'Scraping' Reddit posts for academic research? Addressing some blurred lines of consent in growing internet-based research trend during the time of Covid-19.

ADAMS, N.N.

2024

Routledge
Taylor & Francis Group

'Scraping' Reddit posts for academic research? Addressing some blurred lines of consent in growing internet-based research trend during the time of Covid-19

Nicholas Norman Adams ⬤

School of Nursing, Midwifery and Paramedic Practice, Robert Gordon University, Aberdeen, UK

ABSTRACT

The global scale of Covid-19 has constrained academics from conducting much person-facing research. Reactively, trend is increasing for digital-based methodologies capturing already existing online data. Scholars often 'scrape' user-postings from internet forums using coding algorithms and text capture tools, before analysing data, drawing conclusions and publishing findings. The online social news aggregation and discussion website Reddit is a particularly rich source of data for researchers. The public nature of Reddit materials may suggest rationale for user-data to be replicated, analysed and archived; indefinitely and in multiple locations, for scholarly research. However, this position overlooks several key ethical considerations. This paper presents an overview and explanation of Reddit, followed by an exploration of studies that use Reddit-acquired data. Arising ethical issues are discussed, and solutions to salient dilemmas presented. This is to enhance awareness of potential problems and improve protections for those whose data is unknowingly used for research.

Introduction

Numerous studies suggest that human actors now live out many aspects of their lives online (Blackburn et al., 2005; Gibson et al., 2005; Scott et al., 2017), and that the amount of overall time human actors spend online is ever increasing as we move further into a digital age (Odgers & Jensen, 2020; Twenge & Farley, 2021). Time spent online has also increased further during time of the Covid-19 outbreak, with areas of teaching and learning, shopping, socialising, news and gaming all showing increased online activity impacted by the pandemic (Burnett et al., 2021; Kar et al., 2020; Singh et al., 2020; Zamboni et al., 2021).

Reactive methodological approaches are necessary for the challenging research climate of Covid-19, where many scholars have been unable to conduct much direct face-to-face research with participants. Such shifts in research methods are evident by increases in studies utilising data collected online from publically available internet discussion forums. Parallels are seen between this approach as naturally reactive to the research limitations of a Covid-19 climate, and the investigatory topics of studies. Notably, many recent studies focus on negative impacts of the Covid-19 pandemic, and utilise individual's online discussions to explore salient and important areas of social, psychological and interpersonal affect. Conducting such research, and the use of new methodologies is important. Novel online-based methods allow for a wide spectrum of research topics to be investigated, leading to new areas of academic discovery and inquiry. However, and despite the relevance and importance of some studies, it can at times appear that little consideration is

given to the protection of unaware and non-consenting online 'participants', from whom online data is gathered. Some studies are absent of ethical discussions regarding the natural absence of participant consent in online research. At other times, considerations are underexplored for how the downloading, analysis and (re)publication of conversational data materials – clearly for purposes other than 'participants' originally intended – may risk potentially harmful outcomes for these individuals.

Reddit: 'the front page of the internet'

The online discussion platform Reddit – as of 2021 – is ranked as the 18th most visited website in the world, and 7th most visited in the U.S. (Clement, 2021) [Reddit is commonly stylised as 'reddit'. For quotations, 'reddit' has been left in lowercase. For discussions, I have chosen to capitalise Reddit to remain cannon with much existing research]. Anyone with an internet connection is free to create an account and post on Reddit, and anyone with internet access can read Reddit postings. The primary purpose of Reddit is to facilitate the open discussion of almost any topic, and to bring people together into globally reaching forums to achieve this. As of 2021, there are over 137'000 subreddit spaces on Reddit. These 'subreddits' represent dedicated online topic spaces for the discussion of specific subjects. For example, R/Coronavirus is a subreddit forum for the discussion of Covid-19. Another subreddit forum is R/Anxiety and is dedicated to the 'Discussion and support for sufferers and loved ones of any anxiety disorder' (Reddit.com, 2021a). Reddit users typically post opinions, experiences, thoughts and discussion points relating only to the designated R/subreddit subject, often discussing lengthy and detailed personal experiences, sometimes asking for help, support and information, and sometimes simply sharing their thoughts and opinions. It is not uncommon for Reddit users to post to multiple different R/topic threads over a short period. The front page of any subreddit typically shows a hierarchical list of postings contained within the subreddit, arranged by recency, voted popularity, or volume of comments contained within each post. The name and title of the post, the username of the original poster (highlighted as 'OP'), and the number of subsequent discussion posts (or comments) contained within the subreddit discussion thread are also shown. For context, Figure 1 shows an annotated screenshot of the front page of the subreddit R/Anxiety page.
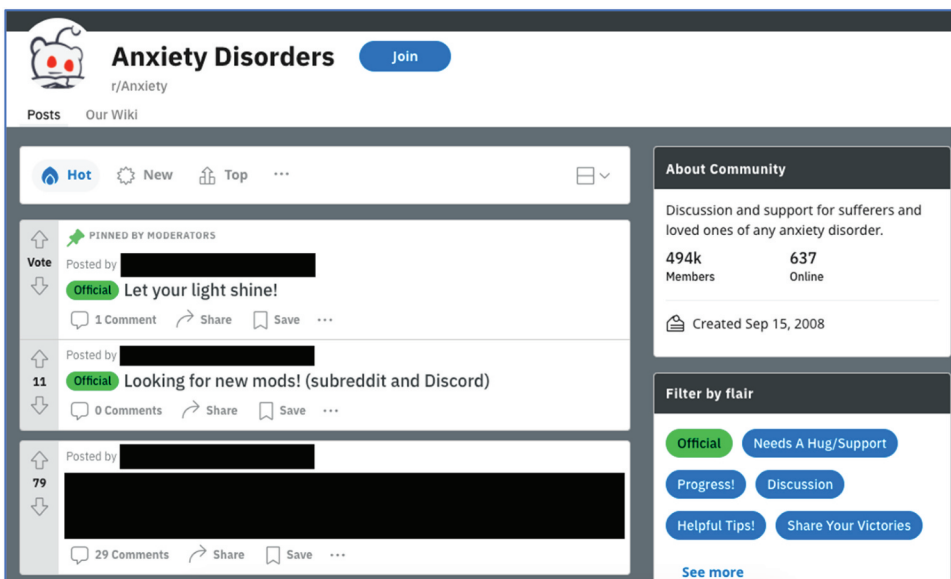
Figure 1. R/Anxiety subreddit. Image was captured on 27 October 2021. Note: *Any user identifying information has been blanked. Reddit layout, design, UI, content, logo and all elements seen within screenshot © 2021 Reddit. Single Reddit screenshot used for illustrative purposes, to provide context of a specific Reddit 'R/' page layout and appearance (fair use/fair dealing, single screenshot used for academic discussion/publication: non-commercial research purposes). Screenshot captured from: www.reddit.com/r/Anxiety/* (Reddit, 2021a).

The almost infinite-seeming nature of subreddit topics hosted on Reddit, and the volume of within-topic threads and large global user-base, lend to frequent framing of the Reddit platform as a highly rich resource of readily available 'real world' data for researchers, particularly for scholars interested in sociological and psychological study of human intentions, behaviours, reactions and processes relating to a wide and diverse range of subject matters (Amaya et al., 2021; Jung et al., 2021). Such thinking is reflected in the recent volume of academic studies that draw upon Reddit forums as the primary source of data for investigations. The following section discusses some of these studies in detail.

Studies that use data drawn from Reddit forums

Numerous recent studies have been conducted based upon online data harvested from Reddit. The methodologies by which data is acquired differs per individual study protocol. Reddit is coded using the Python programming language. Some studies use Python code to automate the capture of large datasets of user postings, before downloading and analysing such postings using specific tools. This method is frequently referred to as 'scraping' (see, Baumgartner et al., 2020; Jung et al., 2021). Another method is to combine Python coding with existing Google BigQuery (GbQ) datasets. GbQ represents existing stored datasets of complete user postings from a given subreddit (or multiple subreddits). Some studies use GbQ to indirectly retrieve large-scale Reddit datasets for analysis (see, Garg et al., 2021).

One of the most fascinating recent studies using Reddit data is the 'Natural Language Processing Reveals Vulnerable Mental Health Support Groups and Heightened Health Anxiety on Reddit During COVID-19: Observational Study' (Low et al., 2020). The goal of the research was to "[analyse and characterise] changes in 15 of the world's largest mental health support groups [...] found on the website Reddit (p. 1). Postings were analysed using Natural Language Processing (NLP); employing both regression and supervised machine-learning. This was to ascertain how individuals living with different mental health problems were uniquely impacted in the initial time of Covid-19 outbreak. The research paints a relevant and in-depth picture of online mental health support groups and online posting language and mental health, during a period of majority global lockdown due to Covid-19. Low et al. downloaded user postings from '15 of the world's largest mental health support groups' and '11 non–mental health groups' hosted on Reddit (p. 1). This was achieved by using the 'Pushshift' application programming interface (API) (for a detailed explanation see, Baumgartner et al., 2020). Broadly, Pushshift represents a method of 'scraping': capturing and analysing large quantities of Reddit data relating to a specific data range. Notably, the Pushshift API stores a copy of Reddit objects (a copy of all Reddit user postings). Data is copied into the Pushshift database from the moment it is posted to Reddit (see, Reddit.com, 2021b). For Low et al.'s study:

> Reddit] posts were extracted from fifteen subreddits focused on specific mental health communities (r/EDAnonymous, r/addiction, r/alcoholism, r/adhd, r/anxiety, r/autism, r/BipolarReddit, r/bpd, r/depression, r/healthanxiety, r/lonely, r/ptsd, r/schizophrenia, r/socialanxiety, and r/SuicideWatch) (p. 4)

This was in addition to:

> two broad mental health subreddits (r/mentalhealth and r/COVID19_support), and 11 nonmental health subreddits (r/conspiracy, r/divorce, r/fitness, r/guns, r/jokes, r/legaladvice, r/meditation, r/parenting, r/personalfinance, r/relationships, and r/teaching) (p. 4)

The work is notable for the breadth of data gathered and the volume of data analysed to develop outcomes. Especially salient of the research is that the full study dataset has been made publically available online: 'We created and released the Reddit Mental Health Dataset including posts from 826,961 unique users from 2018 to 2020' (p. 1).

Outcomes of the research are important and raise several key points with regard to mental health changes. Namely, the authors highlight spikes in several subreddit groups during the Covid-19 outbreak, linked to specific themes of anxiety and stress characterised by negative semantic shifts. Perhaps most saliently, the authors illustrate that sampled Reddit topic postings thematically linked to the contexts of loneliness and suicidality increased more than double during the initial outbreak of the Covid-19 pandemic, with increases also found in self-harm-related analysis clusters in some subreddit posting language. The authors conclude their findings '[uncover] patterns of how specific mental health problems manifest in language, identified at-risk users, and revealed the distribution of concerns across Reddit, which could help provide better resources to its millions of users' (p. 2). Additionally, they suggest study outcomes demonstrate:

> That textual analysis is sensitive to uncover mental health complaints as they appear in real time, identifying vulnerable groups and alarming themes during COVID-19, and thus may have utility during the ongoing pandemic and other world-changing events such as elections and protests (p. 2)

A further fascinating and important research study on similar topic to Low et al. is 'Reddit users' experiences of suicidal thoughts during the CoViD-19 [sic] pandemic: a qualitative analysis of r/Covid19_support posts' by Slemon et al. (2021). This research also responds to the impacts of the Covid-19 pandemic upon mental health. Specifically, the authors focus on analysing – using an iterative 'open coding' (p. 4), thematic analysis approach – 83 postings from the 'Covid19_support' subreddit. This is to examine how users posting within this community describe their experiences of suicidal thoughts. Data was collected directly from Reddit. This was achieved initially using 'a search of the subreddit using the term "suicide," which includes variations of the term (e.g. "suicidal")' (p. 3). The authors also discuss additional protocols with regard to searching. They highlight:

> Testing of search terms such as "kill myself" did not lead to the capture of additional posts; therefore, further terms were not used" [. . .] "Posts were included if the original poster identified experiencing suicidal thoughts, feelings, or behaviours during the COVID-19 pandemic (p. 4)

Importantly, research outcomes focus on representing the range of different responses involving the discussion of suicidal thoughts and how deteriorations in mental health are linked to the progression of the pandemic. Notably, factors such as financial pressures, continued social isolation, anxieties surrounding exposure to the Covid-19 virus, stressors over employment status, and inadequate access to mental health supports are highlighted as contributors to negative mental health state and related linguistic discussions on Reddit. An examination of coping strategies and self-constructed supports are also presented (see, Slemon et al., 2021).

Another study – by Nutley et al. (2021) – explores the impact of the Covid-19 pandemic on disordered eating behaviours. The research acknowledges the impact of the pandemic upon mental health, before suggesting existing studies point to the continued pressures of the pandemic holding negative – anxiety-mediated – impacts over those already suffering from a lifetime history of eating disorders. The authors downloaded user postings from several popular subreddits where discussion of eating disorders was the main topic. Like other studies, an inductive, thematic-based analysis framework was used to code and make sense of user-posting-materials, and develop these into linked thematic outcomes. Six primary themes were identified. These include changes in eating disorder (ED) symptoms, impacts of quarantine upon daily life, emotional wellbeing, changes in exercising routines, help-seeking behaviours, and wellbeing. Findings highlight that users voiced that the pandemic – and associated public health measures – contributed to negative mental health impacts. Notably, and similar to the findings of Low et al. (2020) and Slemon et al. (2021), feelings of frustration, anxiety and loneliness were shown to increase in presence within linguistic discussions by Reddit users, posting in eating disorder subreddit forums. In concluding, Nutley et al. make the tentative suggestion that Reddit ED discussion forums (largely) provide a therapeutic community for the sharing of information, experiences and supports to posters suffering an increased period of psychiatric distress.

Deconstructing existing research approaches to develop three salient ethical considerations for researchers planning to use data 'scraped' from Reddit

The following section draws upon the above three sampled studies to present three salient considerations for researchers considering an online-based approach to research. The following points highlight some specific considerations regarding the important issues of consent, context, and maintenance of 'participant' anonymity that are relevant to scholars planning to draw upon data collected from Reddit. These points are also salient for many other public, online-based discussion and message-board research contexts that may be used for scholarly study.

Considering participant consent and autonomy over data

A primary consideration when considering 'scraping' user data from Reddit – for the purpose of scholarly analysis and publication – is the use of such 'participant' data without explicit consent or permission from forum posters. An argument for collecting online-based data is often that materials lie in the public domain, and are therefore 'free game' for scholars wishing to copy, analyse and replicate such postings. For example, Slemon et al. (2021) state: 'All Reddit posts included in this analysis are publicly available data. Ethical approval was therefore not required for this study . . .' (p. 4). As such, 'participants' posting to Reddit are routinely not asked for their consent in the use, replication, or repurposing of their data. Nor, is such protected consent covered by the usual ethical stipulations attached to research institutions. An important concern for academics using Reddit-collected data is that persons posting to Reddit and subreddits chose to do so in the knowledge that the information they share is bound by a specific local user-agreement. Notably, for Reddit, this extends to user-copyright protection over content submitted and shared on the platform: 'Reddit is designed and supported for personal use only [. . .] You [the Reddit user] retain the rights to your copyrighted content or information that you submit to Reddit ("user content") . . .' (Reddit.com, 2021c). However, the user-agreement also makes clear exceptions; allowing the 'royalty-free', 'unrestricted' use and replication of any content posted by users:

> By submitting user content to reddit, you grant us a royalty-free, perpetual, irrevocable, non-exclusive, unrestricted, worldwide license to reproduce, prepare derivative works, distribute copies, perform, or publicly display your user content in any medium and for any purpose, including commercial purposes, and to authorize others to do so (Reddit.com, 2021c)

While such data replication *may* be permissible from a Reddit policy perspective, replication – or 'scraping' removes any user-autonomy in the handling and control of such user data in a manner that operates differently, and much more restrictively, than the Reddit user-agreement. Crucially, if a Reddit user becomes uncomfortable with something they have posted – for example, concerns that the posting may identify them or others, reflect an outdated opinion or feeling, or cause distress, they (almost always) have the option to return and remove this data from public access later (Reddit.com, 2021c). By replicating data without a user's permission, any control that lies with the original poster is lost. This perspective is important for academic research ethics, and has implications for the use of verbatim quotations, and for the depositing of user postings into shared and public datasets by scholars. Postings utilised in these manners are de-localised from their original source and may exist in a permanent – replicable – archive, where the user now has no influence or control over their temporalities, nor how this content may be used in intended or unintended ways that could impact them negatively. While Reddit users consent to the replication and use of their posted content, this consent does not represent a de facto agreement for a permanent relinquishment of their autonomy over any later content controls. Most importantly, the Reddit user-agreement rationalises that by posting to Reddit, users consent to a wider-use of their posted data. If a user then chooses to delete their original Reddit posting, as some users later

do, this should be taken as rationale for a *removal of consent* for user data to be used more widely, given that this data no longer exists in the original location for which rationale for duplication – and by extension, consent for replication – was first established. To clarify against Slemon et al.'s comments of public domain posting equating to a lack of requirement for ethical approval – then should a user delete their Reddit posting; removing this from the public domain, scholars should assume it unethical to continue to use, replicate, host or store this data. The knowledge for a user that even after deletion of their original Reddit posting or thread, a permanent record of this now-deleted material exists (and from which they may be identifiable) is a risk-factor to participant harm that warrants serious consideration for researchers. Further conversations are required surrounding how informed consent may be sought for specific cases and the above concerns mitigated. These considerations are particularly relevant for studies that archive user's online posting data into online – publically available datasets. For example, Low et al.'s Reddit Mental Health Dataset which '[includes] posts from 826,961 unique users from 2018 to 2020' (Low et al., 2020, p. 1).

Considering research context and acknowledging the vulnerabilities of participants

At present, there is a clear scholarly trend for some studies to focus on examining especially vulnerable groups of individuals by using data 'scraped' from Reddit. Trend has increased during the time of Covid-19, with many studies focussing on the important topic of how the pandemic has influenced mental health. Study can involve the 'scraping' of Reddit postings from subreddit topic boards focussing on suicide, eating disorders, and mental health among others. This is evident in the above studies discussed (e.g. Low et al., 2020; Nutley et al., 2021; Slemon et al., 2021). Importantly, future researchers should consider the role of context when selecting user quotations to present as research findings, before removing these postings from their wider contextual Reddit message-board narrative. To clarify, many Reddit forums position themselves as 'safe spaces' where users can discuss various struggles. Users posting on these forums do so in the knowledge that postings are contextualised within a wider, local topic board conversation: i.e. the 'safe space', which is policed by local online moderators. For example, Nutley et al. (2021) focus on downloading postings from the subreddit R/AnorexiaNervosa. The welcome banner for this subreddit reads: 'Anorexia Nervosa is a real, serious illness that affects thousands upon millions of people daily. [...] This is a safe place for those with this illness, and for those that are in recovery ...' (Reddit.com, 2021d). Similarly, the r/EDAnonymous subreddit welcome banner – one of the sources of data used for Low et al.'s study – reads: 'Eating Disorders Anonymous: A safe space to discuss our struggles with eating disorders' (Reddit.com, 2021e). Removing postings from this context – or selecting partially represented quotations – serves to isolate specific narratives, topic threads and conversation fragments as stand-alone discourse. While users may comment contextually on a specific thematic thread or discussion as part of a wider understanding that such discussions occur in the advertised 'safe space', no such guarantee of any 'safe space' dynamics can be offered when user comments are replicated and displayed in an isolated manner without full context in a publication or a dataset archive, as has occurred in practice with research discussed in this publication. While Reddit users may consent – as discussed earlier – to their postings existing in the wider Reddit context, the isolation of postings to represent an unalterable permanent record, may cause significant anxiety for already vulnerable individuals and groups should they recognise that their comments are now presented outside their original context, immortalised and uneditable; viewable to the public, and potentially used to support arguments or findings for which users did not consent to (and – completely conceivably – may not necessarily agree with). This topic requires further scholarly focus and open discourse to assess possible risks and approach possible solutions regarding the scholarly use of raw public Reddit data, to explore and adequately consider the possible impacts upon non-consenting research 'participants'.

Considering the risk of participant identification

Perhaps the most serious consideration to come to light within this publication, is the possibility that non-consenting participants are not sufficiently anonymised when their Reddit posting data is used for scholarly research. A hallmark of much in-person research study is to maintain the anonymity of research participants (Bos et al., 2009; Brewer, 2000). The possible identification of participants is a legitimate – and most significant – concern when working with data acquired from online discussion forums such as Reddit. Most saliently, replication of Reddit user postings – verbatim – in scholarly publications can often lend to internet reverse-searching. In some cases, this could allow the original Reddit threads to be easily and rapidly located online, therefore risking invalidation of any assumed 'participant' anonymity and allowing the linking of specific isolated comments used in publications to specific user accounts and postings [this is unless the Reddit user has posted using a 'throwaway' account – 'throwaway' accounts are somewhat common on Reddit. The descriptor depicts a new account that a user has created specifically to post to, or respond to a topic of sensitivity within which they wish to remain anonymous, i.e. not using their regular Reddit account]. A similar concerning trend is sometimes present in data-archives created using Reddit data, where the poster's original Reddit username can be left displayed. These concerns are immediately upheld when considering some of the research studies discussed earlier. For online-based studies that are conducted at UK institutions, ready guidance is available that helps safeguard the anonymity of online users during scholarly internet-based research. The British Psychological Society (BPS), and the British Sociological Association (BSA) both provide substantial guidance on the practices of ethics when using data gathered from online sources. Regarding anonymity in the absence of informed consent, the BPS guidelines outline:

> Where it is reasonable to argue that there is likely no perception and/or expectation of privacy [. . .] use of research data without gaining valid consent may be justifiable. However, particular care should be taken in ensuring that any data which may be made accessible as part of the research remains confidential [. . .] often achieved by ensuring anonymity [. . .] (British Psychological Society [BSA], 2017, p. 8)

Similarly, the BSA guidelines print: 'For web data [. . .] personal details should be disguised; however, quotes may be used to evidence any findings' (BSA, 2016, p. 8). However, for some data, using verbatim quotations can allow the original Reddit threads to be found, therefore risking the invalidation of any user anonymity and allowing the linking of specific isolated comments to wider user accounts and postings. Notably, scholars in some of the studies above have attempted steps to secure the anonymity of user data. For example, Slemon et al. (2021) acknowledge:

> [. . .] analysis of Reddit and other social media data is an emerging area of inquiry, and both researchers and institutional review boards have grappled with ethical engagement with these forms of public data, including possible strategies for mitigating risks related to confidentiality and consent. Given the personal and sensitive nature of the posts and comments and following from other researchers' use of Reddit posts as a qualitative data source for understanding mental health experiences [. . .], we do not report specific dates of posts and we have replaced all usernames with numerical identifiers (p. 4)

Slemon et al. highlight recognition for the sensitivity of the data they are working with, and pinpoint the dates of postings and Reddit usernames as the most salient points of identification for user-data. Similar points of discussion are highlighted in other research discussing online ethics (see, Bos et al., 2009). However, for Slemon et al.'s study, illustrative quotations themselves used within their research are readily reverse-searchable, lending to discovery of the original research materials, Reddit user-name, and links to other postings. Figure 2 shows the results of a single Google reverse-search for a randomly selected quotation from Slemon et al.'s research study.

Similarly, reverse-quotation searches for a randomly selected illustrative quotation from Low et al.'s (2020) study and another from Nutley et al.'s (2021) study reveal direct links to original user Reddit data – again, invalidating the anonymity of the original poster by linking to their user name. Figures 3 and 4 below show Google reverse-search results from Low et al.'s and Nutley et al.'s studies, respectively.
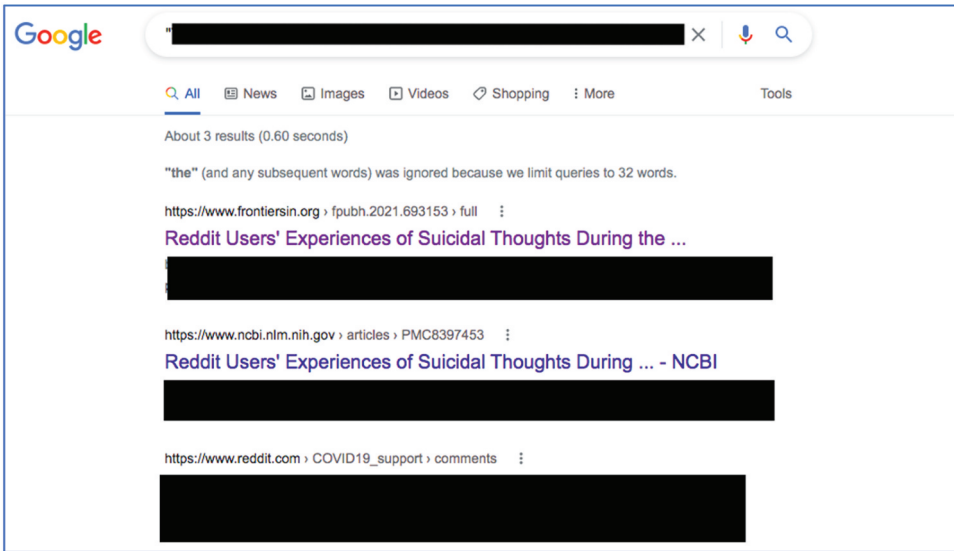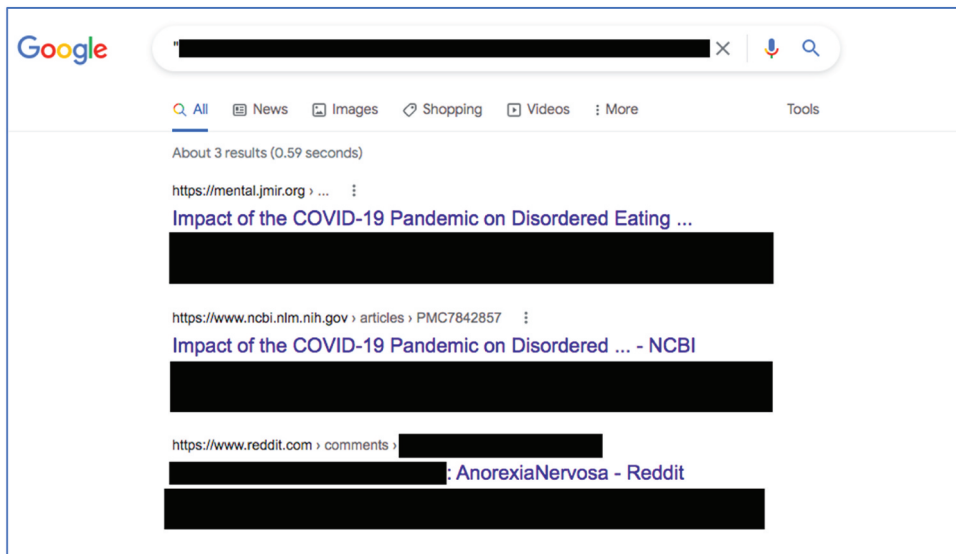
Figure 2. Results of a single (non-advanced) google reverse-search for a randomly selected quotation in Slemon et al. (2021). Results show the publication in question (blanked), an alternative hosting location (blanked), and the original Reddit thread complete with original user details visible (blanked). The link leads to the original source posting, including user-name and date of posting (partially blanked to preserve anonymity for this publication; Google.com, 2021a). Google screenshot(s) used for illustrative purposes, to provide context to Google Search Discussion. Google and the Google logo are trademarks of Google LLC.
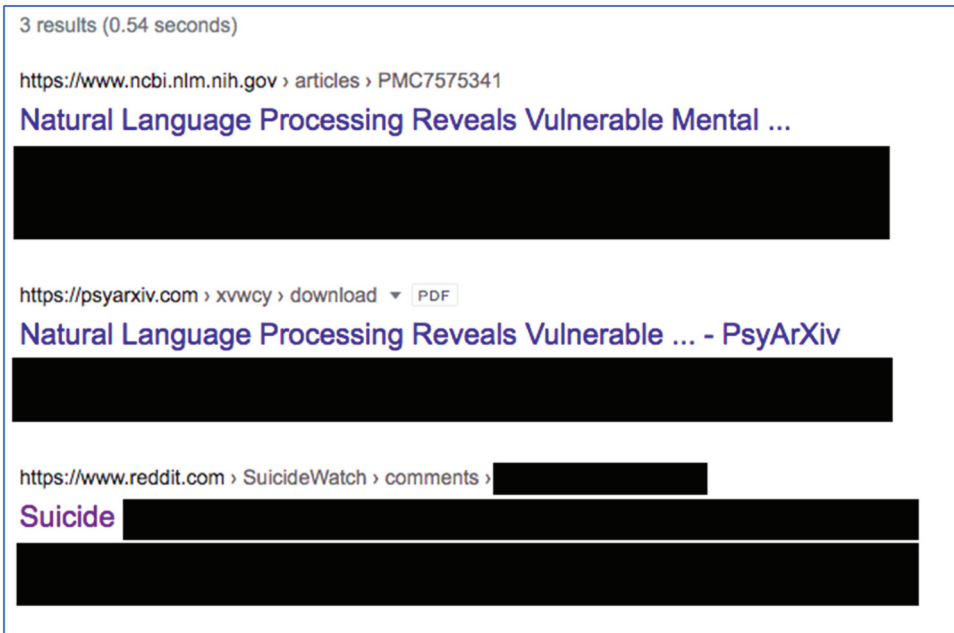
Figure 3. Results of a google reverse-search for a randomly selected quotation from Low et al.'s study. Results show the publication in question (blanked), an alternative hosting location (blanked), and the original Reddit thread link – complete with original user details visible (partially blanked; Google.com, 2021b). Google and the Google logo are trademarks of Google LLC.

Evaluating the ease by which research 'participant's' Reddit accounts may be readily identified is concerning and marks a valid point of consideration for the design of future research studies drawing on user-data collected from Reddit. Where possible, scholars should seek to employ

Figure 4. Results of a google reverse-search for a randomly selected quotation from Nutley et al.'s study. Results show the publication in question (blanked), an alternative hosting location (blanked), and the original Reddit thread complete with original user details visible (partially blanked; Google.com, 2021c). Google and the Google logo are trademarks of Google LLC.

additional practices to make every effort to anonymise Reddit collected data, in a comparable manner to what they would deem appropriate with in-person participant research. This is to avoid inadvertently causing any potential harm to 'participants' who discover that their comments now exist in a permanent archive, for which they have no control over the ways in which such comments are used; no autonomy and decision over the deletion of these materials, nor access to any procedure from which to de-associate these comments with their Reddit username.

Discussion

Given the issues raised in the sections above, the possible impacts upon participants may be significant. The most salient concerns can be split into three areas: *replication, consent*, and *identification*. Replication of data 'scraped' from Reddit forums is of concern for users due to firstly: the decontextualized nature in which data is held, and secondly: the nature of this data now representing a permanent record of which the original originators of this data have no control over. To address the first point: unlike in-person research, where conversations, comments, occurrences, and observations can be considered as fleeting and dynamic; always anchored in their raw-data within a wider contextual conversation or occurrence, data 'scraped' from Reddit and represented as verbatim quotations in publications (or in a dataset archive) exists in a unidimensional bubble. Users, particularly those suffering mental health conditions like anxiety, depression, and those experiencing suicidal ideation may post comments with little consideration for how these may be displayed devoid of context or in a location other than embedded within a wider contextual narrative of Reddit and associated subreddits. As discussed earlier, these online contexts are frequently framed as 'safe spaces', and they document assurances to users regarding their rights over data, via both signposting on the various subreddits, and via the Reddit user agreement. Importantly, should users regret or experience later concerns or anxiety over any materials posted to Reddit, they also have the opportunity to remove these, yet not-so if these postings are replicated and re-located into a new location for which users have no editorial control. In this context, the second point of absent participant consent is particularly relevant. A central characteristic of in-

person gathered research materials is that data collected is typically institutionally held securely (as opposed to publically held openly), remains confidential, and assurances of protection are given to participants (see, Brewer, 2000). An additional hallmark of much in-person participant focussed data is the right for participants to withdraw data from the study – at any time – without incurring any penalty, or being asked why they wish to do so. A parallel to this protection is evident when considering the original nature of participant Reddit postings. As mentioned earlier, if public-posting is taken as rationale for a lack of requirement for informed consent (and sometimes, absence of a need for institutional ethical approval) then considerations need to be approached for how publically-held data will be revisited should original Reddit posting materials be later removed. At a minimum, this suggests that the Reddit user instigating removal is no longer comfortable with the content of the posting. This may be for concerns over identification, harm, impact, recovery from illness, or a change in perception or beliefs. Allowing such posting materials to remain discoverable via a web-search, particularly where the 'participants' user name is also visible, is an under-discussed, yet potentially harmful research practice, which must be addressed in additional ethical considerations for online-based research studies moving forward.

The issue of participant identification interlinks with the above two points. In a growing digital era where much of contemporary life is lived-out online, it is not uncommon for individuals to share usernames between web-based environments. Nor is it uncommon for users to post information to online forums which can identify them – both their virtual presence, and their identity in real-life. Even less uncommon is it for prospective employers, partners, friends and acquaintances to Google-search usernames and real-names. When verbatim quotations from Reddit users are easily discoverable, or when user-posted data are held in a permanent archive and interlinked with identifiable information, and where originators of data have not actively consented to the use of their data – users faced with the discovery of their Reddit postings online – and their sensitive subject-matter, may be damaged, hurt, or placed in circumstances that force them to reveal or discuss topics they previously felt they were protected from through digital anonymity or – most saliently – via the protections afforded and advertised by Reddit for discussions as confined to within a digital 'safe space'. Such concerns are further amplified in the context of the studies discussed earlier, where posters often self-identify as young and vulnerable, and suffering from anxiety, depression, suicidal thoughts, eating disorders and other illnesses, and have been adversely affected by the Covid-19 pandemic.

Importantly, this article has largely dealt with some of the possible ethical pitfalls of conducting online research using Reddit. However, it must also be clearly acknowledged that all the studies discussed in this article investigate timely and relevant research questions, and present relevant and beneficial conclusions that lend to further study of mental health, the impact of the Covid-19 pandemic and the possible positive benefits of users networking on Reddit as a mechanism of support.

Complementing the discussion of the above studies that highlight some points of note regarding future ethical considerations, a further, notable study by Lyons et al. (2021) also explored mental health effects of the Covid-19 pandemic using data collected from Reddit. Scholars focussed on evaluating the impact of the pandemic – either positive or negative – upon persons already suffering experience of psychosis. Like others, the authors turned to Reddit to identify 'participants', focussing on gathering user-posting data from 'those who self-identified as having psychosis via Reddit discussion forum posts' (p. 1). Also, like other studies, an inductive, thematic analysis methodology was used to analyse postings, which consisted of 65 postings from a single subreddit, that contained over 16,600 total users who have experience of psychosis (see p. 2). Postings were selected from the subreddit using a keyword search for the phrases 'COVID, corona, virus, and pandemic' (p. 4). The authors note that the search was concluded when all relevant postings matching keyword search were downloaded; totalling 65 postings. Analysis of user postings highlighted five important themes, including declining mental health, changed experience of psychosis, and personal coping experiences among others. Notably, research concluded that 'participant' experience of the

pandemic was variable, highlighting that while self-identified sufferers of psychosis posting on Reddit voiced negative experiences, thoughts and feelings – for example, increases in the use of medications and fears of becoming dependent and running out of medications, increased social isolation and low social supports (p. 5) – others voiced positive narratives that upheld pandemic experiences of increased understanding and stigma reduction (p. 7), alongside motifs of resilience (p. 7).

Perhaps most notable of the research is the attention to ethical considerations for non-consenting 'participants'. Firstly, the research documents a substantial ethical review process, consultation of various ethical guidelines for online research, and openly addresses the issue of absent consent. On this subject, Lyons et al. state:

> When designing and conducting the study and reporting our findings we consulted ethical guidelines, previously published research, and available guides [...] In particular, we considered the public or private nature of the information shared, the potential for benefit or harm, and the feasibility of seeking informed consent [...] For example, users may feel distressed about their posts being analysed for a scientific study without their consent ... (p. 2)

Scholars then go on to discuss the sensitive nature of the information collected from Reddit and how any concerns surrounding the exposure of personal information in the context of absent 'participant' consent and data replicability may be softened in severity, by introducing several countermeasures:

> As researchers, we are fully aware that the posts contain sensitive information, and the forum users are not aware that their discussions are used for research. However, our aim is not to analyse personal characteristics of individuals, but to gain a broader understanding of the issues that affect people who self-identify as experiencing psychosis during the pandemic (p. 2)

Importantly, Lyons et al. also outline a plan to preserve the anonymity of the postings (and user identities) that they download and classify as research data. They discuss:

> In mitigating potential harm, we adopted a number of measures in accordance with professional body guidelines (e.g., BPS, 2017) and advice from discussion forum researchers (Smedley & Coulson, 2021) to protect the anonymity of the forum users. For example, we do not reveal online usernames, and have allocated a number to each username to preserve that anonymity. In addition, we altered the wording of the quotes, and entered the quotes in google search engine and the subreddit site to make sure the quote cannot be traced back (p. 2)

The study represents an example of a thoughtful and careful approach to conducting online-based research that also 'scrapes' user-postings from Reddit. The research demonstrates a considerate approach to the handling, replication and anonymization of user Reddit data that lends to this data being used in a manner that concludes important research outcomes surrounding mental health in the time of Covid-19, yet also actively protects the unaware and non-consenting research 'participants' whose data is under examination, from any possible harm through accidental identity reveal.

Applying ethical considerations to online platforms beyond Reddit

Salient of the considerations raised in this research is to evaluate whether the issues highlighted vis-à-vis Reddit affect research conducted using material gathered from other social networking and comment aggregation sites. While Reddit as a platform is arguably unique in some of its properties, for example, the volume and grouping of threads and users, the particulars of the ways in which Reddit content is stored and is openly accessible to scholars wishing to 'scrape' such content for research purposes is similar to some other platforms. For instance, postings on the social question-and-answer-based knowledge sharing website Quora are often likened to Reddit threads. This is in part due to the growing popularity of Quora and its user base. On Quora, users typically post a question or comment, which others are then free to answer, debate or discuss by posting a response. Like Reddit, users can respond and quote specific user comments to engage in a 'back-

and-forth' discussion. All comments are visible (for an overview of Quora and the social connections between postings and users, see, Wang et al., 2013). Evaluating the ethical specifics of 'mining' data from Quora is complicated by several factors, notably, Quora previously (up until April 2021) required individuals to register an account with their real name, rather than the internet-based pseudonyms commonplace on Reddit. As such, some Quora accounts still display this. Quora user accounts are also designed to contain photographs of users, rendering an additional layer of concern over scholars using data from this platform for research, due to the gathering of this data representing directly identifiable participant information, for which users should be asked to explicitly consent to the storage and usage of. Users should be offered guarantees of data protection and preservation of anonymity, in keeping with the traditions of in-person research (Bos et al., 2009; Brewer, 2000). Also, according the Quora terms of user policy, some search-engines may index Quora user profiles into their search databases (Quora, 2022). This suggests that user profiles, possibly complete with user-photo and the user's real name may be discoverable when user-posted information is searched using even a basic internet search query, for example, in Google. These particulars are confounded further when considering these factors against some of the ethical pitfalls already outlined in this research. Namely, as with Reddit, the searching of quotations via Google could return information that reveals Quora users' identifying details. Also –as applied to Reddit – the possibility that researchers may archive collected user-postings as datasets into publicly accessible repositories to represent a 'snapshot' of user commentaries, is a significant ethical consideration for data collected from Quora, as this 'scraped' information may also contain users' identifying details that Quora attaches to postings. Further, it is reasonable to expect that the potential for user-identifying data (i.e. names and photographs) to be collected is greater where automated 'scraping' tools are used to capture large quantities of information without any manual – human – checking over whether such information can identify participants, and where the subsequent manual deletion of such sensitive data is not factored as a necessary step during research ethics and data management considerations.

While the volume of academic works that use data drawn from Quora appears significantly less than studies drawing on data 'scraped' from Reddit, some notable scholarship exists (e.g. Lima et al., 2014). Enquiries also sit alongside research that combines data drawn from both Reddit and Quora (e.g. Roundtree, 2018). suggesting that the similarity of the platforms is recognised by scholars focussed on using online-acquired data to develop scholarship. As discussed of studies that use data from Reddit, and considering the points and similarities outlined above, the use of data from Quora appears interlinked with much of the same pitfalls as data collected from Reddit, and the key ethical considerations outlined in this research (data replication, consent, and user identification) are equally salient for scholars drawing data from Quora.

Important also is to consider the ethical considerations of this research much more generally, in the context of using *any* data collected online from Reddit-like platforms. In exploring this point, recent online research appears to draw from a wide range of online locales. For example, scholars have studied addiction and support-seeking on platforms such as Mumsnet (Lee & Cooper, 2019). Mumsnet has also been the focus of research studying the topics of motherhood and entrepreneurship (Makola, 2022). Scholars have also studied non-suicidal-self-injury (Eliseo-Arras et al., 2019), suicidal ideation (Ali & Gibson, 2019; Cavazos-Rehg et al., 2017), and eating disorders (Gies & Martino, 2014) on the public picture and comment sharing platform Tumblr.

Others have 'scraped' user-data from online platforms including the picture sharing site Instagram (Sharabi & Hopkins, 2021), bodybuilding web and mobile applications (Molina & Myrick, 2021;Molina & Sundar, 20182018), Twitter (Kusumasari & Prabowo, 2020), and Facebook (Laor, 2019) among many others. However, and despite the range of online locales from which data is collected, and diversities in the topics such data is used to examine, most online platforms share the same basic premise built around the public display of commentaries and user-names, similar to the model of Reddit (and Quora). Correspondingly, most qualitative scholarly studies share a common paradigm of utilising – and positioning – publicly available

online data as a rich source of easily accessible material for scholarly study. Some studies take great care to anonymise data and paraphrase quotations. However, these cautions cannot be overstated when using online data, as this manuscript makes clear. For these reasons, it is logical to suggest that while this manuscript has focussed on systematically highlighting potential ethical pitfalls when using data drawn from Reddit for the purposes of academic study, it is expected that these points of ethical consideration are equally salient, and directly translatable with regard to scholars gathering data from many other online-based social and comment aggregation platforms that function in similar ways. While the use of Reddit – and subsequent data-collection of Reddit user postings from users by academics – appears particularly popular, likely due to Reddit's global user-base, international reach and volume of regular users, gathering data from Twitter, Facebook, Instagram, and other comment and message board-type online locales likely attach the same ethical dilemmas discussed throughout this work. Scholars will benefit from considering the points outlined in this manuscript when considering 'scraping' data from Reddit, but also more universally, involving the platforms mentioned above and on any like internet space where user comments are freely posted; visible and accessible to anyone with internet access.

Conclusion

In the time of Covid-19 – and looking beyond this – online-based research is increasing. In contemporary society – and particularly in the wake of Covid-19 – many aspects of human life are increasingly lived out and shared online. Reactive to a restrictive research climate, a scholarly trend where researchers 'scrape': download and analyse, rich and wide-ranging message-board or 'thread' conversations posted by users on Reddit, has become a popular way of conducting meaningful, impactful and relevant research. At times, however, internet-based research appears to be associated with less overt considerations for ethical review, consent, anonymity, and data-protection. As this publication demonstrates, additional considerations regarding enhancing the safety and wellbeing of 'participants', and minimising any ethical risks, can and should be approached. More specifically, this publication clearly discusses three of the most salient areas where additional ethical considerations should be actively factored in the wake of highlighted risks. These observations are raised with a view to enhance the protection of individuals posting on online forums, and most specifically Reddit, who are unaware of their non-consenting role in participating in research study and the overall wider use, storage and dissemination of their original Reddit discussion materials, which can – at times – be used to identify their original posting materials, and Reddit username. While this publication does in no way suggest that using such online-based materials, and using Reddit as a primary data-source is wholly problematic and must be ceased, careful handling and anonymization of such materials is of paramount importance for maximising ethical research practice going forward. This careful and considerate perspective is demonstrated by exampling one such study that records such an approach. In concluding, scholarly conversation should immediately be widened regarding engagement on the topic of online research and ethical mindfulness. This is to proactively consider some potential ethical limitations regarding this methodology that are presently under discussed. Relevance is particularly due to increased – and rapid – uptake in such online methods following the Covid-19 outbreak, the continuing impact of the pandemic on research, and rising research trend for studying already vulnerable groups of individuals turning to Reddit online discussion forums in search of support during this difficult time.

Disclosure statement

Funding

Notes on contributor

Nicholas Norman Adams is a Chartered Psychologist; a member of the British Psychological Society, and a Science Council Registered Scientist. His academic interests are interdisciplinary and draw from applied psychology and sociology. His research focus is most centralised upon ethnographically examining the influences of societal gender constructs upon human behaviour; with particular interest in men and masculinities, the cultures of industrial workplaces, and the linkages between identity, expectations and safety/risk behaviours. He also has strong research interests in mental health at work, and men's mental health and health-seeking practices, alongside interests in established – and emerging – research methodologies.

ORCID

Nicholas Norman Adams http://orcid.org/0000-0003-1999-1134

References

Ali, A., & Gibson, K. (2019). Young people's reasons for feeling suicidal: An analysis of posts to a social media suicide prevention forum. *Crisis: The Journal of Crisis Intervention and Suicide Prevention*, *40*(6), 400. https://doi.org/10.1027/0227-5910/a000580

Amaya, A., Bach, R., Keusch, F., & Kreuter, F. (2021). New data sources in social science research: Things to know before working with reddit data. *Social Science Computer Review*, *39*(5), 943–960. https://doi.org/10.1177/0894439319893305

Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., & Blackburn, J. (2020, May). The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media* (Vol. 14, pp. 830–839).

Blackburn, C., Read, J., & Hughes, N. (2005). Carers and the digital divide: Factors affecting Internet use among carers in the UK. *Health & Social Care in the Community*, *13*(3), 201–210. https://doi.org/10.1111/j.1365-2524.2005.00547.x

Bos, N., Karahalios, K., Musgrove-Chávez, M., Poole, E. S., Thomas, J. C., & Yardi, S. (2009). Research ethics in the Facebook era: Privacy, anonymity, and oversight. In *CHI'09 extended abstracts on human factors in computing systems* (pp. 2767–2770).

Brewer, J. (2000). *Ethnography*. McGraw-Hill Education (UK).

British Psychological Society. (2017). *Ethics guidelines for Internet-mediated research* (INF206/04.2017). Retrieved March 1st, 2020. from https://beta.bps.org.uk/sites/beta.bps.org.uk/files/Policy%20-%20Files/Ethics%20Guidelines%20for%20Internet-mediated%20Research%20%282017%29.pdf

British Sociological Association. (2016). *Researching Online forums*. Ethics Case Study. (1). Retrieved March 1st, 2020. from https://www.britsoc.co.uk/media/24834/j000208_researching_online_forums_-cs1-_v3.pdf

Burnett, H., Olsen, J. R., Nicholls, N., & Mitchell, R. (2021). Change in time spent visiting and experiences of green space following restrictions on movement during the COVID-19 pandemic: A nationally representative cross-sectional study of UK adults. *BMJ Open*, *11*(3), e044067. https://doi.org/10.1136/bmjopen-2020-044067

Cavazos-Rehg, P. A., Krauss, M. J., Sowles, S. J., Connolly, S., Rosas, C., Bharadwaj, M., Grucza, R., & Bierut, L. J. (2017). An analysis of depression, self-harm, and suicidal ideation content on Tumblr. *Crisis: The Journal of Crisis Intervention and Suicide Prevention*, *38*(1), 44. https://doi.org/10.1027/0227-5910/a000409

Clement, J. (2021, January 28) *Regional distribution of desktop traffic to Reddit.com as of December 2020, by country*. Statista. Retrieved August 25th, 2021, from. https://www.statista.com/aboutus/our-research-commitment/408/j-clement>

Eliseo-Arras, R. K., Brous, R., & Sheppard, S. M. (2019). A content and thematic analysis of Tumblr posts related to non-suicidal self-injury. *Journal of Ethnographic & Qualitative Research*, *13*(3).

Garg, S., Taylor, J., El Sherief, M., Kasson, E., Aledavood, T., Riordan, R., Kaiser, N., Cavazos-Rehg, P., & De Choudhury, M. (2021). Detecting risk level in individuals misusing fentanyl utilizing posts from an online community on Reddit. *Internet Interventions*, *26* [December 2021] , 100467. https://doi.org/10.1016/j.invent.2021.100467

Gibson, R. K., Lusoli, W., & Ward, S. (2005). Online participation in the UK: Testing a 'contextualised'model of Internet effects. *The British Journal of Politics and International Relations*, *7*(4), 561–583. https://doi.org/10.1111/j.1467-856x.2005.00209.x

Gies, J., & Martino, S. (2014). Uncovering ED: A qualitative analysis of personal blogs managed by individuals with eating disorders. *The Qualitative Report*, *19*(29), 1.

Google.com. (2021a). [Google search screenshot image]. Retrieved November 11th, 2021a, from <Google.com/ >

Google.com. (2021b). [Google search screenshot image]. Retrieved January 23, 2021b, from <Google.com/>

Google.com. (2021c). [Google search screenshot image]. Retrieved November 11th, 2021c, from <Google.com/ >

Jung, Y., Lee, Y. K., & Hahn, S. (2021). Web-scraping the Expression of Loneliness during COVID-19. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 43, No. 43).

Kar, S. K., Arafat, S. Y., Sharma, P., Dixit, A., Marthoenis, M., & Kabir, R. (2020). COVID-19 pandemic and addiction: Current problems and future concerns. *Asian Journal of Psychiatry*, *51* [June 2020] , 102064. https://doi.org/10.1016/j.ajp.2020.102064

Kusumasari, B., & Prabowo, N. P. A. (2020). Scraping social media data for disaster communication: How the pattern of Twitter users affects disasters in Asia and the Pacific. *Natural Hazards*, *103*(3), 3415–3435. https://doi.org/10.1007/s11069-020-04136-z

Laor, T. (2019). Hello, is there anybody who reads me? *Radio Programs and Popular Facebook Posts*.

Lee, E., & Cooper, R. J. (2019). Codeine addiction and internet forum use and support: Qualitative netnographic study. *JMIR Mental Health*, *6*(4), e12354. https://doi.org/10.2196/12354

Lima, M. C., Namaci, L., & Fabiani, T. (2014). A netnographic study of entrepreneurial traits: Evaluating classic typologies using the crowdsourcing algorithm of an online community. *Independent Journal of Management & Production*, *5*(3), 693–709. https://doi.org/10.14807/ijmp.v5i3.171

Low, D. M., Rumker, L., Talkar, T., Torous, J., Cecchi, G., & Ghosh, S. S. (2020). Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on reddit during covid-19: Observational study. *Journal of Medical Internet Research*, *22*(10), e22635. https://doi.org/10.2196/22635

Lyons, M., Bootes, E., Brewer, G., Stratton, K., & Centifanti, L. (2021). " COVID-19 spreads round the planet, and so do paranoid thoughts". A qualitative investigation into personal experiences of psychosis during the COVID-19 pandemic. *Current Psychology (New Brunswick, N.J.)* [October 2021] , 1–10. https://doi.org/10.1007/s12144-021-02369-0

Makola, Z. S. (2022). Mumpreneurs' experiences of combining motherhood and entrepreneurship: A netnographic study. *Journal of Contemporary Management*, *19*(si1), 1–20. https://doi.org/10.35683/jcm20129.139

Molina, M. D., & Myrick, J. G. (2021). The 'how'and 'why'of fitness app use: Investigating user motivations to gain insights into the nexus of technology and fitness. *Sport in Society*, *24*(7), 1233–1248. https://doi.org/10.1080/17430437.2020.1744570

Molina, M. D., & Sundar, S. S. (2018). Can mobile apps motivate fitness tracking? A study of technological affordances and workout behaviors. *Health Communication*, *35*(1), 65–74 .

Nutley, S. K., Falise, A. M., Henderson, R., Apostolou, V., Mathews, C. A., & Striley, C. W. (2021). Impact of the COVID-19 pandemic on disordered eating behavior: Qualitative analysis of social media posts. *JMIR Mental Health*, *8*(1), e26011. https://doi.org/10.2196/26011

Odgers, C. L., & Jensen, M. R. (2020). Annual research review: Adolescent mental health in the digital age: Facts, fears, and future directions. *Journal of Child Psychology and Psychiatry*, *61*(3), 336–348. https://doi.org/10.1111/jcpp.13190

Quora. (2022). [Quora privacy policy]. Retrieved June 14th, 2022, from https://www.quora.com/about/privacy

Reddit.com. (2021a). [Reddit R/Anxiety homepage]. Retrieved October 27, 2021a, from <https://www.reddit.com/r/Anxiety/>

Reddit.com. (2021b). [Pushshift FAQ]. Retrieved November 3rd, 2021b, from <https://www.reddit.com/r/pushshift/comments/bcxguf/new_to_pushshift_read_this_faq/>

Reddit.com. (2021c). [Reddit user agreement]. Retrieved January 13, 2021c, from <https://www.reddit.com/wiki/useragreement>

Reddit.com. (2021d). [Reddit R/AnorexiaNervosa homepage]. Retrieved November 11th, 2021d, from. <https://www.reddit.com/r/AnorexiaNervosa/>

Reddit.com. (2021e). [Reddit R/EDAnonymous homepage]. Retrieved November 11th, 2021e, from <https://www.reddit.com/r/EDAnonymous>

Roundtree, A. K. (2018, July). From engineers' tweets: Text mining social media for perspectives on engineering communication. In *2018 IEEE international Professional Communication Conference (ProComm)* (pp. 6–15). IEEE.

Scott, C. F., Bay-Cheng, L. Y., Prince, M. A., Nochajski, T. H., & Collins, R. L. (2017). Time spent online: Latent profile analyses of emerging adults' social media use. *Computers in Human Behavior*, *75* [October 2017] , 311–319. https://doi.org/10.1016/j.chb.2017.05.026

Sharabi, L. L., & Hopkins, A. (2021). Picture perfect? Examining associations between relationship quality, attention to alternatives, and couples' activities on Instagram. *Journal of Social and Personal Relationships*, *38*(12), 3518–3542. https://doi.org/10.1177/0265407521991662

Singh, P., Cumberland, W. G., Ugarte, D., Bruckner, T. A., & Young, S. D. (2020). Association between generalized anxiety disorder scores and online activity among US adults during the COVID-19 pandemic: Cross-sectional analysis. *Journal of Medical Internet Research*, *22*(9), e21490. https://doi.org/10.2196/21490

Slemon, A., McAuliffe, C., Goodyear, T., McGuinness, L., Shaffer, E., & Jenkins, E. K. (2021). Reddit users' experiences of suicidal thoughts during the CoViD-19 pandemic: A qualitative analysis of r/Covid19_support posts. *Frontiers in Public Health*, *1175* [August 2021] .

Smedley, R. M., & Coulson, N. S. (2021). A practical guide to analysing online support forums. *Qualitative Research in Psychology*, *18*(1), 76–103. https://doi.org/10.1080/14780887.2018.1475532

Twenge, J. M., & Farley, E. (2021). Not all screen time is created equal: Associations with mental health vary by activity and gender. *Social Psychiatry and Psychiatric Epidemiology*, *56*(2), 207–217. https://doi.org/10.1007/s00127-020-01906-9

Wang, G., Gill, K., Mohanlal, M., Zheng, H., & Zhao, B. Y. (2013, May). Wisdom in the social crowd: An analysis of quora. In *Proceedings of the 22nd international conference on World Wide Web* (pp. 1341–1352).

Zamboni, L., Carli, S., Belleri, M., Giordano, R., Saretta, G., & Lugoboni, F. (2021). COVID-19 lockdown: Impact on online gambling, online shopping, web navigation and online pornography. *Journal of Public Health Research*, *10*(1). https://doi.org/10.4081/jphr.2021.1959