

Analyzing Mastodon Social Media Data for Patterns about AI Technology Conversations

Siddharth Axaykumar Mehta
CSE 472: Social Media Mining
Arizona State University
Email: smehta74@asu.edu

Abstract—This report presents the results of Project I for the course CSE472: Social Media Mining, under the guidance of Prof. Huan Liu. The primary objective of this project is to collect, analyze, and visualize Mastodon discussions related to artificial intelligence (AI) and related technologies. To accomplish this, two datasets—posts and users—were constructed, and corresponding networks representing information diffusion and friendship were developed. Additionally, network visualizations were generated to better understand the patterns and dynamics of these discussions.

I. INTRODUCTION

On social media, there is a lot of discussion about artificial intelligence and technology, with a range of viewpoints from optimism about innovation to concerns about ethics and job displacement. Mastodon was chosen as the data source for this project because of its open and user-friendly API. The study advances our knowledge on user interaction patterns, the concepts that predominate in online discourse, and how communities discuss AI.

II. METHODOLOGY

A. Step 1: API Credentials

We created a Mastodon application then from Mastodon website obtained API keys. These credentials were hard-coded into a simple test script using the `Mastodon.py` library to verify that the API connection worked correctly.

B. Step 2: Data Collection

Two datasets were collected:

- **Keyword-based posts:** At least 600 posts were crawled using hashtags such as #ai, #machinelearning, #ChatGPT, #deeplearning etc.
- **User-based network:** A minimum of 180 users were gathered starting from seed accounts (e.g., @machinelearning, @techcrunch), expanding through follower/followee lists.

Data was stored in JSON format. Python scripts automated the crawling process, ensuring deduplication for posts.

C. Step 3: Network Construction and Visualization

Two networks were constructed:

- **Information Diffusion Network:** Nodes represent posts; edges represent replies and boosts.

- **Friendship Network:** Nodes represent users; edges represent mutual follower/followee relationships.

Graphs were built in `networkx` and exported to Gephi (`.gexf` format). Gephi was used to visualize communities.

D. Step 4: Network Measures

For the Friendship Network:

- **Average Path Length:** Average distance from every to node all other nodes.
- **Clustering Coefficient:** Average value computed to evaluate local cohesion.
- **PageRank:** Top users ranked to identify influential accounts.

E. Step 5: Content Analysis

We applied LLM-based keyword extraction using Deepseek R1 with the OpenRouter API. For each post, three concise keywords were generated. A word cloud visualization was created from keyword frequencies to highlight dominant AI themes.

III. RESULTS

A. Data Collection

- Posts collected: **608**
- Users collected: **200**

B. Network Visualizations

Figures 1 and 2 present the two networks constructed from Mastodon data. The **information diffusion network** (Figure 1) captures how posts propagate through replies and boosts. The **friendship network** (Figure 2) illustrates user connections based on follow relationships. In this visualization, red edges indicate one-way relations (follower or followee), while green edges represent mutual relations (both users follow each other). Node colors in the diffusion network indicate the number of replies associated with each post.

C. Network Measures

- Degree distribution histogram: See Figure 3.
- Average clustering coefficient: **0**.
- Average Path Length: **1.97**.
- Top 5 PageRank (usernames): @Techcrunch, @machinelearning, @TedUnderwood, @lea_bei, @nic221.

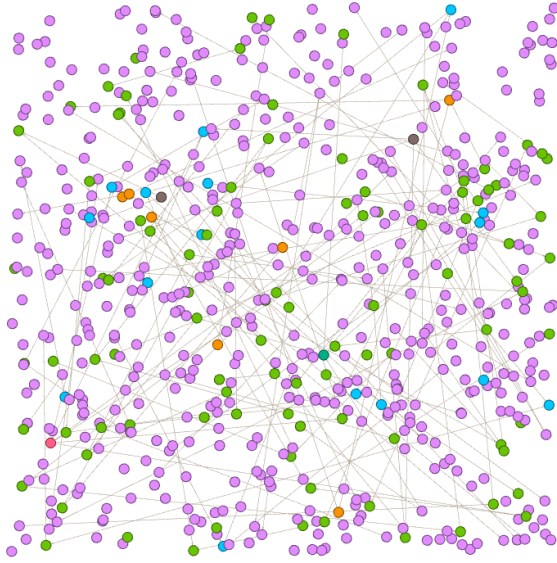


Fig. 1. Information diffusion network (Gephi). Node colors correspond to the number of replies a post received.

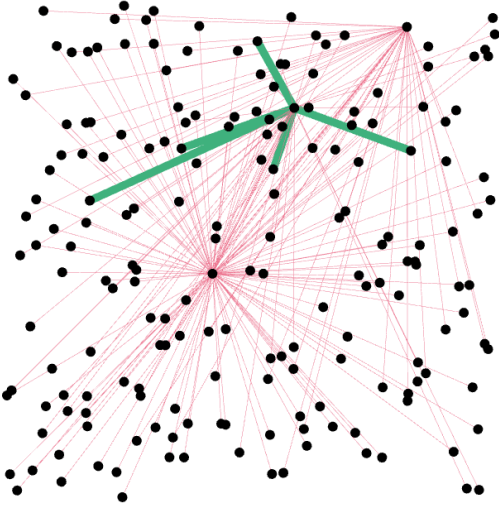


Fig. 2. Friendship network (Gephi). Red: one-way relation
Green: mutual relation.

D. Content Analysis

We applied LLM-based keyword extraction using Deepseek R1 with the OpenRouter API. For each post, three concise keywords were generated to capture the central theme. Example extracted keywords include ["technology", "artificialintelligence", "llm"]. A word cloud visualization (Figure 5) was then constructed to highlight dominant topics. The most frequently occurring terms were **ChatGPT**, **machine learning**, **ethics**, and **automation**, consistent with current AI discourse.

To better contextualize the content dynamics, we also examined properties of the **friendship network**. The average

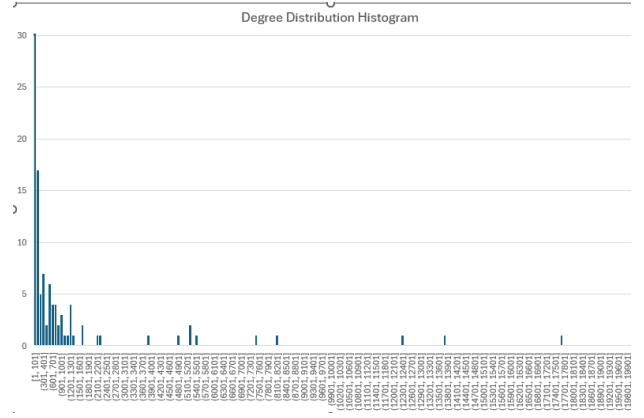


Fig. 3. Degree distribution of the friendship network.

clustering coefficient was found to be **0**, suggesting that users in this dataset rarely form tightly-knit groups or triadic closures. However, the average path length was relatively short at **1.97**, indicating that information can spread efficiently across the network despite limited clustering. Furthermore, PageRank analysis revealed the top five most influential accounts as **@Techcrunch**, **@machinelearning**, **@TedUnderwood**, **@lea_bei**, and **@nic221**, all of which act as central hubs in the AI discourse. These network-level results complement the content analysis by showing how highly visible users shape conversations around a concentrated set of recurring themes.

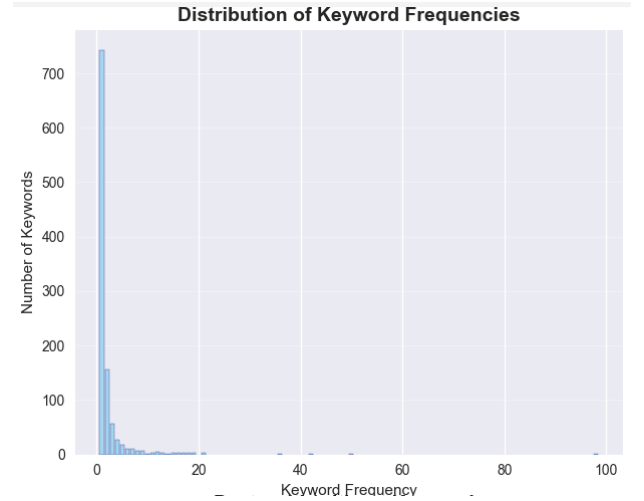


Fig. 4. Keyword frequency distribution, exhibiting a power law pattern.

Interestingly, the keyword frequency distribution itself follows a power law pattern, similar to the degree distribution observed in the network analysis. This suggests that a small subset of terms dominates the discussion, while a long tail of less common but diverse keywords captures the breadth of perspectives in Mastodon conversations on AI. Figure 4 illustrates this phenomenon, where a handful of highly frequent terms sharply contrast with many infrequent ones.



Fig. 5. Word cloud of keywords extracted from posts.

IV. DISCUSSION

Our analysis suggests that Mastodon conversations on AI are diverse, with recurring themes of Technology, LLM, and AI. Network analysis revealed that a small subset of users (high PageRank/Betweenness) act as central hubs for disseminating AI content. The diffusion network indicates active reposting of ChatGPT-related content, while the friendship network exhibits moderate clustering, suggesting topical communities.

V. CONCLUSION

This project demonstrated the application of social media mining techniques to Mastodon data in order to uncover patterns of conversation around artificial intelligence and related technologies. Through the collection of over 600 posts and 200 users, we constructed both information diffusion and friendship networks, which allowed us to analyze how AI-related content spreads and which actors play influential roles in the discourse.

The network analysis revealed several key findings. The average path length of 1.97 indicates that information can traverse the network efficiently, despite the absence of strong clustering (average clustering coefficient = 0). This suggests that Mastodon’s decentralized structure still enables rapid dissemination of AI-related discussions across loosely connected communities. PageRank results further identified a small set of highly influential accounts, such as @Techcrunch and @machinelearning, which act as central hubs in amplifying AI-related content. These results highlight how a limited number of users can disproportionately shape the conversation space.

From the content analysis, keyword extraction and word cloud visualizations revealed that terms like **ChatGPT**, **machine learning**, **ethics**, and **automation** dominate discussions. The keyword frequency distribution exhibited a power law pattern, mirroring the degree distribution observed in the network. This reinforces the insight that a small subset of highly recurrent topics captures most of the attention, while a long tail of less frequent keywords reflects the diversity of perspectives on AI. Together, these results illustrate both the concentrated and distributed nature of online AI discourse.

Overall, this project not only provided hands-on experience in API-based data collection, network construction, and LLM-based text analysis, but also generated meaningful insights into how online communities engage with emerging AI technologies. The combination of structural network properties and thematic content analysis paints a holistic picture: Mastodon serves as a platform where a few influential hubs shape the global flow of AI discourse, while broader participation ensures that ethical concerns, technological optimism, and critical debates remain visible. Future work could expand the dataset, incorporate temporal dynamics, and compare Mastodon conversations with those on other social media platforms to better understand cross-platform differences in AI discourse.

ACKNOWLEDGMENTS

I thank Prof. Huan Liu and TA Alimohammad Beigi for their guidance in CSE 472.

REFERENCES

- 1) Zafarani, R., Abbasi, M. A., & Liu, H. (2014). *Social Media Mining: An Introduction*. Cambridge University Press. <https://socialmediamining.info/>
- 2) Mastodon API documentation: <https://docs.joinmastodon.org/methods/>
- 3) Mastodon.py documentation: <https://mastodonpy.readthedocs.io/>
- 4) Gephi project: <https://gephi.org/>
- 5) NetworkX documentation: <https://networkx.org/documentation/stable/>
- 6) Matplotlib: <https://matplotlib.org/>
- 7) WordCloud for Python: https://amueller.github.io/word_cloud/
- 8) OpenRouter API docs: <https://openrouter.ai/docs>
- 9) xAI Grok 4 Fast (OpenRouter model): <https://openrouter.ai/models/x-ai/grok-4-fast>
- 10) DeepSeek R1 (OpenRouter model): <https://openrouter.ai/models/deepseek/deepseek-r1>

TOOLS AND ASSISTANCE DISCLOSURE

Portions of the literature write-up and phrasing were drafted with ChatGPT and then edited by the authors. Cursor AI was used during debugging and for code formatting/refactoring in the Python scripts. All code was executed by the authors, model outputs were verified against the dataset, and no generative tool was used to fabricate data or results.