Multi-Platform Aggregated Dataset of Online Communities (MADOC)

Marija Mitrović Dankulov[1], Aleksandar Tomašević[2], Slobodan Maletić[3], Miroslav Anđelković[3], Ana Vranić[1], Darja Cvetković[1], Boris Stupovski[1], Dušan Vudragović[1], Sara Major[2], Aleksandar Bogojević[1]

[1]Institute of Physics Belgrade, University of Belgrade, Belgrade, Serbia
[2]Faculty of Philosophy, University of Novi Sad, Novi Sad, Serbia
[3]Vinča Institute of Nuclear Sciences, University of Belgrade, Belgrade, Serbia

Abstract

The Multi-platform Aggregated Dataset of Online Communities (MADOC) is a comprehensive dataset that facilitates computational social science research by providing a unified, standardized dataset for cross-platform analysis of online social dynamics. MADOC aggregates and standardizes data from four distinct platforms: Bluesky, Koo, Reddit, and Voat, spanning from 2012 to 2024. The dataset includes 18.9 million posts, 236 million comments, and data from 23.1 million unique users across all platforms, with a particular focus on understanding community dynamics, user migration patterns, and the evolution of toxic behavior across platforms. By providing standardized data structures and FAIR-compliant access through Zenodo and corresponding Python and R packages, MADOC enables researchers to conduct comparative analyses of user behavior, interaction networks, and content sentiment across diverse social media environments. The unique value of the dataset lies in its cross-platform scope, standardized structure, and rich metadata, making it particularly suitable for studying societal phenomena such as community formation, toxic behavior propagation, and user migration patterns in response to platform moderation policies.

Dataset — https://zenodo.org/records/14637314

Introduction

The proliferation of social media platforms has created diverse digital spaces where users interact, share content, and form communities. Understanding these interactions and their societal impact requires comprehensive datasets that span multiple platforms and enable comparative analyses. The Multi-platform Aggregated Dataset of Online Communities (MADOC) addresses this need by providing a standardized collection of user interactions, content, and sentiment data from four distinct platforms: Reddit, Bluesky, Koo, and Voat. Comprising 18.9 million posts and 236 million comments from 23.1 million unique users, MADOC represents one of the largest cross-platform datasets available for social media research. In an era where platform API access has become increasingly restricted—what some researchers call the "post-API era" (Poudel and Weninger 2024)—MADOC leverages existing public datasets (Baumgartner et al. 2020; Mekacher and Papasavva 2022; Failla and Rossetti 2024; Mekacher, Falkenberg, and Baronchelli 2024) to enable cross-platform research.

The dataset is structured to support various computational social science research directions:

- **Cross-platform User Behavior:** Standardized interaction data enables comparative analysis of user engagement patterns across different platform architectures and community structures (Wang, Koneru, and Rajtmajer 2024; Vranić et al. 2023).
- **Community Dynamics:** Comprehensive community-level data allows researchers to study how communities respond to major events like content moderation, user influx, or platform-wide policy changes, though individual user migration cannot be tracked across platforms due to privacy protections (Papasavva and Mariconti 2023).
- **Content and Sentiment Analysis:** Textual content, toxicity and sentiment scores facilitate research on how discourse and sentiment vary across platforms and communities.
- **Moderation Impact:** Historical data from banned communities provides insight into the effectiveness of platform moderation policies and their impact on user behavior (Cima et al. 2024).

MADOC adheres to FAIR principles (Findable, Accessible, Interoperable, and Reusable) through the following key features:

- **Findable:** The dataset is distributed through Zenodo [1] with persistent identifiers and comprehensive metadata for discovery.
- **Accessible:** All data and metadata are openly accessible through standard protocols with clear access procedures. We provide Python [2]and R [3] packages for seamless access to the dataset.
- **Interoperable:** Data is provided in standardized Apache Parquet format with a consistent schema across platforms, enabling seamless integration.
- **Reusable:** Detailed documentation of collection methodologies, processing steps, and ethical guidelines supports research reproducibility.

[1]https://zenodo.org/records/14637314
[2]https://pypi.org/project/pymadoc/
[3]https://github.com/atomashevic/rMADOC

MADOC's primary contribution is its methodologically rigorous approach to cross-platform data alignment and standardization. The dataset construction begins with a careful selection of 12 Reddit communities, with corresponding communities later identified on Voat. The first set consists of six general-interest communities (*r/funny, r/gaming, r/pics, r/videos, r/gifs, and r/technology*) that focus on mainstream topics. The second set comprises controversial communities, including four that were banned from Reddit for violating platform policies: *r/fatpeoplehate* (banned in 2015 for systematic harassment), *r/GreatAwakening* (banned in 2018 for inciting violence and promoting conspiracy theories), *r/MillionDollarExtreme* (banned in 2018 for promoting hate speech and alt-right content), and *r/CringeAnarchy* (banned in 2019 following increased hate speech after the Christchurch mosque shootings). Two additional controversial communities, while not banned, have been subjected to increased scrutiny and moderation: *r/KotakuInAction* (heavily moderated since 2014 due to its role in harassment campaigns during Gamergate) and *r/MensRights* (subject to content restrictions due to repeated incidents of coordinated harassment and anti-feminist rhetoric). These communities were chosen specifically because they represent different types of content moderation challenges: from organized harassment campaigns and hate speech to conspiracy theories and extremist content. This selection enables researchers to study both typical online interactions and potentially harmful social dynamics, particularly in communities that have faced different levels of platform enforcement (Jhaver, Chan, and Bruckman 2017). The selection is strategic: it allows us to analyze how community-level behavior, content patterns, and discourse evolve before and after significant events such as platform bans, content restrictions, or influxes of new users. While individual users cannot be tracked across platforms due to privacy protections and ethical considerations, the dataset captures aggregate changes in community characteristics, sentiment, and content patterns during periods of user migration and community restructuring.

Data Collection and Processing

Data Sources

MADOC aggregates and standardizes data from four distinct social media platforms: Reddit, Voat, Bluesky, and Koo.

For Reddit, we utilized the Pushshift.io dataset (Baumgartner et al. 2020), which encompasses submissions and comments from 2006 to 2020. From this extensive collection, we focused on 12 subreddits: six in the general-interest category, and six controversial.

The Voat dataset (Mekacher and Papasavva 2022), spanning from 2013 to the platform's shutdown in 2020, provides a complete archive of the platform's content. We extracted data from twelve subverses that directly correspond to our selected Reddit communities, maintaining the same naming convention (e.g., v/funny corresponding to r/funny).

For Bluesky, we incorporated a comprehensive dataset containing over 4 million accounts (81% of registered users) and their complete posting activity (235M posts) from March 2023 to March 2024 (Failla and Rossetti 2024). Similarly, we used the existing Koo dataset (Mekacher, Falkenberg, and Baronchelli 2024) which encompasses 1.4 million users, 72 million posts, and 75 million comments, spanning from January 2020 to September 2023 . This platform's inclusion is particularly valuable as it represents perspectives from the Global South, with a predominantly Indian user base.

Data Processing and Standardization

A key challenge in creating MADOC was aligning content across platforms with different organizational structures. While Reddit and Voat organize content into topic-specific communities (subreddits and subverses), Bluesky and Koo lack such explicit topic segregation. To address this, we developed a systematic content alignment process using Latent Dirichlet Allocation (LDA) topic modeling (Blei, Ng, and Jordan 2003; Maier et al. 2018). The detailed methodology of this process is described in the next section.

The dataset standardization process consists of four main components. First, we implemented comprehensive data cleaning procedures. These include deduplication of entries while preserving temporal precedence, conversion of all date and time related fields to UNIX timestamps, standardization of text encoding and character sets, and cleaning of URLs to remove formatting artifacts. For Bluesky and Koo, we implemented additional URL extraction from content fields, separating embedded links from the main text while preserving both components. To account for the presence of bots on Reddit and Voat—programmed to post, comment automatically, or respond to specific prompts — we implemented a "bot removal" procedure. These bots often include the phrase "I am a bot" in their posts or comments. Therefore, we screened users whose interactions (posts and comments) contained this phrase. If more than 70% of a user's interactions included "I am a bot," they were removed from our dataset. The 70% threshold was chosen after analyzing the number of users flagged as bots across various detection thresholds (ranging from 40% to 100%) for all communities. We observed a sharp decrease in flagged users at 70% for in every community, indicating it as the most suitable threshold for identifying bots across all communities.

Second, we standardized interaction types across platforms. All interactions were classified as either posts, comments, or reposts (the latter only for Bluesky and Koo). Posts are submissions on Reddit and Voat and tweet-like posts on Koo and Bluesky. Comments include submission comments on Reddit and Voat and replies to posts on Koo and Bluesky.

For Voat comments, we had only parent information linking posts and first-level comments (excluding comments on comments). This is because the original Voat dataset does not contain complete information about comment-to-comment relationships. To maintain a consistent network structure across platforms, we preserved only first-level replies and comments, excluding nested conversations. This standardization required careful validation of parent-child relationships in threaded discussions, particularly for platforms like Reddit and Voat where deep conversation trees are common. We kept comments that had no parent informa-

Metric	Reddit	Voat	Bluesky	Koo
Time span	2014-2020	2013-2020	2023-2024	2020-2023
Total interactions	247.6M	1.2M	2.8M	4.3M
Posts	14.5M	0.4M	0.9M	3.1M
Comments	233.1M	0.8M	0.9M	1.2M
Unique users	22.6M	0.1M	0.2M	0.2M
Avg. posts per user	2.7	8.5	14.4	23.0
Avg. comments per user	11.6	7.8	6.2	10.6
Mean VADER sentiment	0.063	0.011	0.088	0.054
% with URLs	10.3%	31.8%	4.5%	11.6%

Table 1: Platform-level statistics for MADOC. The dataset covers different time periods for each platform, reflecting their operational histories and data availability.

tion, because they relate to posts which have been deleted or banned from the platform. While we acknowledge that the depth of reply trees on Reddit and Voat is a distinctive feature of these platforms compared to Bluesky and Koo, our decision to include only level-1 comments was driven by the limitations of the Voat source data and the need for cross-platform standardization. Researchers specifically interested in examining conversation depth and structure should refer to the original datasets, as this aspect of platform comparison is outside MADOC's primary focus on content and community-level analysis.

The absence of parent information is noticeable on Reddit where, due to increased content-moderation, bans, and user migrations, original post-to-comment relationships are not always preserved. The most extreme case is r/fatpeoplehate subreddit, where all posts are missing because the content was banned.

For sentiment analysis, we applied two complementary approaches using lexicon and rule-based models: VADER (Hutto and Gilbert 2014) and TextBlob. VADER is specifically attuned to social media text and provides context-aware sentiment scoring, which accounts for elements like punctuation, capitalization, and modifiers. TextBlob offers a more generalized polarity assessment based on pattern analysis and provides a useful alternative perspective. Both sentiment scores range from -1 (negative) to 1 (positive), allowing researchers to compare and validate sentiment patterns using multiple methodologies. Additionally, we included TextBlob's subjectivity score, which ranges from 0.0 (objective) to 1.0 (subjective), measuring the degree to which text expresses personal opinions, emotions, or judgments versus factual information.

Additionally, we used the ToxiGen RoBERTa model (Hartvigsen et al. 2022) to detect subtle forms of implicit toxicity and hate speech in the content. Although Perspective API is widely considered the standard for toxicity detection, its rate limits were a significant limiting factor given the size of our dataset, and thus utilizing a fully open-source model like ToxiGen better supports reproducibility and aligns with open science principles. The ToxiGen model was specifically trained on adversarial examples of implicit hate speech targeting minority groups, making it particularly well-suited for detecting toxic content that might evade traditional detection methods. The toxicity_toxigen field provides a normalized score from 0.00 (non-toxic) to 1.00 (toxic), enabling researchers to analyze toxicity patterns across different platforms and communities.

For the initial release, we restricted the dataset to English-language content. For Koo, which serves a predominantly multilingual user base (Mekacher, Falkenberg, and Baronchelli 2024), we implemented additional language filtering steps to ensure reliable sentiment and toxicity analysis results.

Fourth, we implemented privacy protection measures. All post and user identifiers across all platforms are pseudonymized using UUID-based hashing, which maintains referential integrity while preventing casual re-identification. The only personally identifiable information is the content of the posts along with the timestamp of the post. However, this information was already present in the original datasets, so our pseudonymization makes it harder to reconstruct the original content in comparison with the original datasets. We deliberately chose not to attempt cross-platform profile matching (identifying the same users across different platforms), both because some source datasets already contained pseudonymized usernames making such matching technically infeasible, and because such matching would raise significant privacy concerns by potentially enabling re-identification of users across platforms.

This approach maintains the network structure and enables research on user behavior patterns while protecting user privacy through pseudonymization.

The resulting dataset follows a standardized schema across all platforms, with the structure detailed in Table 2.

To facilitate efficient data access and analysis, we store the dataset in Apache Parquet format, with separate files for each community (subreddit/subverse) while Koo and Bluesky are stored in a single file. This modular organization enables researchers to easily load specific subsets of the data while maintaining the ability to perform cross-platform analyses. The standardized schema ensures that files can be seamlessly combined (concatenated) across platforms and communities, allowing flexible dataset construction based on research needs. For example, researchers can combine all controversial communities across platforms, merge specific platform pairs for comparative analysis, or create custom subsets based on temporal or content criteria. The standardized structure also allows for the construction of interaction networks where nodes represent users and edges represent their interactions through posts and replies, enabling com-

parative analysis of community structures across platforms.

Topic Analysis and Selection of Posts

To enable cross-platform content analysis, we developed a systematic approach for identifying thematically similar content across platforms with different organizational structures. While Reddit and Voat organize content into topic-specific communities (subreddits and subverses), Koo and Bluesky lack explicit topic segregation. We employed Latent Dirichlet Allocation (LDA) topic modeling (Blei, Ng, and Jordan 2003; Maier et al. 2018) to extract characteristic keywords from Reddit and Voat communities, which were then used to identify related content on Koo and Bluesky.

For each community pair (subreddit and corresponding subverse), we created a combined corpus by merging posts and their first-level comments from both platforms. Each document in the corpus represents a post with its comments, preprocessed by removing mentions, hyperlinks, non-alphanumeric characters, converting to lowercase, removing stopwords, and lemmatizing. We excluded words appearing in more than 99% of documents following (Maier et al. 2018) and filtered-out documents with fewer than 20 or more than 2000 words to ensure LDA performance (Tang et al. 2014).

Due to platform size disparities, we implemented a balanced sampling strategy based on document volume:

- For communities with similar document counts ($<$10,000): Created nine balanced samples of 20,000 documents (10,000 from each platform)
- For communities with 10,000-20,000 Voat documents: Created five samples matching Voat's volume with Reddit documents
- For communities with $<$10,000 Voat documents: Created five samples of 20,000 Reddit documents plus all Voat documents

We determined the optimal number of topics for each community by maximizing coherence scores, then extracted the top 20 words per topic across all samples. The resulting keyword lists were filtered to retain only the least frequent third of English words (Norvig 2009) to improve topic specificity.

For content alignment, we applied two filtering criteria to Koo and Bluesky posts:

- Basic filter: Documents containing at least one community keyword
- Strict filter: Documents containing at least two different community keywords

It is important to note that our topic modeling approach does not attempt to align the structural differences between platforms (as Reddit and Voat have explicit community structures while Koo and Bluesky do not), but rather ensures topical equivalence across platforms. Without this filtering method, we would be left with massive, unstructured Bluesky and Koo datasets with no meaningful way to conduct comparative analysis with the community-structured data from Reddit and Voat. This approach allows researchers to compare content discussing similar themes regardless of the underlying platform architecture, maintaining consistent thematic scope while acknowledging the fundamental differences in how these platforms organize user interactions.

The topic modeling analysis revealed distinct linguistic patterns characteristic of each community, providing strong validation for our cross-platform content alignment approach. For general-interest communities, the extracted keywords closely matched their intended focuses. For example, r/gaming exhibited gaming-specific terminology (e.g., "pokemon", "hack"), while r/technology contained science and technology terms (e.g., "solar", "nasa", "electricity").

The topic modeling also revealed distinct linguistic patterns in communities focused on specific social discussions. For instance, r/MensRights showed a concentration of legal and family-related terminology ("father", "parent", "judge"). Communities focused on political discourse showed distinctive patterns. The r/GreatAwakening community, which centered on conspiracy theories before its ban in 2018, featured high frequencies of governance-related terms ("congress", "campaign", "election", "truth"). Similarly, r/KotakuInAction, which emerged during the GamerGate controversy (Jhaver, Chan, and Bruckman 2017) and focused on media criticism, developed its own specialized vocabulary around content moderation and journalistic practices. These distinctive vocabulary patterns proved particularly valuable for identifying thematically similar content on platforms lacking explicit community structures.

The topic modeling analysis also revealed patterns of problematic language usage across platforms. We observed systematic variations in the frequency and context of inflammatory rhetoric, derogatory terms, and exclusionary language. These patterns were particularly pronounced in communities that were eventually subject to platform moderation actions. This shows that the sampling strategy we used is effective in recognizing the specific language employed by communities on Voat.

These results demonstrate the robustness of our topic modeling and keyword filtering methodology for cross-platform content alignment. The approach successfully captures both explicit topical focus (through technical and domain-specific terminology) and implicit community characteristics (through platform-specific linguistic patterns), enabling meaningful comparative analyses across diverse social media environments.

Dataset Statistics

MADOC provides a comprehensive view of user interactions across four distinct platforms, with data spanning multiple years and interaction types. Table 1 presents the key statistics for each platform, including temporal coverage, interaction counts, and user activity metrics. [4]

The platform-level statistics reveal significant differences in scale and user behavior across platforms. Reddit dominates in terms of total interactions (247.6M) and user base

[4]The most recent version of the key statistics table for each platform, including TextBlob sentiment and subjectivity as well as ToxiGen toxicity scores is available at https://zenodo.org/records/14637314

(22.6M), being orders of magnitude larger than the other platforms. However, smaller platforms show higher per-user engagement rates, with Koo users averaging 23.0 posts per user compared to Reddit's 2.7. Voat shows notably higher URL sharing (31.8% of posts) compared to other platforms (4.5-11.6%), suggesting a stronger focus on external content sharing. Sentiment analysis reveals that Bluesky has the most positive average sentiment (0.088), while Voat shows the lowest (0.011), potentially reflecting differences in community norms and content moderation approaches.

For Reddit and Voat, where content is organized into topic-specific communities, we provide detailed statistics for each community pair in Table 3. These statistics enable direct comparisons between equivalent communities across the two platforms.

The community-level comparison between Reddit and Voat reveals several interesting patterns. First, there is a clear scale difference, with Reddit communities typically being 100-1000 times larger than their Voat counterparts. For instance, r/funny has 5.5M users compared to v/funny's 20K. Second, sentiment patterns show that general-interest communities (e.g., gaming, pics) tend to maintain positive sentiment scores on both platforms, with Reddit generally showing more positive values. In contrast, controversial communities exhibit negative sentiment scores across both platforms, with some communities like MensRights showing particularly negative values (Reddit: -0.076, Voat: -0.162). The absence of posts in Reddit's r/fatpeoplehate while having 1.4M comments reflects the community's ban during the data collection period, providing an interesting case study of community dynamics around platform moderation actions.

For Bluesky and Koo, which lack explicit community structures, we provide statistics based on the keyword filtering approach described in previous section. Using the basic filtering criteria (one keyword match), we identified 910,376 posts and 932,545 comments from Bluesky, and 3,083,191 posts and 1,180,449 comments from Koo that align with the topics of the Reddit/Voat communities. With the strict filtering criteria (two keyword matches), these numbers reduce to 114,431 posts and 322,193 comments for Bluesky, and 249,983 posts and 138,418 comments for Koo, respectively.

It is important to note that the MADOC dataset is constantly being updated and expanded with new metrics. Some metrics, particularly for larger platforms like Reddit, are still being processed and will be available only in the most recent versions of the dataset. As our analysis progresses, we incorporate additional sentiment analysis models, toxicity measures, and other relevant metrics. These continuous updates ensure that researchers have access to the most comprehensive and up-to-date data for their analyses. While this paper presents a snapshot of its current state, we encourage researchers to refer to the latest version of the dataset, including the most recent metrics and statistics, at the Zenodo repository: https://zenodo.org/records/14637314.

Limitations

Our dataset production approach has several limitations. First, the choice of Reddit communities, especially the non-controversial ones, was arbitrary and based on the 20 most active general-interest communities which existed on both Reddit and Voat. In the next versions of the dataset, we aim to include more Reddit-Voat community pairs.

Second, the LDA topic modeling approach was not cross-validated by performing topic modeling again on the filtered communities to verify alignment between the communities. It is important to note that traditional cross-validation is not applicable in our case due to the fundamental structure of our methodology. The content from Koo and Bluesky was filtered based on a union of keywords extracted from multiple community-specific topics from Reddit and Voat, resulting in inherently overlapping topics and combined keyword sets that cannot be meaningfully validated through conventional methods. We also limited our evaluation to selecting models based on coherence scores without exploring alternative topic modeling configurations or parameters. This lack of systematic evaluation may affect the quality and representativeness of the extracted keywords used for cross-platform content alignment. Future versions of the dataset will address these limitations by developing specialized validation techniques better suited to our cross-platform alignment approach.

Third, we introduce multiple text analysis metrics in this version of the dataset: VADER sentiment scores, TextBlob sentiment polarity, and ToxiGen toxicity scores. This combination provides researchers with complementary perspectives on content sentiment and toxicity. The ToxiGen model is specifically designed to detect subtle forms of implicit toxicity and hate speech that might evade traditional detection methods. Future versions will provide an even more comprehensive overview of posts' sentiment and toxicity using additional approaches (other RoBERTa-based models, transformer-based sentiment analysis, etc.).

Fourth, the time spans of collected data vary significantly across platforms (2014-2020 for Reddit, 2013-2020 for Voat, 2023-2024 for Bluesky, and 2020-2023 for Koo), which introduces challenges in studying long-term trends across all platforms simultaneously. However, this temporal misalignment is unavoidable given the historical reality of platform lifecycles—Bluesky did not exist during the period of Reddit-Voat migrations, making perfect temporal alignment impossible. Rather than viewing this as a limitation, researchers can leverage these different time periods to study how similar communities and content patterns manifest in different platform eras, from early alternative platforms like Voat to newer decentralized networks like Bluesky. The dataset particularly excels at enabling thematic comparisons across platforms regardless of temporal overlap, focusing on how similar discourses evolve in different platform environments and moderation contexts.

Finally, while we have aggregated a substantial amount of data, there may be gaps in our collection due to the limitations of archival sources and API restrictions. Some content might have been deleted or modified before collection, and we cannot guarantee complete coverage of all interactions that occurred on the platform.

Conclusion

In this work, we present MADOC, a comprehensive cross-platform dataset comprising 18.9 million posts and 236 million comments from 23.1 million unique users across four distinct social media platforms: Reddit, Voat, Bluesky, and Koo. The dataset spans from 2012 to 2024 and represents one of the largest standardized collections of cross-platform social media interactions. We combine data from multiple sources and implement standardization procedures to ensure consistency and comparability across platforms.

The dataset's unique value lies in its standardized structure and rich metadata, which facilitate various types of computational social science research. The inclusion of both mainstream and alternative platforms, along with comprehensive historical data on banned communities, makes MADOC particularly valuable for studying platform governance, content moderation, and user migration patterns. Our careful curation process and adherence to FAIR principles ensure the dataset's accessibility and reusability while maintaining privacy protections through pseudonymization protocols.

We envision MADOC enabling several important research directions: First, researchers can use it to study how content moderation policies affect user behavior and community dynamics across different platform architectures. Second, the dataset's temporal breadth supports longitudinal analyses of how online discourse and community norms evolve over time and across platforms. Third, the standardized sentiment scores and content features facilitate comparative studies of how platform design influences user interactions and content toxicity.

Beyond these immediate applications, we believe that MADOC will contribute to a broader understanding of online social phenomena. The dataset can help answer questions about community formation, user activity patterns, and the propagation of both harmful and benevolent narratives across platform boundaries. For computational researchers, it provides a rich testbed for developing and evaluating natural language processing models, particularly in areas like toxic content detection, cross-platform user behavior prediction, and community dynamics modeling.

Through this work, we aim to support both quantitative and qualitative research into online social dynamics. We encourage researchers to use MADOC responsibly, considering the ethical implications discussed in this paper, and to build upon our methodology for future cross-platform dataset development. To facilitate easy adoption, we provide Python (pyMADOC) and R (rMADOC) packages with command-line interfaces that enable selective downloading of data by platform or community. As social media continues to evolve and fragment across multiple platforms, we believe standardized cross-platform datasets like MADOC will become increasingly valuable for understanding and improving online social spaces.

Acknowledgments

This research was financially supported by the Science Fund of the Republic of Serbia, Prizma program (grant No. 7416). Data processing was performed on the PARADOX-IV supercomputing facility at the Scientific Computing Laboratory, National Center of Excellence for the Study of Complex Systems, Institute of Physics Belgrade.

Claude 3.5 Sonnet LLM was used in preparation of this manuscript, for table generation, text editing and LaTeX formatting.

References

Baumgartner, J.; Zannettou, S.; Keegan, B.; Squire, M.; and Blackburn, J. 2020. The Pushshift Reddit Dataset. *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1): 830–839.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3: 993–1022.

Cima, L.; Trujillo, A.; Avvenuti, M.; and Cresci, S. 2024. The Great Ban: Efficacy and Unintended Consequences of a Massive Deplatforming Operation on Reddit. *arXiv [cs.SI]*.

Failla, A.; and Rossetti, G. 2024. "I'm in the Bluesky Tonight": Insights from a year worth of social data. *PloS one*, 19(11): e0310330.

Hartvigsen, T.; Gabriel, S.; Palangi, H.; Sap, M.; Ray, D.; and Kamar, E. 2022. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. *arXiv [cs.CL]*.

Hutto, C.; and Gilbert, E. 2014. VADER: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1): 216–225.

Jhaver, S.; Chan, L.; and Bruckman, A. 2017. The view from the other side: The border between controversial speech and harassment on Kotaku in Action. *arXiv [cs.OH]*.

Maier, D.; Waldherr, A.; Miltner, P.; Wiedemann, G.; Niekler, A.; Keinert, A.; Pfetsch, B.; Heyer, G.; Reber, U.; Häussler, T.; Schmid-Petr, H.; and Adam, S. 2018. Applying LDA topic modeling in communication research: Toward a valid and reliable methodology. *Communication Methods and Measures*, 12: 93–118.

Mekacher, A.; Falkenberg, M.; and Baronchelli, A. 2024. The Koo dataset: An Indian microblogging platform with global ambitions. *arXiv [cs.SI]*.

Mekacher, A.; and Papasavva, A. 2022. "I can't keep it up." A dataset from the defunct Voat.Co news aggregator. *arXiv [cs.SI]*.

Norvig, P. 2009. Natural Language Corpus Data. In Segaran, T.; and Hammerbacher, J., eds., *Beautiful Data*, 219–242. O'Reilly Media, Inc. ISBN 9780596157111.

Papasavva, A.; and Mariconti, E. 2023. Waiting for Q: An Exploration of QAnon Users' Online Migration to Poal in the Wake of Voat's Demise. arxiv:2302.01397.

Poudel, A.; and Weninger, T. 2024. Navigating the Post-API dilemma | Search Engine Results Pages present a biased view of social media data. *arXiv [cs.IR]*.

Tang, J.; Meng, Z.; Nguyen, X.; Mei, Q.; and Zhang, M. 2014. Understanding the Limiting Factors of Topic Modeling via Posterior Contraction Analysis. In Xing, E. P.; and Jebara, T., eds., *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, 190–198. Bejing, China: PMLR.

Vranić, A.; Tomašević, A.; Alorić, A.; and Mitrović Dankulov, M. 2023. Sustainability of Stack Exchange Q&A communities: the role of trust. *EPJ Data Science*, 12(1).

Wang, X.; Koneru, S.; and Rajtmajer, S. 2024. The failed migration of academic Twitter. *arXiv [cs.SI]*.

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, our research advances science by providing a standardized dataset for studying online social dynamics while implementing strong privacy protections and ethical safeguards as detailed in the Ethical Considerations section.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope?
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? This is a dataset paper, it does not make specific claims.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? We have addressed questions like these in Ethical Considerations section.
   (e) Did you describe the limitations of your work? Yes, we discuss limitations including arbitrary community selection, LDA topic modeling validation, sentiment analysis scope, and data collection gaps in the Limitations section.
   (f) Did you discuss any potential negative societal impacts of your work? We have addressed this in the Ethical Considerations section.
   (g) Did you discuss any potential misuse of your work? We have addressed this in Ethical Considerations section.
   (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, we have implemented extensive measures including pseudonymization, clear documentation, usage tracking, and FAIR principles as detailed in the Ethical Considerations section.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them?
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results?
   (b) Have you provided justifications for all theoretical results?
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results?
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study?
   (e) Did you address potential biases or limitations in your theoretical framework?
   (f) Have you related your theoretical results to the existing literature in social science?

(g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain?

3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results?
   (b) Did you include complete proofs of all theoretical results?
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)?
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)?
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)?
   (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)?
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made?
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance?
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, we cited Pushshift.io, Voat dataset, Bluesky dataset, and Koo dataset creators.
   (b) Did you mention the license of the assets? Yes, we mentioned CC0 for Reddit Pushshift and CC BY 4.0 International for Voat, Bluesky, and Koo datasets.
   (c) Did you include any new assets in the supplemental material or as a URL? Yes, we provided the Zenodo URL for the dataset.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? No, because we aggregated publicly available datasets.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes, we explicitly discuss both PII handling and the presence of offensive content in the Ethical Considerations section.
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? Yes, we discuss FAIR principles implementation including findability through Zenodo, accessibility through standard protocols, interoperability through standardized formats, and reusability through documentation.
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? Partially. The paper itself serves as a comprehensive datasheet covering motivation, composition, collection process, preprocessing, uses, distribution, and maintenance. Additionally, these is a detailed README file in the Zenodo repository.
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?
   (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals?
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation?
   (d) Did you discuss how data is stored, shared, and deidentified?

Ethical Considerations

The MADOC dataset raises several important ethical considerations that we have addressed throughout its development and release. Our approach to these considerations is guided by the principles of responsible data science and research ethics, with particular attention to privacy protection, content handling, and potential misuse prevention.

The dataset is created through aggregation of existing, publicly available datasets, and we have taken extensive measures to enhance privacy protections beyond those present in the source data. We have not introduced any new sources of personally identifiable information (PII). Instead, all user identifiers are pseudonymized using UUID-based hashing, which maintains referential integrity while preventing casual re-identification. Platform-specific identifiers have been standardized and pseudonymized to prevent cross-platform user tracking. The only retained potentially identifying information is post content and timestamps, which were already public in the source datasets.

The aggregation of data complies with the existing licenses of previously published datasets, where licensing information is available (Reddit Pushshift under CC0; Voat, Bluesky, and Koo under CC BY 4.0 International). The MADOC dataset is released under CC BY 4.0 International License to ensure broad accessibility while maintaining attribution requirements.

The dataset includes offensive content present in post and comment text, URLs attached to posts, and community names and descriptions, which may be harmful to some users. However, the inclusion of this content is necessary for several important research purposes: enabling cross-platform studies of toxic and hateful content propagation, understanding the effectiveness of different moderation approaches, studying community dynamics around controversial topics, and analyzing patterns of harmful behavior across platforms. This content is particularly valuable for researchers studying the evolution and spread of harmful narratives across different social media environments.

We acknowledge several potential risks of misuse of this dataset, including: use of the data to train harmful AI mod-

els or content generation systems, analysis aimed at identifying vulnerable communities or users, attempts to re-identify users across platforms, and exploitation of toxic content patterns for harassment. To mitigate these risks, we have implemented several safeguards. These include clear documentation of appropriate use cases and research guidelines, strict pseudonymization protocols that make re-identification significantly more difficult, release through established academic repository (Zenodo) with usage tracking, and detailed metadata and documentation following FAIR principles.

Despite these risks, we believe the dataset provides significant value for legitimate research purposes. It enables understanding of cross-platform content moderation effectiveness, studying community migration patterns and their impact, developing better detection systems for harmful content, and improving platform design to promote healthier online interactions. We encourage researchers using MADOC to carefully consider these ethical implications in their work and to implement appropriate safeguards in their research designs. Following standard ethical guidelines and making no attempt to re-identify users, we have focused on creating a resource that balances research utility with responsible data stewardship.

Field	**Description**
post_id	Unique identifier for the interaction, anonymized to prevent reconstruction of original URLs
publish_date	UNIX timestamp of the interaction (seconds since epoch)
user_id	Anonymized identifier of the content creator, consistent across all interactions from the same user
parent_id	Identifier of the parent post for comments/replies, NA for original posts
parent_user_id	Identifier of the parent post's creator, NA for original posts
content	Textual content of the interaction, with URLs extracted
url	External URLs referenced in the content, multiple URLs separated by '|'
language	Detected language code (currently 'English' for all entries)
interaction_type	Type of interaction: 'POST', 'COMMENT', or 'REPOST'
platform	Source platform: 'reddit', 'voat', 'bluesky', or 'koo'
community	Community identifier (subreddit/subverse name), NA for Bluesky/Koo
sentiment_vader	VADER sentiment score ranging from -1 (negative) to 1 (positive)
sentiment_textblob	TextBlob sentiment polarity score ranging from -1 (negative) to 1 (positive)
subjectivity_textblob	TextBlob subjectivity score ranging from 0.0 (objective) to 1.0 (subjective)
toxicity_toxigen	ToxiGen toxicity score ranging from 0.00 (non-toxic) to 1.00 (toxic)
strict_filter	Whether content matches strict keyword filtering criteria (TRUE/-FALSE)

Table 2: Structure of the MADOC dataset. Each row represents a single interaction (post, comment, or repost) with the fields standardized across all platforms.

Community	Reddit				Voat			
	Posts	**Comments**	**Users**	**Avg. Sent.**	**Posts**	**Comments**	**Users**	**Avg. Sent.**
funny	4.2M	57.6M	5.5M	0.056	47K	97K	20K	0.001
gaming	2.9M	41.4M	4.1M	0.114	29K	41K	12K	0.093
pics	2.4M	49.5M	4.9M	0.075	15K	24K	9K	0.037
videos	3.1M	33.7M	3.2M	0.054	45K	41K	14K	-0.037
gifs	449K	20.6M	2.8M	0.051	8K	14K	6K	0.029
technology	848K	10.9M	1.3M	0.055	35K	49K	19K	0.029
fatpeoplehate	0	1.4M	79K	0.011	75K	279K	22K	0.002
GreatAwakening	67K	848K	26K	0.079	103K	222K	12K	0.016
MillionDollarExtreme	69K	1.2M	28K	0.008	7K	14K	3K	0.000
CringeAnarchy	195K	6.3M	342K	-0.031	1K	2K	1K	-0.038
KotakuInAction	128K	6.7M	146K	-0.029	3K	5K	2K	-0.028
MensRights	148K	3.0M	185K	-0.076	2K	2K	1K	-0.162

Table 3: Community-level statistics for Reddit and Voat. The first six rows show general-interest communities, while the last six show controversial communities. Average sentiment scores range from -1 (negative) to 1 (positive).