Health Behavior Research

Volume 8
Number 4 *Mentorship in Health Behavior Research*

Article 5

October 2025

Exploring the unique challenges of social media mining and analysis: A case-study of physical activity-related content on Reddit recovery forums

Aakash Parekh
Indiana University School of Public Health, Bloomington, aakparek@iu.edu

Danny Valdez
Indiana University School of Public Health, Bloomington, danvald@iu.edu

Alex M. Russell
Recovery Research Institute, Massachusetts General Hospital and Harvard Medical, arussell11@mgh.harvard.edu

Recommended Citation

Parekh, Aakash; Valdez, Danny; and Russell, Alex M. (2025) "Exploring the unique challenges of social media mining and analysis: A case-study of physical activity-related content on Reddit recovery forums," *Health Behavior Research*: Vol. 8: No. 4. https://doi.org/10.4148/2572-1836.1336

Exploring the unique challenges of social media mining and analysis: A case-study of physical activity-related content on Reddit recovery forums

Abstract

Introduction: Recovery from substance use is a lifelong process, and physical activity can serve as a valuable support strategy. People in, or who are seeking, recovery may turn to online forums – including Reddit – to exchange experiences and offer mutual support. Efforts to extract physical activity-related content from these forums may encounter challenges due to ambiguous and metaphorical language inherent to social media, which warrants further consideration.

Objective: This study examines how people discuss physical activity within SUD-recovery subreddits while also highlighting the challenges in extracting content relevant to the topic.

Methods: We used Reddit PRAW, to scrape r/RedditorsInRecovery, r/Addiction, r/OpiatesRecovery, and r/AlcoholicsAnonymous for posts referencing types of physical activity. After allocated N=4946 posts, we used Term-Frequency Inverse Document Frequency to create embeddings with our data, k-means clustering to extract themes within the data, and Word2Vec to identify common word associations.

Results: Our k-means clustering returned an optimum number of 7 clusters, or themes, including themes comprising sober meetings, spirituality, difficulties with withdrawal, and physical activity. A Word2Vec of physical activity related queries yielded few words associated with physical activity, though diverse facets of recovery were expressly mentioned.

Discussion and Conclusion: Our findings illustrate the difficult nature of identifying physical activity content in Reddit recovery spaces due to inconsistent literal and metaphorical language. These challenges can lead to misinterpretations in population health surveillance. Improved tools and analysis strategies can better support research aimed to understand how recovery-supporting behaviors, including physical activity, are discussed and promoted online.

Keywords

Social Media; Natural Language Processing; Education; Mentorship

Acknowledgements/Disclaimers/Disclosures

We have no conflicts of interest to disclose.

Exploring the unique challenges of social media mining and analysis: A case-study of physical activity-related content on Reddit recovery forums

Aakash Parekh
Danny Valdez, PhD*
Alex M. Russell, PhD

Abstract

Recovery from substance use is a lifelong process, and physical activity can serve as a valuable support strategy. People in or who are seeking recovery may turn to online forums—including Reddit—to exchange experiences and offer mutual support. Efforts to extract physical activity-related content from these forums may encounter challenges due to ambiguous and metaphorical language inherent to social media, which warrants further consideration**.** This study examines how people discuss physical activity within SUD-recovery subreddits while also highlighting the challenges in extracting content relevant to the topic. We used Reddit PRAW to scrape r/RedditorsInRecovery, r/Addiction, r/OpiatesRecovery, and r/AlcoholicsAnonymous for posts referencing types of physical activity. Within the allocated N=4946 posts, we used Term-Frequency Inverse Document Frequency to create embeddings with our data, k-means clustering to extract themes within the data, and Word2Vec to identify common word associations. Our k-means clustering returned an optimum number of seven clusters, or themes, including themes comprising sober meetings, spirituality, difficulties with withdrawal, and physical activity. A Word2Vec of physical activity-related queries yielded few words associated with physical activity, though diverse facets of recovery were expressly mentioned. Our findings illustrate the difficult nature of identifying physical activity content in Reddit recovery spaces due to inconsistent literal and metaphorical language. These challenges can lead to misinterpretations in population health surveillance. Improved tools and analysis strategies can better support research aimed to understand how recovery-supporting behaviors, including physical activity, are discussed and promoted online.

Keywords: recovery, addiction, social media, natural language processing

*Corresponding author may be reached at danvald@iu.edu

Introduction

Previous research highlights the iterative and sometimes messy nature of social media mining and analysis (Brooker et al., 2016), including in a substance use and recovery context (Nasralah et al., 2020). While there is no shortage of content across social media platforms (Ashford et al., 2020a), efforts to meaningfully extract insights can be muddled by "noise," or otherwise incongruous content, unidentifiable slang, and similar irrelevancies untethered to a topic (Zimmerman, 2022).

This challenge may stem from two design-specific considerations: (1) the analytic approach, and (2) the specificity of the inclusion criteria. Whether qualitative or computational—both common analytic approaches in social media research with respective limitations—noisy data can impact the clarity of results. By refining our

inclusion criteria—in this case by using a targeted sampling strategy from content-specific channels—data may in turn be more representative of a topic. However, additional considerations remain, highlighting the need for additional data-driven insights.

In this brief report, we consider a search for physical activity-related content extracted from recovery-related forums on Reddit, a semi-anonymous social networking website that organizes user-generated content into topical areas known as subreddits. We aim to illustrate how even refined search strategies are subject to social media's unique vernacular, emphasizing how even messy data, when engaged with thoughtfully, can nonetheless yield important insights.

Addiction Recovery and Physical Activity as Expressed on Reddit: A Data-Driven Example

Individuals in or who are seeking SUD recovery may benefit from participation in alternative rewarding activities outside of substance use to support their recovery journeys (e.g., physical activity) (Moos, 2006). Additionally, physical activity has been consistently shown to aid recovery through reductions in cravings (Patterson et al., 2022) and improvements in psychological symptoms (More et al., 2017).

Beyond the physical and psychological health benefits of physical activity on recovery outcomes, regular movement may also promote social connection. Studies have found that adults with SUD tend to be lonelier than their peers (Hosseinbor et al., 2014). Adaptive social network changes (i.e., increasing connections with recovery-oriented peers through physical activity and other events) are a strong facilitator of positive recovery-related outcomes, including reductions in substance use (Kelly et al., 2011).

Reddit can be used as a tool to increase one's social connection with others in a shared community. Within these communities, or subreddits, individuals may express how they have used physical activity to get/stay on their substance use recovery journey, though limited data science work has been done to examine the intersection of physical activity and recovery on these outlets.

Purpose

This study illustrates how mining highly specific content from recovery-related subreddits (i.e., physical activity) can improve the clarity and precision of classic NLP models and/or manual verification thereafter. Our research is guided by the following questions (RQs).

1. What themes emerge from a collection of Reddit posts pertaining to addiction recovery and physical activity?
2. What does the distribution of themes tell us about the prevalence of physical activity in Reddit recovery forums?
3. What terms are most associated with physical activity specific terms (e.g., run, swim, bike, and others)?

Methods

Data

Subreddit Selection. Data for this study were collected between 2009 and 2025 from the social networking website Reddit using the Python Reddit Application Programming Interface Wrapper (PRAW)—a third-party application that allows users to collect Reddit content from specific channels (i.e., subreddits) using specific queries (i.e., search terms).

As part of a mentored research project, the mentee was tasked with briefly reviewing physical activity content in the r/RedditorsInRecovery, one of the largest and most active subreddits pertaining to recovery

from substance use. Upon identifying such content, the mentee was tasked with identifying additional recovery-specific subreddits using the "Related Subreddits" feature to bolster data collection. To be included in the study, each new subreddit had to meet the following inclusion criteria: (1) originate from a subreddit specific to addiction recovery, (2) have daily posting activity, (3) have a user base greater than 500 people, and (4) contain physical activity content even if minimal. The following subreddits meeting this criteria include: r/RedditorsInRecovery, r/Addiction, r/OpiatesRecovery, and r/AlcoholicsAnonymous.

Query Selection. Once we identified appropriate subreddits, we used simple Google and Google Trends searches to identify popular physical activity options for people in recovery from substance use. Some of the most popular activities include aerobic activities (i.e., running, biking, swimming), anaerobic activities (i.e., weightlifting), and holistic activities (i.e., yoga, hiking), all of which are widely supported in the literature (Castillo-Viera et al., 2022).

The lead author created a composite list of activities—defined as tasks involving physical exertion—that corresponded to aerobic, anaerobic, and holistic physical activities. The following queries resulted from this list: (1) running (ran a 5k, ran a 10k, first run, runner, running shoes, for a run, marathon), (2) biking (bike, riding a bike, biking, peloton), (3) swimming (swim, swimming, swam, pool, swimmer), (4) weightlifting (gym, weight training, lift weights, heavy lifting, strength training, bodybuilding, powerlifting), (5) workout (exercise, athletics, fitness, workout, activity), and (6) yoga/meditation (yoga, meditation, hiking). After filtering duplicates, we retained N=4946 reddit posts across all queries. Note that non-physical activity queries for specific activities (i.e., mindfulness as a reference to yoga) were excluded.

Procedure

We used RedditPRAW to scrape the included subreddits with all queries. All data were compiled into a single .csv file for further analysis. Stop words (parts of speech with no semantic meaning), URLs, and contractions were removed at this stage. We then generated unigram (single-word), bigram (two-word), and trigram (three-word) word clouds to examine general term frequencies in our data. We then ran several unsupervised NLP analyses to bolster our initial observations, including a Term-Frequency Inverse Document Frequency (TF-IDF), k-means clustering, and Word2Vec Embedding. In brief, TF-IDF is an embedding technique that calculates how important each word is in a single post relative to all posts within a dataset. This process is an extension of word frequency and assigns weights to each term while allowing us to determine the most important words within the data (Singh et al., 2024). K-means clustering is also an NLP technique that uses TF-IDF embeddings to identify clusters or themes—in other words, similar embeddings calculated by TF-IDF may share similar meaning, and k-means clustering is an approach that groups those embeddings together to help uncover latent meaning (Na et al., 2010). Finally, Word2Vec is an NLP technique that uses embeddings and vectors to identify highly similar words—for example, with Word2Vec we can query our dataset for the term "run" and extract all other dataset-specific terms most synonymous with the term "run" (Ma & Zhang, 2015). Mentee and mentor then met to discuss findings and adjust the pipeline if needed.

Figure 1

Summary of Computational Output

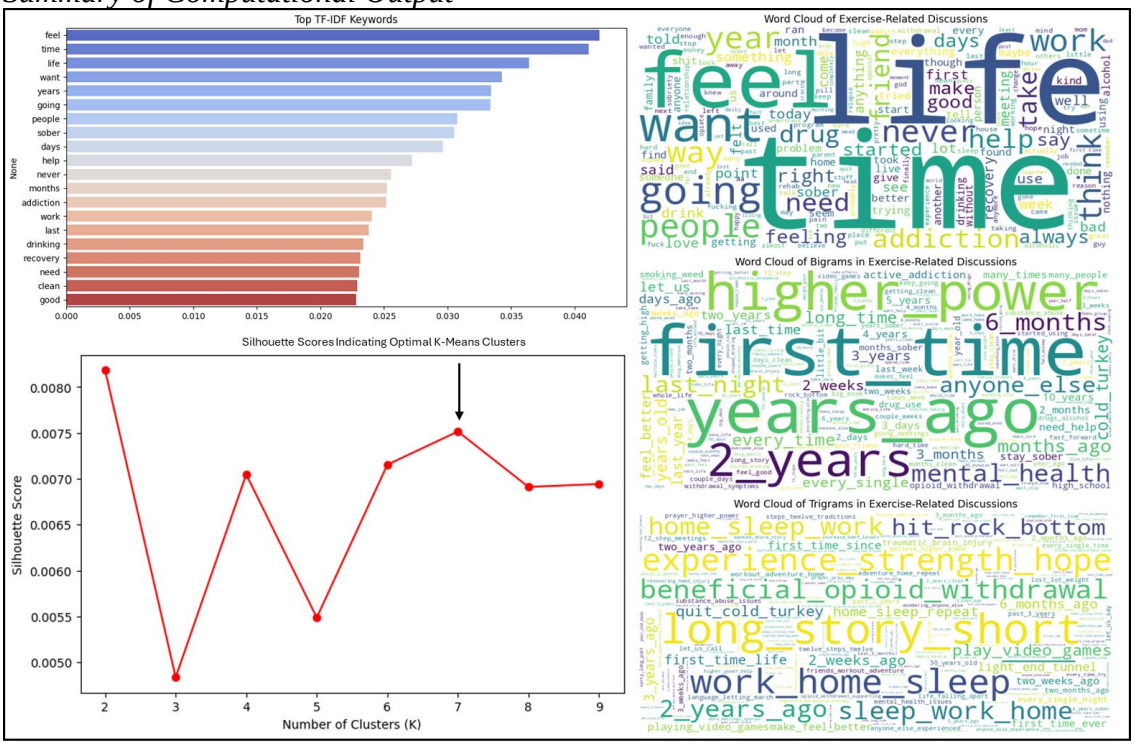


Figure 1. Summary of Computational Output, including the most important words in our data (top left), unigram, bi-gram, and tri-gram word clouds (right), and optimal number of clusters (bottom left).

Results

Our findings illustrate the diverse array of content on these subreddits, which collectively illustrate the broad and complex nature of addiction recovery. We explain our findings below:

What themes emerge from a collection of Reddit posts pertaining to substance-use recovery and physical activity?

Figure 1 illustrates a unigram, bigram, and trigram, giving us a visual of how often certain words appear in the posts. The TF-IDF analysis shows us the words that appear to be the most important throughout the posts. The k-means clustering shows us that 7 clusters is the optimal number of clusters within this dataset.

Results from the k-means clustering identified seven optimal clusters, each pertaining to a unique theme in our data. After reviewing the cluster groupings, as well as posts related to each cluster, AP and DV collaboratively named the seven clusters accordingly (Table 1). Quotations highlight our query terms, which in some cases clearly reference physical activity, but in others were used metaphorically in other contexts.

What does the distribution of themes tell us about the prevalence of physical activity content in Reddit recovery forums?

Figure 2 illustrates the distribution of topics: Sober Meetings (n=465), Addiction, Physical Activity, and the Human Body (n=206), General Addiction & Recovery (n=1536), Alcohol Use Discussions (n=426), Withdrawal Discussions (n=1004), Starting Sobriety (n=949), and Spirituality and Higher Powers (n=246).

https://newprairiepress.org/hbr/vol8/iss4/5
DOI: 10.4148/2572-1836.1336

Table 1
K-Means Clustering Output

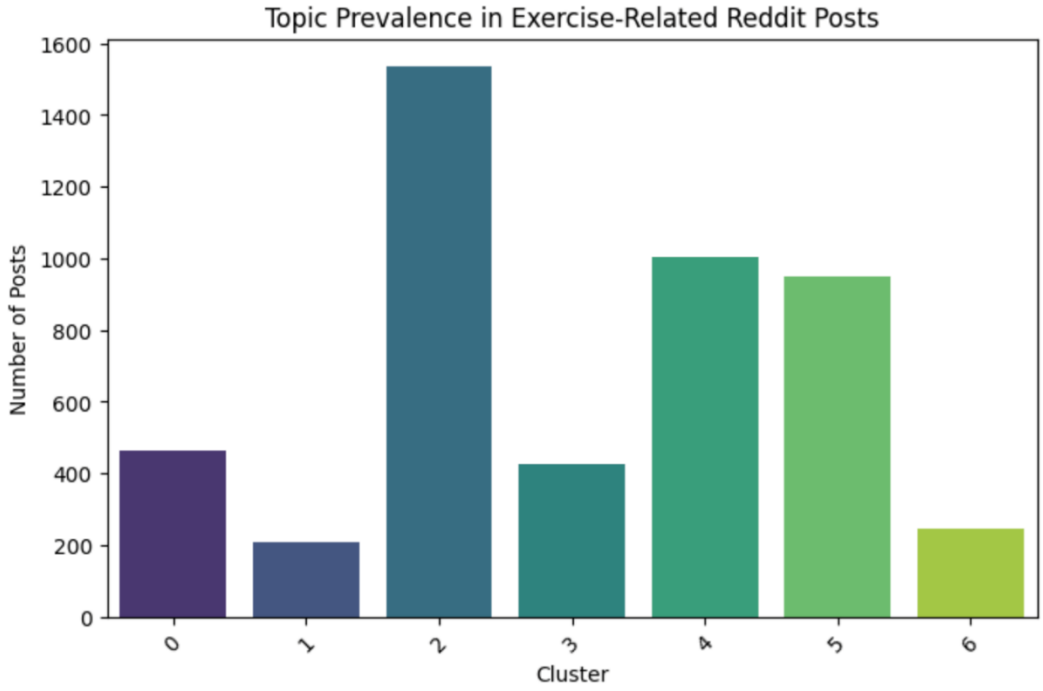
Clusters	Cluster names	Cluster Terms	Representative Quotations
Cluster 0	Sober Meetings	sponsor, aa, meetings, meeting, steps, people, sober, program, group	Example • Started 'meditating' at least half an hour a day. Started 'exercising' at least an hour every single day. Going to meetings. • I don't want to drink, I'm working on a good program. I'm sick as a dog with the flu. And my head is 'swimming'. I know I'm still sick and I need my sponsor and this way of life and the steps because I like how my head feels right now. • Aside from lingering insomnia & stomach issues, I laced up my 'running' shoes and ran 5 miles (jog is more accurate). 'Exercise is key', even if feeling lethargic & tweaked.. I'm going to 2-3 meetings a day, before & after work
Cluster 1	Addiction, Physical Activity, and the Human Body	weight, gain, gained, eating, lose, pounds, anyone, eat, lost	• I gained like 40 lbs after getting sober. Anyone have any advice...I enjoy 'biking, swimming', and can tolerate 'jogging' when I need to be quick. 'Lifting' sounds fun too. Maybe it'd be cool to have a healthy variety? • I am far healthier and have lost a lot of weight as I am throwing myself in to 'exercise' and healthy eating.
Cluster 2	General Addiction & Recovery	addiction, recovery, people, life, want, feel, years, time, help	• What were some of the things that helped you in early recovery? I've thinking about supplements get my mind and better in better shape, intermittent fasting, HIIT, cold showers, 'meditation' and 'yoga' • I need help. Addiction 'runs' in my family very thick and as all kids believe as they grow up
Cluster 3	Alcohol Use Discussions	drinking, drink, alcohol, drunk, sober, time, feel, alcoholic, years	• I had a massive alcohol craving last night. Got seconds away from relapsing. Instead, I drove to a 'gym' I just joined recently and ended up LOVING it! Crushed my cravings and I ended up making a couple new late-night gym goer friends along the way :)

			• I woke up after my first day of not drinking (daily drinking 10-15 drinks a day) and I went to the ‘gym’ because i felt motivated to ‘work out’.
Cluster 4	Withdrawal Discussions	feel, days, last, withdrawal, sleep, pain, take, feeling, time	• When stopping using, especially after a longer period, those feelings you ‘run’ away from will go back. It will be a whole lot of them. It will be confusing, and painful, and with withdrawal symptoms combined, it might even feel unbearable. But it will pass. Just allow yourself to be you. • My acute symptoms seem to have pretty much passed at this point, but I feel like I ‘ran a marathon’ after every time I get up and do anything. • I know the answer is just to white knuckle it but it's hard. It's SOO HARD. When I feel like a relapse is coming it's like I can't think of anything but relapsing. Yesterday, I felt it coming, I fought with my mind and ‘went for a swim and a long walk’, and I thought I defeated the urge.
Cluster 5	Starting Sobriety	time, life, going, never, years, want, started, feel, told	• Life does get better. It may not seem that way. But stay committed and life will be worth it, and your hard work will pay off. Just keep ‘swimming • Where do I start? What things are necessary for spiritual growth? What do I do on a daily basis? Journal? ‘Meditate’? Spend more time in nature? ‘Exercise’? Read more? • Now today is day 11 and things are getting better! This is for people detoxing now I am entering into my next phase of sobriety. I am hopeful and happy. I want to stress how important ‘exercise’ is!!
Cluster 6	Spirituality & Higher Powers	us, power, god, higher, prayer, others, god, meditation, spiritual, today	• I have been in and out of the joint since 2014 and finally I surrendered to God and asked for his help to Change and live better. It took me getting beat down so bad to realize that I could be different I ‘ran’ everyone one around me through the gutters I was a user”

			• Prayer, ‘meditation’, and dedication to my program has saved my life. God gave me a second chance. • Maintaining my spiritual condition is like ‘working out every day, planning for the marathon, swimming laps, jogging’. It's staying in good shape spiritually, and that requires ‘prayer and meditation’. The single most important way for me to improve my conscious contact with a Higher Power is to pray and meditate

*Topic names and representative quotations have been abridged slightly to uphold the privacy of Reddit users by removing personally identifiable information and graphic depictions of substance use.

Figure 2
Topic Prevalence in Exercise-Related Reddit Posts



What terms are most associated with physical activity specific terms (e.g., run, swim, bike, and others)?

We used Word2Vec to identify the words that most commonly appeared with our query terms. Table 2 provides a summary which includes a physical activity, query terms associated with that physical activity, and words that commonly co-occurred with these queries.

Discussion

Our study supports prior work showing that individuals in or seeking recovery are using online communities to support their efforts (Ashford et al., 2020b; Parker et al., 2023; Russell et al., 2022; Valdez & Patterson, 2022).

However, while our findings align strongly with previous work, content pertaining to physical activity was only minimally represented in our findings despite querying for this content, specifically. For example, only one topic in our k-means clustering analysis (Addiction, Physical Activity, and the Human Body) comprised exclusive physical activity-related content. Still, all other topics and associated terms were contextually relevant to other facets of recovery, again highlighting the contextual and nuanced nature of recovery discussions on these subreddits (Vigdal et al., 2022).

We propose two reasons for limited representation, and both reflect the challenges of social media surveillance and value of the data notwithstanding: (1) physical activity is not a universal recovery strategy, and (2) query terms often have non-literal uses. First, although many consider physical activity a sustainable outlet that promotes positive recovery outcomes (Theodorakis et al., 2024), not all adults are willing or able to engage with physical activity, and not all people in recovery use physical activity as a recovery driver despite

Table 2

Word2Vec Output illustrating word associations with select query terms

Queries	Associated terms
Biking [biking, bike, riding a bike, peloton]	Explained, killed, want lose, science, seriously, safe, monster, friend, inside, refused, super, involved, dark, pool, standing, run, train, research, negative, attempt, gym, rest, unable, play, attend, field, method, essential, church, chronic, homeless, insurance, stroke
Yoga/Meditation [yoga, meditation, Zen]	Training, early, discovered college, overdose, running, water, diet, food, treatment, helpful, spiritual, focus, information, therapist, helps, gods, pray, sharing, suffering, creator, manipulation, recovering addicts, every meeting, intelligence, insecurity, weakness, young people
Swimming [swim, swimming, swimmer, swam, pool]	Regular, play, order, action, stupid, dying, painful, practice, journey, aware, touch, dying, mental health, anger, helping, suffering, serious, alcoholism, practice, opportunity, jump, long-term, trouble, decision-making, complex, psych ward, drug dealer, prayed, yoga, dark, train, beast, activity, training, gained, walking, alcoholics anonymous, rest, test
Weightlifting [Gym, weightlifting, strength training, bodybuilding, powerlifting]	Training, drive, forced, super, becoming, pool, dreams, diet, quiet, eating train, health, walking, quickly, yoga, dark, many times, finding, caused, eating, beast, practice, physical, health, drug use, activities, run, notice, painful, angry, stronger, works, positive, partner, forgive, shame, actions, suffering, super, gym, activity, run, rest, regular, study, experiences, faith, positive, purpose, gods, helpful
Workout [Workout, activity, fitness, athletics, exercise]	Desperate, play, stupid, exercise, order, medication, complete, calm, present, running, symptoms, hot, mental, pool, effects, rest, dreams, legs, struggled, swimming, building, whole life, alcoholism, alcoholics anonymous, abuse, death, severe
Running [ran a 5k, ran a 10k, first run, marathon, running, runner, for a run]	Smokes, relaxing, madness, adhd, successfully, alcohol_cocaine, drinker, fitness, crack, amphetamine, expensive, psychosis, blood, unhealthy, death, law, meditating, pleasure, hits, dopamine, praying, rest, super, training, party, tolerance, gym, gained_weight, memory, cutting, depression, ball, marijuana, booze, mental

its established psychosocial benefits (Abrantes et al., 2011). In fact, numerous entities and scholars advise physical activity as a recovery outlet (Humphreys, 2003; Smith & Lynch, 2012), notably as a facilitator of social connection, which is now recognized as a crucial component of long-term recovery (Martino et al., 2015; Patterson et al., 2020). Aspects of social connection, as an extension of true physical exertion, were highly present in our findings, with evidence of others offering encouragement and support by sharing their own narratives in and out of physical activity contexts.

We also contend our queries contained diverse uses, making effective surveillance of physical activity and recovery content challenging but insightful. Diverse uses for search terms (i.e., uses of words and phrases that do not align with their intended meaning) are not uncommon in social media research (Rao et al., 2024). In our study, terms including "run," "swim," and "exercise" were more commonly referenced metaphorically. For example, "I was swimming in my drug addiction" or "recovery is a daily physical activity in patience" illustrate how language on social media may diverge from literal meanings yet carry relevance to recovery-related themes.

We further propose that, despite misaligned word use, our findings still pertained to recovery because they originated from dedicated recovery subreddits, which are concentrated conversations about the topic. It is likely that had we used open social media platforms (i.e., Instagram, TikTok), our queries would contain noise entirely unrelated to addiction and recovery. Further sections of this paper outline recommendations moving forward.

Conclusions, Implications, and Recommendations for Health Behavior Research from a Mentor-Mentee Perspective

An undergraduate student and faculty member at a research-intensive university identified mutual academic interests in addiction, recovery, and social media mining and analysis. From a mentor perspective, it was crucial to work with this mentee on the foundations of social media research, including the well-documented difficulties of procuring relevant data (Stieglitz et al., 2018). From a mentee perspective, it was important to understand how online forums can be used to discuss and facilitate addiction recovery through social connection. Collectively, our work highlights the benefit of team science in which formal mentor and mentee pairs work collaboratively to pursue projects of advancing both careers (Bacsu et al., 2024; Behar-Horenstein & Prikhidko, 2017; Smith et al., 2019).

Our work has several implications for health behavior research and mentor-mentee pairs. From a research perspective, researchers should align their data collection strategies with the focus of their study, research design, or question(s). For population surveillance, insights from truly open social media platforms (i.e., X, YouTube, Instagram, and others) are appropriate, regardless of noise, but only if data is available (Valdez et al., 2020). However, for more targeted research questions, more tailored outlets (i.e., subreddits or mobile apps) paired with domain-specific queries, dictionaries, and Boolean logic may more suitably yield relevant data. Regardless, researchers should approach social media research with the expectation that the process is iterative. For more information on cleaning social media data, see Calisir & Brambilla, 2018.

From an education perspective, the ability to use third-party resources to procure data, regardless of noise levels, creates cost-effective data streams that teach students and mentees about how people express health behaviors online. These tools provide low-cost entry to engage with real-world data and challenges therein. Incorporating these methods into curricula can foster digital surveillance skills valuable among those entering clinical or public-facing careers. As mentees pursue careers in and out of academia (e.g., medicine), learning about health from a social perspective affords exposure to diverse viewpoints that they otherwise would have been oblivious to. This exposure increases knowledge about how people in differing communities act/react to various situations, which can be directly applied in career-specific settings.

Limitations

Our study is subject to limitations. First, we did not perform a full qualitative review of our data. We acknowledge that further qualitative review of our resulting clusters and outcomes may potentially add additional insights into our findings. Second, we used classic NLP embedding methods to visualize our data, which rely on calculations that are not context aware. We assume more sophisticated analysis pipelines harmoniously utilizing qualitative inquiry with neural network or large-language-model architecture can address these limitations. We also acknowledge that, as with all research undertaken with social media data, there is a risk of collecting automated or bot-related content. To the best of our knowledge, there are no publicly accessible tools that can filter bot activity from Reddit. However, prior research suggests that because our topic is not contentious, political, or polarizing, the presence of bots would likely be minimal (Hurtado et al., 2019). Finally, we acknowledge that our data collection may not be representative of all recovery-related subreddits or recovery narratives on Reddit. Data collection was kept generally constrained to fit the parameters of a mentored research study. However, we strongly encourage such uses with our corpus for educational and research-related endeavors—including with additionally collected data. Our data is available by reasonable request.

Discussion Questions

Our results highlight both the role of physical activity in addiction recovery and the difficult nature of curating topical social media content. These findings support the need for further research into refined analytic procedures leveraging recent tech advances to better account for context and the need for improved search criteria to identify relevant social media content.

What strategies can health behavior researchers adopt to control more effectively for noise in text-based data in all forms, including open-ended response data, qualitative interviews, or social media content?

Is noise in text-based data (i.e., data that may contain irrelevant content) inherently negative in all circumstances? In which circumstances can noisy data be beneficial?

References

Abrantes, A. M., Battle, C. L., Strong, D. R., Ing, E., Dubreuil, M. E., Gordon, A., & Brown, R. A. (2011). Exercise preferences of patients in substance abuse treatment. *Mental Health and Physical Activity*, *4*(2), 79–87. https://doi.org/10.1016/j.mhpa.2011.08.002

Ashford, R. D., Bergman, B. G., Kelly, J. F., & Curtis, B. (2020a). Systematic review:

Digital recovery support services used to support substance use disorder recovery. *Human Behavior and Emerging Technologies*, *2*(1), 18–32. https://doi.org/10.1002/hbe2.148

Ashford, R. D., Bergman, B. G., Kelly, J. F., & Curtis, B. (2020b). Systematic review: Digital recovery support services used to support substance use disorder recovery. *Human Behavior and Emerging Technologies*, *2*(1), 18–32. https://doi.org/10.1002/hbe2.148

Bacsu, J.-D., Sefcik, J. S., Rahemi, Z., Petrovsky, D., Baker, Z. G., Ma, K., & Smith, M. L. (2024). Supporting Early Career Faculty in Geriatrics Through a New Mentorship Model: Development, Implementation, and Evaluation. *Canadian Geriatrics Journal*, *27*(3), 390–391.

Behar-Horenstein, L. S., & and Prikhidko, A. (2017). Exploring mentoring in the context of team science. *Mentoring & Tutoring: Partnership in Learning*, *25*(4), 430–454. https://doi.org/10.1080/13611267.2017.1403579

Brooker, P., Barnett, J., & Cribbin, T. (2016). Doing social media analytics. *Big Data & Society*, *3*(2), 2053951716658060. https://doi.org/10.1177/2053951716658060

Calisir, E., & Brambilla, M. (2018). The Problem of Data Cleaning for Knowledge Extraction from Social Media. In C. Pautasso, F. Sánchez-Figueroa, K. Systä, & J. M. Murillo Rodríguez (Eds.), Current Trends in Web Engineering (pp. 115–125). Springer International Publishing. https://doi.org/10.1007/978-3-030-03056-8_10

Castillo-Viera, E., Gago-Valiente, F.-J., Giménez-Fuentes-Guerra, F.-J., Abad-Robles, M.-T., & Moreno-Sánchez, E. (2022). Physical Activity Programmes in the Treatment of Addictions: A Systematic Review. *Applied Sciences, 12*(18), 9117. https://doi.org/10.3390/app12189117

Hosseinbor, M., Ardekani, S. M. Y., Bakhshani, S., & Bakhshani, S. (2014). Emotional and Social Loneliness in Individuals With and Without Substance Dependence Disorder. International Journal of High Risk Behaviors & Addiction, 3(3), e22688. https://doi.org/10.5812/ijhrba.22688

Humphreys, K. (2003). Circles of Recovery: Self-Help Organizations for Addictions. Cambridge University Press.

Hurtado, S. (2019). Bot Detection in Reddit Political Discussion. Proceedings of the Fourth International Workshop on Social Sensing - SocialSense'19. https://doi.org/10.1145/3313294.3313386

Kelly, J. F., Hoeppner, B., Stout, R. L., & Pagano, M. (2011). Determining the Relative Importance of the Mechanisms of Behavior Change within Alcoholics Anonymous: A Multiple Mediator Analysis. Addiction (Abingdon, England), 107(2), 289. https://doi.org/10.1111/j.1360-0443.2011.03593.x

Ma, L., & Zhang, Y. (2015). Using Word2Vec to process big text data. 2015 IEEE International Conference on Big Data (Big Data), 2895–2897. https://doi.org/10.1109/bigdata.2015.7364114

Martino, J., Pegg, J., & Frates, E. P. (2015). The Connection Prescription: Using the Power of Social Interactions and the Deep Desire for Connectedness to Empower Health and Wellness. *American Journal of Lifestyle Medicine*, *11*(6), 466. https://doi.org/10.1177/1559827615608788

Moos, R. H. (2006). Theory-Based Processes that Promote the Remission of Substance Use Disorders. *Clinical Psychology Review*, *27*(5), 537. https://doi.org/10.1016/j.cpr.2006.12.006

More, A., Jackson, B., Dimmock, J. A., Thornton, A. L., Colthart, A., & Furzer, B. J. (2017). Exercise in the Treatment of Youth Substance Use Disorders: Review and Recommendations. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.01839

Na, S., Xumin, L., & Yong, G. (2010). Research on k-means Clustering Algorithm: An Improved k-means Clustering Algorithm. *2010 Third International Symposium on Intelligent Information Technology and Security Informatics*, 63–67. https://doi.org/10.1109/IITSI.2010.74

Nasralah, T., El-Gayar, O., & Wang, Y. (2020). Social Media Text Mining Framework for Drug Abuse: Development and Validation Study With an Opioid Crisis Case Analysis. *Journal of Medical Internet Research*, *22*(8), e18350. https://doi.org/10.2196/18350

Parker, M. A., Valdez, D., Rao, V. K., Eddens, K. S., & Agley, J. (2023). Results and Methodological Implications of the Digital Epidemiology of Prescription Drug References Among Twitter Users: Latent Dirichlet Allocation (LDA) Analyses. *Journal of Medical Internet Research*, *25*, e48405. https://doi.org/10.2196/48405

Patterson, M. S., Russell, A., Spadine, M. N., Prochnow, T., & Heinrich, K. M. (2020). Impact of social networks, mental health, and sobriety on exercise within a collegiate recovery community. *Health Behavior Research*, *3*(1). https://doi.org/10.4148/2572-1836.1074

Patterson, M. S., Spadine, M. N., Graves Boswell, T., Prochnow, T., Amo, C., Francis, A. N., Russell, A. M., & Heinrich, K. M. (2022). Exercise in the Treatment of Addiction: A Systematic Literature Review. *Health Education & Behavior: The Official Publication of the Society for Public Health Education*, 10901981221090155. https://doi.org/10.1177/10901981221090155

Rao, V. K., Valdez, D., Muralidharan, R., Agley, J., Eddens, K. S., Dendukuri, A., Panth, V., & Parker, M. A. (2024). Digital Epidemiology of Prescription Drug References on X (Formerly Twitter): Neural Network Topic Modeling and Sentiment Analysis. *Journal of Medical Internet Research*, *26*, e57885. https://doi.org/10.2196/57885

Russell, A. M., Valdez, D., Chiang, S. C., Montemayor, B. N., Barry, A. E., Lin, H.-C., & Massey, P. M. (2022). Using Natural Language Processing to Explore "Dry January" Posts on Twitter: Longitudinal Infodemiology Study. *Journal of Medical Internet Research*, *24*(11), e40160. https://doi.org/10.2196/40160

Singh, A., Arora, V., & Chen, Y.-P. P. (2024). An efficient TF-IDF based Query by Example Spoken Term Detection. *2024 IEEE Conference on Artificial Intelligence (CAI)*, 170–175. https://doi.org/10.1109/CAI59869.2024.00039

Smith, M. A., & Lynch, W. J. (2012). Exercise as a Potential Treatment for Drug Abuse: Evidence from Preclinical Studies. *Frontiers in Psychiatry*, *2*, 82. https://doi.org/10.3389/fpsyt.2011.00082

Smith, M., Reitzel, L., Rath, J., Scott, L. B., & Seal, D. (2019). Advancing academic careers through formal professional mentorship: The Research Scholars Mentorship Program (RSMP). *Health Behavior Research*, *2*(4). https://doi.org/10.4148/2572-1836.1071

Stieglitz, S., Mirbabaie, M., Ross, B., & Neuberger, C. (2018). Social media analytics – Challenges in topic discovery, data collection, and data preparation. *International Journal of Information Management*, *39*, 156–168.

https://doi.org/10.1016/j.ijinfomgt.2017.12.002

Theodorakis, Y., Hassandra, M., & Panagiotounis, F. (2024). Enhancing Substance Use Disorder Recovery through Integrated Physical Activity and Behavioral Interventions: A Comprehensive Approach to Treatment and Prevention. *Brain Sciences*, *14*(6), Article 6. https://doi.org/10.3390/brainsci14060534

Valdez, D., & Patterson, M. S. (2022). Computational analyses identify addiction help-seeking behaviors on the social networking website Reddit: Insights into online social interactions and addiction support communities. *PLOS Digital Health*, *1*(11), e0000143. https://doi.org/10.1371/journal.pdig.0000143

Valdez, D., Thij, M. ten, Bathina, K., Rutter, L. A., & Bollen, J. (2020). Social Media Insights Into US Mental Health During the COVID-19 Pandemic: Longitudinal Analysis of Twitter Data. *Journal of Medical Internet Research*, *22*(12), e21418. https://doi.org/10.2196/21418

Vigdal, M. I., Moltu, C., Bjornestad, J., & Selseng, L. B. (2022). Social recovery in substance use disorder: A metasynthesis of qualitative studies. *Drug and Alcohol Review*, *41*(4), 974–987. https://doi.org/10.1111/dar.13434

Zimmerman, T. (2022). Social noise: The influence of observers on social media information behavior. *Journal of Documentation*, *78*(6), 1228–1248. https://doi.org/10.1108/JD-08-2021-0165