

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

J'Andra Antisdel, Wendy R Miller, Doyle Groves

Submitted to: JMIR Infodemiology
on: April 04, 2025

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript..... 5

Supplementary Files..... 19

 Figures 20

 Figure 1..... 21

 Figure 2..... 22

 Figure 3..... 23

 Figure 4..... 24

 Figure 5..... 25

 Figure 6..... 26

 Multimedia Appendixes 27

 Multimedia Appendix 1..... 28

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

J'Andra Antisdel^{1, 2*} RN, PhD; Wendy R Miller^{1*} RN, PhD; Doyle Groves^{1*} BS

¹Center for Enhancing Quality of Life in Chronic Illness School of Nursing Indiana University Indianapolis Indianapolis, US

²1001 Regina Hall Center for Integrated Healthcare Education Saint Mary's College Notre Dame US

*these authors contributed equally

Corresponding Author:

J'Andra Antisdel RN, PhD

Center for Enhancing Quality of Life in Chronic Illness

School of Nursing

Indiana University Indianapolis

600 Barnhill Drive,

Indianapolis,

US

Abstract

Background: Cyber victimization exposes teens to numerous risks. Their developmental stage often leaves them unaware of potential dangers, making them susceptible to psychological distress. Despite this vulnerability, methods for identifying teens at risk of cyber victimization within healthcare settings are limited, as is research that explores their experiences of cyber victimization. The purpose of this study was to analyze how teens describe experiences of cyber victimization on the social media platform Reddit using data mining.

Objective: This study aimed to analyze and describe how teens 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 2012-2023 from subreddits r/cyberbullying and r/bullying. GPT-4, 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: 13,381 posts from 3,283 Reddit were analyzed, with 5.07% originating between 2012 and 2018 and 94.93% 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 data sets for qualitative research. These methods can inform future studies on risky internet behavior, victimization, and assessment strategies in healthcare settings.

(JMIR Preprints 04/04/2025:75493)

DOI: <https://doi.org/10.2196/preprints.75493>

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain visible to the public.

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <https://www.jmir.org/preprint/75493>, the full text will be available to the public.

No. Please do not make my accepted manuscript PDF available to anyone. I understand that if I later pay to participate in <https://www.jmir.org/preprint/75493>, the full text will be available to the public.



Original Manuscript

Original Paper

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

Abstract

Background: Cyber victimization exposes teens to numerous risks. Their developmental stage often leaves them unaware of potential dangers, making them susceptible to psychological distress. Despite this vulnerability, methods for identifying teens at risk of cyber victimization within healthcare settings are limited, as is research that explores their experiences of cyber victimization. The purpose of this study was to analyze how teens describe experiences of cyber victimization on the social media platform Reddit using data mining.

Objective: This study aimed to analyze and describe how teens 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 2012-2023 from subreddits r/cyberbullying and r/bullying. GPT-4, 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: 13,381 posts from 3,283 Reddit were analyzed, with 5.07% originating between 2012 and 2018 and 94.93% 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 data sets for qualitative research. These methods can inform future studies on risky internet behavior, victimization, and assessment strategies in healthcare settings.

Keywords: Cyber Victimization, Word Adjacency Graphing, Cyberbullying, AI-Assisted

Introduction

Cyber victimization is a broad term used to describe harmful experiences that occur through the internet, social media, or communication devices and refers to the psychological and physiological distress individuals face due to these experiences. These experiences are important health issues that largely affect teens [1]. In 2022, the Pew Research Center reported that approximately 50% of American teens had experienced some form of cyber victimization [2]. In 2023, the FBI's Internet Crime Complaint Center reported that victims under 20 years old accounted for \$40 million in losses due to cybercrime [3].

Cyber victimization has also been linked to signs of psychological distress [4], such as depression [5], anxiety [6], self-harm behavior, and suicidal ideation [7,8]. The prevalence of cyber victimization and harmful health effects in teens has led political and health leaders to call for the development of interventions to reduce risky online behaviors and cyber victimization [9,10]. However, there are limitations in the research, including a lack of studies exploring cyber victimization from the perspectives of teens who experience it. Also limited are validated instruments designed to detect and study cyber victimization. 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 teens' perception of their experience. For example, in one study, users provided more reliable responses when the term "cyber victimization" was used rather than "cyberbullying [12]." The variations and inconsistencies in the measurement and reporting of cyber victimization among teens further emphasize the need for more studies designed to understand the experiences of this population.

Qualitative studies have been used to understand the specific challenges and needs of teens who have experienced cyber victimization [13,14], and, while beneficial, traditional qualitative study methods are limited by sample size and researcher bias in comparison to methods used in this study [15]. 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 in teens. This method has been successfully used to study various public health issues such as substance abuse [16], epilepsy [17], and intimate partner violence [18].

By mining data from social media sites, researchers can access large-scale data that reflects subjects' 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 aimed to analyze and explore how teens describe and discuss their experience of cyber victimization on the social media platform Reddit using data mining and computational analysis of unsolicited, publicly available data. The findings of this study will influence interventions and strategies to detect experiences of cyber victimization and provide relevant support and treatment for those affected. Additionally, our use of data mining and computational qualitative analysis will provide researchers with new and innovative techniques for collecting information about the mental health experiences of teens.

Methods

Study Design Overview

This qualitative computational analysis used data mining and Word Adjacency Graph (WAG) modeling [19] to examine cyber victimization narratives shared by teens on Reddit. Data was collected from two subreddits, r/cyberbullying and r/bullying, over an 11 period (2012-2013) to capture relevant trends and patterns in discussions surrounding cyber victimization. A systematic data extraction process was conducted using Reddit's API and a custom web-scraping tool to gather posts and comments. After data cleaning to remove irrelevant duplicate and bot-generated content, WAG modeling was applied to identify patterns and thematic clusters within the text. Following Cluster identification, GPT-4 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 commonly used by teens. 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 [20]. 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 [21].

Although it is estimated that approximately 21% of users are between the ages of 10 and 19 [22], 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 [23–25]. 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 [19] to examine cyber victimization discussions on Reddit. 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 the process was in alignment with the aim of the study, which was to analyze how teens on Reddit describe and discuss their experience of cyber victimization. Following a systematic approach [26], data mining comprised six stages: collecting requirements, data investigation, data collection, modeling, assessment, and presentation.

Collection Requirements and Data Investigation

Collecting requirements and data investigation involved defining the needs and the aim of the study and understanding Reddit's structure to identify challenges and limitations in data analysis [26]. This study identified the potential for irrelevant data, or "noise," to influence and potentially mislead the interpretation of results. Examples of noise include stop words such as "with," "the," and "a," which would not contribute meaningfully to the analysis [27].

Another concern was bot-generated content that does not reflect human experiences. Bot-generated content was filtered out by identifying patterns in posting behavior, such as frequency, standardized language, and bot-generated usernames. Bot-generated content refers to posts, comments, or other forms of content that are automatically generated and posted by software applications, known as

bots, instead of human users [28,29].

Data Collection

Data were extracted using access to Reddit's application programming interface (API). Reddit's API grants researchers access to the platform's data, allowing for the extraction of user posts and comments. In addition, a custom web scraping tool was created and specifically designed to collect data relevant to the study from the r/cyberbullying and r/bullying subreddits. The search parameters were set to extract data from 2012-2023 with emphasis on the most upvoted posts within this time frame, as these posts were the most relevant and reflective of the community's experience. This time frame also allows for the inclusion of a diverse range of teenage experiences with the perspectives of older teens who were active on the subreddits during the earlier years. Access to the API was achieved in adherence to Reddit's API access rules.

Modeling

A member of our team (author D.G.) originally developed WAG modeling to identify relationships in text and focus on how words in a text connect and how frequently words and concepts appear near each other [19]. The WAG modeling process constructs a network graph in which words are represented as nodes in a network. Words that are adjacent to each other in the text form connecting edges, allowing for visualization of how words are interconnected, revealing patterns and relationships between different words and concepts that form clusters. These clusters identify common themes within the text [19]. Text from posts and comments are organized into word pairs after removing stop words. The resulting pairs of adjacent words are used to construct the nodes and edges of the WAG model, where nodes represent distinct words or phrases, and edges represent the adjacency relationship between nodes.

Building upon this foundation, this study applies WAG modeling to cyber victimization narratives on Reddit, leveraging the ability to reveal patterns in large sets of unstructured data. The Leiden algorithm [30], a mathematical method for detecting topics or themes in large networks, was used to arrange nodes into clusters. When the Leiden algorithm is applied to these data, it systematically organizes the text entries into clusters based on the frequency and patterns of word adjacency, identifying words and phrases that are commonly used in the context of each other and ultimately revealing underlying themes in the data.

Assessment

The nodes were manually reviewed in the assessment phase to identify relevance to the assigned cluster. Irrelevant topics assigned to a cluster were removed from the model; for example, the term "troll" might initially appear in a cluster due to its common association with harmful online behaviors. However, it may become clear that "troll" referred to fantasy creatures in a gaming context and was irrelevant to the cluster or cyber victimization. A sample of Reddit posts (10%) from each cluster was manually reviewed for relevance.

Cluster Labeling

GPT-4 was used to facilitate cluster labeling once the clusters were identified. GPT-4 is an advanced language model developed by OpenAI [31]. In this study, its role was to summarize posts and comments and suggest potential labels for each cluster. In this process, 10-15 percent of posts from each cluster were uploaded into GPT-4.

Following the initial AI-assisted labeling, a manual review of the cluster data and their respective labels was conducted to ensure accuracy and relevance and to adjust where necessary. The combination of AI-driven and manual review allowed for accurate labeling and understanding of the

thematic clusters within the data. This approach ensured that the final labels reflected the contextual content of the posts and that the narratives were present in the online discussions related to cyber victimization.

Thematic Analysis of Clusters

Following labeling, clusters were thematically analyzed with MAXQDA 2022, a qualitative data management software program [32], to organize clusters into overlapping themes. The full data set 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.

Keyword Searching

To account for the evolving teen language, we used focused searching to capture variations, abbreviations, and related expressions. MAXQDA's lemmatization feature identified word variants for semantically related terms [32]. For example, "thick skin" led to searches for similar phrases like "not taking things too seriously" or "brushing off comments." In addition, the study acknowledged the prevalence of abbreviations in teen communication, especially in online platforms like Reddit. To address this, the methodology included searches for common abbreviations that might signify cyber victimization. For example, the abbreviation "KYS" (kill yourself), often used in cyberbullying contexts, was specifically searched within the data set.

Demographic data were also identified through keyword searches. Age-related keywords included "Born in [year]"; "years old"; "I am a [grade level]"; and "in [grade level]"; "high school"; "college student." Geographic locations were identified by phrases such as "I am from"; "I live in" and mentions of specific countries with focused searches included all 50 U.S. states. Race and Ethnicity were identified using keywords including "Asian," "biracial," "black," "Caucasian," "Pacific Islander," "skin color," "skin colour," and "white." Additionally, the analysis included searches for phrases like "My ethnicity is"; "My family is from"; and for the word "identify." Gender and sexual identity searches included keywords such as "coming out"; "I identify as [gender identity]"; "straight ally"; and "androgyny," "fluid," "gay," "gender," "homosexual," "lesbian," "poly," "queer," "sexual," "trans," "two-spirit," "FTM" (female to male), and "MTF" (male to female).

Data Presentation

Findings are presented through data visualization using Gephi, an open-source graph modeling tool for visualizing data [33]. Nodes and edges are color-coded and sized according to their weight, providing a clear and concise visual understanding of the trends and patterns related to cyber victimization. The results and the identified patterns and topics are discussed with examples of specific quotes for emphasis. Covert signs, symptoms, and narratives will be presented, linking to the original aim to analyze how teens on Reddit describe and discuss their experience of cyber

victimization. An overview of the full methodological process is provided in (Figure 1)

Ethical Considerations

Prior to data collection, this study was submitted for review to the Institutional Review Board (IRB) at Indiana University. The study was deemed exempt. This exemption was based on the criteria that the study did not involve direct interaction with human subjects and relied on publicly available data. Usernames associated with Reddit posts will not be used or included in the study. Access to Reddit's API was achieved in adherence to Reddit's API access policies.

Results

This study successfully applied data mining, WAG modeling, GPT-4 assisted labeling, and thematic analysis to examine cyber victimization narratives shared by teens on Reddit. The extracted data set comprised 13,381 posts and comments from 3,283 unique Reddit users. Approximately 5.07% of the posts, amounting to 678 in total, were posted between 2012 and 2018. The remaining 94.93% of the posts were posted to Reddit from 2019-2023.

The WAG modeling process identified 38 clusters, including two overflow clusters containing weakly matched or unmatched material (Figure 2). Of the 38 clusters, 35 were relevant to cyber victimization experiences, while two clusters were excluded due to irrelevance. Additionally, the WAG modeling process placed the word pairs "law enforcement" and "legal action" into two clusters, but due to their overlapping content, they were merged into a single cluster (Figure 3).

The distribution of posts and comments across clusters varied, with some clusters containing a high volume of discussion while others had minimal engagement (Figure 4). 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 Appendix 1.

Cluster Labeling and Validation

Clusters were further 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 data set. 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 cyberbullying and victimization, GPT-4 was initially used to label the resulting 38 (see Figures 4, 5, and Appendix 1). GPT-4, a language model, synthesized 10-15 posts and comments within each cluster to generate suggestive labels. Additionally, a thorough manual review of the content within each cluster was conducted to confirm the accuracy and relevance of the labels. The manual review process led to 12 labels with data that either matched the concepts that were suggested with minimal changes in wording, 12 labels that generally matched the concept of the data but needed more refining with wording to reflect the data accurately, and ten labels that had to be completely revised to match the content reflected in the data (Figure 5). Two clusters were merged as they contained a clear overlap of concepts, and two clusters were excluded due to irrelevance. Overall, GPT-4 was able to summarize the content of 26 clusters accurately.

Focused Keyword Searching

Focused keyword searching was used to expand the analysis of clustered data. This approach involved a review of the narratives within each cluster to identify key concepts, expressions, or themes. These identified concepts led to the generation of related terms or alternative expressions, which were then used as the basis for focused keyword searching. This targeted search aimed to locate similar narratives that might have been weakly clustered or overlooked due to the absence of common anchor words. 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" (Table 1). These terms were then used to search for and uncover additional narratives that shared similar experiences but may not have been strongly associated with the original cluster due to variations in language or expression. This method was effective in revealing a broad range of experiences and narratives related to cyber victimization.

Thematic Analysis

Following labeling, clusters were thematically analyzed with MAXQDA 2022, a qualitative data management software program [32], 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 (Figure 6).

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) Discussing or Seeking Understanding of Experience, which includes philosophical discussions about the nature of victimization. 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.

Table 1. Example analysis and narrative linkages

Example narrative	Alternative expression	Focused searches	Similar experiences in the narrative outside the cluster
C6-49: ... He tells me he still has all the photos and he plans on keeping them to look at and to send to my family and friends.	blackmail, catfish	"blackmail" "catfish"	CW-8270: Need advice about [...] blackmail) ... he took some pics / video of her [...] now he is threatening to publish them on social media
C0-104: I have been traumatized by this topic to the point of asking my mother every day and every night if I am beautiful and if I deserve to live just because some idiots told me every day that I deserve to die, no one likes me and that I deserve the abuse	intrusive thoughts unable to stop thinking about abuse	"stop thinking"	CW-7294: I can't stop thinking about the way this kid made me feel ... I couldn't defend myself I just sat there and took it. I think about it all the time the shame and anger and inferiority.
C6-30: They made contact with me 3 months after the falling out and started doxing me, They have personal videos/pictures of myself. They have made Tik tok accounts, dating sites using my government name.	invasion of privacy identity theft	"dox" "privacy" "identity"	CW-5260: Someone doxxed my entire life story on instagram-full name, where i live, where i used to work, where i go to college, even relatives (named one who is deceased) ... I am however very creeped out and angry.
C0-88: And hammer yourself down so you don't stick out."	blending in conforming staying under the radar	"blend" "conform" "radar"	CW-8011: I keep to myself, don't stare at people and do my best to just blend in, but it never seems to work.

Discussion

Principal Results

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 limits the breadth and depth of the narratives. The anonymity of Reddit also encouraged users to share sensitive information without fear of stigma. By leveraging data mining, WAG modeling, and GPT-4 assisted labeling, this approach facilitated the identification

of hidden 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.

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 [34]. In addition, social media platforms and online forums 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 [35,36], while TikTok, Facebook, and Pinterest users are more likely to be women [36].

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 [37]. The anonymous nature of Reddit may also lead to false answers or misinformation, as online platforms can be prone to exaggerated or false claims.

An important limitation of this research is the potential inclusion of bot-generated content [38]. 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 data set, potentially influencing the results.

Implications for Further Research

Future research could expand on 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 two 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 online communities.

To explore how cyber victimization experiences vary across online spaces with different user demographics and privacy structures, future research could compare narratives from different platforms (TikTok, Discord, Bluesky, Instagram, and Tumblr). Each platform has unique privacy settings, user demographics, and moderation policies, which may influence how users discuss and experience cyber victimization.

Conclusions

This study aimed to analyze how teens on Reddit describe their experience of cyber victimization using data mining and computational analysis of unsolicited data. This approach identified themes in discussions of cyber victimization. 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. The methodological approach demonstrated in this study will be valuable to data scientists and healthcare 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 healthcare settings.

Acknowledgements

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.

Conflicts of Interest

None declared

Abbreviations

API: Application programming interface

GPT-4: Generative pre-trained transformer

WAG: Word adjacency graph

References

1. David-Ferdon C, Hertz MF. Electronic media, violence, and adolescents: An emerging public health problem. *Journal of Adolescent Health* 2007 Dec 1;41(6, Supplement):S1–S5. PMID:18047940
2. Vogels EA, Gelles-Watnick R, Massarat N. Teens, social media and technology 2022. Pew Research Center; 2022 Aug p. 1–30. Available from: <https://www.pewresearch.org/internet/2022/08/10/teens-social-media-and-technology-2022/>
3. Federal Bureau of Investigation. Internet Crime Complaint Center(IC3) | Home Page. 2023. Available from: <https://www.ic3.gov/> [accessed Nov 19, 2023]
4. Cénat JM, Smith K, Hébert M, Derivois D. Cybervictimization and suicidality among French undergraduate students: A mediation model. *J Affect Disord* 2019 Apr 15;249:90–95. PMID:30769296
5. Li Y, Li D, Li X, Zhou Y, Sun W, Wang Y, Li J. Cyber victimization and adolescent depression: The mediating role of psychological insecurity and the moderating role of perceived social support. *Children and Youth Services Review* 2018 Nov;94:10–19. doi: 10.1016/j.chilyouth.2018.09.027
6. Rose CA, Tynes BM. Longitudinal associations between cybervictimization and mental health among US adolescents. *J Adolesc Health* 2015 Sep;57(3):305–312. PMID:26115909

7. John A, Glendenning AC, Marchant A, Montgomery P, Stewart A, Wood S, Lloyd K, Hawton K. Self-harm, suicidal behaviours, and cyberbullying in children and young people: systematic review. *Journal of Medical Internet Research* 2018;20(4):e129. PMID:29674305
8. 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 Jul;9(7):1–9. doi: 10.1371/journal.pone.0102145
9. Pallone remarks at hearing on child safety in an increasingly digital age. 2021. Available from: <https://energycommerce.house.gov/newsroom/press-releases/pallone-remarks-at-hearing-on-child-safety-in-an-increasingly-digital-age> [accessed Mar 11, 2021]
10. Sasson H, Mesch G. Parental mediation, peer norms and risky online behavior among adolescents. *Computers in Human Behavior* 2014 Apr;33:32–38. doi: 10.1016/j.chb.2013.12.025
11. Menesini E, Nocentini A. Cyberbullying definition and measurement: Some critical considerations. *Zeitschrift für Psychologie / Journal of Psychology* 2009 Jan;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. Radovic A, Gmelin T, Stein BD, Miller E. Depressed adolescents' positive and negative use of social media. *Journal of Adolescence* 2017 Feb;55:5–15. PMID:27997851
14. St. Surin O, Blankenship RJ. Silent voices: The perception of cyberbullying among at-risk middle school students. *International Journal of Cyber Behavior, Psychology and Learning IGI Global*; 2019 Oct;9(4):1–21. doi: 10.4018/IJCBPL.2019100101
15. Queirós A, Faria D, Almeida F. Strengths and limitations of qualitative and quantitative research methods. *Zenodo*; 2017 Sep 7; doi: 10.5281/ZENODO.887089
16. Lu J, Sridhar S, Pandey R, Hasan MA, Mohler G. Investigate transitions into drug addiction through text mining of Reddit data. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining Anchorage AK USA: ACM*; 2019. p. 2367–2375. doi: 10.1145/3292500.3330737
17. 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 & Behavior* 2017 Sep;74:149–153. doi: 10.1016/j.yebeh.2017.06.023
18. Sivagurunathan M, Walton DM, Packham T, Booth RG, MacDermid JC. Discourses around male IPV related systemic biases on Reddit. *J Interpers Violence SAGE Publications Inc*; 2021 Jul 12;08862605211030015. doi: 10.1177/08862605211030015
19. Miller WR, Groves D, Knopf A, Otte JL, Silverman RD. Word Adjacency Graph Modeling: Separating signal from noise in big sata. *West J Nurs Res* 2017 Jan;39(1):166–185. doi: 10.1177/0193945916670363

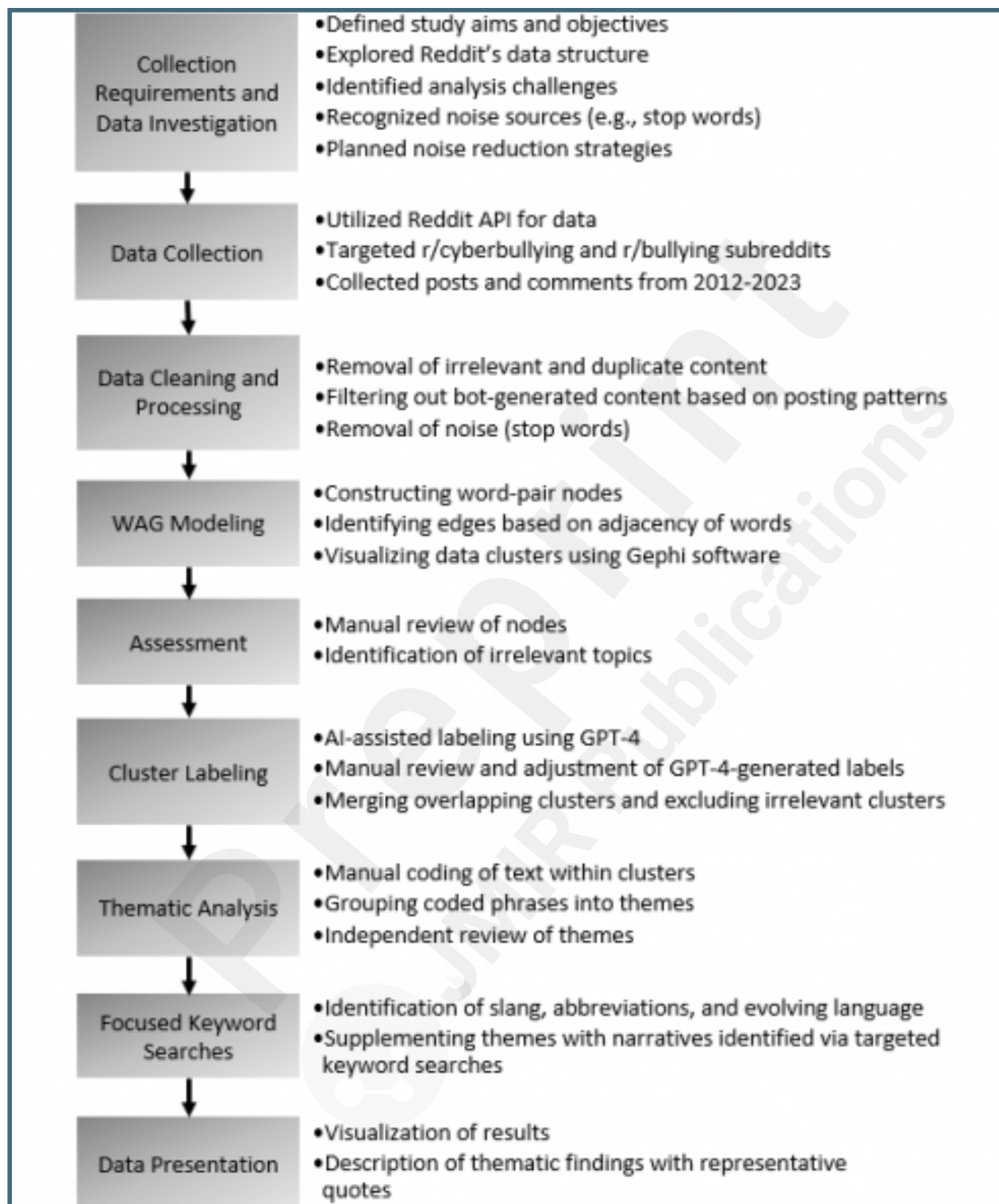
20. Reddit Inc. Reddit content policy. 2018. Available from: <https://www.redditinc.com/policies/content-policy> [accessed Nov 22, 2023]
21. Reddit Inc. Reddit homepage. 2023. Available from: <https://www.redditinc.com/> [accessed Nov 22, 2023]
22. Dixon S. U.S. Reddit app users by age 2021. Statista. 2022. Available from: <https://www.statista.com/statistics/1125159/reddit-us-app-users-age/> [accessed Jan 1, 2023]
23. Sowles SJ, McLeary M, Optican A, Cahn E, Krauss MJ, Fitzsimmons-Craft EE, Wilfley DE, Cavazos-Rehg PA. A content analysis of an online pro-eating disorder community on Reddit. *Body Image* 2018 Mar 1;24:137–144. doi: 10.1016/j.bodyim.2018.01.001
24. Arya S, Nagappala S, Krawczyk N, Gu 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. PMID:36106770
25. Overbeek D, Janke A. 360 Characteristics of posts of opioid users on Reddit, an online social media forum, an area for improved harm reduction. *Annals of Emergency Medicine Elsevier*; 2018 Oct 1;72(4):S142. doi: 10.1016/j.annemergmed.2018.08.365
26. Ogunleye JO. The concept of data mining. *Data mining : Concepts and applications United Kingdom: IntechOpen*; 2022. Available from: <https://directory.doabooks.org/handle/20.500.12854/90223> [accessed Jul 31, 2023] ISBN:978-1-83969-267-3
27. Manning CD, Raghavan P, Schütze H. Introduction to information retrieval. Illustrated edition. New York: Cambridge University Press; 2008. ISBN:978-0-521-86571-5
28. Jhaver S, Birman I, Gilbert E, Bruckman A. Human-machine collaboration for content regulation: The case of Reddit automoderator. *ACM Trans Comput-Hum Interact* 2019 Jul 19;26(5):31:1-31:35. doi: 10.1145/3338243
29. Lebeuf C, Storey M-A, Zagalsky A. Software bots. *IEEE Software* 2018 Jan;35(1):18–23. doi: 10.1109/MS.2017.4541027
30. Waltman L, van Eck NJ. A smart local moving algorithm for large-scale modularity-based community detection. *Eur Phys J B* 2013 Nov 13;86(11):471. doi: 10.1140/epjb/e2013-40829-0
31. OpenAI. ChatGPT [Large language model]. 2023. Available from: <https://chat.openai.com/chat>
32. VERBI Software. MAXQDA 2020 [computer software]. Berlin, Germany: Berlin, Germany: VERBI Software; 2019. Available from: Available from maxqda.com
33. Bastion M, Heymann S, Jacomy M. Gephi: An open source software for exploring and manipulating networks. *International AAAI Conference on Weblogs and Social Media*; 2009. Available from: <https://gephi.org/>

34. Proferes N, Jones N, Gilbert S, Fiesler C, Zimmer M. Studying reddit: A systematic overview of disciplines, approaches, methods, and ethics. *Social Media + Society* 2021 Apr 1;7(2):20563051211019004. doi: 10.1177/20563051211019004
35. Amaya A, Bach R, Keusch F, Kreuter F. New data sources in social science research: things to know before working with Reddit data. *Social Science Computer Review* SAGE Publications Inc; 2021 Oct 1;39(5):943–960. doi: 10.1177/0894439319893305
36. Pew Research Center. Social media fact sheet. 2021. Available from: <https://www.pewresearch.org/internet/fact-sheet/social-media/> [accessed Dec 3, 2023]
37. Richard B, Sivo SA, Ford RC, Murphy J, Boote DN, Witta E, Orłowski M. A guide to conducting online focus groups via reddit. *International Journal of Qualitative Methods* SAGE Publications Inc; 2021 Jan 1;20:16094069211012217. doi: 10.1177/16094069211012217
38. Storozuk A, Ashley M, Delage V, Maloney EA. Got bots? Practical recommendations to protect online survey data from bot attacks. *The Quantitative Methods for Psychology* 2020;16(5):10. doi: 10.20982/tqmp.16.5.p472

Supplementary Files

Figures

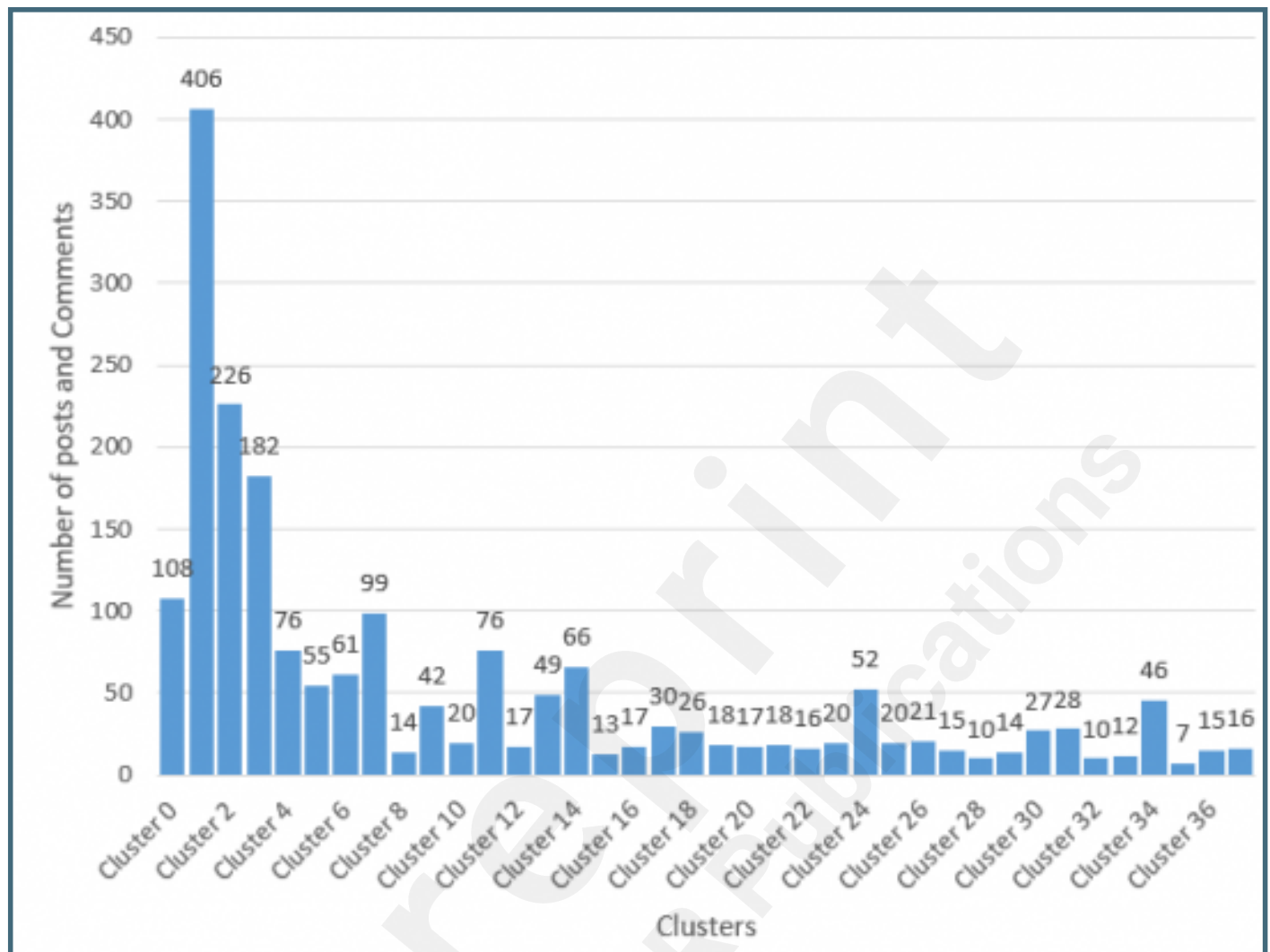
Overview of the methodological process.



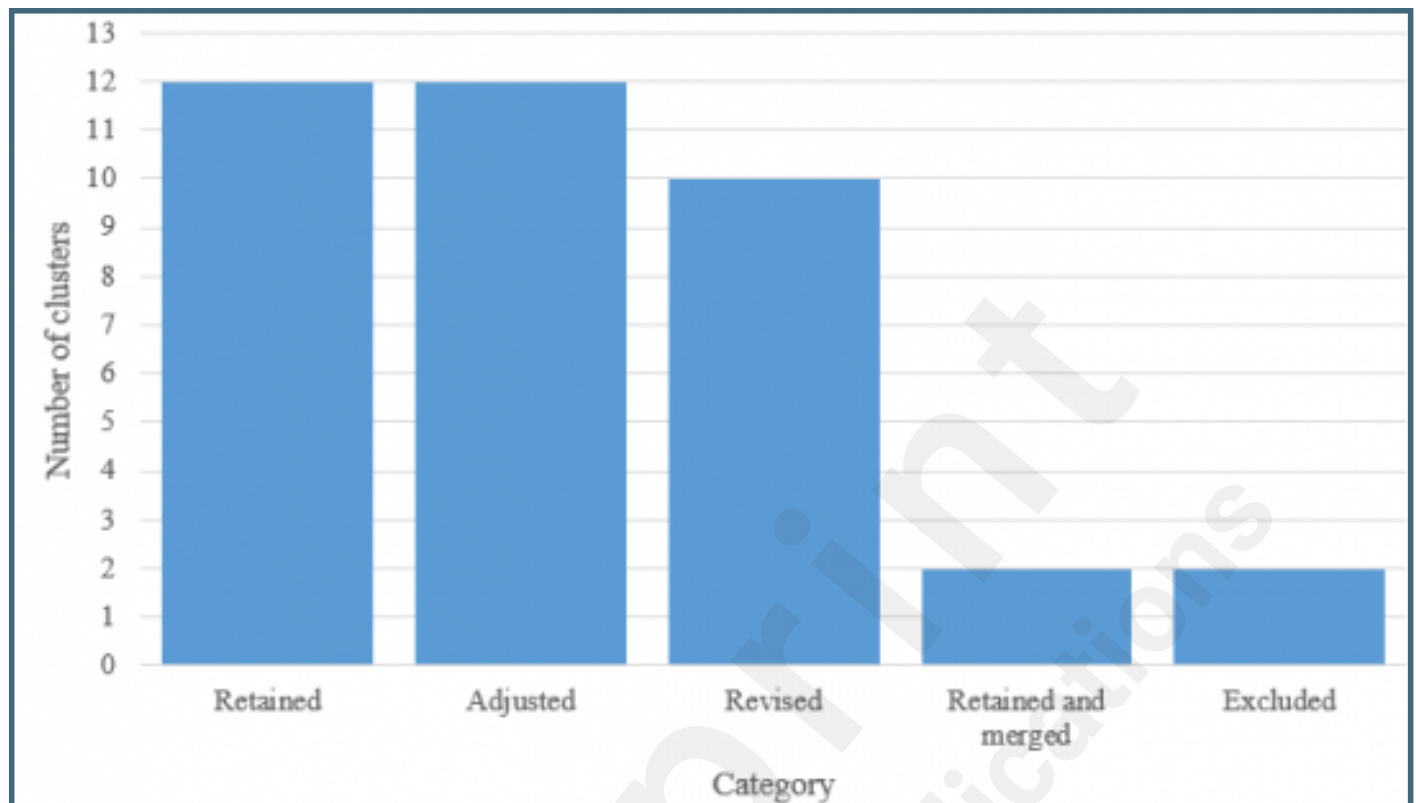
WAG results cluster labels.



Number of posts and comments per cluster.



Outcome of cluster label review.



[illegible]

Multimedia Appendixes

List of clusters labels 0-37.

URL: <http://asset.jmir.pub/assets/727ca03b6c523e9d219230c4a2e7639b.docx>

