12^{th} International Conference on Computational Social Science IC^2S^2
July 28-31, 2026, Burlington, VT, United States

SocialPulse: An Open-Source Subreddit Sensemaking Toolkit

Keywords: Social media analysis; online communities; Reddit; sensemaking; exploratory data analysis

Extended Abstract

Understanding how online communities discuss and make sense of complex social issues is a central challenge in social media research. Despite the scale of user-generated content online, prior work shows that analyses of this content often overrepresent highly active users, obscuring the perspectives of long-tail contributors [1]. Tools that support analysis across participation levels are therefore critical for accurately characterizing online discourse. At the same time, there is growing demand for fine-grained analysis of online discourse across domains, including public opinion monitoring, policy-making, and crisis response [2], [3]. Researchers and practitioners increasingly seek to identify latent themes, shifts in discussion, and patterns of participation in online conversations [4]. Exploratory data analysis (EDA) tools that support rapid *sensemaking* are particularly valuable in these settings:

> *"Sensemaking is a motivated, continuous effort to understand connections (which can be among people, places, and events) in order to anticipate their trajectories and act effectively."* - [5]

Reddit is a widely studied platform for social science research, offering large-scale, community-structured discussions across diverse topics [6], [7], [8], [9], [10]. Topic modeling has been central to analyzing such discourse, enabling the identification of latent themes and their temporal dynamics, with applications to gender norms [11], sociopolitical conflict [12], and more. These works highlight that in online discourse, meaning emerges not only from topical content but also from participation structures, community norms, and interaction patterns. However, practical analysis remains challenging due to social bots, spam, and fragmented analytical systems [13], [14], motivating the need for open-source, adaptable sensemaking tools that integrate topical, user-, and community-level perspectives.

To address these challenges, we introduce **SocialPulse, an open-source subreddit sensemaking toolkit for interactive, exploratory analysis of Reddit discussions**. SocialPulse supports exploratory sensemaking within Reddit communities through an integrated analysis pipeline (Figure 1) and an interactive dashboard for exploration and comparison (Figure 2). The architecture comprises four sequential stages – data ingestion, optional bot filtration, analytical modeling, and interactive visualization – which together enable researchers to move fluidly between aggregate patterns and individual content while accounting for participation dynamics and data quality concerns. **We open-source the SocialPulse system to support large-scale social media sensemaking research** at `https://anonymous.4open.science/r/SocialPulse-7D28`.

The system ingests Reddit data based on user-specified subreddits and configurable collection strategies, supporting both high-traffic and long-tail communities. An optional bot filtration stage with the BotBuster tool [13] allows users to identify and filter automated accounts, enabling analysis of human-driven discourse when desired. SocialPulse then applies a suite

of complementary analytical methods to extract thematic, temporal, and behavioral patterns in Reddit discussions. Topic modeling is performed using BERTopic [15], which leverages transformer-based embeddings and clustering to identify latent themes in text, while VADER sentiment analysis [16] is used to characterize the polarity of posts. The system further allows cross-community comparison, enabling the analysis of shared themes, overlapping users, and duplicated content. Analysis results are presented via an interactive dashboard, which enables users to compare discourse across communities, examine topic structure and sentiment over time, and move from aggregate summaries to individual posts and comments.

Online discussions of conspiracy theories present a challenging setting for sensemaking, as they often involve a heterogeneous mix of genuine belief, skepticism, and play [17]. We demonstrate how SocialPulse supports sensemaking within and across `r/conspiracy`. Over the week of January 4, 2026, BERTopic identifies 20 distinct topics with a high degree of thematic diversity as seen in **Figure 2a**: the topic labels range from long-standing conspiracy topics (e.g., aliens, intelligence agencies, and consciousness) to discussions about current events. Topic frequencies are relatively evenly distributed, and most topics are well-separated with little overlap, suggesting clear thematic boundaries. **Figure 2b** shows posting activity peaks on Saturdays and is lowest on Mondays, with the highest volume occurring from 15:00-16:00 UTC. Sentiment polarity across collected posts is slightly negative (421 negative posts, 140 neutral posts, 336 positive posts). Finally, user activity over the week is dominated by long-tail participation, where most users post less than once a week. As shown in **Figure 2c**, specifically examining Topic 0 reveals that topic-level sentiment polarity is largely neutral (score = 0.47) within this topic. Examining individual topics (**not shown**), we see that topics centered on current events exhibit a more negative sentiment, while topics associated with long-standing conspiracy theories such as consciousness, religion and cults, and historical theories tend to be more neutral or positive. Extending this analysis to a *cross-subreddit comparison* with `r/politics` in **Figure 2d** reveals measurable overlap in both users and content, including two shared users, two identical external links posted in both subreddits, and multiple duplicated posts; all of the duplicated posts discussed current events. Our analysis yields three exploratory insights:

Exploratory Insight 1: Topic-Dependent Sentiment Patterns Topic-level analysis reveals systematic differences in sentiment between discussion types within `r/conspiracy`. Topics centered on current events tend to exhibit more negative sentiment, while long-standing conspiracies remain relatively stable and are often more neutral or positive in tone.

Exploratory Insight 2: Temporal Variation in Community Engagement Posting activity within `r/conspiracy` is unevenly distributed across the week, with high activity levels on Saturday and lower activity levels on Monday. The shift in activity suggests that engagement with conspiracy-related discourse fluctuates over time. This fluctuation highlights the importance of temporal context when interpreting engagement, as changes may reflect shifts in attention or availability rather than stable interest.

Exploratory Insight 3: Cross-Community Content Duplication A cross-subreddit comparison between `r/conspiracy` and `r/politics` reveals instances of content duplication. The identical content appears in both communities within a short time window and corresponds to current events, indicating that politically relevant posts circulate rapidly across otherwise distinct subreddits.

SocialPulse provides an interactive, open-source pipeline for exploratory analysis of Reddit communities, enabling rapid sensemaking through integrated modeling, filtering, and visualization. The system empowers researchers to conduct rapid, exploratory sensemaking of online community discourse by providing an integrated pipeline that unifies data collection, analysis, and interactive visualization.

References

[1] L. Oswald, W. Schulz, R. Hertwig, D. Lazer, and S. Stier, "The tip of the iceberg: How the social media production–consumption gap distorts public opinion for citizens and researchers," *SocArXiv*, 2025, Preprint.

[2] Y. Zhu, E.-U. Haq, G. Tyson, et al., "A study of partisan news sharing in the russian invasion of ukraine," in *ICWSM*, vol. 18, 2024, pp. 1847–1858.

[3] K. Yin et al., "DisastIR: A comprehensive information retrieval benchmark for disaster management," in *Findings of the Association for Computational Linguistics: EMNLP 2025*, C. Christodoulopoulos, T. Chakraborty, C. Rose, and V. Peng, Eds., Suzhou, China: Association for Computational Linguistics, Nov. 2025, pp. 1836–1867. DOI: 10.18653/v1/2025.findings-emnlp.97.

[4] T. Islam and D. Goldwasser, "Discovering latent themes in social media messaging: A machine-in-the-loop approach integrating llms," in *ICWSM*, vol. 19, 2025, pp. 859–884.

[5] G. Klein, B. Moon, and R. R. Hoffman, "Making sense of sensemaking 1: Alternative perspectives," *IEEE intelligent systems*, vol. 21, no. 4, pp. 70–73, 2006.

[6] M. De Choudhury and S. De, "Mental health discourse on reddit: Self-disclosure, social support, and anonymity," in *ICWSM*, vol. 8, 2014, pp. 71–80.

[7] S. Kumar, W. L. Hamilton, J. Leskovec, and D. Jurafsky, "Community interaction and conflict on the web," in *The Web Conference*, 2018, pp. 933–943.

[8] X. Dong, C. Li, and J. D. Choi, "Transformer-based context-aware sarcasm detection in conversation threads from social media," in *Proceedings of the Second Workshop on Figurative Language Processing*, B. B. Klebanov et al., Eds., Online: Association for Computational Linguistics, Jul. 2020, pp. 276–280. DOI: 10.18653/v1/2020.figlang-1.38.

[9] Y. Guo, X. Dong, M. A. Al-Garadi, A. Sarker, C. Paris, and D. M. Aliod, "Benchmarking of transformer-based pre-trained models on social media text classification datasets," in *Proceedings of the 18th Annual Workshop of the Australasian Language Technology Association*, M. Kim, D. Beck, and M. Mistica, Eds., Virtual Workshop: Australasian Language Technology Association, Dec. 2020, pp. 86–91.

[10] C. Davidson, *Use of Reddit for Social Science Research: A Review of Current Use, Exploration of Potential Sampling Error, and Practical Demonstration Using Reddit to Study Post-Pandemic Teacher Resignation*. Western Michigan University, 2023.

[11] M. Teleki, X. Dong, H. Liu, and J. Caverlee, "Masculine defaults via gendered discourse in podcasts and large language models," in *ICWSM*, vol. 19, 2025, pp. 1893–1912.

[12] E. Steffen, "More than memes: A multimodal topic modeling approach to conspiracy theories on telegram," in *ICWSM*, vol. 19, 2025, pp. 1831–1844.

[13] L. H. X. Ng and K. M. Carley, "Botbuster: Multi-platform bot detection using a mixture of experts," in *ICWSM*, vol. 17, 2023, pp. 686–697.

[14] M. Mendoza, E. Providel, M. Santos, and S. Valenzuela, "Detection and impact estimation of social bots in the chilean twitter network," *Scientific reports*, vol. 14, no. 1, p. 6525, 2024.

[15] M. Grootendorst, "Bertopic: Neural topic modeling with a class-based tf-idf procedure," *arXiv preprint arXiv:2203.05794*, 2022.

[16] C. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *ICWSM*, vol. 8, 2014, pp. 216–225.

[17] M. Samory and T. Mitra, "Conspiracies online: User discussions in a conspiracy community following dramatic events," in *ICWSM*, vol. 12, 2018.

Figures

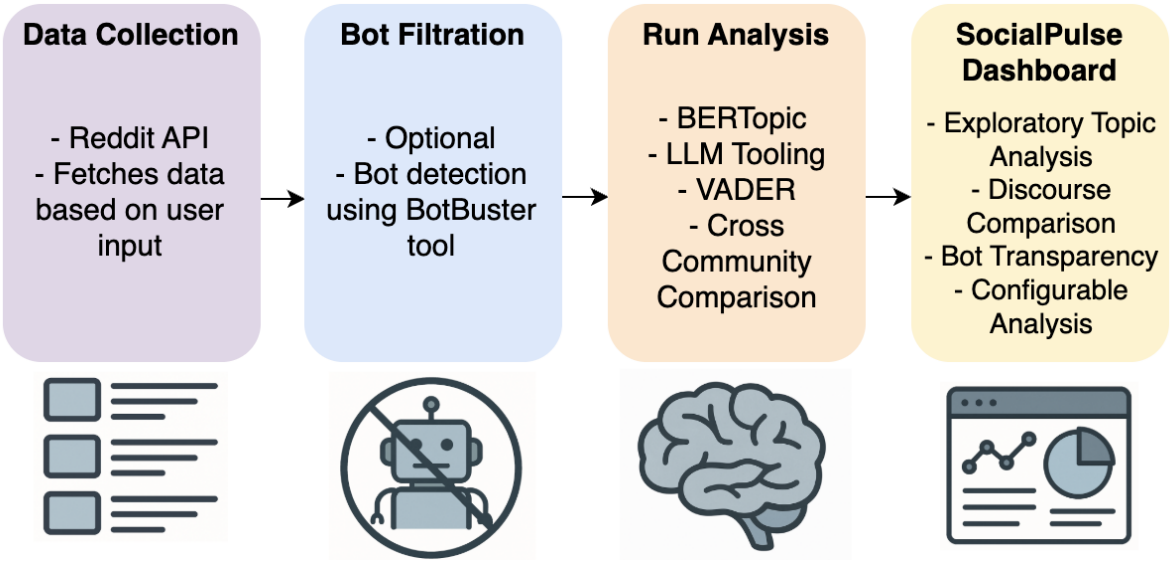

Figure 1: **The SocialPulse pipeline supports rapid exploratory data analysis and sense-making of Reddit communities**; this is done in four stages: data ingestion, optional bot filtration, analytical modeling, and an interactive visualization interface.

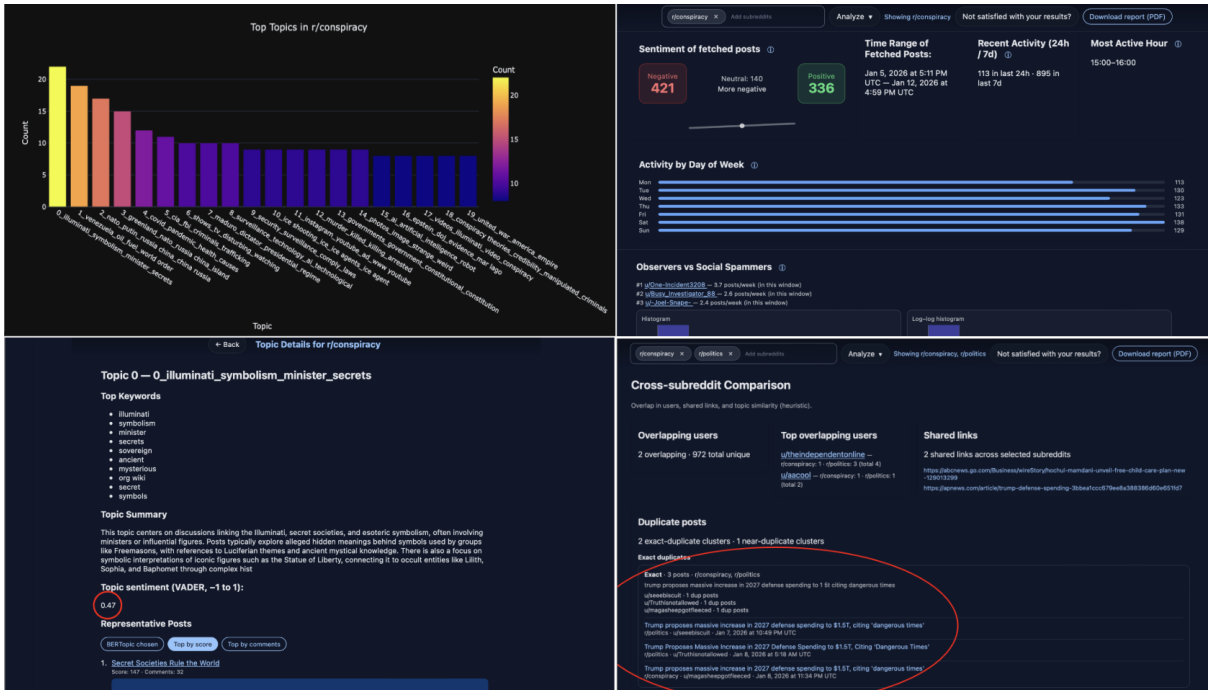

Figure 2: **The SocialPulse analytics interface supports rapid, multi-level sensemaking of Reddit discourse.** The interface enables (**a**, *top left*) interactive topic exploration, (**b**, *top right*) temporal and sentiment analysis, (**c**, *bottom left*) topic-specific analysis, and (**d**, *bottom right*) cross-subreddit comparison, illustrated here for `r/conspiracy` and `r/politics`.