

3.3 Image Category Understanding

Given that the social media images in our dataset are richly diverse, we analyze these visual contents using large language models (LLMs) to identify potential semantic categories. Specifically, we use `qwen2.5-vl-32b-instruct` to clarify the images based on given categories. In line with [22], we consider 26 categories of

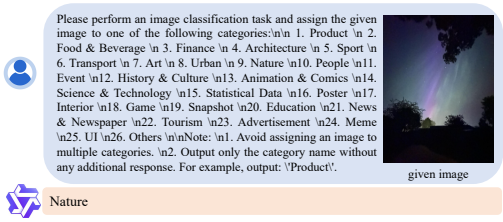
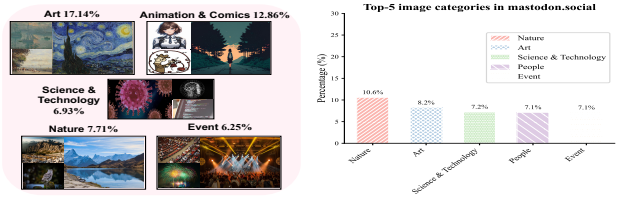


Figure 7: An example prompt for image classification.



(a) FediData **(b) Mastodon.social**

Figure 8: The top-5 image categories within FediData and that of *mastodon.social* instance, respectively.

images in our analysis. Figure 7 shows an example prompt with an image and the corresponding output. Figure 8(a) displays the top-5 image categories in our dataset, including art, animation, nature, science, and event. We also analyze the top-5 image categories in *mastodon.social*, as depicted in Figure 8(b). These findings provide a foundation for further exploration of topic distribution and content preferences across various instances.

4 Conclusion and Future Work

In this work, we constructed and open-sourced the FediData dataset. To our knowledge, it is the first public and comprehensive multi-modal dataset from Mastodon. Our comprehensive dataset contains users' metadata, posts that contain text and images, and user social relationships. We have aligned the text and image content and provided a high-quality annotation of user attributes, indicating whether they are social bots or not. Moreover, we have designed several tasks to evaluate the practical applications of our dataset for studies in sentiment analysis, topic extraction, social bot detection, and image category understanding. We hope that FediData will become a fundamental resource for a wide range of decentralized web-related research.

We envision several possible directions for future work. Firstly, researchers could conduct studies on FediData, such as social network analysis, UGC content understanding, temporal pattern analysis, and uncover differences between DOSNs and centralized platforms. Secondly, we plan to delve further into the potential biases of the emotion analysis outcomes produced by LLMs. Finally, researchers could identify potential content moderation problems [3, 12] based on our dataset and cooperate with instance administrators to solve these problems to manage their instances for a safe and inclusive social environment.

Acknowledgments

This work is sponsored by National Natural Science Foundation of China (No. 62072115, No. 62472101), Shanghai Science and Technology Innovation Action Plan Project (No. 22510713600).

Usage of Generative AI

In this paper, we employ AI tools to enhance the robustness and efficiency of the crawling code while utilizing them to refine and clarify the text, maintaining the original meaning. All authors have reviewed and confirmed the final text.

References

[1] Roel Roscam Abbing, Cade Diehm, and Warreth Shahed. 2023. Decentralised social media. *Internet Policy Review* 12, 1 (2023).

[2] Roel Roscam Abbing and Robert W Gehl. 2024. Shifting your research from X to Mastodon? Here's what you need to know. *Patterns* 5, 1 (2024).

[3] Ishaku Hassan Anaobi, Aravindh Raman, Ignacio Castro, Haris Bin Zia, Damilola Ibosiola, and Gareth Tyson. 2023. Will Admins Cope? Decentralized Moderation in the Fediverse. In *Proceedings of the ACM Web Conference 2023*. 3109–3120.

[4] Haris Bin Zia, Aravindh Raman, Ignacio Castro, Ishaku Hassan Anaobi, Emiliano De Cristofaro, Nishanth Sastry, and Gareth Tyson. 2022. Toxicity in the decentralized web and the potential for model sharing. *Proceedings of the ACM on Measurement and Analysis of Computing Systems* 6, 2 (2022), 1–25.

[5] Björn Brembs, Adrian Lenardic, and Leslie Chan. 2023. Mastodon: a move to publicly owned scholarly knowledge. *Nature* 614, 7949 (2023), 624–624.

[6] Ying-Ying Chang, Wei-Yao Wang, and Wen-Chih Peng. 2024. SeGA: preference-aware self-contrastive learning with prompts for anomalous user detection on Twitter. In *Proceedings of AAAI*, Vol. 38. 30–37.

[7] Lianhua Chi and Xingquan Zhu. 2017. Hashing techniques: A survey and taxonomy. *Comput. Surveys* 50, 1 (2017), 1–36.

[8] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20, 1 (1960), 37–46.

[9] Shangbin Feng, Zhaoxuan Tan, Herun Wan, Ningnan Wang, Zilong Chen, Binchi Zhang, Qinghua Zheng, Wenqian Zhang, Zhenyu Lei, Shujie Yang, et al. 2022. Twibot-22: Towards graph-based Twitter bot detection. *Advances in Neural Information Processing Systems* 35 (2022), 35254–35269.

[10] Shangbin Feng, Herun Wan, Ningnan Wang, and Minnan Luo. 2021. BotRGCN: Twitter bot detection with relational graph convolutional networks. In *Proceedings of ASONAM*. 236–239.

[11] Min Gao, Qiang Duan, Boen Liu, Yu Xiao, Xin Wang, and Yang Chen. 2025. Higher-Order Information Matters: A Representation Learning Approach for Social Bot Detection. In *Proceedings of ACM CIKM*.

[12] Anaobi Ishaku Hassan, Aravindh Raman, Ignacio Castro, Haris Bin Zia, Emiliano De Cristofaro, Nishanth Sastry, and Gareth Tyson. 2021. Exploring Content Moderation in the Decentralised Web: The Pleroma Case. In *Proceedings of the 17th International Conference on Emerging Networking EXperiments and Technologies*. 328–335.

[13] Jiahui He, Haris Bin Zia, Ignacio Castro, Aravindh Raman, Nishanth Sastry, and Gareth Tyson. 2023. Flocking to Mastodon: Tracking the Great Twitter Migration. In *Proceedings of the 2023 ACM on Internet Measurement Conference*. 111–123.

[14] Ujun Jeong, Alimohammad Beigi, Anique Tahir, Susan Xu Tang, H Russell Bernard, and Huan Liu. 2025. FediverseSharing: A Novel Dataset on Cross-Platform Interaction Dynamics between Threads and Mastodon Users. In *Proceedings of ASONAM*.

[15] Ujun Jeong, Paras Sheth, Anique Tahir, Faisal Alatawi, H Russell Bernard, and Huan Liu. 2024. Exploring platform migration patterns between Twitter and Mastodon: A user behavior study. In *Proceedings of ICWSM*, Vol. 18. 738–750.

[16] Seeun Kim, Li Zeng, Sijia Ma, and Giulia Sturlese. 2025. Sentiment Dynamics and Shifts across Instances on Mastodon. In *Proceedings of ICWSM Workshop*.

[17] Kai Kupferschmidt. 2022. As Musk reshapes Twitter, academics ponder taking flight. *Science* 378, 6620 (2022), 583–584.

[18] Lucio La Cava, Sergio Greco, and Andrea Tagarelli. 2021. Understanding the growth of the Fediverse through the lens of Mastodon. *Applied Network Science* 6, 64 (2021).

[19] Wei Li, Jiawen Deng, Jiali You, Yuanyuan He, Yan Zhuang, and Fuji Ren. 2025. ETS-MM: A Multi-Modal Social Bot Detection Model Based on Enhanced Textual Semantic Representation. In *Proceedings of the ACM on Web Conference 2025*. 4160–4170.

[20] Shaojie Min, Shaobin Wang, Yaxiao Luo, Min Gao, Qingyuan Gong, Yu Xiao, and Yang Chen. 2025. FediLive: A Framework for Collecting and Preprocessing Snapshots of Decentralized Online Social Networks. In *Companion Proceedings of the ACM on Web Conference 2025*. 765–768.

[21] Matthew N Nicholson, Brian C Keegan, and Casey Fiesler. 2023. Mastodon rules: characterizing formal rules on popular Mastodon instances. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing*. 86–90.

[22] Yusu Qian, Hanrong Ye, Jean-Philippe Fauconnier, Peter Grasch, Yinfei Yang, and Zhe Gan. 2025. MIA-Bench: Towards better instruction following evaluation of multimodal LLMs. In *Proceedings of ICLR*.

[23] Boyu Qiao, Wei Zhou, Kun Li, Shilong Li, and Songlin Hu. 2024. Dispelling the fake: Social bot detection based on edge confidence evaluation. *IEEE Transactions on Neural Networks and Learning Systems* (2024), 7302–7315.

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of ICML*. 8748–8763.

[25] Tianyi Wang, Yang Chen, Zengbin Zhang, Tianyin Xu, Long Jin, Pan Hui, Beixing Deng, and Xing Li. 2011. Understanding graph sampling algorithms for social network analysis. In *Proceedings of the 31st International Conference on Distributed Computing Systems Workshops*. 123–128.

[26] Kai-Cheng Yang, Onur Varol, Pik-Mai Hui, and Filippo Menczer. 2020. Scalable and generalizable social bot detection through data selection. In *Proceedings of AAAI*, Vol. 34. 1096–1103.

[27] Haris Bin Zia, Ignacio Castro, and Gareth Tyson. 2024. Mastodoner: A Command-line Tool and Python Library for Public Data Collection from Mastodon. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 5314–5317.

[28] Matteo Zignani, Sabrina Gaito, and Gian Paolo Rossi. 2018. Follow the "Mastodon": Structure and Evolution of a Decentralized Online Social Network. In *Proceedings of the Twelfth International Conference on Web and Social Media*. 541–551.