SocialPulse: An Open-Source Subreddit Sensemaking Toolkit

Abstract

Understanding how online communities discuss and make sense of complex social issues is a central challenge in social media research, yet existing tools for large-scale discourse analysis are often closed-source, difficult to adapt, or limited to single analytical views. We present SocialPulse, an open-source subreddit sensemaking toolkit that unifies multiple complementary analyses – topic modeling, sentiment analysis, user activity characterization, and bot detection – within a single interactive system. SocialPulse enables users to fluidly move between aggregate trends and fine-grained content, compare highly active and long-tail contributors, and examine temporal shifts in discourse across subreddits. This work presents end-to-end exploratory workflows that allow researchers and practitioners to rapidly surface themes, participation patterns, and emerging dynamics in large Reddit datasets. A case study using SocialPulse highlights how exploratory analysis of the `r/conspiracy` subreddit can reveal findings on topic-dependent sentiment patterns, temporal variation in community engagement, and cross-community content duplication with other subreddits such as `r/politics`. By offering an extensible and openly available platform, SocialPulse provides a practical and reusable foundation for transparent, reproducible sensemaking of online community discourse.

Code —
https://anonymous.4open.science/r/SocialPulse-7D28

Introduction

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 over-represent highly active users, obscuring the perspectives of long-tail contributors (Oswald et al. 2025). Tools that support analysis across participation levels are therefore critical for accurate characterization of online discourse. At the same time, there is growing demand for fine-grained analysis of web discourse across domains, including public opinion monitoring, policy-making, and crisis response (Zhu et al. 2024; Yin et al. 2025). Researchers and practitioners increasingly seek to identify latent themes, shifts in discussion, and patterns of participation in online conversations (Islam and Goldwasser 2025). 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."* - Klein, Moon, and Hoffman (2006)

Reddit is a widely studied platform for social science research, offering large-scale, community-structured discussions across diverse topics (De Choudhury and De 2014; Kumar et al. 2018; Dong, Li, and Choi 2020; Guo et al. 2020; Davidson 2023). Topic modeling has been central to analyzing such discourse, enabling the identification of latent themes and their temporal dynamics, with applications to gender norms (Teleki et al. 2025), sociopolitical conflict (Steffen 2025), 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 (Ng and Carley 2023; Mendoza et al. 2024), motivating the need for open-source, adaptable sensemaking tools that integrate topical, user-, and community-level perspectives (Table 2).

To address these challenges, we present **SocialPulse**, an open-source subreddit sensemaking toolkit for interactive,

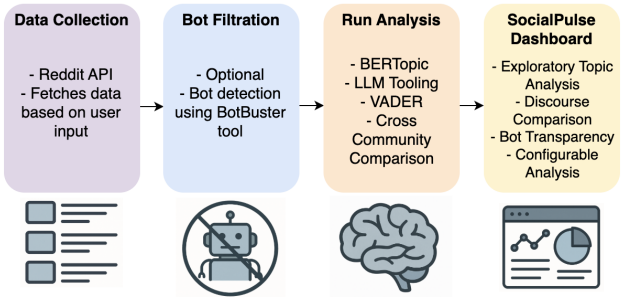

Figure 1: **The SocialPulse pipeline supports rapid exploratory data analysis and sensemaking of Reddit communities**; implementation details are provided in Table 1.

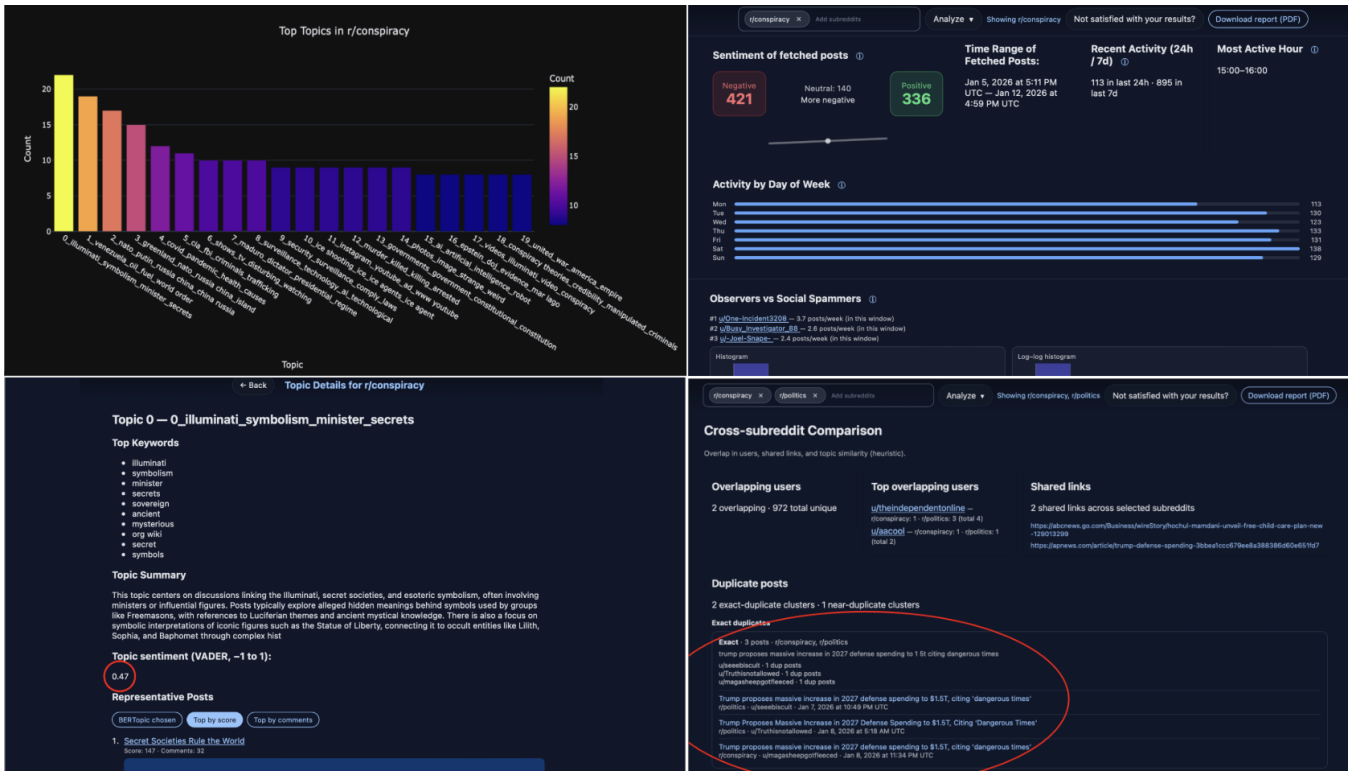

Figure 2: **The SocialPulse analytics interface supports rapid, multi-level sensemaking of Reddit discourse.** The interface enables (**a**, *top left*) interactive topic exploration, (**b**, *top right*) temporal and sentiment analysis, (**c**, *bottom left*) topic-specific analysis, and (**d**, *bottom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

exploratory analysis of Reddit discussions. We provide (i) an **overview of the analysis pipeline** (Figure 1), (ii) **implementation details** (Table 1), and (iii) a **case study** yielding three findings.

SocialPulse: System Architecture and Implementation

SocialPulse supports exploratory sensemaking of Reddit communities through an integrated analysis pipeline and an interactive dashboard for discovery and comparison (Figures 1-2; Table 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 fine-grained content while accounting for participation dynamics and data quality concerns. Figure 1 provides an overview of the pipeline, Figure 2 illustrates the interactive dashboard interface, and Table 1 lists the detailed resource links associated with each stage.

(i) Data Collection and Statistics

SocialPulse leverages the Reddit API to fetch public data based on user input. Users may analyze one or multiple subreddits and configure data collection by either specifying a fixed number of posts retrieved via the Reddit API using the *Best, Hot* and *Recent* sorting categories or by constraining the data to a particular time interval. This enables a flexible exploration of both "high-traffic" and "long-tail" community dynamics. Consider `r/AskHistorians`, for example. Although posts to this subreddit often receive only a few responses, the responses are more in-depth and higher-effort compared with other subreddits. As is typical in niche communities, interactions between community members may be less frequent, and so a longer time interval may be needed to capture sufficient network activity. Tailoring the data collection method towards the community of interest based on group sizes and time intervals is important (Panek et al. 2018).

(ii) Bot Filtration

To ensure the integrity of social media analysis, the system includes an optional bot detection stage. We utilize the BotBuster tool (Ng and Carley 2023), a mixture-of-experts neural network architecture designed for multi-platform bot detection. BotBuster analyzes information pillars such as user metadata and posts to estimate the probability of an account being a bot. This allows researchers to isolate human-driven discourse from automated amplification.

(iii) Analytical Engine

Once the data is cleaned, SocialPulse runs a suite of complementary analyses to extract thematic and behavioral patterns:

Table 1: **Implementation details for SocialPulse processing pipeline shown in Figure 1,** detailing the tools and stages used for collecting Reddit data, filtering bot activity, extracting topics and sentiment, and presenting results via an interactive dashboard.

Tool	Purpose	Link
(i) Input & Data Sources		
Reddit API	Used to collect publicly available posts, comments, and metadata from user-specified subreddits for analysis.	https://www.reddit.com/dev/api/
(ii) Bot Filtration		
BotBuster (Ng and Carley 2023)	A neural network-based framework that leverages user metadata and posting behavior to estimate bot likelihood in social media accounts.	https://github.com/quarbby/BotBuster-Universe
(iii) Run Analysis		
BERTopic (Grootendorst 2022)	A transformer-based framework that utilizes Sentence-BERT embeddings and clustering (HDBSCAN) to discover latent themes in unstructured Reddit text.	https://maartengr.github.io/BERTopic/index.html
LLM Tooling	LLMs are integrated to refine topic representations, generate concise labels, and provide higher-level summaries of complex threads. We use `gpt-4.1-mini`.	https://platform.openai.com/docs/api-reference/introduction
VADER Sentiment Analysis (Hutto and Gilbert 2014)	A lexicon-based sentiment analysis tool that computes the polarity of posts and comments.	https://pypi.org/project/vaderSentiment/
(iv) SocialPulse Dashboard		
Flask	A lightweight web framework used to manage the backend routing and data exchange of the interactive SocialPulse dashboard.	https://flask.palletsprojects.com/en/stable/

- BERTopic Analysis (Grootendorst 2022): A transformer-based framework that utilizes Sentence-BERT embeddings and clustering (HDBSCAN) to discover latent themes in unstructured Reddit text.
- LLM Tooling: Large Language Models are integrated to refine topic representations, generate concise labels, and provide higher-level summaries of complex threads.
- VADER Sentiment Analysis (Hutto and Gilbert 2014): A lexicon-based sentiment analysis that computes the polarity of posts and comments.
- Cross-Community Comparison: The system supports comparative analysis across multiple user-specified subreddits that examines shared themes, duplicate posts, and overlapping users.

(iv) SocialPulse Dashboard

The final output is a user-centralized interactive dashboard. It provides a suite of features for rapid sensemaking:

- Exploratory Topic Analysis: Interactive visualizations of the BERTopic clusters present topic labels, representative keywords, and example posts, allowing users to drill down from aggregate trends to individual posts and raw comments for contextual analysis.
- Discourse Comparison: Side-by-side comparisons of sentiment polarity distributions, user activity levels, and topic frequencies across subreddits support analysis of differences in discourse tone and thematic emphasis among communities.
- Bot Transparency: Visual indicators of bot probability and flagged content, enabling users to evaluate the extent to which automated users contribute to observed discourse patterns.
- Configurable Analysis: The interface allows users to customize the analysis by adjusting BERTopic hyperparameters, specifying one or multiple subreddits, setting the minimum threshold required to flag an account as a bot, and selecting between a fixed-size or time-based data collection.

SocialPulse bridges the gap between raw data collection and deep qualitative understanding. By open-sourcing the toolkit, we aim to provide the ICWSM community with a reproducible foundation for studying social media dynamics.

Case Study: Sensemaking in `r/conspiracy`

Online discussions of conspiracy theories present a challenging setting for sensemaking, as they often involve a heterogeneous mix of genuine belief, skepticism, and play (Samory and Mitra 2018). In this section, we demonstrate how SocialPulse supports sensemaking within and across `r/conspiracy`, yielding three findings.

Applying SocialPulse to `r/conspiracy` over the week of 1/4/26, 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, war, 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 is at a high on Saturdays and at a low on Mondays, with the highest volume occurring from 15:00-16:00 UTC. Sentiment polarity across collected posts skews 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 fewer 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. Looking individually at each topic (**unpictured**), 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 (e.g. the Illu-

Table 2: **Comparison With Related Tools**: ✓indicates yes, △indicates partial, and ×indicates no.

Tool	Open-Source	Reddit	Bots	Topic Modeling	Summarization	Sentiment	Link
Hootsuite	×	△	×	×	✓	✓	https://www.hootsuite.com/
SproutSocial	×	✓	×	×	✓	✓	https://sproutsocial.com/
Meltwater	×	✓	✓	✓	✓	✓	https://www.meltwater.com/en
Sanjaya	✓	×	×	×	×	✓	https://github.com/Sanjaya-OSSMM/Sanjaya

minati or the moon landing) tend to be more neutral or positive. Adjusting the data collection window to include posts from earlier weeks further reveals differences in topic sentiment over time: current event topics become more generalized and overlapping, while long-standing conspiracy topics remain relatively semantically stable. 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. Furthermore, there are more social spammers in `r/conspiracy` than in `r/politics`, and the sentiment scores for both subreddits are predominantly negative.

▷ **Finding 1: 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 and greater instability over time, while long-standing conspiracies remain relatively stable and are often more neutral or positive in tone. This pattern is consistent with the overall sentiment distribution across the collected posts, but becomes more pronounced when analyzed at the topic level, where certain topics deviate from the aggregate trend. Together, these differences indicate that distinct topics of discussion are associated with various emotional dynamics within the community.

▷ **Finding 2: Posting activity within `r/conspiracy` is unevenly distributed across the week, with high activity levels on Saturday and lower activity levels on Monday.** This temporal pattern is further highlighted in hourly trends, where posting peaks between 15:00-16:00 UTC, suggesting habitual engagement periods and fluctuations in conspiracy-related discourse 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. Furthermore, analyses based on a more narrow collection window may miss or misrepresent broader participation patterns.

▷ **Finding 3: 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 window of time and corresponds to current events. Additionally, overlapping users and a higher percentage of social spammers suggests that some of this cross-posting behavior may be driven by coordinated diffusion. These observations indicate that politically relevant posts circulate rapidly across otherwise distinct subreddits, highlighting how analyzing subreddits in isolation may overlook important pathways of information dissemination.

Broader Impact

Computational social science research increasingly relies on large-scale social media data, yet the analytical workflows required to interpret this data remain fragmented across disparate tools and scripts. SocialPulse aims to support researchers by providing an integrated, interactive environment that connects data collection, modeling, and analysis within a single workflow to improve exploration and hypothesis generation.

A key contribution of SocialPulse is its support for examining participation structure alongside topical and sentiment-based analyses. By explicitly distinguishing between highly active users and long-tail contributors, the toolkit helps researchers avoid over-reliance on highly active users and encourages more nuanced interpretations of online discourse. This is particularly valuable for studies concerned with community norms and the dynamics of attention and engagement.

More broadly, SocialPulse is intended to improve transparency and reproducibility in social media analysis. By offering an open-source pipeline with clearly defined analytical components, the system allows researchers to inspect and adjust parameters, and replicate exploratory workflows across datasets and communities. In doing so, SocialPulse supports more replicable sensemaking practices in computational social science, rather than replacing or automating interpretation.

Conclusion

SocialPulse provides an interactive, open-source pipeline for analyzing 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.

Future Work

Currently, SocialPulse is optimized for Reddit's nested comment structure. Future iterations will include:

- **Cross-Platform Support:** Extending the pipeline to support other social media platforms.
- **LLM-in-the-Loop:** Integrating LLMs to provide automated "Sensemaking Summaries" of complex thread hierarchies.
- **Real-time Notifications:** Enabling practitioners to set thresholds for "sentiment spikes" or "bot activity" in monitored communities.

References

Davidson, C. 2023. *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.

De Choudhury, M.; and De, S. 2014. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *ICWSM*, volume 8, 71–80.

Dong, X.; Li, C.; and Choi, J. D. 2020. Transformer-based Context-aware Sarcasm Detection in Conversation Threads from Social Media. In Klebanov, B. B.; Shutova, E.; Lichtenstein, P.; Muresan, S.; Wee, C.; Feldman, A.; and Ghosh, D., eds., *Proceedings of the Second Workshop on Figurative Language Processing*, 276–280. Online: Association for Computational Linguistics.

FORCE11. 2020. The FAIR Data principles. https://force11.org/info/the-fair-data-principles/.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Grootendorst, M. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

Guo, Y.; Dong, X.; Al-Garadi, M. A.; Sarker, A.; Paris, C.; and Aliod, D. M. 2020. Benchmarking of Transformer-Based Pre-Trained Models on Social Media Text Classification Datasets. In Kim, M.; Beck, D.; and Mistica, M., eds., *Proceedings of the 18th Annual Workshop of the Australasian Language Technology Association*, 86–91. Virtual Workshop: Australasian Language Technology Association.

Hutto, C.; and Gilbert, E. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *ICWSM*, volume 8, 216–225.

Islam, T.; and Goldwasser, D. 2025. Discovering latent themes in social media messaging: A machine-in-the-loop approach integrating llms. In *ICWSM*, volume 19, 859–884.

Klein, G.; Moon, B.; and Hoffman, R. R. 2006. Making sense of sensemaking 1: Alternative perspectives. *IEEE intelligent systems*, 21(4): 70–73.

Kumar, S.; Hamilton, W. L.; Leskovec, J.; and Jurafsky, D. 2018. Community interaction and conflict on the web. In *The Web Conference*, 933–943.

Mendoza, M.; Providel, E.; Santos, M.; and Valenzuela, S. 2024. Detection and impact estimation of social bots in the Chilean Twitter network. *Scientific reports*, 14(1): 6525.

Ng, L. H. X.; and Carley, K. M. 2023. Botbuster: Multi-platform bot detection using a mixture of experts. In *ICWSM*, volume 17, 686–697.

Oswald, L.; Schulz, W.; Hertwig, R.; Lazer, D.; and Stier, S. 2025. The Tip of the Iceberg: How the Social Media Production–Consumption Gap Distorts Public Opinion for Citizens and Researchers. *SocArXiv*. Preprint.

Panek, E.; Hollenbach, C.; Yang, J.; and Rhodes, T. 2018. The effects of group size and time on the formation of online communities: Evidence from Reddit. *Social Media+ Society*, 4(4): 2056305118815908.

Samory, M.; and Mitra, T. 2018. Conspiracies online: User discussions in a conspiracy community following dramatic events. In *ICWSM*, volume 12.

Steffen, E. 2025. More than Memes: A Multimodal Topic Modeling Approach to Conspiracy Theories on Telegram. In *ICWSM*, volume 19, 1831–1844.

Teleki, M.; Dong, X.; Liu, H.; and Caverlee, J. 2025. Masculine Defaults via Gendered Discourse in Podcasts and Large Language Models. In *ICWSM*, volume 19, 1893–1912.

Yin, K.; Dong, X.; Liu, C.; Huang, L.; Xiao, Y.; Liu, Z.; Mostafavi, A.; and Caverlee, J. 2025. DisastIR: A Comprehensive Information Retrieval Benchmark for Disaster Management. In Christodoulopoulos, C.; Chakraborty, T.; Rose, C.; and Peng, V., eds., *Findings of the Association for Computational Linguistics: EMNLP 2025*, 1836–1867. Suzhou, China: Association for Computational Linguistics. ISBN 979-8-89176-335-7.

Zhu, Y.; Haq, E.-U.; Tyson, G.; et al. 2024. A Study of Partisan News Sharing in the Russian invasion of Ukraine. In *ICWSM*, volume 18, 1847–1858.

Paper Checklist

1. For most authors...

   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes, this work advances understanding of online discourse using publicly available data without violating social contracts.

   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes, see Abstract and Introduction.

   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, see System Architecture and Implementation.

   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, see Introduction and Case Study.

   (e) Did you describe the limitations of your work? Yes, see Future Work.

   (f) Did you discuss any potential negative societal impacts of your work? No, we did not explicitly discuss potential negative societal impacts because SocialPulse is an only intended to be an exploratory data analysis tool that uses public data.

   (g) Did you discuss any potential misuse of your work? No, we did not explicitly discuss potential misuse because SocialPulse is only intended for transparent research and sensemaking.

   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, see Conclusion, Bot Filtration, and SocialPulse Dashboard.

(i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes

2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? NA
   (b) Have you provided justifications for all theoretical results? NA
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA
   (e) Did you address potential biases or limitations in your theoretical framework? NA
   (f) Have you related your theoretical results to the existing literature in social science? NA
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? NA
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? NA
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? NA
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? NA
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? NA
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? NA
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? Yes, see References and Table 1.
   (b) Did you mention the license of the assets? No, the paper does not explicitly list the licenses of external software assets used in SocialPulse.
   (c) Did you include any new assets in the supplemental material or as a URL? Yes, see Table 1.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? NA
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? No, we do not explicitly discuss the presence of personally identifiable information or offensive content because SocialPulse retrieves publicly available data and analyzes content at an aggregate level.
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? NA
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? NA
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
   (a) Did you include the full text of instructions given to participants and screenshots? NA
   (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA
   (d) Did you discuss how data is stored, shared, and deidentified? NA