Check for updates

Original Article

Research Ethics
1–18
© The Author(s) 2023
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/17470161231210542
journals.sagepub.com/home/rea
Sage

Passive data collection on Reddit: a practical approach

Tiago Rocha-Silva
Universidade de Psicologia e Ciências de Educação da Universidade do Porto, Portugal

Conceição Nogueira
Liliana Rodrigues
Faculdade de Psicologia e de Ciências da Educação da Universidade do Porto, Portugal

Abstract

Since its onset, scholars have characterized social media as a valuable source for data collection since it presents several benefits (e.g. exploring research questions with hard-to-reach populations). Nonetheless, methods of online data collection are riddled with ethical and methodological challenges that researchers must consider if they want to adopt good practices when collecting and analyzing online data. Drawing from our primary research project, where we collected passive online data on Reddit, we explore and detail the steps that researchers must consider before collecting online data: (1) planning online data collection; (2) ethical considerations; and (3) data collection. We also discuss two atypical questions that researchers should also consider: (1) how to handle deleted user-generated content; and (2) how to quote user-generated content. Moving on from the dichotomous discussion between what is public and private data, we present recommendations for good practices when collecting and analyzing qualitative online data.

Keywords

Online data collection, online research, Reddit, research ethics, social media, social science research

Corresponding author:
Tiago Rocha-Silva, Universidade de Psicologia e Ciências de Educação da Universidade do Porto, Rua Alfredo Allen, Porto 4200-135, Portugal.
Email: rochasilva.t@gmail.com

Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (https://creativecommons.org/licenses/by-nc/4.0/) which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (https://us.sagepub.com/en-us/nam/open-access-at-sage).

Introduction

Since the last decade, the Internet has provided unique opportunities to researchers, allowing them to easily collect data from different sources (Reips, 2002). Studies have systematically posited that the Internet should be considered an invaluable resource for data collection (Buhrmester et al., 2011; Rodham and Gavin, 2006). This is especially relevant for studies that require data from specific communities that could be harder to reach via traditional methods (Rodham and Gavin, 2006). The website Reddit is an example of an online platform where researchers can collect data, since Reddit's structure allows researchers to quickly collect data from comments and/or posts (Jamnik and Lane, 2017). However, this data collection method still presents several constraints since scholars significantly disagree over several methodological and ethical questions, such as what constitutes public and private content, anonymity, informed consent, and searchability, among others (Stommel and Rijk, 2021; Vitak et al., 2016). These constraints can be regarded as the major drawbacks of online data collection since the lack of familiarity with these methods and ethical questions can influence the researchers' perception.

Previously, several articles provided a framework regarding the characteristics and methods of data collection on Reddit (Amaya et al., 2021; Richard et al., 2021; Shatz, 2017). In this article, we aim to provide a practical approach to passive online data collection. We elaborate upon our planning for this data collection method by presenting the questions that we had to consider to understand if this method was appropriate for our research objectives (e.g. is the data available online? Where? What is the quality of the data? How will the data be collected?). Since we sought research ethics committee approval, we also present the ethical questions that we had to consider when developing our research project (e.g. informed consent and auto-determination, anonymity and pseudonymization of participants, deleted content, compliance with the EU General Data Protection Regulation, devolution to participants, storage of data). We also present the steps we had to take during our data collection process. Lastly, we discuss several recommendations (e.g. the non-inclusion of deleted user-generated content, obtaining consent and validation from the participants when reporting unaggregated data) to promote the adoption of good practices when collecting online data. It is our aim that this article contributes to better informing future researchers about the questions that must be considered before embarking on the challenge of passive online data collection.

What is Reddit?

Reddit (https://reddit.com) was created in 2005 and is defined as "a vast network of communities that are created, run, and populated" by Redditors (users that use

Table 1. Methods of online data collection.

Data collection method	Strategy of data collection	Contact with participants
Active data collection	Posting a survey	Direct passive contact
	Sampling	Direct active contact
Passive data collection	Extracting posts	Indirect passive contact

Reddit are called Redditors) and is characterized as "the front page of the Internet" (Reddit, 2022). As of August 2022, Reddit counted more than 1.5 billion registered accounts, and 430 million were active that month (Statista, 2022). Regarding the number of communities (called subreddits), as of May 2022, there were 3.5 million subreddits. In greater detail, this social media website is composed of multiple subreddits that cover various topics (e.g. subreddits related to specific countries, specific video games, specific mental health issues). In these subreddits, registered users can interact with each other by creating posts (e.g. links, images, texts, videos) or by commenting on and voting on other users' content (Kumar et al., 2022). Compared with other social media platforms, Reddit encourages user interactions since the text length for commenting is nearly unlimited (up to 40,000 characters) (Alsinet et al., 2021).

One key aspect of Reddit is its users' potential anonymity since Reddit norms tend to discourage disclosing the users' real names while creating an account and interacting with other users (Ammari et al., 2018; Leavitt, 2015). This anonymity is facilitated in the registration process since new users only need a valid e-mail address to create an account. They are then greeted with a pop-up page that allows them to choose a randomly generated username (e.g. "Routine_Syrup_8642," "ConversationSorry985") or to customize their username. Then, the user sets up a password, and the account is created. Since the registration process is relatively straightforward, it is not uncommon for its users to have "throwaway" accounts. Throwaway accounts tend to be commonly used by users who wish to avoid associating their main account with the purpose that made them create the throwaway account (Ammari et al., 2018; Proferes et al., 2021). This is particularly relevant because the users' participation history on Reddit (e.g. comments and posts) is publicly available on their profile page, and anyone can browse it.

Methods of data collection in Reddit. Online data collection on Reddit tends to occur through active or passive methods of data collection (Table 1).

Active data collection occurs when there is direct contact between the researchers and the participants (Schwab-Reese et al., 2018). As such, the participants are informed that the data they will be producing (e.g. answering a survey, participating in a digital text focus group) will be used for data analysis (Facca et al., 2020). Following this rationale, this direct contact can be either active or passive. By

active direct contact, the researchers actively try to identify and recruit participants with specific characteristics or experiences, such as the most prevalent data collection method for qualitative studies. By passive direct contact, researchers simply create a post to recruit participants and passively wait for them to complete a survey or to participate in the study, such as the most prevalent data collection method for quantitative studies.

A recent article presented three strategies for active data collection on Reddit (Amaya et al., 2021). The first two strategies can be characterized as active data collection methods with passive contact between the researchers and the participants. The first strategy consists of posting a survey on r/SampleSize (https://www.reddit.com/r/SampleSize). Its admins characterize this subreddit as "*a place for surveys and pools to be posted*" (Reddit, 2022), and as of September 2023, there were 210,000 registered users. The premise of this subreddit is that any registered user can create a post to recruit participants (e.g. to complete an online survey). Although the premise of r/SampleSize and the number of registered users can be appealing to researchers (Luong and Lomanowska, 2022), it is essential to keep in mind that a considerable number of surveys are posted daily, which can contribute to certain studies being overlooked. For example, a recent study that resorted to r/SampleSize collected data from 277 participants in 1 month (Luong and Lomanowska, 2022). In contrast, another study posted their survey 279 times (in r/SampleSize and other subreddits) and only managed to obtain 75 completed surveys in 15 days (Amaya et al., 2021).

The second strategy consists of posting on specific subreddits. Compared with the first strategy, targeting specific subreddits can be a more efficient approach to recruiting participants (Amaya et al., 2021). This strategy was used in several studies focused on specific research questions that could only be answered by specific populations. For example, Jarrett's (2021) study focused on understanding League of Legends players' motivations to spend money on in-game purchases (e.g. cosmetic skins for playable characters). To gather data, the researcher created a post on r/leagueoflegends asking players why they spent money on a free-to-play game. With that post, the researcher obtained 49 direct responses and 37 comments (Jarrett, 2021).

The third strategy for active data collection consists of purposive sampling (Amaya et al., 2021). With this strategy, researchers try to identify relevant subreddits and then try to identify adequate participants (by analyzing users' posts and comments) to invite them to participate in their study. This strategy differentiates itself from the previous two because there is active contact between the researchers and the participants since the researchers are actively trying to find individuals that fit their criteria.

In contrast, researchers can also resort to methods of passive data collection. It is important to note that these methods contrast significantly with the three

methods of active data collection. Firstly, researchers typically collect and analyze user-generated data (e.g. posts, comments) that was created by "unintentional participants" (Facca et al., 2020; Schwab-Reese et al., 2018). Secondly, this data collection method does not presume direct contact between the researchers and the participants, especially during the data collection phase. These two factors characterize and differentiate this data collection method since they deviate from what tends to be expected when conducting research. One significant benefit of this data collection method is that it allows researchers to collect natural data (data produced without the researchers' influence) (Spiti et al., 2022). Furthermore, this method allows researchers to quickly collect data from participants with diverse demographic backgrounds (Kierniesky, 2005). This is especially relevant in qualitative studies since it facilitates collecting data related to topics that tend to be considered complex to research (Mason and Singh, 2022). For example, previous studies that resorted to passive data collection on Reddit have analyzed topics such as experiences of suicidal thoughts during the COVID-19 pandemic (Slemon et al., 2021) and experiences of males disclosing intimate partner victimization (Sivagurunathan et al., 2021). Despite these benefits, this data collection method presents several constraints regarding ethical issues (Schwab-Reese et al., 2018). Since the aim of this section was to present the different methods of online data collection, these ethical issues will be explored in a subsequent section of this article.

Strengths of data collection on Reddit. As previously mentioned, Reddit counts over 3.5 million subreddits and 430 million active users. These statistics can be one of the main strengths of data collection on Reddit since the structure of Reddit allows researchers to efficiently target, recruit, or gather data from specific populations that would be of difficult access in real life (Amaya et al., 2021). Additionally, depending on the data collection strategy (e.g. studies that resort to passive data collection methods will typically gather more data), a large sample size can be collected in a relatively short period (Shatz, 2017). This strength is even more relevant for studies that collect passive data since the data is readily available. Thus, the researcher only needs to identify relevant subreddits and verify the rules of the subreddits (e.g. some subreddits explicitly prohibit data collection or recruitment without the consent of the moderation team) to start collecting data from posts and/or comments.

A second strength is related to the demographic diversity of participants. This strength is particularly relevant in the field of psychology since psychological research still tends to predominantly resort to college-based samples (Hanel and Vione, 2016). In comparison, online data collection tends to provide greater demographic diversity and data quality (Luong and Lomanowska, 2022). For example, a previous study compared survey responses between participants recruited

through Reddit and participants recruited in college, and the responses from the Reddit users offered greater diversity and greater valid data (Jamnik and Lane, 2017). Furthermore, the authors verified that the Reddit sample exhibited greater diverse demographics than the college sample. Lastly, online data collection is relatively low time-consuming, and cost-free for researchers. This strength can be appealing and relevant for students and independent researchers who do not have funding (Kierniesky, 2005) since most participants do not tend to expect any kind of monetary compensation (Luong and Lomanowska, 2022; Shatz, 2017).

Limitations of data collection on Reddit. Even though there are multiple strengths to data collection through Reddit, there are also several limitations. Although the demographic diversity of participants tends to be considered a strength, in some cases, this diversity can be hard to achieve because, depending on the data collection method, it can be arduous to gather data regarding the participants' demographics (Kumar et al., 2022). In studies that resort to methods of active data collection, this constraint is easily tackled because the participants will provide the necessary demographic data to the researchers, such as by answering the demographic questions of the surveys. In studies that resort to methods of passive data collection, this issue can be considered a considerable constraint because gathering demographic data from thousands of users can be nearly impossible to achieve. In some cases, this issue can be easily circumvented because some subreddits require that users post demographic data regarding their age and gender/sex when they create a new post. It can be argued that the age and sex of the participants can be considered the bare minimum of socio-demographic information. Still, it prevents cases of unknowingly collecting data from minors (Kumar et al., 2022). Additionally, there can also be constraints regarding the representativeness of the data since the socio-demographic characteristics of the users cannot be comparable to the general population (Amaya et al., 2021).

Another limitation regards the truthfulness of the data. This limitation was extensively discussed in Smith et al. (2017) study. In this study, the authors analyzed online discussion sites where parents discussed the administration of eye drops to their children. Since the authors resorted to a passive data collection method, critics challenged the quality and truthfulness of the data (Smith et al., 2017). However, as presented by the authors, the anonymity provided by digital contexts tends to reinforce honesty rather than deception. Additionally, since the data is generated naturally through the interaction of users, response bias is significantly reduced since it is not influenced by the researchers' participation (Smith et al., 2017). Nonetheless, anonymous interactions can also lead to less accountability, so researchers must resort to criteria to understand if there is any truthfulness to the data that is being collected (Record et al., 2018).

Even though we did not present an exhaustive list of limitations regarding data collection on Reddit, we enunciated some to exemplify that this method also presents several limitations that can impact the planning and the data collection process. Since every data collection method presents its advantages and disadvantages, it is up to the researchers to understand how their study objective and methodology fit the online context of Reddit. Nonetheless, through proper methodological planning and rigor, most of these limitations can be answered (Shatz, 2017).

Digital dating abuse in long-distance romantic relationships

Digital dating abuse (DDA) is defined as "a pattern of behaviors that control, pressure, or threaten a dating partner using a cell phone or the Internet" (Reed et al., 2016). Following this definition, DDA can be conceptualized as a triadic phenomenon composed by: (i) a digital element (e.g. encompassing all means of digital communication); (ii) a dating element (e.g. the abusive behaviors occur in a current or former intimate relationship); and (iii) an abuse element (e.g. the existence of behavioral patterns that harm an intimate partner) (Reed et al., 2016). Our primary research project aimed to develop a grounded theory of DDA in long-distance romantic relationships to understand how individuals experience DDA. Although grounded theory studies typically resort to interviews to collect data, we decided to collect online data to bridge the constraints that COVID-19 imposed during the beginning of this decade. Additionally, we conducted several interviews in the early stages of the project, and we realized that most interviews were characterized by a high degree of social desirability (e.g. participants minimizing their behaviors of DDA). Since we aimed to gather data from perpetrators and victims, we started considering different approaches to data collection to tackle the limitations we were facing. After analyzing various studies and methodological articles, we considered the strategy of online data collection. In the following sections, we will explore our planning for data collection, the ethical questions we had to consider when we sought research ethics committee approval, and the different phases of data collection that we went through.

Planning online data collection

As previously mentioned, in our primary study we decided to resort to the strategy of passive data collection. Nevertheless, before settling for this strategy, we had to consider several methodological questions to understand if this strategy would be adequate. The questions we considered were the following: (i) is the data available online? If so, where?; (ii) what is the quality of data?; (iii) how will the data be collected?

Is the data available online? If so, where? The first question we considered was, "is the data available online? If so, where?." The first part of this question is the simplest since it could be argued that any kind of data will be available online. The "where?" is the main question since researchers need to identify the digital platforms that contain the data they seek. Since our research team was familiar with Reddit and previous studies that collected data on Reddit, we considered this platform as our main target for data collection. We were also acquainted with the subreddit r/LongDistance, so we had a preconceived notion of where we could identify and gather data. From our initial analysis, which consisted of reading and categorizing every post from December 2021, we concluded that this subreddit would be a relevant platform since its members tended to frequently post about their experiences of DDA. Additionally, it is relevant to mention that the users of this subreddit could categorize their posts with pre-established "flairs" (e.g. breakup, venting, need support, need advice). Although this feature can be considered helpful for members of the subreddit, we noticed that most users did not choose the option to flair their posts. Consequently, we decided that we would not resort to those flairs to identify posts since that option could imply the non-inclusion of several posts.

We also considered gathering data from different subreddits, such as r/relationship_advice, and r/NarcissisticAbuse since those subreddits are mainly populated by individuals in romantic relationships. Still, we realized that it would be arduous to filter through thousands of daily posts (especially in the case of r/relationship_advice). Nonetheless, we noticed that users typically tend to cross-post. Cross-posting consists of a user posting in a specific subreddit and later posting the same content on a different subreddit. This is a common practice on Reddit since users who seek advice tend to cross-post in subreddits with thematic commonalities.

What is the quality of data? The second question was, "what is the quality of the data?" We considered this question because it is important to understand if the available data would be relevant to our research questions. This question can be highly contextual and implies the development of exclusion and inclusion criteria. In our case, we established the following inclusion criteria: (i) posts by users that mention being aged 18 or older; (ii) posts in which the user mentions adopting behaviors of DDA (e.g. monitoring the partners' social media pages; (iii) posts in which the user mentions that their partner adopted behaviors of DDA; (iv) the users' account must be older than 3 months. As exclusion criteria, we established the following: (i) posts in which the user does not mention their age; (ii) posts in which the user mentions being aged 17 or younger; (iii) posts with a length inferior to 10 lines. We established this last criterion because we noticed that some users posted experiences of DDA without providing contextual information that could help the

reader understand what was being shared. Exemplifying, posts such as "my bf left me on read for five days, and I don't know what to do" were excluded. Nonetheless, from our experience, those types of posts tend to be uncommon since users tend to provide detailed contextual information about what they are experiencing.

Lastly, questions related to the veracity of the data could also be raised. In the context of Reddit, the veracity of the posts tends to be more questionable when users are trying to "farm karma." Farming karma consists of posting fabricated content to influence the sympathy of other users to gain their upvotes and awards. Considering r/LongDistance, upvoted posts tended to be related to success stories (e.g. posts about closing the distance and being proposed). However, critics could argue that the posts were not meant to farm karma but to deceive other users (e.g. posting a fabricated story to "troll" other users). From the data we initially analyzed, it was possible to understand that the users seemed genuinely truthful while sharing their experiences. Hence, it appeared that their motivation to post was to validate their experiences.

How will the data be collected? Regarding this final question, it is important to mention that in comparison with other social media websites (e.g. Twitter), Reddit has a limit for the posts that can be viewed when browsing specific subreddits (capped at 1000 posts). So, the strategy of browsing and collecting data that spans several months is nearly impossible without the use of the Reddit API, Pushshift API, or Pushshift-based websites (e.g. Redditsearch). The Reddit API and Pushshift API tend to be the most practical, but researchers must possess engineering skills to fully understand how to use them. In comparison, Pushshift-based websites are search services that allow their users to perform specific searches. Additionally, those websites tend to offer a simplified and user-friendly interface. Since we intended to collect every post that had been posted in r/LongDistance between January 2021 and December 2021, we decided to resort to a Pushshift-based website (Camas). Lastly, it is important to mention that resorting to these types of data repositories can be characterized as a limitation since the data sets might be incomplete (Gaffney and Matias, 2018). Thus, researchers should consider how this limitation can impact their research.

Ethical considerations

In this section, we present the ethical considerations that we had to review when we sought ethical approval from our institutional research ethics committee (REC) for our research project. The ethical considerations we had to review were the following: (i) REC approval or not?; (ii) informed consent and auto-determination; (iii) anonymity and pseudonymization of participants; (iv) deleted content; (v)

general data protection regulation; (vi) devolution to participants; and (vii) data storage.

Rec approval or not? The first question we had to consider was whether we should submit our project for research ethics committee (REC) approval. Although seeking REC approval is a common international practice, a recent systematic review verified that in 727 studies, only 13.9% sought REC approval (Proferes et al., 2021). Another review also verified that in 500 studies, more than half did not disclose any kind of ethical considerations (Coates, 2021). Despite these results, it is essential to mention that some studies sought ethical approval, but their RECs exempted those requests (Mason and Singh, 2022; Sivagurunathan et al., 2021; Slemon et al., 2021). Nonetheless, it is essential to reiterate that there is a considerable difference between a study being exempted by an REC and the researchers claiming that there was no need for REC approval because the data they wanted to collect was publicly available (Proferes et al., 2021). In our case, we decided to seek ethics approval and our research project was approved by the REC from Faculdade de Psicologia e Ciências da Educação da Universidade do Porto (Refº 2022/01-04 – 13 January 2022).

Informed consent and auto-determination. The informed consent and participants' auto-determination was the second question we had to consider. Until this point, we had already established that our data collection method would be passive, and our strategy would consist of extracting posts, hence having indirect passive contact with the participants. Since most studies need to provide and obtain informed consent from their participants, our method and data collection strategy made it difficult because we would be collecting data from thousands of posts. Hypothetically, we could try to contact every user from whom we collected data. Still, the logistics involved in messaging thousands of users could compromise the research's conduct. Since we intended to use text excerpts to exemplify the users' experiences of DDA, we decided to follow the recommendations provided by the Association of Internet Researchers (AoIR) (Franzke et al., 2020). Thus, we only sought informed consent from participants who shared experiences that we wanted to quote. As such, we redacted a form of informed consent to share with those participants. It is worth mentioning that when we contacted those participants, we asked them if they authorized us to modify their text excerpts as a measure to try to preserve their anonymity and to try to prevent the trackability of their content. Lastly, we also asked them if they would be available to read and approve our text excerpt modifications. Regarding the participants' auto-determination, we recognize that our research project did not guarantee that every participant would have auto-determination. However, we ensured that the participants who would provide text excerpts would exercise their right to auto-determination without any kind of personal prejudice.

Anonymity and pseudonymization of participants. Another topic we had to consider was the pseudonymization and the possible risk of de-identification of the participants. In a previous study, it was reported that researchers managed to de-identify participants of several studies (Vitak et al., 2016). This de-identification was mainly achieved because some researchers did not pseudonymize information that could contribute to de-identifying their participants. This question is particularly relevant because some users might have expectations regarding their privacy so researchers must try to adopt best practices to preserve their participants' identity (Dym and Fiesler, 2020). In our case, we followed the general guidelines of pseudonymization in qualitative studies. Therefore, any information that could contribute to de-identifying participants (e.g. usernames, workplace) or de-identifying the participants' relational and social networks (e.g. partners' names) was altered or removed. In addition, we followed the principle proposed by the AoIR (Franzke et al., 2020) that text excerpts should be changed or worked upon to guarantee that those excerpts could not be tracked through search engines (e.g. Google). This principle was also recommended by a mixed-method study with members of a vulnerable online community (Dym and Fiesler, 2020). As such, the text excerpts were slightly modified through synonyms. To further guarantee that the excerpts could not be tracked, we tried tracking the original posts by searching the modified excerpts in search engines (e.g. Google) and on Pushift-based websites. This practice was recently analyzed and promoted by a study that verified that only resorting to the strategy of modifying text excerpts could prove insufficient to prevent the data from being tracked (Reagle, 2022).

Deleted content. The fourth topic we had to consider was how to manage deleted content. This topic is particularly relevant for researchers who resort to Pushshift-based websites since those websites tend to index user-deleted content. As an example, a Redditor can write a post and decide to delete it after a few days. If the Redditor deletes their post it will not appear on Reddit but, since Pushshift tends to be updated in real-time (Baumgartner et al., 2020), it will appear on those websites. In our case, we noticed that several posts that appeared in our searches had been deleted. Drawing upon previous research that also faced this constraint (Ravn et al., 2020), we decided that during our phase of data collection (January 2022 until March 2022) we would not collect indexed data that had been deleted from Reddit. Additionally, we decided to verify if the posts that we had collected ($n = 1292$) were still available on Pushshfit and Reddit before we started the process of data analysis. In May 2022, we verified that in 2 months, only 44 posts (3.52%) had been deleted. One explanation for this low number of deletions could be that most deletions tend to happen between the first day and the first week of the submission (Reagle, 2023).

Following Zimmer's (2010) rationale, passive methods of data collection tend to ignore their participants' intentions regarding the accessibility of their

user-generated content. Since those Redditors choose to delete their content, we consider that researchers must respect their participants' decisions regarding the availability of their user-generated content. That is why we also chose a considerable timelapse for data collection. Since we were conducting a grounded theory, we did not adopt additional measures to verify the availability of the data because it could compromise our data analysis process.

General Data Protection Regulation (GDPR). The fifth topic we had to consider was the GPDR. This topic and legislation typically only apply to researchers conducting research in the European Union (EU) or that involve participants based in the EU. Succinctly, this legislation aims to harmonize data protection laws across EU Member States regarding EU citizens' personal data. According to the GDPR, we were able to process this data for research purposes. This said, researchers still need to ensure that their research complies with, among other provisions, the safeguards specified under GDPR Article 89. Succinctly, this article entails practices that researchers must adopt when processing data for research purposes. As an example, researchers are required to take technical measures (e.g. the use of strategies of pseudonymization) to ensure the principle of data minimization if the intended end of the research can be achieved that way. Since we would resort to practices of pseudonymization and would adopt the strategy of modifying the users' text excerpts to prevent the data from being tracked, our research project was in line with the GDPR. It is also worth noting that the terms of service of Reddit (and r/LongDistance) also do not prohibit analyzing user-generated content for academic and scientific purposes.

Devolution to participants. Another topic we considered was how the results would be shared with the participants. In our case, we proposed that the results would be shared at an individual and a community level. On an individual level, we would share our research through private messages with the users who signed the informed consent. The community return would happen by creating a post on the subreddit to share our research with the community.

Data storage. The last topic we considered regards data storage. Since we were conducting a qualitative study, the data was stored in text files and we resorted to a letter and number encoding system (e.g. A001, A002) to pseudonymize and organize the text files. This encoding system was also thought to allow the research team to track the original post since, at a later point, we would need to contact several users. Furthermore, any data that could contribute to the identification of the posts was replaced with pseudonyms (e.g. username) or removed entirely (e.g. date of the post). Regarding data storage, the files were stored locally in the primary researcher's personal computer and digitally through a cloud storage and

file-sharing service. Those files could only be accessed by the members of the research team, and a two-step verification process protected access to the cloud storage. We also proposed that the research data would only be stored for 5 years after the conclusion and publication of our primary research results.

Data collection

Regarding our data collection method, we decided to use a Pushshift-based website (Camas) to collect data (e.g. posts) from the targeted subreddit. We also decided that we would collect data from posts between January 2021 and December 2021. Since we were using a Pushshift-based website the data collection process was considerably straightforward. In greater detail, this Pushshift-based website allows individuals to search posts and/or comments that were posted on Reddit. Additionally, it also allows to refine the search queries (e.g. search by subreddit, search by karma). In our case, we only resorted to the options of: (i) searching by subreddit (r/LongDistance); (ii) searching for posts (since we were only interested in collecting data from posts); or (iii) searching by timeframe (e.g. 1 January 2021 until 2 January 2021). It is important to mention that this Pushshift-based website has a maximum number of 100 returned queries. In practical terms, this means that it is not possible to choose extended timeframes to collect data.

The process of data collection was also relatively straightforward since we only had to copy and save the data from which individual post. Even though we had established exclusion and inclusion criteria, we did not contrast the content of the posts with our criteria during this phase. The only triage that we were obligated to do was related to posts of images and videos since we would only analyze text-based content. At the end of our data collection process, we had gathered data from 4966 posts. Contrasting those posts with our exclusion and inclusion criteria, we remained with 1248 posts (25.13%) for data analysis. In greater detail, 3301 posts (66.47%) were excluded because those posts were unrelated to experiences of digital dating abuse, and 417 (8.39%) were excluded because: (i) the user did not mention their age; (ii) the user mentioned being in a long-distance romantic relationship with a minor; (iii) the post length was inferior to 10 lines; or (iv) the user had deleted the post. Since the targeted subreddit theme does not revolve exclusively around abusive experiences in romantic relationships, we consider that we managed to gather data from a considerable number of posts.

Since we had no contact with the participants, the phase of data collection was relatively straightforward because the process could be controlled by the research team. Studies that resort to methods of active data collection (e.g. posting a survey) still tend to face constraints related to the identification and recruitment of participants. Nonetheless, we must reiterate that this method might not be suitable for every research (mostly suitable for qualitative studies) so, researchers must

balance the advantages and disadvantages of this method. Lastly, it is important to mention that as of July 2023, access to Reddit API is restricted. This is due to the Reddit CEO's decision to charge for access to their API. In practical terms, this means that most Pushshift-based websites are currently offline. Although these changes were heavily criticized by Reddits' communities, the policy change seems to remain. In the meantime, researchers should focus on alternative Pushshift services and/or strategies for passive data collection.

Conclusions and recommendations

Two decades ago, Rodham and Gavin (2006) mentioned that conducting studies through online means was in its infancy, so there were no clear rules of conduct and ethical guidelines to support researchers. Nearly two decades later, these constraints remain predominantly unanswered and vary greatly between academics and researchers. Nonetheless, we do believe that there is enough empirical and theoretical data to support the establishment of good practices. Comparing the questions that we had to review during the phase of data collection planning (is the data available? If so, where? What is the quality of data? How will the data be collected?), it is possible to equate these questions with the "traditional" questions of qualitative research in social sciences. The questions we had to review when we sought institutional REC approval can also be considered standard for qualitative research. The only ethical considerations that deviated from the standards were: (1) how we would deal with deleted user-generated content; and (2) how we would quote the participants' user-generated content to prevent data de-identification. Since we were collecting data from a public subreddit, we could have argued that those questions did not apply because the data was publicly available. Although that rationale can be appealing, we decided to find solutions that exhibited some degree of responsibility toward the collection and analysis of user-generated content.

In Table 2, we present several recommendations that we consider to be good practices when collecting and analyzing qualitative online data. Since we already fully discussed the first topic, we will elaborate only on the remaining ones. Regarding the second topic, we decided that we would not include user-generated content that had been deleted during the phase of data collection. The rationale for this decision was to respect the individuals' free will regarding the availability of their user-generated content. Since the deletion of user-generated content can suggest that some individuals might feel uncomfortable with the availability of their data (Reagle, 2022), researchers should be mindful of the implications of collecting and analyzing data that was previously deleted. Regarding the third topic, we decided that we would follow the general guidelines of pseudonymization for qualitative studies. We also decided to follow the guidelines proposed by the Association of Internet Researchers regarding the modification of text excerpts.

Table 2. Summary of recommendations for good practices.

1. Researchers should always seek REC/research ethics committee approval for their research projects. If such approval is not required in the researcher's jurisdiction or host institution, researchers should conceptualize their research according to the general principles of research ethics and consider principles such as:
   - Participants informed consent and auto-determination.
   - Participants' anonymity and pseudonymization.
   - How the data will be stored.
   - How the research results will be shared with the participants.
   - Compliance with relevant data protection law (e.g. General Data Protection Regulation).
2. Researchers should consider how to handle deleted user-generated content. We suggest that researchers refrain from collecting deleted content since the individuals are manifesting that they do not want it to be available.
   - An adequate time frame for data collection should be established to allow individuals the possibility of deciding whether they want their content available or not.
3. Researchers should also consider how to quote user-generated content and should resort to strategies of disguise (e.g. altering word expressions) to try to prevent the quotes from being tracked and/or their participants de-identified.
   - Researchers should test their modified quotes to verify if they can be traced to the original source.
4. Researchers should try to contact the participants who will be quoted to obtain their informed consent.
   - Researchers can also try to understand if those participants are available to verify and approve the modified quote.

The rationale for these decisions is that even though the data is publicly available, it is the researchers' responsibility to try to guarantee that their participants' anonymity and to maintain confidentiality. As a previous study posited, the strategy of modifying text excerpts reduces the possibility of participants being tracked and can also help protect them from possible negative consequences (Mancosu and Vegetti, 2020). Researchers should also consider that individuals might have expectations regarding their user-generated content being used for secondary purposes. This is particularly relevant in studies with vulnerable populations since those individuals tend to have protective beliefs toward their data (Dym and Fiesler, 2020). Regarding the last topic, we recommend that researchers try to contact the participants who will provide quotes to try to obtain their informed consent. Since the quotes will be presented as unaggregated data, we consider that researchers should make efforts to understand if those participants are comfortable with having their experiences shared in scientific research. Additionally, researchers should also try to understand if those participants are available to read and approve the modified quote. Since qualitative studies have a strong epistemological standpoint, this step is particularly relevant for researchers following a constructivist standpoint since the participants can validate the accuracy of the researchers' second-order construct (e.g. the quote that the researchers modified).

With this article, we hope we contributed to bridging some of the gaps between the methodological literature and the empirical literature. By fully elaborating on our methodological and ethical considerations, we also hope that we contributed to provide a practical example of the process of methodological and ethical decision-making when conducting studies that resort to methods of passive data collection. Lastly, although the methods of passive data collection exhibit multiple characteristics that permit tackling the most common constraints in traditional research, researchers must always keep in mind how these strategies fit their research objectives. Although this method exhibits great potential, researchers must remember that some constraints are considerably harder to tackle (e.g. lack of in-depth socio-demographic information). Furthermore, some ethical questions have only started to be considered and debated in the last few years, such as the means of appropriately de-identifying participants in passive online data collection. We believe that acknowledging these constraints should not deter researchers from conducting research through passive data collection methods. On the contrary, this acknowledgment should be regarded as a principle guiding researchers to develop ethical good practices for collecting and analyzing online data.

Funding

All articles in Research Ethics are published as open access. There are no submission charges and no Article Processing Charges as these are fully funded by institutions through Knowledge Unlatched, resulting in no direct charge to authors. For more information about Knowledge Unlatched please see here: http://www.knowledgeunlatched.org

Ethical approval

This study was reviewed and approved by the Faculdade de Psicologia e Ciências da Educação da Universidade do Porto (Refº 2022/01-04).

ORCID iD

Tiago Rocha-Silva https://orcid.org/0000-0003-3564-4423

References

Alsinet T, Argelich J, Béjar R, et al. (2021) Discovering dominant users' opinions in Reddit. In: Mateu V, Teresa A and Cèsar F (eds) *Frontiers in Artificial Intelligence and Applications*. Amsterdam: IOS Press, pp.113–122.

Amaya A, Bach R, Keusch F, et al. (2021) New data sources in social science research: Things to know before working with Reddit data. *Social Science Computer Review* 39(5): 943–960.

Ammari T, Schoenebeck S and Romero DM (2018) Pseudonymous parents: Comparing parenting roles and identities on the Mommit and Daddit Subreddits. In: *Proceedings of the 2018 CHI conference on human factors in computing systems*, pp.1–13.

Baumgartner J, Zannettou S, Keegan B, et al. (2020) The Pushshift Reddit dataset. In: *Proceedings of the international AAAI Conference on Web and Social Media*, vol. 14, pp.830–839.

Buhrmester M, Kwang T and Gosling SD (2011) Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? *Perspective on Psychological Science* 6(1): 3–5.
Coates A (2021) How often are basic details of the research process mentioned in social science research papers? *Learned Publishing* 34(2): 128–136.
Dym B and Fiesler C (2020) Ethical and privacy considerations for research using online fandom data. *Transformative Works and Cultures* 33: 1–19.
Facca D, Smith MJ, Shelley J, et al. (2020) Exploring the ethical issues in research using digital data collection strategies with minors: A scoping review. *PLoS One* 15(8): e0237875.
Franzke AS, Bechmann A, Zimmer M, et al. (2020) Internet research: Ethical guidelines 3.0. Available at: https://aoir.org/reports/ethics3.pdf (accessed 1 October 2023).
Gaffney D and Matias JN (2018) Caveat emptor, computational social science: Large-scale missing data in a widely-published Reddit corpus. *PLoS One* 13(7): e0200162.
Hanel PH and Vione KC (2016) Do student samples provide an accurate estimate of the general public? *PLoS One* 11(12): e0168354.
Jamnik M and Lane D (2017) The use of Reddit as an inexpensive source for high-quality data. *Practical Assessment, Research, and Evaluation* 22(5): 1–10.
Jarrett J (2021) Gaming the gift: The affective economy of League of Legends 'fair' free-to-play model. *Journal of Consumer Culture* 21(1): 102–119.
Kierniesky NC (2005) Undergraduate Research in small psychology departments: Two decades later. *Teaching of Psychology* 32(2): 84–90.
Kumar N, Corpus I, Hans M, et al. (2022) COVID.19 vaccine perceptions in the initial phases of US vaccine roll-out: An observational study on reddit. *BMC Public Health* 22(1): 446.
Leavitt A (2015) 'This is a throwaway account': Temporary technical identities and perceptions of anonymity in a massive online community. In: *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*, pp.317–327.
Luong R and Lomanowska AM (2022) Evaluating Reddit as a crowdsourcing platform for psychology research projects. *Teaching of Psychology* 49(4): 329–337.
Mancosu M and Vegetti F (2020) What you can scrape and what is right to scrape: A proposal for a tool to collect public Facebook data. *Social Media + Society* 6(3): 1–11.
Mason S and Singh L (2022) Reporting and discoverability of "Tweets" quoted in published scholarship: Current practice and ethical implications. *Research Ethics* 18(2): 93–113.
Proferes N, Jones N, Gilbert S, et al. (2021) Studying Reddit: A systematic overview of disciplines, approaches, methods, and Ethics. *Social Media + Society* 7(2): 1–14.
Ravn S, Barnwell A and Barbosa Neves B (2020) What is "publicly available data"? Exploring blurred public–private boundaries and ethical practices through a case study on Instagram. *Journal of Empirical Research on Human Research Ethics* 15(1–2): 40–45.
Reagle J (2022) Disguising Reddit sources and the efficacy of ethical research. *Ethics and Information Technology* 24(3): 41–41.
Reagle J (2023) Even pseudonyms and throwaways delete their Reddit posts. *First Monday* 28(6): 1–16.
Record RA, Silberman WR, Santiago JE, et al. (2018) I sought it, I *Reddit*: Examining health information engagement behaviors among *Reddit* users. *Journal of Health Communication* 23(5): 470–476.
Reddit (2022). http://reddit.com, (accessed 1 October 2023).
Reed LA, Tolman RM and Ward LM (2016) Snooping and sexting: Digital media as a context for dating aggression and abuse among college students. *Violence Against Women* 22(13): 1556–1576.

Reips UD (2002) Standards for internet-based experimenting. *Experimental Psychology* 49: 243–256.

Richard B, Sivo SA, Ford RC, et al. (2021) A guide to conducting online focus groups via Reddit. *International Journal of Qualitative Methods* 20: 1–9.

Rodham K and Gavin J (2006) The ethics of using the Internet to collect qualitative research data. *Research Ethics* 2(3): 92–97.

Schwab-Reese LM, Hovdestad W, Tonmyr L, et al. (2018) The potential use of social media and other internet-related data and communications for child maltreatment surveillance and epidemiological research: Scoping review and recommendations. *Child Abuse & Neglect* 85: 187–201.

Shatz I (2017) Fast, free, and targeted: Reddit as a source for recruiting participants online. *Social Science Computer Review* 35(4): 537–549.

Sivagurunathan M, Walton DM, Packham T, et al. (2021) "Punched in the balls": Male Intimate Partner Violence Disclosures and replies on Reddit. *American Journal of Men's Health* 15(4): 15579883211039666–15579883211039714.

Slemon A, McAuliffe C, Goodyear T, et al. (2021) Reddit users' experiences of suicidal thoughts during the COVID-19 pandemic: A qualitative analysis of r/Covid19_support posts. *Frontiers in Public Health* 9: 693153–693212.

Smith H, Bulbul A and Jones CJ (2017) Can Online discussion sites generate quality data for research purposes? *Frontiers in Public Health* 5: 156–164.

Spiti JM, Davies E, McLiesh P, et al. (2022) How social media data are being used to research the experience of mourning: A scoping review. *PLoS One* 17(7): e0271034.

Statista (2022) Reddit - Statistics & facts. Available at: https://statista.com/topics/5672/reddit/ (accessed 1 October 2023).

Stommel W and Rijk LD (2021) Ethical approval: None sought. How discourse analysts report ethical issues around publicly available online data. *Research Ethics* 17(3): 275–297.

Vitak J, Shilton K and Ashktorab Z (2016) Beyond the Belmont Principles: Ethical Challenges, Practices, and Beliefs in the Online Data Research Community. In: *Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing*, pp.941–953.

Zimmer M (2010) "But the data is already public": On the ethics of research in Facebook. *Ethics and Information Technology* 12: 313–325.