Original Paper

Data Mining Trauma: AI-Assisted Qualitative Study of Cyber Victimization on Reddit

J'Andra Antisdel[1,2*], PhD, RN; Wendy R Miller[1*], PhD, RN; Doyle Groves[1*], BS

[1]Center for Enhancing Quality of Life in Chronic Illness, School of Nursing, Indiana University Indianapolis, Indianapolis, IN, United States
[2]Center for Integrated Healthcare Education, Saint Mary's College, Notre Dame, IN, United States
*all authors contributed equally

Corresponding Author:

J'Andra Antisdel, PhD, RN
Center for Enhancing Quality of Life in Chronic Illness, School of Nursing
Indiana University Indianapolis
600 Barnhill Drive
Indianapolis, IN 46202
United States
Phone: 1 574-703-4472
Email: jalantis@iu.edu

Abstract

Background: Cyber victimization exposes individuals to numerous risks. Developmental and psychological factors may leave some users unaware of the potential dangers, increasing their susceptibility to psychological distress. Despite this vulnerability, methods for identifying those at risk of cyber victimization within health care settings are limited, as is research that explores their experiences of cyber victimization. The purpose of this study was to analyze how users describe experiences of cyber victimization on the social media platform Reddit (Reddit, Inc) using data mining.

Objective: This study aimed to analyze and describe how users on Reddit describe and discuss their experience of cyber victimization using data mining and computational analysis of unsolicited data.

Methods: This computational qualitative study used data mining, Word Adjacency Graph (WAG) modeling, and thematic analysis to analyze discussions of Reddit users surrounding cyber victimization. Inclusion criteria included posts from 2012 to 2023 from subreddits r/cyberbullying and r/bullying. GPT-4 (OpenAI), an advanced artificial intelligence language model, summarized posts and assisted in cluster labeling. Posts were reviewed to remove irrelevant content and duplicates. User anonymity was maintained throughout the study.

Results: A total of 13,381 posts from 3283 Reddit were analyzed, with approximately 5.1% (n=678) originating between 2012 and 2018 and 94.9% (n=12,703) from 2019 to 2023. The WAG modeling approach identified 38 clusters, with 35 deemed to be relevant to cyber victimization experiences. Two clusters containing irrelevant material were excluded. Six overarching themes emerged: (1) psychological impact, (2) coping and healing, (3) protecting yourself online, (4) protecting yourself offline, (5) victimization across various settings, and (6) seeking meaning and understanding.

Conclusions: The study highlights the effectiveness of data mining and AI in analyzing large public datasets for qualitative research. These methods can inform future studies on risky internet behavior, victimization, and assessment strategies in health care settings.

JMIR Infodemiology 2025;5:e75493; doi: 10.2196/75493

Keywords: cyber victimization; word adjacency graphing; cyberbullying; artificial intelligence; data mining; thematic analysis

Introduction

Cyber victimization refers to harmful experiences that occur through the internet, social media, or communication devices. These experiences are often psychologically distressing and have been linked to depression, anxiety, self-harm, and suicidal ideation [1-5].

Although well-documented as a public health concern [6,7], there are limitations in the research, including a lack of studies exploring cyber victimization from the perspectives of those who experience it. Traditional qualitative studies

have contributed valuable insights into cyber victimization experiences [8,9] but are often limited by small sample sizes and researcher bias [10]. In addition, instruments to detect cyber victimization often have different methods for operationalization, using specific terms or providing brief descriptions of acts or experiences [11]. The terminology researchers use to define cyber victimization also may not align with the individual perceptions of the experience. For example, in 1 study [12], users provided more reliable responses when the term "cyber victimization" was used rather than "cyberbullying," suggesting that language choices influence how participants relate to research prompts.

Data mining and computational qualitative analysis, which involves using computer algorithms and software to collect and analyze qualitative data, is a novel method for understanding cyber victimization. This method has been successfully used to study various public health issues such as substance abuse [13], epilepsy [14], and intimate partner violence [15].

By mining data from social media sites, researchers can access large-scale data that reflect participants' genuine, organic thoughts, feelings, and discussions. Analyzing this user data can reveal patterns and trends that are not apparent using traditional research methods. This information can then be used to inform the development and implementation of interventions tailored to a specific population's unique experiences and needs.

In this study, we applied data mining and computational qualitative analysis to explore how individuals describe and discuss their experience of cyber victimization on the social media platform Reddit (Reddit, Inc). Our aim was to identify patterns and themes in unsolicited narratives. The findings of this study will inform future interventions and improve methods for identifying and supporting individuals who experience cyber victimization.

Methods

Study Design Overview

This qualitative computational analysis used data mining and Word Adjacency Graph (WAG) modeling [16] to examine cyber victimization narratives shared by users on Reddit. Data were collected from 2 subreddits, r/cyberbullying and r/bullying, over an 11-year period (2012-2023) to capture relevant trends and patterns in discussions surrounding cyber victimization. A systematic data extraction process was conducted using Reddit's application programming interface (API) and a custom web-scraping tool to gather posts and comments. After data cleaning to remove irrelevant duplicates and bot-generated content, WAG modeling was applied to identify patterns and thematic clusters within the text. Following cluster identification, GPT-4 (OpenAI) was used to generate preliminary labels and summaries, which were then manually reviewed for accuracy. A keyword searching process was also conducted to account for evolving language, slang, and abbreviations. Finally, a thematic analysis was performed to refine clusters into relevant themes.

Study Setting and Population

Reddit is a widely used social media and news aggregate platform with over 50 million active users [17]. It is known for its anonymous membership, allowing users to share their thoughts and opinions without revealing their identity. The platform is divided into over 10,000 "subreddits" with a wide range of topics, from current events and politics to hobbies and interests [18].

Although there are approximations regarding the age demographics of Reddit [19], Reddit does not collect demographic data, making it impossible to determine the exact demographic of users. Despite this limitation, Reddit has been used in previous research studies, providing valuable insights into mental health topics [20-22]. As demographic information cannot be verified, this study uses the term "users" to describe individuals who authored posts and comments. The study targeted subreddits r/cyberbullying and r/bullying for data mining, as these communities encourage personal discussions of cyber victimization. Data extraction focused on titles, post bodies, and comments from 2012-2023.

Data Mining and Computational Analysis

This study used data mining and WAG modeling [16] to examine discussions on Reddit about cyber victimization. Data mining was first conducted, followed by WAG modeling, to reveal patterns and relationships between words and concepts and identify common themes within the text. Data mining consultations were conducted to enhance the validity and to confirm that the process was in alignment with the aim of the study, which was to analyze how users on Reddit describe and discuss their experience of cyber victimization. Following a systematic approach [23], data mining comprised 6 stages: collecting requirements, data investigation, data collection, modeling, assessment, and presentation [24-30]. These are outlined in Multimedia Appendix 1. A visual overview of the full methodological process is provided in Figure 1, and the topic detection process using WAG modeling is illustrated in Figure 2.

Figure 1. Overview of the methodological process. AI: artificial intelligence; API: application programming interface; WAG: Word Adjacency Graph;

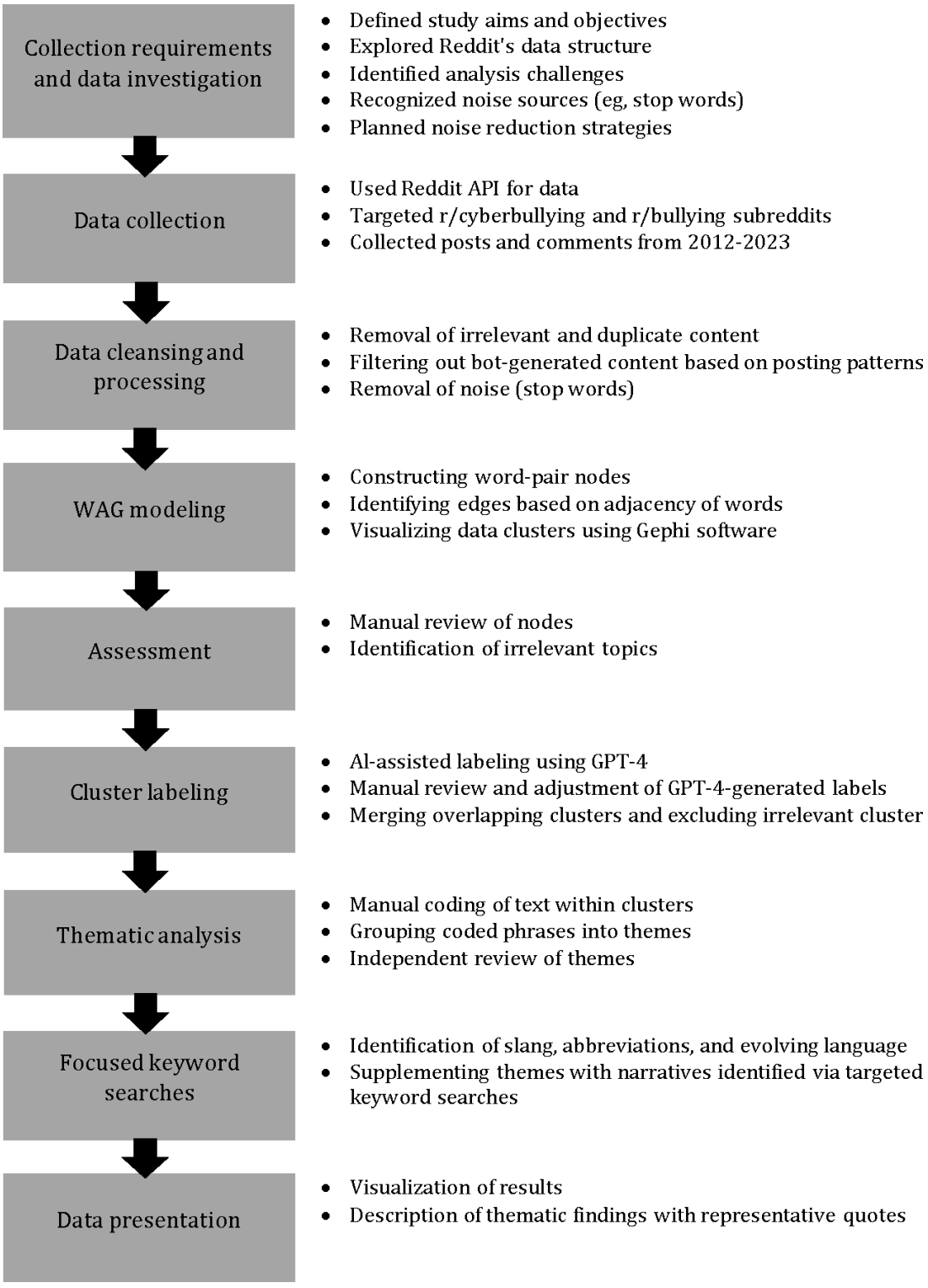
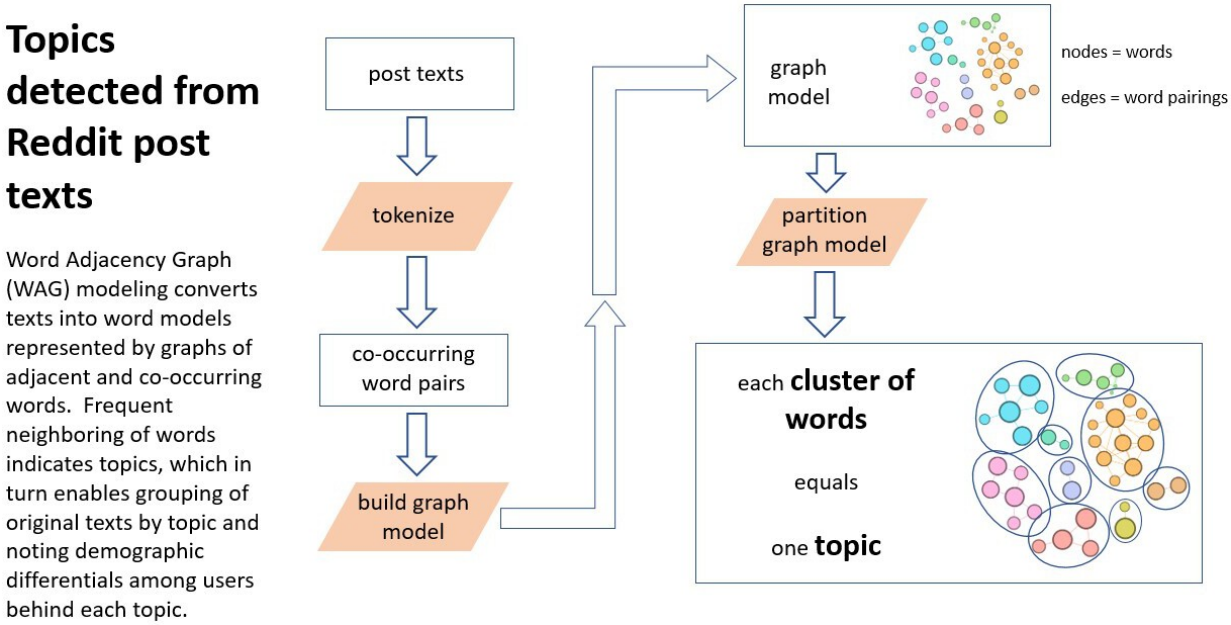


Figure 2. Topic detection via Word Adjacency Graph modeling.



Thematic Analysis of Clusters

Following labeling, clusters were thematically analyzed with MAXQDA 2022 (VERBI Software GmbH), a qualitative data management software program [29], to organize clusters into overlapping themes. The full dataset was organized and prepared in a Microsoft Excel document, arranged by cluster, as well as into categories of weak clusters and those not fitting into any cluster. Each post or comment within the clusters was assigned an individual identification number. For example, when a post or comment is the 61st data point within cluster 4, it would be labeled as "C4-61."

Posts and comments organized by cluster were transcribed into MAXQDA and coded with key phrases based on the content and context. Phrases that were similar in context were grouped and organized into themes. Following the categorization of the clusters into themes, another researcher independently reviewed these themes to ensure accuracy and consistency. This review process involved a thorough examination of how the themes were derived, ensuring that they accurately reflected the key phrases and context from the original posts and comments. The reviewer also assessed the alignment of the themes with the overall objectives of the study, adjusting where necessary to address any discrepancies or oversights.

Ethical Considerations

This study was reviewed by the Institutional Review Board at Indiana University and determined to be exempt under category 4(i): publicly available information or specimens (Protocol #18415; initial approval February 28, 2023). The exemption was granted because the research involved analysis of publicly accessible Reddit posts without direct interaction with human subjects. Informed consent was not sought, as the data were unsolicited, publicly available, and collected in accordance with established guidelines for internet-based research on publicly accessible content without user interaction, as outlined by Eysenbach and Till [31]. To protect privacy and confidentiality, no usernames, profile information, or other potentially identifying details were stored or reported, and example quotes were paraphrased when necessary to minimize traceability via search engines. All electronic data were collected and stored on encrypted devices. Data collection complied with Reddit's API access policies [32]. Funding for this research was made possible (in part) by Grant Number 5H79SM080386-05 from the Substance Abuse and Mental Health Services Administration (SAMHSA).

Results

Overview

This study successfully applied data mining, WAG modeling, GPT-4–assisted labeling, and thematic analysis to examine cyber victimization narratives shared by users on Reddit. The extracted dataset comprised 13,381 posts and comments from 3283 unique Reddit users. Approximately 5.1% (n=678) of the posts were posted between 2012 and 2018. The remaining 94.9% (n=12,703) of the posts were posted to Reddit from 2019-2023.

To construct the WAG model, only words and word pairs appearing more than 10 times were included, resulting in 150 unique words and 123 word pairs for analysis. As a result, 15% (n=1965) of posts and comments were strongly categorized, 62% (n=8290) were weakly categorized, and 23% (n=2984) were not fitted into the model. The posts that were strongly categorized formed the basis of the 38 clusters (Figure 3), which were then visualized using Gephi (Gephi Consortium). In this visualization, each node represents a word pair, with node size reflecting word frequency and proximity to similar words. Of the 38 clusters, 35 were relevant to cyber victimization experiences, while 2 clusters were excluded due to irrelevance. In addition, the WAG

modeling process placed the word pairs of “law enforcement” and “legal action” into 2 clusters, but due to their overlapping content, they were merged into a single cluster (Figure 4).

The distribution of posts and comments across clusters varied, with some clusters containing a high volume of discussion while others had minimal engagement (Multimedia Appendix 2). Clusters 1 and 2 had the highest number of posts and comments, followed by clusters 3 and 0. Some clusters contained fewer than 20 posts and comments, indicating less frequent discussion of those themes. A full list of cluster labels and their thematic categorization is provided in Multimedia Appendix 3.

Figure 3. Word Adjacency Graph results cluster numbers.

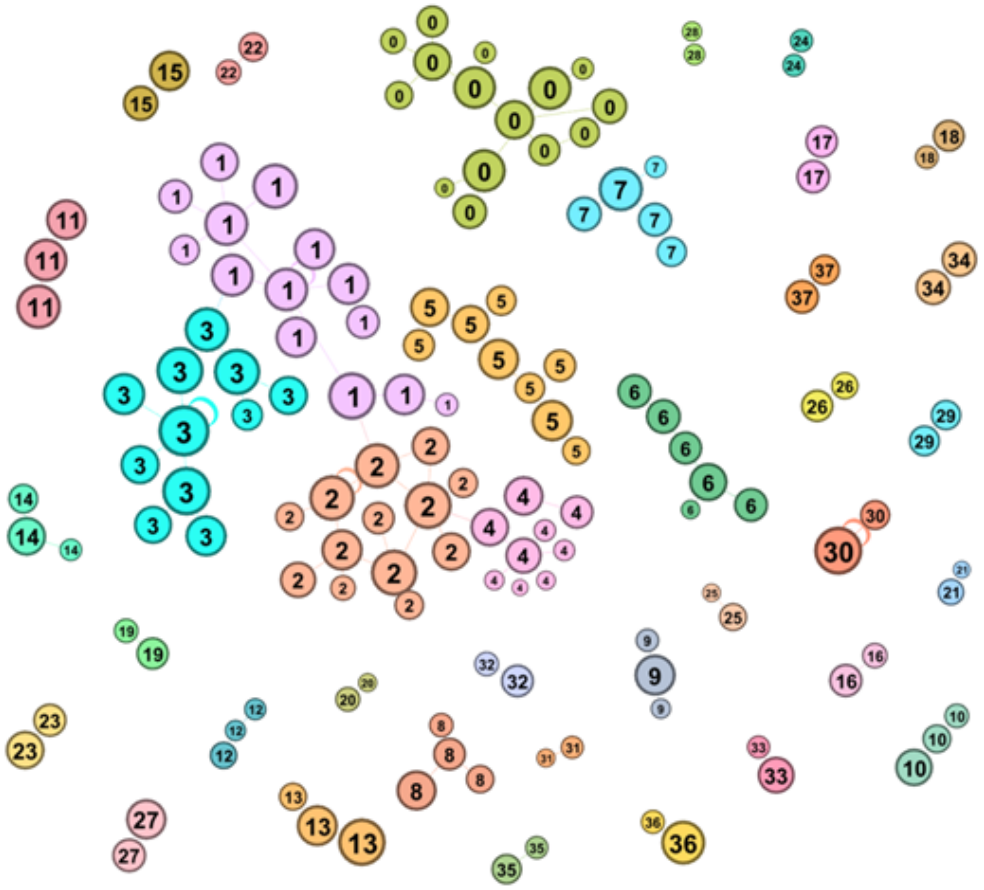
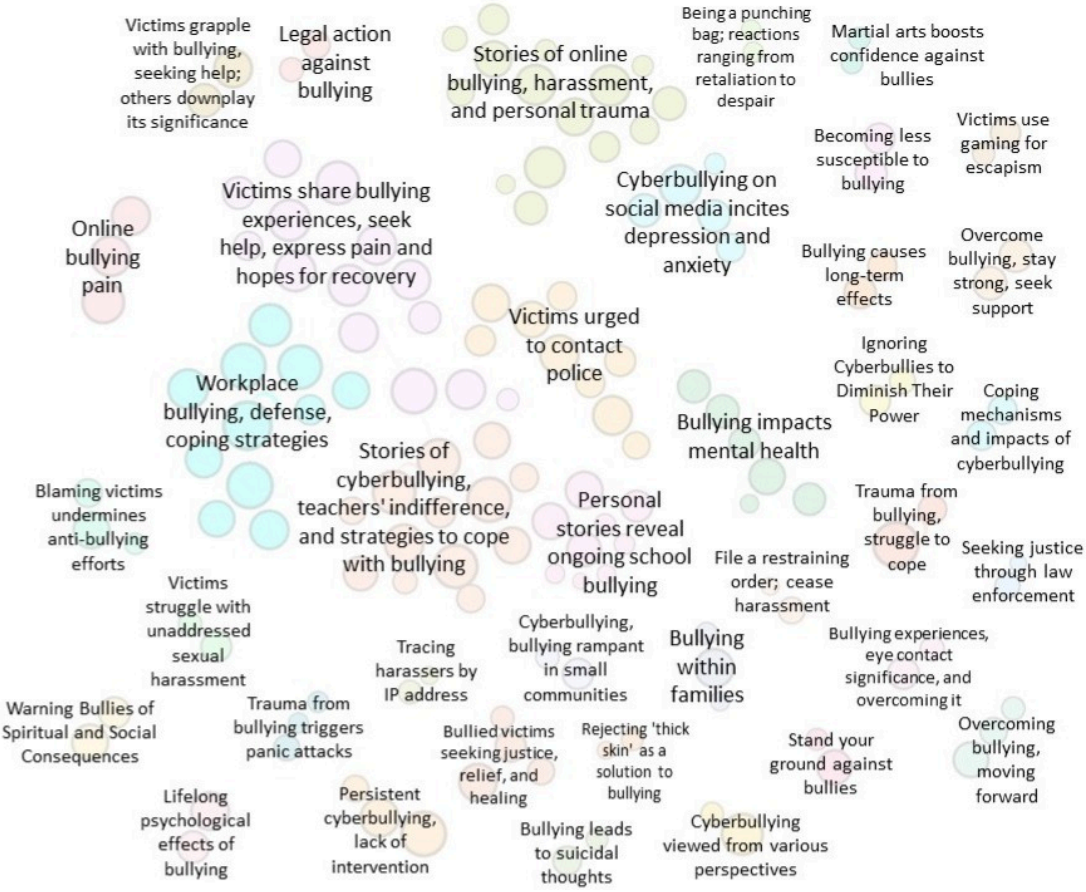


Figure 4. Word Adjacency Graph results cluster labels.



Cluster Labeling and Validation

All Clusters were manually reviewed and excluded from the model in instances in which users were identified as automated bots or where posts consisted of research recruitment by researchers. The analysis involved screening for data skewness caused by the overrepresentation of individual users. This occurred in situations where users repeatedly posted similar content, which artificially inflated the presence of certain themes or topics in the dataset. Such repetitive postings, not indicative of genuine user interactions, were identified and excluded to ensure the integrity and representativeness of the data.

In analyzing online interactions and narratives surrounding cyber victimization, GPT-4 was initially used to label the resulting 38 clusters (see Multimedia Appendices 2–4). GPT-4 synthesized a subset of posts and comments from each cluster to generate suggestive labels. The mean number of posts per cluster was 22 (SD 1) , with a typical range of 13-30 posts; the smallest cluster included 7 posts, and the largest contained 44. These sample sizes were shaped by technical limitations at the time of the study (2023) when the GPT-4 model's context window restricted how much data could be processed at once. At that time, approximately 819 (40%) posts of the total cluster data were used in GPT-4–assisted label generation.

Each suggested label was subsequently evaluated by a human reviewer. The manual review process relied on five validation categories: (1) retained, (2) adjusted, (3) revised, (4) merged, and (5) excluded, which were applied based on how closely the GPT-4 suggested label aligned with the cluster data. As a result, 12 labels were retained with only minimal word changes, 12 labels were adjusted, meaning the suggested label generally matched the content but required rewording to improve clarity or reflect the data more accurately, and 10 labels were completely revised to align with the content of the posts. Two clusters were merged due to thematic overlap, and 2 were excluded from the analysis due to irrelevance (see Multimedia Appendix 4).

To enhance interpretive rigor, a secondary researcher reviewed all cluster labels and confirmed their alignment with the underlying post content and resulting themes. While formal interrater reliability metrics were not calculated, consensus was achieved through collaborative review and discussion.

Overall, GPT-4 provided a usable starting point for 36 out of 38 clusters (94.7%), with 26 clusters (68%) requiring some degree of human refinement. Definitions and illustrative examples of each validation category are provided in Multimedia Multimedia Appendix 5.

Focused Keyword Searching

Focused keyword searching was used to expand the analysis of clustered data and improve the identification of themes. This approach involved reviewing the narratives within each cluster to identify key concepts, expressions, or themes. From

these, related terms and alternative phrasings were generated and used as targeted search terms across the dataset. For example, a narrative discussing a specific form of cyberbullying, like blackmail, would lead to the identification of related terms such as "blackmail" and "catfish" (Multimedia Appendix 6). This strategy allowed for the identification of narratives that were weakly clustered or not clustered at all due to variations in words or phrasing. The total number of posts retrieved via focused keyword searching was not tracked. However, the process did lead to the identification of additional narratives that were thematically aligned but not strongly categorized. These supplemental posts helped confirm existing themes and broader representation of experiences.

Thematic Analysis

Following labeling, clusters were thematically analyzed with MAXQDA 2022, a qualitative data management software program [29], to organize clusters into overlapping themes. Posts and comments organized by cluster were transcribed into MAXQDA and read and coded with key phrases based on the content and context. Phrases that were similar in context were grouped and organized into themes.

As a result of this analysis, six themes emerged: (1) psychological impact, which examines the symptoms of cyber victimization; (2) coping and healing, focusing on healing and overcoming cyber victimization and seeking support; (3) protecting yourself online, highlighting methods for preventing or stopping cyber victimization; (4) protecting yourself offline, detailing methods to decrease the risk of being targeted in the physical world; (5) victimization across various settings, exploring the dynamics of victimization in different environments; and (6) seeking meaning and understanding, which includes philosophical discussions about the nature of victimization. A summary table outlining each theme with supporting subtopics is provided in Multimedia Appendix 7. A detailed thematic analysis of the qualitative findings will be presented in a separate publication where the themes will be explored in-depth, along with direct quotes and case examples.

Discussion

Principal Findings

This study identified 6 overarching themes in Reddit posts related to cyber victimization: psychological impact, coping and healing, protecting oneself online, protecting oneself offline, victimization across various settings, and seeking meaning and understanding.

By following a structured, hybrid analytic process [Figure 1] combining data mining, WAG modeling, and GPT-4–assisted labeling, this study demonstrates the effectiveness of computational qualitative methods in analyzing large-scale, unsolicited data on cyber victimization. This approach addressed the limitations of traditional qualitative research, especially in the context of handling large amounts of unstructured data. In addition, traditional qualitative research methods are often limited by participant selection biases and social desirability biases, which limit the breadth and depth of the narratives [10]. The anonymity of Reddit also encouraged users to share sensitive information without fear of stigma. This approach facilitated the identification of patterns and emerging themes that may not have been captured through manual coding alone.

This study provided a novel methodological approach for examining cyber victimization experiences and highlights the potential for AI-assisted qualitative analysis. GPT-4 was used in initial cluster labeling, which, when combined with manual review, improved the accuracy of thematic categorization. This builds on previous qualitative research by demonstrating how computational tools can be applied to analyze unsolicited narratives, reduce researcher bias, and identify themes across a vast dataset.

While hybrid computational-qualitative methods have been applied to topics such as substance abuse [13], epilepsy [14], and intimate partner violence [15], this study extends that work by applying similar techniques to cyber victimization. By doing so, it demonstrates the adaptability of WAG modeling and GPT-4–assisted labeling to new areas of public health. Recent studies support this hybrid approach. For example, Piper and Wu [33] found that large language models (LLMs) performed well in narrative topic labeling, while Castellanos et al [34] demonstrated that although GPT-4 generated themes aligned with human coding in over 79% of cases, human coders were still required for accuracy. Our integration of GPT-4–assisted labeling with manual review aligns with these findings and demonstrates the need for human oversight.

The findings have practical implications for health care settings. Recognizing the diverse ways users describe psychological impacts and coping strategies could inform the development of educational resources, screening instruments, and assessment strategies that reflect the language and experiences of victims. These resources would be valuable to health care professionals in identifying individuals at risk in primary care or mental health settings where cyber victimization may go unreported.

Limitations

While this study has many strengths, it is not without limitations. Our data consisted of anonymous user narratives from Reddit, making it challenging to determine the generalizability of the sample to a wider population. Reddit users also may have specific characteristics, interests, or behaviors that are not reflective of a broader population [35]. In addition, social media platforms have distinct user demographics and cultures, which might influence the nature and extent of cyber victimization experienced by users. For example, Reddit users are more likely to be men [36,37], while TikTok (ByteDance Ltd), Facebook (Meta Platforms), and Pinterest (Pinterest, Inc) users are more likely to be women [37].

Due to the inability to follow up with users for explanation and clarification and the inability to meet users face-to-face

to assess body language, vocal tones, and facial expressions, there was a potential for misinterpretation of context [38]. The anonymous nature of Reddit may also lead to false answers or misinformation, as internet-based platforms can be prone to exaggeration, false claims, or recall bias. In addition, demographic inferences cannot be made, as users are not required to disclose personal information.

An important limitation of this research is the potential inclusion of bot-generated content [39]. Despite efforts to identify and exclude bots based on patterns in posting behavior, timing, frequency, language use, and identifiable usernames, the sophisticated nature of some bots may have allowed them to bypass detection. It is possible that bot-generated posts, which do not reflect human experiences, were incorporated into the dataset, potentially influencing the results.

Another limitation relates to the evolving nature of LLMs. GPT-4, the model used at the time of this study (2023), had a significantly smaller context window, which limited the number of posts that could be processed per cluster. As a result, only a subset of cluster content was used for label identification. Newer versions of the model support much larger input sizes, which may produce different results, affecting replication in future studies.

Finally, while GPT-4 accelerated the qualitative analysis process, several limitations must be acknowledged. LLMs are prone to selective summarization and misrepresentation, known as hallucination [40]. LLMs may also simplify content while being overconfident in tone, which can influence researchers' judgment by making inaccurate or biased content appear more credible than it is. These limitations may have affected the accuracy of cluster label identification. Manual validation was used to mitigate these risks. However, the reliance on GPT-4 for suggestive labeling remains a methodological limitation worth noting.

Implications for Further Research

Future research could expand this methodology to further explore and deepen the understanding of cyber victimization experiences and further refine computational qualitative analysis techniques. While this study focused on cyber victimization experiences in 2 subreddits, other communities may provide further insight into cyber victimization. Future studies could extend data mining and WAG modeling to specific types of cyber victimization-related subreddits such as r/stalking, r/cyber security, and r/scams, which focus on distinct aspects of harmful experiences. Cyber victimization is broad; these specific subreddits could provide a more nuanced understanding of how different forms of cyber victimization are discussed within internet-based communities.

To explore how cyber victimization experiences vary across internet-based spaces with different user demographics and privacy structures, future research could compare narratives from different platforms (TikTok [ByteDance Ltd], Discord [Discord Inc], Bluesky [Bluesky PBLLC], Instagram [Meta Platforms], and Tumblr [Automattic]). Each platform has unique privacy settings, user demographics, and moderation policies, which may influence how users discuss and experience cyber victimization.

Conclusions

This study used a hybrid methodological approach to analyze how users on Reddit describe their experience of cyber victimization using data mining and computational analysis of unsolicited data. By leveraging data mining and WAG modeling, this study demonstrated the effectiveness of computational methods in qualitative analysis. GPT-4-assisted labeling and focused keyword searching further refined thematic identification, resulting in 6 themes: psychological impact, coping and healing, protecting oneself online, protecting oneself offline, victimization across various settings, and seeking meaning and understanding. The methodological approach demonstrated in this study will be valuable to data scientists and health care researchers seeking to analyze social media data on mental health issues. These methods can inform future studies on risky internet behavior, victimization, and assessment strategies in health care settings.

Acknowledgments

Funding for this research was made possible (in part) by Grant Number 5H79SM080386-05 from SAMHSA. The views expressed in written training materials or publications and by speakers and moderators do not necessarily reflect the official policies of the Department and Human Services; nor does mention of trade names, commercial practices, or organizations imply endorsement by the U.S. Government.

The author would like to acknowledge the contributions of Indiana University School of Nursing for its support in facilitating this research. Special thanks to Dr. Ukamaka Oruche for their guidance and feedback throughout the development of this study.

Data Availability

The datasets generated or analyzed during this study are not publicly available due the sensitive personal narratives of cyber victimization but are available from the corresponding author on reasonable request.

Authors' Contributions

JA took the lead in conceptualization, with equal contributions from WRM. Data curation was primarily carried out by JA, with equal contributions from DG. Formal analysis was led by JA, with equal contributions from DG and WRM. Funding acquisition was led by JA, with supporting contributions from WRM. Investigation was led by JA, with supporting contributions from WRM and DG. Methodology was led by JA, with equal contributions from DG and WRM. Project administration was led by WRM, with equal contributions from JA. Resources were provided by WRM, with equal contributions from JA.

Software was not applicable. Supervision and validation were carried out by WRM. Visualization was completed by DG. The original draft was written primarily by JA, with supporting contributions from WRM and DG. Review and editing were led by JA, with equal contributions from WRM and DG.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Detailed methodological process for data mining.
[DOCX File (Microsoft Word File), 22 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Number of posts and comments per cluster.
[PNG File (Portable Network Graphics File), 27 KB-Multimedia Appendix 2]

Multimedia Appendix 3

List of cluster labels and thematic categorization.
[DOCX File (Microsoft Word File), 19 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Manual Review Outcomes for GPT-4 Generated Labels
[DOCX File (Microsoft Word File), 17 KB-Multimedia Appendix 4]

Multimedia Appendix 5

Definitions and examples of GPT-4 label validation categories.
[DOCX File (Microsoft Word File), 15 KB-Multimedia Appendix 5]

Multimedia Appendix 6

Example analysis and narrative linkages.
[DOCX File (Microsoft Word File), 18 KB-Multimedia Appendix 6]

Multimedia Appendix 7

Summary of identified themes from thematic analysis.
[DOCX File (Microsoft Word File), 17 KB-Multimedia Appendix 7]

References

1. Cénat JM, Smith K, Hébert M, Derivois D. Cybervictimization and suicidality among French undergraduate Students: A mediation model. J Affect Disord. Apr 15, 2019;249:90-95. [doi: 10.1016/j.jad.2019.02.026] [Medline: 30769296]
2. Li Y, Li D, Li X, et al. Cyber victimization and adolescent depression: The mediating role of psychological insecurity and the moderating role of perceived social support. Child Youth Serv Rev. Nov 2018;94:10-19. [doi: 10.1016/j.childyouth.2018.09.027]
3. Rose CA, Tynes BM. Longitudinal associations between cybervictimization and mental health among U.S. adolescents. J Adolesc Health. Sep 2015;57(3):305-312. [doi: 10.1016/j.jadohealth.2015.05.002] [Medline: 26115909]
4. John A, Glendenning AC, Marchant A, et al. Self-harm, suicidal behaviours, and cyberbullying in children and young people: systematic review. J Med Internet Res. Apr 19, 2018;20(4):e129. [doi: 10.2196/jmir.9044] [Medline: 29674305]
5. Sampasa-Kanyinga H, Roumeliotis P, Xu H. Associations between cyberbullying and school bullying victimization and suicidal ideation, plans and attempts among Canadian schoolchildren. PLoS ONE. 2014;9(7):1-9. [doi: 10.1371/journal.pone.0102145] [Medline: 25076490]
6. House Committee on Energy and Commerce Subcommittee on Consumer Protection and Commerce (Committee on Energy and Commerce). Kids online during COVID: Child safety in an increasingly digital age.. 2021. URL: https://docs.house.gov/Committee/Calendar/ByEvent.aspx?EventID=111298 [Accessed 2025-08-13]
7. Sasson H, Mesch G. Parental mediation, peer norms and risky online behavior among adolescents. Comput Human Behav. Apr 2014;33:32-38. [doi: 10.1016/j.chb.2013.12.025]
8. Radovic A, Gmelin T, Stein BD, Miller E. Depressed adolescents' positive and negative use of social media. J Adolesc. Feb 2017;55:5-15. [doi: 10.1016/j.adolescence.2016.12.002] [Medline: 27997851]
9. Blankenship RJ, St. Surin O. Silent voices: the perception of cyberbullying among at-risk middle school students. Int J Cyber Behav Psychol Learn. Oct 2019;9(4):1-21. [doi: 10.4018/IJCBPL.2019100101]

10. Queirós A, Faria D, Almeida F. Strengths and limitations of qualitative and quantitative research methods. Zenodo; Sep 7, 2017. [doi: 10.5281/ZENODO.887089]
11. Menesini E, Nocentini A. Cyberbullying definition and measurement: Some critical considerations. J Psychol. Jan 2009;217(4):230-232. [doi: 10.1027/0044-3409.217.4.230]
12. Akbulut Y, Sahin YL, Eristi B. Development of a scale to investigate cybervictimization among online social utility members. Contemp Educ Technol. 2010;1(1). [doi: 10.30935/cedtech/5961]
13. Lu J, Sridhar S, Pandey R, Hasan MA, Mohler G. Investigate transitions into drug addiction through text mining of Reddit data. Presented at: KDD '19: The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining; Aug 4-8, 2019:2367-2375; Anchorage AK USA. [doi: 10.1145/3292500.3330737]
14. Miller WR, Gesselman AN, Garcia JR, Groves D, Buelow JM. Epilepsy-related romantic and sexual relationship problems and concerns: indications from internet message boards. Epilepsy Behav. Sep 2017;74:149-153. [doi: 10.1016/j.yebeh.2017.06.023] [Medline: 28756337]
15. Sivagurunathan M, Walton DM, Packham T, Booth RG, MacDermid JC. Discourses around male IPV related systemic biases on Reddit. J Interpers Violence. Oct 2022;37(19-20):NP17834-NP17859. [doi: 10.1177/08862605211030015] [Medline: 34251276]
16. Miller WR, Groves D, Knopf A, Otte JL, Silverman RD. Word adjacency graph modeling: separating signal from noise in big data. West J Nurs Res. Jan 2017;39(1):166-185. [doi: 10.1177/0193945916670363] [Medline: 27655959]
17. Reddit Inc. Reddit content policy. 2018. URL: https://www.redditinc.com/policies/content-policy [Accessed 2023-11-22]
18. Reddit Inc. Reddit homepage. 2023. URL: https://www.redditinc.com/ [Accessed 2023-11-22]
19. Dixon S. US reddit app users by age 2021. Statista; 2022. URL: https://www.statista.com/statistics/1125159/reddit-us-app-users-age/ [Accessed 2023-01-01]
20. Sowles SJ, McLeary M, Optican A, et al. A content analysis of an online pro-eating disorder community on Reddit. Body Image. Mar 2018;24:137-144. [doi: 10.1016/j.bodyim.2018.01.001] [Medline: 29414146]
21. Arya S, Nagappala S, Krawczyk N, Gi Y, Meacham MC, Bunting AM. Fentanyl in pressed oxycodone pills: a qualitative analysis of online community experiences with an emerging drug trend. Subst Use Misuse. 2022;57(13):1940-1945. [doi: 10.1080/10826084.2022.2120365] [Medline: 36106770]
22. Overbeek D, Janke A. 360 characteristics of posts of opioid users on Reddit, an online social media forum, an area for improved harm reduction. Ann Emerg Med. Oct 2018;72(4):S142. [doi: 10.1016/j.annemergmed.2018.08.365]
23. Ogunleye JO. The concept of data mining. In: Data Mining: Concepts and Applications. IntechOpen; 2022. URL: https://directory.doabooks.org/handle/20.500.12854/90223 [Accessed 2023-07-31] ISBN: 978-1-83969-267-3
24. Manning CD, Raghavan P, Schütze H. Introduction to Information Retrieval. Cambridge University Press; 2008. ISBN: 978-0-521-86571-5
25. Jhaver S, Birman I, Gilbert E, Bruckman A. Human-machine collaboration for content regulation: The case of Reddit automoderator. ACM Trans Comput-Hum Interact. Jul 19, 2019;26(5):1-35. [doi: 10.1145/3338243]
26. Lebeuf C, Storey MA, Zagalsky A. Software bots. IEEE Softw. Jan 2018;35(1):18-23. [doi: 10.1109/MS.2017.4541027]
27. Waltman L, van Eck NJ. A smart local moving algorithm for large-scale modularity-based community detection. Eur Phys J B. Nov 2013;86(11):471. [doi: 10.1140/epjb/e2013-40829-0]
28. ChatGPT [large language model]. OpenAI. 2023. URL: https://chat.openai.com/chat [Accessed 2023-08-11]
29. MAXQDA 2020 [computer software]. VERBI Software. 2019. URL: https://www.maxqda.com/ [Accessed 2025-08-06]
30. Bastian M, Heymann S, Jacomy M. Gephi: An open source software for exploring and manipulating networks. ICWSM. 2009;3(1):361-362. URL: https://gephi.org/ [doi: 10.1609/icwsm.v3i1.13937]
31. Eysenbach G, Till JE. Ethical issues in qualitative research on internet communities. BMJ. Nov 10, 2001;323(7321):1103-1105. [doi: 10.1136/bmj.323.7321.1103] [Medline: 11701577]
32. Reddit Inc. User agreement. 2020. URL: https://www.redditinc.com/policies/user-agreement-october-15-2020 [Accessed 2023-11-22]
33. Piper A, Wu S. Evaluating large language models for narrative topic labeling. Presented at: Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities; Albuquerque, USA. 2025.URL: https://aclanthology.org/2025.nlp4dh-1 [doi: 10.18653/v1/2025.nlp4dh-1.25]
34. Castellanos A, Jiang H, Gomes P, Vander Meer D, Castillo A. Large language models for thematic summarization in qualitative health care research: comparative analysis of model and human performance. JMIR AI. Apr 4, 2025;4(1):e64447. [doi: 10.2196/64447] [Medline: 40611510]
35. Proferes N, Jones N, Gilbert S, Fiesler C, Zimmer M. Studying Reddit: a systematic overview of disciplines, approaches, methods, and ethics. Soc Media Soc. Apr 2021;7(2):20563051211019004. [doi: 10.1177/20563051211019004]
36. Amaya A, Bach R, Keusch F, Kreuter F. New data sources in social science research: things to know before working with Reddit data. Soc Sci Comput Rev. Oct 2021;39(5):943-960. [doi: 10.1177/0894439319893305]

37. Pew Research Center. Social media fact sheet. 2021. URL: https://www.pewresearch.org/internet/fact-sheet/social-media/ [Accessed 2023-12-03]
38. Richard B, Sivo SA, Ford RC, et al. A guide to conducting online focus groups via Reddit. Int J Qual Methods. Jan 2021;20:16094069211012217. [doi: 10.1177/16094069211012217]
39. Storozuk A, Ashley M, Delage V, Maloney EA. Got bots? Practical recommendations to protect online survey data from bot attacks. TQMP. 2020;16(5):472-481. [doi: 10.20982/tqmp.16.5.p472]
40. Chelli M, Descamps J, Lavoué V, et al. Hallucination rates and reference accuracy of ChatGPT and Bard for systematic reviews: comparative analysis. J Med Internet Res. May 22, 2024;26:e53164. [doi: 10.2196/53164] [Medline: 38776130]

Abbreviations

AI: artificial intelligence
API: application programming interface
GPT-4: Generative pre-trained transformer
LLM: Large language model
SAMHSA: Substance Abuse and Mental Health Services Administration
WAG: Word Adjacency Graph

Edited by Michael Haupt; peer-reviewed by Ankit Gupta, Ravi Teja Potla, Sadhasivam Mohanadas, Song-Bin Guo; submitted 04.04.2025; final revised version received 07.07.2025; accepted 20.07.2025; published 03.09.2025

Please cite as:
Antisdel J, Miller WR, Groves D
Data Mining Trauma: AI-Assisted Qualitative Study of Cyber Victimization on Reddit
JMIR Infodemiology 2025;5:e75493
URL: https://infodemiology.jmir.org/2025/1/e75493
doi: 10.2196/75493