WORKING PAPER

Using social media and text analytics to investigate marginal phenomena

Author: Simon Caton

Research managers: Gijs van Houten and Franz Ferdinand Eiffe

Eurofound reference number: WPEF21052

Research carried out prior to the UK's withdrawal from the European Union on 31 January 2020, and published subsequently, may include data relating to the 28 EU Member States. Following this date, research only takes into account the 27 EU Member States (EU28 minus the UK), unless specified otherwise.

The European Foundation for the Improvement of Living and Working Conditions (Eurofound) is a tripartite European Union Agency established in 1975. Its role is to provide knowledge in the area of social, employment and work-related policies according to Regulation (EU) 2019/127.

European Foundation for the Improvement of Living and Working Conditions

Telephone: (+353 1) 204 31 00

Email: information@eurofound.europa.eu

Web: www.eurofound.europa.eu

Contents

Introduction

Eurofound orchestrate multiple regularly repeated research projects that make use of surveys as well as other research methods. Specifically, these include the European Company Survey (ECS), the European Quality of Life Survey (EQLS), and the European Working Conditions Survey (EWCS). Noted too is the recent Living, working and COVID eSurvey. A specific major challenge in the orchestration of these surveys is maintaining sufficient representational power of participant sample populations across the pan-European area. With all sampling methods, there are no guarantees of representation within the sample population; it is also difficult to measure how representative a given sample is beyond what is known a priori. Or in other words, is it hard to measure the representation of small populations, and also study marginal phenomena. Thus, the objective of this working paper is to explore to what extent social media could facilitate the study of marginal phenomena, specifically, where traditional methods (e.g. surveys, interviews etc.) suffer from poor representational power. Whilst there have been many studies of marginal populations in the literature, these are often (naturally) quite self-contained, and do not capture the many different considerations needed to leverage social media as a data source in diverse scenarios. Yet, there are many technical, operational, methodological and ethical challenges associated to this "scale" and "assessibility" of data (see [Hall et al., 2018]) which will be explored and contextualised in this paper.

The motivation for social media is that it has become an essential day-to-day communication platform over the last decade [Lenhart et al., 2015, Pew Research, 2017], and correspondingly researchers have turned to it as a source of data. Whether part of academic or business research a wide range of methods both qualitative and quantitative have been employed to study specific social contexts and processes through the lens of social media and the user generated content it encompasses. The scale of the underpinning network structure allows social media researchers to answer questions that perhaps more traditional methods cannot due to underpinning method, sample, response rate or other biases [Hughes et al., 2012, Kahneman et al., 1999, Krosnick, 1999, Schwarz et al., 1985, Podsakoff et al., 2003].

This working paper seeks to review and discuss the potential for social media mining methodologies as a mechanism to augment and/or complement traditional quantitative research methodologies within the context of marginal phenomena; specifically within (European) labour markets, a key area of interest for Eurofound. Where examples include but are not limited to: "bogus" self-employment, migrant or marginalised workers, labour exploitation and their related socio-cultural effects or observations. To date, social media has been used in a wide variety of scenarios as a lens to study modern day society. A key question for this working paper is how transferable (and correspondingly reliable, ethical and legally permissible) social media mining techniques could be in meaningfully representing worker communities that would otherwise be overlooked or not satisfactorily captured. To address this, a significant review of literature is conducted of the domains of Natural Language Processing, Information Retrieval, and Social Network Analysis, but also techniques used for curating and assembling social media data for analysis. As such, it is structured as follows: first some needed scope the paper is provided, outlining key definitions, and challenges, and thus introduces some of the key trade-offs researchers need to consider. Following this, it introduces the means through which social media is typically leveraged as a source of data for researchers, and

builds on existing studies to act as frames of reference and exemplars. This part of the discussion introduces the methodological basis, i.e. methods, through which social media can be leveraged to answer different types of research questions using existing studies to anchor these approaches. Finally, the main findings are summarised with an outlook on the suitability of social media as a means for Eurofound to study marginal phenomena. This discussion centres around the main considerations of ethics, cost, legality, the skill set needed by analysts, and potential method biases that may arise that are discussed throughout the working paper.

Background and Key Challenges

In order to provide key scope and context, some definitions are needed to give structure to the discussion in this working paper, which are provided in this section. Social media as a set of research methods, also comes with several ethical and methodological considerations that can generally be applied to any study that relies on one or more social media platforms. These are (briefly) presented to raise awareness and provide some insights for how they affect the research process. It should be noted that it may not be possible to address all these challenges in every study.

Key Definitions

(Sub)Population / Group: a community or group of individuals that share one or more unique social socio-cultural identity markers, geographic space(s) or both [McInroy, 2016].

Marginalised (Sub)Populations / Groups: often defined as populations outside of "mainstream" society [Schiffer and Schatz, 2008] and "highly vulnerable populations that are systemically excluded from national or international policy making forums" [Siddiqui, 2014]. Common examples include, but are not limited to: the homeless, drug users, sex workers, refugees, ethnic minorities, and members of the LGBTQ community [O'Donnell et al., 2016].

Marginal Phenomena: Building on Markey's definition of social phenomena [Markey, 1926], marginal phenomena are considered as including all behaviour which influences or is influenced by marginalised (sub)populations or groups, i.e. social phenomena within the context of one or more marginalised (sub)populations or groups.

Social Media: As defined by Carr and Hayes [Carr and Hayes, 2015], "Social Media are Internet-based channels that allow users to opportunistically interact and selectively self-present, either in real-time or asynchronously, with both broad and narrow audiences who derive value from user-generated content and the perception of interaction with others." However, specifically for this working paper, a slight restriction on this definition is needed, as there needs to be a mechanism to interact with users, their user generated content (in specific contexts), or both without needing to be "friended" or otherwise formally connected to them. This, for instance, would exclude more "closed" platforms such as Snapchat, Kik, Telegram, and in some scenarios also Facebook and Instagram.

Considerations for Social Media-based Research

There is significant literature that discusses the challenges of representation in studies. [Yancey et al., 2006] summarise many of these as well as methods often employed by the scientific community to better access marginalised populations. They cite for example the need to leveraging the community to promote participation, have significant follow-up and incentives (monetary as well as social), maintaining consistency of research staff and points of contact (among many other factors). They also note that many scholars have highlighted that trust (or a lack thereof) in the research process acts as a barrier to participation. With [Morgan et al., 2013] noting that promoting the study through well-known organisations can alleviate this somewhat. Moving forwards, and of specific value for this working paper, is the suggestion to allow sufficient resources (time and funding) for the conduct of feasibility studies.

Whilst concerns over excluding potential respondents that lack internet connectivity [Andrews, 2012] exist, many studies have shown that this concern is rapidly waning. In the context of marginal phenomena, social technologies (which include social media) are often essential for marginalised communities [Fox and Ralston, 2016, Gonzales, 2017, Devito et al., 2019], often acting as a "safe place" due to aspects of anonymity and control over the intended audience [Nemer, 2016]. The general availability of social technologies is also high among marginalised groups [Nemer, 2016, Bender et al., 2014].

Social media data should be seen as communicative data, i.e. data that is produced as a side effect of some communicative act between users. This communicative act also intrinsically captures a specific context and setting of the interaction, which can be hard to observe but still encodes special meaning that can be easily overlooked. Often, social media data is used to model some underlying activity and correspondingly answer some set of research questions. Yet, social media activities are not always representative of a "reality" beyond an overarching desire to be a part of something [Caton et al., 2012, Nemer, 2016]. This can create key challenges in the use of social media as a vehicle for the analysis of marginal phenomena through the presence of specific ethical concerns and/or other biases as discussed at length in the literature (for example [Nosek et al., 2002, Andrews, 2012, Hall et al., 2018, Zwitter, 2014]). Here, there are several different dilemmas which could present themselves depending on the modality of social media usage, i.e. as a source of data, or mechanism to recruit respondents. These dilemmas briefly discussed below outline key considerations and scenarios where they might arise. In short, they arise from: 1) the context of the communicative act being "lost" in the analysis; 2) a disparity between data being available and permission to process it; 3) difficulty in completely ensuring there is anonymity in the data, and thus not exposing users without their permission or knowledge; 4) fake, false or otherwise unverified information, which can be quite prolific on social media, being integrated into the research; 5) other informational biases such as social posturing, overtly positive presentation of data, and the presence of sarcasm and figurative language use.

Context: The source of many ethical concerns around social media often stems from the ability to gather and therefore analyse large quantities of data [Zwitter, 2014]. When processing large amounts of (social media) data, data is often reduced to what can fit into a (mathematical) model. Yet, taken out of context, this data may lose its meaning [Boyd and Crawford, 2012] or be otherwise warped beyond its original communicative intent. This is key, as social media is often leveraged for information of the now, but when used historically, it can be difficult to access representative data [Gaffney and Matias, 2018]. Many studies (e.g. [Boyd and Crawford, 2012, Gaffney and Matias, 2018, Rost et al., 2013, González-Bailón et al., 2014]) have discussed this at length. Essentially, key takeaways here are that data may be missing, and that changing the sample of data used (e.g. due to it being missing, or platforms being inconsistent in what data is available to reseachers) can fundamentally change the findings and the interpretation of the data. This maps to two different forms of potential bias. Firstly, that the coverage and representativeness of an event, entity or phenomenon can be insufficient, and secondly, that the accuracy (in the sense of truly presenting phenomena) of network and other effects is compromised. Or in other words, that the sample of the data is incomplete and missing completely at random, thus drawing a different (random) sample from the true distribution would change the results of the analysis; as was shown by [González-Bailón et al., 2014]. The implication here is that the researcher may miss key aspects of the

phenomenon they are studying; potentially under or over representing specific aspects, which may distort its perceived definition and defining concepts. Many of the consequences of these effects go unnoticed, but they may have important theoretical implications depending on the question(s) being asked. In the context of marginal phenomena, this is important as there may in general be a lack of suitable data further amplifying these concerns.

Availability vs. Permission(s): A severe grey area in social media research is the issue of data availability – generally it is available, which gave rise to many controversial studies. Many problems stemmed from data that was publicly available (on social media platforms) and freely "browsable" but where researchers had permission from the platform, but not the users themselves to process the data. This became a battleground for privacy campaigners [Zimmer, 2020]. Essentially coming down to just because the data is online doesn't mean that "informed" consent has been given, or that users consider themselves as subjects in a study. In this sense, there may be ethical tensions in the use of social media content as a means to study marginal phenomena especially if the population is vulnerable or otherwise belong to an at-risk group.

Exposure: Even though data is often anonymised, or at least is supposed to be, anonymity can be quite hard to achieve in practice. As noted by [Zimmer, 2020], it can be difficult to truly anonymise data. Aspects of the content (e.g. handles, tags, etc.) alone can compromise the anonymity of data. As such users may be exposed unintentionally to the lens of the researcher(s). An example where data might be difficult to anonymise is when there are indirect references to users in the text content. For example, when the user is named, but not tagged, there is some spelling error in their name, a pronoun, nickname or other short-form name is used to refer to them etc. It can also be difficult to predict potential harms from analyses. As will be discussed later, the placing of Facebook ads to recruit participants may lead to toxic behaviour (such as derogatory comments towards the marginalised population) on social media platforms. Even when the research design may have specific mechanisms to ensure anonymised data, other risks and privacy concerns can arise quickly. Specifically, when large amounts of potentially personally identifiable data (username, profile data etc.) are collected, stored, processed outside the original context, and/or made available to persons not properly or specifically authorised to have access [Boyd and Crawford, 2011, Zimmer, 2020, Smith et al., 1996, Buchanan, 2012].

Misrepresentation and "fake" information: Social media, like many forms of media, is a central platform for daily discourse. Yet, with this comes manipulation and other more nefarious uses. Specifically for studies that seek to leverage social media as a source of data for scientific study, there is an abundance of additional challenges related to the quality and/or veracity of information communicated. Key here is fake news and the spread of misinformation. Yet, also bots are a key example here that can distort or bias studies in a number of different ways. Firstly, by (re)sharing, (re)posting, or (re)tweeting information without checking the reliability of content, thus amplifying misinformation [Gupta et al., 2013], or having a high probability of posting spam content [Gorwa and Guilbeault, 2020, Cresci et al., 2018]. This is not necessarily always malicious, however, there are many cases where it is, as shown by [Ratkiewicz et al., 2011, Metaxas and Mustafaraj, 2012, Gupta et al., 2013,Cassa et al., 2013,Ferrara et al., 2016]. Regardless, the result is the same: it increases the likelihood of misinformation being included in a study. This means that any study using social media content, needs to consider the potential for bot-based (dis)information, however, as discussed by

[Ferrara et al., 2016, Hurtado et al., 2019, Madahali and Hall, 2020], there are now significant resources for bot detection.

Other information biases: Aside from bots, there are other representational issues that researchers need consider. Where self-representation – the promotion of an often stylised "voice" of the individual catered for the expected audience, e.g. being more positive online that offline – is a key consideration, see [Hall and Caton, 2017] for an overview. Self-representation is akin to a common method bias for social media researchers as there are not many easy-to-use tools to combat it. Whilst not specific to marginal phenomena, it would be naïve to not expect some amount of self-representation in studies of marginal phenomena. Similarly, sarcasm and other forms of figurative content are quite widespread on social media platforms, and emojis especially can quickly warp and change the meaning of content. There has been a lot of work on the detection of sarcasm and irony (e.g. [Amir et al., 2016, Muresan et al., 2016, Schifanella et al., 2016, Reyes et al., 2012, Reyes et al., 2013]), the classification of emojis (e.g. [Kralj Novak et al., 2015, Subramanian et al., 2019, Hasyim, 2019]) and related topics pertaining to figurative language use. The impact of these aspects of language use on social media platforms can vary. For example, in analysis of tone (or sentiment) sarcasm or irony would undermine the approach. Thus, in these scenarios, it is important that we attempt to detect figurative speech, and consider removing potentially problematic content. However, other methods of analysing social media content, for example in topic detection (i.e. what users talk about) the presence of sarcasm is less impactful as such approaches identify the terms and phrases, and are impacted less by the context of their use.

Social Media for Participant Recruitment

A significant number of studies have leveraged social media as a means to recruit marginalised or "hard-to-reach" populations for participation in studies [Carter-Harris et al., 2016, Pedersen et al., 2015]. [Russomanno et al., 2019, Andrews, 2012] comprehensively discuss some key considerations when social media acts as a recruitment platform for research studies, which act as a basis for discussion in this section.

Overview and General Observations

A very common approach visible in the literature is the use of Facebook targeted advertisements, i.e. where the target user profile for the advert is tailored towards specific sampling inclusion criteria for the study. Targeted advertisements allow the researcher to define a specific "profile" that matches their target participant(s). This has been employed across a wide range of populations: smokers [Carter-Harris et al., 2016], veterans [Pedersen et al., 2015], studies of sexuality [Hernandez-Romieu et al., 2014], abortion and early pregnancy [Altshuler et al., 2015, Arcia, 2014], substance use [Ramo and Prochaska, 2012, Lord et al., 2011], exposure to violence [Chu and Snider, 2013], mental health [Morgan et al., 2013], sexual health [Ahmed et al., 2013, Mitchell et al., 2016], low English proficiency [Carlini et al., 2015], and Eurofound's own COVID surveys.[1] [Altshuler et al., 2015] also note the use of Twitter posts and YouTube video content on the study website to increase awareness of the research objectives.

With the general observations that this manner of recruitment is typically more cost effective, quicker, and more efficient than other recruitment processes (e.g. newspaper, radio, flyers, direct community engagement etc.) and has the benefit of convenience for the respondent [Andrews, 2012]. With recruitment costs per participant reported in the range of US$3-15 (€2.85-14.25)[2] per respondent [Altshuler et al., 2015, Chu and Snider, 2013]. Note this cost can rise significantly if specific participation incentives are in place. Similarly, many studies also report that their population samples are "sufficiently diverse" (thus aligning with the objectives of social media-based recruitment), have no discernibly different biases or retention issues, and thus are a valid mechanism for recruiting individuals.

Many studies report high view rates of ads (e.g. tens of millions according to [Arcia, 2014, Ramo and Prochaska, 2012]) but with relatively low click-through rates, often around 1% and a conversion rate similarly small. In general, studies report the costs of an advertising campaign in the region of a few thousand US dollars (few thousand Euro) [Altshuler et al., 2015, Arcia, 2014]. With [Morgan et al., 2013] also comparing these amounts across other channels (e.g. search engine advertising) and observing similar costs over these platforms as well. An added benefit of this recruitment method is that the research team can observe the recruitment process in real-time [Lohse et al., 2013, Chu and Snider, 2013], e.g. modifying ad words or budget as needed. It is worth noting, however, that many of these studies were conducted in the US, and thus Euro costs may differ slightly. However, it is

[1] See: https://www.eurofound.europa.eu/publications/report/2020/living-working-and-covid-19#tab-03

[2] Based on the mid-market conversion rate from xe.com 20th April 2022

noted that Eurofound have adopted this recruitment strategy already in the COVID round 1 and round 2 surveys (see footnote 2), thus there is in house experience in this recruitment method to draw on, which can be augmented by this discussion.

As [Bauermeister et al., 2012] observe, there is the benefit that respondents refer new respondents via social media channels, through activities like tagging. This behaviour can significantly advance a recruitment campaign in a short time frame. [Bauermeister et al., 2012] discuss the formation of a seed population, recruited through (Facebook) adverts, and then outline how selection of the seeds allowed for a referral-based recruitment strategy, i.e. seeds referring individuals from within their own network(s) to participate. Care here is needed in the initial "seed matrix" to ensure sufficient diversity, and that appropriate screening of seeds is also undertaken. However, the authors note how this method resulted in a "large" (n =3,426) and diverse sample. The notion of "large" here is somewhat subjective.

Concerns and Obstacles

A point of potential concern is that recruitment over social media platforms can reach individuals well below the age of consent for studies, and also potentially reach vulnerable and at-risk individuals. Social media platforms allow for accounts from the age of 16 within the EU, however, there is evidence that this is often ignored or not adequately enforced by platforms. [Weeden et al., 2013] note that children as young as 9 misrepresent their age in order to gain access to social media platforms. [Altshuler et al., 2015, Russomanno et al., 2019] both note this in differing ways. [Altshuler et al., 2015] in accessing the age range 13-29, i.e. including under 18s on a sensitive subject matter (abortion). Thus, whilst this approach enables researchers to expand the reach of their research, significant care is needed.

A specific limitation of this method of recruitment, however, is the somewhat difficult to quantify non-participation bias. [Altshuler et al., 2015] discuss this in a similar manner to [Yancey et al., 2006]: that feasibility studies can act as markers for representativeness, and that the goal of these studies are to gather macro level insights. Thus, providing the sample is suitably large, the impacts of non-representation, is hopefully small. There is also the challenge of addressing repeated attempts to participate due to the online and semi-anonymous nature of this recruitment process [Lohse et al., 2013]. There are also self-representation biases (e.g. [Hall and Caton, 2017] for a more expansive discussion on self-representation) where potential respondents have been observed to manipulate their profile(s) in order to meet the eligibility criteria, this can be especially prevalent in studies that pay respondents [Kramer et al., 2014]. Yet, there are means to address some of these challenges, which will be discussed in subsection 3.3.

Subject-specific concerns may also arise depending on the context of the study, as marginalised populations can often be vulnerable, at risk, or otherwise stigmatised. Thus, there is a risk to social media users, and researchers as a consequence of the advertisement campaign. [Russomanno et al., 2019] succinctly highlights many of these possible side-effects. They note the existence of negative engagement in the form of derogatory and degrading comments with discriminatory comments in response to the advertisement on the social media platform (Facebook in this case). This is in line with other observations in the literature of social media in general (e.g. [Myers et al., 2017, Powell et al., 2020]). [Russomanno et al., 2019] also note, however, the potential affect this can have on the

researcher team, who may feel "responsible" for this behaviour and who also ultimately monitor and manage the advertisement campaign. This is even more poignant if a member of the research team identifies as a member of the population under study. Similarly, as with all studies of this nature, there is a risk to researchers of compassion fatigue stemming from an emotional response to the subject matter and potential online responses to the study and advertisement campaign. Thus, measures are needed in the research design to accommodate subject-specific concerns that may arise. Note that each platform has different functionality for controlling engagement with ads; comments in this case. For example, Facebook[3] allows for moderation, Instagram allows for disabling comments, Reddit[4] allows for both moderation and disabling comments, and Twitter[5] provides functionality to control who can respond to Tweets.

Quality Assurance in "Crowd Sourced" Respondents

When social media is used as a recruitment platform, some discussion is needed concerning mechanisms to increase the quality of responses. For example, [Kramer et al., 2014] suggest increasing technical hurdles so that only respondents that really wish to participate can do so. However, this can also be detrimental, as it can limit participation. Other suggestions can be to use data analytic solutions to, for example, identify "unusual" responses, or otherwise anomalous ones. However, it is not always clear what a good approach is, therefore, this section discusses some key quality assurance practices from the crowd sourcing literature and how they (potentially) increase the quality of contributions.

Quality control can mean many things depending on the research method and task(s)[6] that the respondents perform. It is a well-researched area in creative endeavours [Oppenlaender et al., 2020], policy and budget deliberations [Niemeyer et al., 2018, Prpić et al., 2015, Smith et al., 2015] open collaboration platforms [Friess and Eilders, 2015], and the broader (scientific) community [Law et al., 2017]. There are several other factors that may affect the quality of the work as per literature, including the characteristics of the respondent and demographics [Kazai et al., 2011] or personality traits [Kazai et al., 2012]). Key, for this working paper is the notion of (under)performance, which can be linked to the research team; e.g. instructions being unclear and the language used in the description being difficult to understand. Respondents also see task clarity as playing a major role in their performance [Gadiraju et al., 2017]. Task complexity, while being subjective, can be measured by visual appearance and language used in task description [Yang et al., 2016]. The crowd sourcing literature suggests many measures for assuring quality and authenticity which may be of specific relevance to the study of marginal phenomena when sourcing individual respondents via social media. Key approaches are discussed in this section to illustrate potential mechanisms to consider.

Of specific use may be that of pre-selection methods: to pre-select respondents based upon specific requirements or preferences; this is much aligned to the idea of recruitment via targeted adverts, but acts as an additional layer of quality assurance in the presence of self-representation biases.

[3]See for Facebook and Instagram: https://www.facebook.com/business/help/1129470964230971
[4]See: https://advertising.reddithelp.com/en/categories/optimization-management/manage-ads-with-comments-on
[5]See: https://business.twitter.com/en/blog/announcing-conversation-settings-for-ads.html
[6]In this context of this working paper, a task is set of questions or surveys. However, in general it can be any unit of human effort or work, e.g. translation, tagging, annotation etc.

[Geiger et al., 2011] define pre-selection as “a means of ensuring a minimum ex-ante quality level of contributions.” In other words, mitigate the risk of poor-quality solutions by screening, e.g. the completion of some process that demonstrates certain knowledge, skills, or other attributes. Researchers have utilised various techniques to apply pre-selection methods. With [Oleson et al., 2011] examining this process, and illustrating that it is typically performed via multiple-choice tests. For example, if the target respondent should have sufficient language fluency, part of the pre-selection could be completing some language competency test, or providing other evidence (e.g. uploading a DuoLingo certificate). Similarly, if respondents should have specific sociodemographic properties, a multiple-choice test could be used to screen for desired properties combined with specific logic in the survey to direct respondents to appropriate questions or sections.

Closely related to pre-selection methods are qualification tests, which can assess the abilities (or lack thereof) of a respondent and assess basic properties. [Gerber and Krzywdzinski, 2019] and [Chen et al., 2011] are examples here, and provide some guidance. They state that qualification tests can also capture demographic (and similar) properties such as geographical location. Similar to the basic notion of qualification tests are also initial screening questions based on reading attentiveness employed in order to minimize ‘clickthrough’ behaviours [Berinsky et al., 2012]. Such measures aim to ensure that contributors are dedicating significant attention to key elements of information, like the instructions.

Another well-established measure of quality control is the use of in-task (here perhaps the survey itself) quality control measures. Proposed by [Ipeirotis et al., 2010, Sheng et al., 2008] the idea is to infer a level of trust in the respondent via the accuracy of their response(s). There are different possibilities here. [Oleson et al., 2011] propose the use of gold standard questions. These are (sub)questions with known solutions (to the researcher, not the participant) that should not be easily identifiable or have easy to find answers. The presence of these questions enables the accuracy of a given respondent to be estimated. This helps improve the quality of responses solutions by providing an explanation of why something is incorrect.

A key consideration in quality assurance is less about making the “right” decision pertaining to which process of quality assurance is appropriate, rather giving the participant the impression of quality control. [Krause et al., 2019] illustrated that just alluding to the presence of a quality control process is sufficient to increase response quality and by extension reliability of collected data, i.e. it should be visible, but not too obtrusive or cumbersome. Thus, while some approaches have been highlighted here as ‘helpful’ realistically there is a lot of scope to design the quality assurance mechanism specifically for the research design, provided this is well signalled, and that there is feedback given to the respondent.

Summary and suggestions

Many studies on marginal populations have successfully leveraged social media as a platform to recruit respondents in surveys, interviews and similar research methods. Whilst there is a slight bias towards Facebook as the platform for targeted recruitment adverts, there is no reason to not expand to other platforms, where obvious examples (especially for work and employment-based studies) would be LinkedIn and the German Xing. Similarly, platforms like Reddit and Twitter also have similar structures for targeted advertisements, not forgetting search engines like Google too.

Much of the literature comments on such adverts being a cost-effective recruitment mechanism specifically in the study of marginalised populations. Yet, care is needed to not create new biases in the data by being too "precise" in the definition(s) of the targeted user profile. Thus, there is the suggestion to always conduct a small-scale feasibility (or pilot) study first.

To take an example of recruiting marginalised workers via social media adverts the literature suggests a few different strategies. The focus, however, is the formalisation of a set of representation criteria. Here, this would entail returning to past surveys (EWCS, EQLS, etc.) and review the underrepresented proportions of the sample population. Then to use this to tailor sets (plural) of target recipients for the adverts across multiple social media platforms. Once a set of target profiles have been composed, there are two main approaches in the literature in terms of how to use targeted social media adverts. The first, presented by [Bauermeister et al., 2012], comprises constructing an initial population or seed set of respondents who will recruit within their own networks under the assumption of homophily, and expectation that word of mouth referrals will increase a sense of purpose in participation. In fact, the development of this seed set, could also comprise a feasibility study. The alternative suggestion leveraged by many researchers is to use the advert to recruit directly. However, another recommendation in the literature ([Altshuler et al., 2015]) is to create a sense of purpose using social media platforms (e.g. YouTube, a companion website etc.) to raise awareness for the study and its objectives to help potential respondents understand why their participation is valued. [Dosono and Semaan, 2020] suggest that conducting interviews with online social media moderators may reveal additional insights into a community, the challenges they face and help highlight key voices to engage in other research designs. Yet we can make the same comment about specific structurally significant members of a community as well, we shall return to how we might potentially identify these users in the following section.

The main hesitation with this method of recruitment is that there may be specific negative outcomes of the advertisement campaign that will require a corresponding risk and ethical review. Briefly summarised, this can include, but is not limited to: 1) toxic comments towards the marginal population using the advert as a vehicle to promote derogatory content; 2) emotional fatigue for researchers stemming from this kind of response, but also the topic of the research in general. In addition, other methodological concerns may also arise: 1) non-participation biases are hard to quantify; 2) there may be click-through behaviours or other quality concerns stemming from additional anonymity afforded to the respondent, which may require additional quality assurance processes to be implemented; and 3) social media may generate other self-representation, or even echo-chamber-like biases in the sample population. For example, individuals may pose as marginal workers in order to promote their own (perhaps disparaging) views under the pretence of being a marginal worker, or to receive payment for completing a survey.

Social Media Communities as Data

Moving beyond explicitly seeking to recruit via social media, there are several studies that also try to use social media as a platform to study social phenomena and social groups. Many studies did, and still do, turn to Twitter for this, and others made use of Facebook pages, e.g. [Amon et al., 2016, Caton et al., 2015]. Today, things are more challenging, as we are in a "Post-API Age" [Freelon, 2018] where platforms like Twitter and Facebook have restricted access following major data scandals e.g. Facebook-Cambridge Analytica [Isaak and Hanna, 2018].[7] Facebook is now all but off limits to researchers, and to achieve similar results to [Amon et al., 2016, Caton et al., 2015], researcher(s) would now need to be a Facebook page administrator or have an embedded researcher at Facebook. Both scenarios create separate challenges with respect to impartiality, echo chambers, sample biases and related ethical concerns. An alternative, as introduced below, is to use a third-party data (re)seller.

In this section, the discussion focuses on a set of hypothetical research scenarios to discuss opportunities to study marginal phenomena with social media. These scenarios are presented as fictitious research questions. They explore how such a question might be addressed using Twitter and Reddit. This choice is motivated around how these two platforms represent quite different research designs. The goal is to present an outline of how meaningful such approaches might be in the study of marginal phenomenon.

However, first some discussion on the basic research design is needed. Many social media studies follow the same basic high-level research design, and this is somewhat independent of which social media platform(s) are in use and the general objectives of the research. To give an impression this basic design is depicted in Figure 1. Approaches focus on: 1) curating a dataset as well as defining what constitutes data; 2) apply one or more methods of quantitative analysis[8]; and 3) expected solution quality. To keep the methodology simple, KDD [Fayyad et al., 1996] is used as a reference framework. KDD is a methodology for knowledge extraction from data, Figure 1 captures the main aspects of KDD: identify key target data, transform and pre-process it to make it ready for specific techniques, apply these techniques in an attempt to extract useful information, evaluate these models using one or more metrics (e.g. accuracy, precision, recall, etc.), use appropriate domain expertise to interpret the results into actionable insights.

Figure 1: High-level approach to social media analysis comprising the steps: 1) data selection and filtering; 2) pre-processing and data preparation; 3) analysis and modelling; 4) evaluation; and 5) output and dissemination of findings.

[7] An API (application programming interface) is, in this scenario, a web-based endpoint that receives input (e.g. a set of search terms) and returns a response (e.g. a set of tweets). Web APIs operate over standardised communication protocols and have become the most common mechanisms for researchers to engage with, and access social media data.

[8] Note that qualitative methods are also possible, yet due to scale of data, these are omitted for practical reasons.

The remainder of this section is structured as follows: first, we explore what constitutes data, and how researchers typically gain access to it. Here, we introduce notions of data sampling that are key research design considerations. Second, some key research design considerations that impact how to leverage social media are introduced. Third, some key methods for processing social media content are introduced, these include trend and topic discovery (for text processing) and network analysis (for community detection). This discussion is only a sample of possible methods that could be applied to social media data, and thus it acts as a brief introduction. Subsequently, the three research scenarios for studying marginal phenomena with social media are introduced and then discussed with respect to key practicalities and potential hurdles. It is worth noting that the literature is surprisingly sparse on technical approaches to studying marginal phenomena with social media, and as discussed by [Arastoopour Irgens, 2022], there are very few research resources or guidelines available to researchers.

Data Access and Sampling

When considering social media, there are multiple possible definitions of data. Thus, before discussing data access and how to sample social media platforms, we will address the question of what constitutes data? Typically, data is represented as a post, comment, like (or similar), picture, video, share (or similar) or other related meta data (e.g. author, geo-location, tagged user etc.). This gives rise to studies that either look at the text content, and as such the data is user generated text (here a post: initiating a thread of discussion, and a comment: a response to this post, or other comments in the thread), or that uses users or their posting behaviours (and related interactions) to develop a graph or network representation of one or more communities.[9] Thus, the key decision(s) relate to how this text data is selected or discovered either to facilitate forms of content analysis, or to create a set of interactions between platform users to identify communities of users.

To curate a dataset or corpus of text content, researchers typically apply one of two approaches: 1) defining a set of keywords or search vocabulary as a means of data selection to search the platform(s) for content, or 2) select some subset of platform content (subreddits, Facebook pages etc.). For studies leveraging Twitter data, we see a number of consistent strategies: sampling via selecting key users and extracting / accessing their tweet history (e.g. [Patton et al., 2017]), using a set of seed users and sampling from “key” related users (e.g. [Patton et al., 2017, Nartey, 2021, Coe and Griffin, 2020]), focusing around specific hashtags (e.g. [Arastoopour Irgens, 2022,Nartey, 2021]), based on specific locations (e.g. [Murthy et al., 2016])[10], and using online repositories of Twitter archives for key public figures (e.g. [Coe and Griffin, 2020]). It is regardless of the approach, important to ensure sufficient topic representation [Nartey, 2021]. This typically entails ensuring that there is scope to discover sufficient content related to the phenomena of interest. For example, in a study using Reddit, common approaches would be to identify and extract “popular” and thematically relevant subreddits (e.g. [Datta and Adar, 2019, Hada et al., 2021, Webster, 2020, Bunting et al., 2021], perform a keyword based search (e.g. [Hada et al., 2021]), do random sampling of the platform as a whole [Hada et al., 2021], and undertake snowball sampling through cross posting and/or user activity (e.g. [Webster, 2020]). Thus it is key to explore multiple different

[9] Note that it is also possible to analyse multimedia content, images, video etc., but these have been omitted as not (yet) deemed relevant for this working paper to consider.
[10] It should be noted that geo-location data can be unreliable.

avenues of discourse on the platform, i.e. not just undertake a simple search based on a set of keywords. This will often entail a need for significant domain knowledge and experience with the platform(s) being used. It goes without saying that the rarer or harder to define a phenomena is, the more time intensive it will be to refine the means to sample data from one or more social media platforms. Later in this section, some hypothetical scenarios will elucidate this part of the process.

In order to access data, most social media platforms (that would be relevant for studying marginal phenomena) provide APIs. However, it is key to outline here that there is a wide range of privacy considerations and standards among different platforms. We can typically refer to social media APIs as open: generally available, but perhaps with specific limitations; partnered: where some third-party organisation (re)sell access to social media content under certain conditions (e.g. anonymity of users etc.) under varying subscription plans, which can be prohibitive in terms of cost; closed or otherwise restricted: where access is provided on the basis of certain conditions being met. An example here is access to data from Facebook pages, where only the page administrator can use the API to access their page's data.

Under the "open" API scenario, data is often made available by the platforms themselves, and most social media platforms offer a suite of APIs to developers. Yet, the approach to curate a meaningful dataset can be somewhat arduous. With key challenges being the number of API requests per hour, day, etc., the maximum amount of data (e.g. posts, comments, shares, engaged users etc.) that can be retrieved per request, and storage limitations within the terms and conditions of the platform (i.e. that raw data must not be stored externally, but rather only the post-processed data). The latter can have specific impacts regarding how quickly data can be curated as the time needed to process data can become a limitation quite quickly. Often platforms allow up to 1000 requests per hour (or day) and limit each request up to 1000 units of data (posts, comments, etc.). To give some context, [Caton et al., 2015] note that to curate a Facebook dataset required weeks for only a moderately sized sample. Whilst processing capabilities have increased significantly so too has data size and the complexity of models applied.

Under the "partnered" API scenario, data is made available by the platform to an authorised reseller. Some key examples here are Infegy[11] and Radian6[12]. Under this model, the rate at which data can be curated is bound by the terms of the agreement and the willingness to pay of the researcher / host organisation. However, subscription rates can be prohibitive; often in the range of thousands of US$ (thousands of Euro) per user per month, and may still come with quotas on data access.

Once the source(s) data within the platform(s) have been identified, the data itself needs to be made more manageable. It is common to removing things like user handles and other contextual information and meta data (unless this is specifically needed). This may also involve retaining data with specific content/terms [Li et al., 2012] or coming from a specific geo-location [Comito et al., 2019]. Such filtering may also consist of removing uninformative content [Zhou et al., 2020], the

[11] See: https://www.infegy.com
[12] Now rebranded as social studio and offered as a software as a service from SalesForce: https://www.salesforce.com/eu/products/marketing-cloud/social-media-marketing/

tokenization[13], normalization (e.g. stemming or lemmatization[14] [Kim et al., 2019,Tuarob and Tucker, 2015,Ko et al., 2020]) and removal of uninformative words e.g. stop words,[15] URLs, curse words etc. [Singh and Tucker, 2015, Choi et al., 2020, Zhou et al., 2020]. In addition, other forms of pre-processing and transformation may be applied depending on how any text data might be processed. Common steps may include the application of techniques like sentiment analysis, part of speech tagging, named entity recognition etc. It is, however, worth noting that much of the literature that studies marginal populations via social media uses methods more common in social sciences than technical approaches from computer science. These include, but are not limited to, textual analysis (e.g. [Nartey, 2021, Patton et al., 2017]), content analysis (e.g. [Coe and Griffin, 2020, Webster, 2020]), and thematic analysis (e.g. [Bunting et al., 2021, Lundmark and LeDrew, 2019]). Whilst such methods are often used in studying marginal populations, they can struggle to scale for big social data, which may be an inhibitor for specific research designs.

Impactful Research Design Considerations

Once the source(s) of data have been identified, there are quite a few significant research design considerations needed. The role of time in the planned study is a critical aspect of the research design; this is beyond the articulation of the research question. Here, we can essentially categorise the study as historical: it seeks to garner past observations of some phenomena, current / recent: it seeks to get a sense of some phenomena occurring relatively recently, e.g. the past few weeks; and present: it seeks to observe and track developments in (near) real time. This matters because different social media platforms and specifically APIs are better suited to different time scales in the research question(s). For example, it is quite difficult to use open APIs to access historical Twitter data, whereas this is not an issue for platforms like Reddit (e.g. using the PushShift API, which also has an API limit "five times greater" [Baumgartner et al., 2020] than the official Reddit). Twitter is relatively accommodating of allowing access to recent data (previous 2 weeks), however. Accessing real time data is much easier with Twitter (via its streaming API) than it is with Reddit. Another key consideration with time in the research design is whether there are forms of concept drift (for example the definition of the phenomena studied changes within the time frame of the research). Social media have well known echo chamber effects, and as such, it is important to consider that the nature of phenomena changes over time as expressed on social media platforms. In such scenarios, it can be very useful to explore how topic trends (discussed below) change over time to potentially identify how users discuss specific phenomena over time.

[13] Typically splitting text content into words, i.e., tokens.

[14] The aim of stemming is to reduce words to their word stem, e.g. worker, and working have the word stem: work. Typically, by cutting off word endings. Stemming is useful because it reduces the size of the vocabulary in the text content and groups words that are similar: work, working and worker for example. Like stemming, lemmatization uses more sophisticated methods to morphologically analyse words to remove inflections. For a more thorough overview of stemming and lemmatization see: https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html

[15] Stop words are words that are often filtered out of text content prior to processing it. There is no universal set of stop words, but typically they include articles (the, a, an, etc.), filler words (e.g. so, um, eh, etc.), question words (which, what, why etc.), and pronouns (she, her, him, he, they, etc.). Stop words are not always removed (they are not always uninformative), and the set of stop words (or stop list) is usually tailored to the problem at hand.

Another key consideration is whether the phenomenon to be studied is well formed within the platform. Again, we can use Twitter and Reddit here as examples. Twitter structures access to data around key aspects of the data: presence of one or more keywords / hashtags / user handles, geographic location, language, etc. This means the research needs to consider how the dataset will be discovered and curated. Reddit, however, organises content into "subreddits" (discussion forums) and research has stated the benefit of having data organised in this manner capturing data. Platforms like Facebook, Twitter and Instagram are limited in this sense [Jeong et al., 2019]. Thus, if the research theme is clear and suitable subreddit(s) exist, the process of curating the dataset can leverage this in-built structure.

The expressivity of the platform should also not be underestimated. Expressivity here refers to the affordance given to users by the social media platform to articulate themselves. This is important because it controls how users generate content and engage with the content of others. For example, Twitter limits users to 280 characters of text, thus Tweets need to be short and to the point. Reddit and Facebook, however, place no such limitations on their users, and thus users are free to more expansively express themselves. The aspect of consideration here is that when length limitations are not in place, one post can contain the entirety of the content and its intended purpose. However, when length limits are in place, users can chain their posts into threads, and this sequencing may need to be both identified and considered in the analysis. As noted by [Coe and Griffin, 2020], threads can create methodological inconsistencies, and as such care is needed in their handling. Closely related is the use case that platforms have, as this also affects how users express themselves. For example, Facebook captures much more of a user's private sphere verses for example LinkedIn where their focus and content generation has a more professionally aligned stance. In terms of expression, this means that researchers need to consider how users use the platform and consider any methodological implications of this.

The granularity of the research is also key. Approaches to text analysis work at a "document" level, and this means that a key aspect in the research design is to determine the granularity of study, i.e. what constitutes a document. This can be any singular unit of text content: post, comment, post and all its comments, all posts and comments from one hour / day / month etc. Determining the "right" level of granularity can be quite tricky, too fine grained, and there is too much information to sift through, too coarse grained and the analysis will not be nuanced enough. From a computational perspective, the finer the granularity, the more computationally expensive the study is likely to be.

Key Analytical Methods for Social Media Mining

Text as Data

When analysing text data, social media researchers tend to apply one or more modelling techniques to extract information, potentially further transform the data, and/or generate new representations of the data. In the simplest case, this can be to use a bag-of-words method such as word clouds to get a sense of common terms in the community as undertaken by [Arastoopour Irgens, 2022]. This,

coupled with other exploratory data analysis techniques, can also be used for vernacular discovery,[16] which as noted by [Patton et al., 2017] is important as different communities may make use of specific sociolinguistic variations of language as well as slang, shorthand etc. Thus, this type of exploratory data analysis is often helpful in the absence of specific domain expertise. However, as noted by [Patton et al., 2017, Bunting et al., 2021], researcher training in the area of study (to generate domain knowledge) is key. An alternative method to determine a (starting) vocabulary set, as introduced by [Burnap et al., 2017], is to collect anonymised data from relevant Web forums, blogs and microblogs, process the text data into a set of terms or phrases, perform some initial processing to remove obviously superfluous terms (for example, use natural language processing tools to keep only nouns, noun phrases, etc.) and ask human annotators to identify whether remaining content (terms or phrases) contains specific references of use, i.e. are they representative of the phenomena?

As noted by [Coe and Griffin, 2020, Hart et al., 2013], tone is a useful mechanism to understand content context through the means to create specific social impressions. However, as noted by [Qiu et al., 2012] users communicate their positive emotions online more frequently via social posturing, finding that negative emotions are hardly communicated. When negative (and positive) emotions are used, they tend to cluster around user groups [Bollen et al., 2011, Cacioppo et al., 2009]. There are many tools available to researchers to study tone (or sentiment, or valance), with LIWC (Linguistic Inquiry and Word Count) [Pennebaker et al., 2001] being popular among computational social science studies, especially those that use user-generated content via social media, and studies which move beyond English. Key however, is to not rely on only one tool as there can be quite substantial differences between tools in what they define as "positive" and "negative" in tone. Similarly, it is noteworthy that analysis of non-native speakers (which may be quite common in marginal populations and on social media where there is a strong English bias) can be misrepresented by such tools as highlighted by [Zhiltsova et al., 2019, Kiritchenko and Mohammad, 2018, Bolukbasi et al., 2016, Davidson et al., 2019]. Yet, this notwithstanding, as noted by [Coe and Griffin, 2020], research into marginal populations has often been concerned with positive vs. negative views, viewpoints, and discourse. Sentiment analysis tools enable studies on either text content in general, or focused subsets of social media content. For example, sentiment around or towards specific terms, people, concepts, etc. [Hada et al., 2021] illustrate an example of creating a large corpus of Reddit data, and also discuss a large number of inclusion and exclusion criteria for determining which Reddit content to consider for analysis.

Here methods for the identification of keywords or topics that "represent" aspects of the phenomenon being investigated are of specific use. This can also be focused on keywords, key people / things, and/or topics; and other linguistic categories of analysis. We can identify a few

[16] In the study of marginal phenomena, there might be a need to refine the vocabulary used to describe or define specific aspects of importance. The idea here of vernacular discovery is similar to snowball sampling: we look for terms that have a high co-occurrence rate with other terms that are "important". There are quite a few techniques for doing this, but one specific example of relevance (albeit used in product discovery) is to also look for usually rare words that are common in the social media corpus, and which co-occur frequently with previously identified key terms. In doing so, the researcher can identify new terms and even emerging terms over time that can be used to refine and extend the vocabulary set used in the study, i.e. discover new elements of vernacular. A more detailed overview (using Reddit) can be found in [Kilroy et al., 2022].

different "types" of modelling to meet different objectives. In the area of topic detection, i.e. what are the topics of discussion related to the phenomena of study, it is common to find what is currently "popular" or "trending" by detecting "bursts"[17] in activity [Romo-Fernández et al., 2013, Trinquart and Galea, 2015, Kleinberg, 2003, Uchitpe et al., 2016, Krishnamoorthy, 2015]. Often this type of analysis can be coupled with event detection: time-specific instances of significant online activity around a given theme (e.g. the scoring of a goal, outbreak of a riot etc.). When exploring social media content for events, we do not know ex-ante how many events occur: they are unknown. Generally speaking, the detection of unknown events [Allan et al., 1998] can be approached in three ways: the documents themselves are clustered (document-pivot) [Ifrim et al., 2014, Comito et al., 2019, Petrović et al., 2010], the terms from the documents are selected then clustered (feature-pivot) [Aiello et al., 2013,Saeed et al., 2019, Zhang and Qu, 2015] or events are seen as a probability distribution over documents/terms (topic modelling) [Xie et al., 2016]. Key here is that event detection can be used to observe movements, themes or epochs in a phenomena, and that the amount of text content available doesn't significantly hamper the approach. More is obviously better, but as these approaches tend to work using changes in word frequency and absolute frequency, they can be applied to smaller sample datasets and can be applied to find "rare" events.

To derive topics from text content, several methods are common with Latent Dirichlet Allocation (LDA) [Blei et al., 2003], Non-Negative Matrix Factorization [Lee and Seung, 1999] and Latent Semantic Analysis (LSA) [Deerwester et al., 1990] being the most frequently used. Much work on topic modelling in social media focuses on variants of these popular algorithms for short text mining [Cheng et al., 2014], while other works focus purely on real-time over online topic models [Xie et al., 2016]. An additional option would be approaches to phrase detection, for example the "Phrase Detection" model in Gensim (a Python library for topic modelling) could be used in order to identify collocate words, i.e. phrases.[18]

It should be noted that often there is a need to use human annotators to support computational methods. This usually occurs where insufficient or insufficiently labelled data exists. As illustrated by [Burnap et al., 2015, Hada et al., 2021, Burnap et al., 2017], this can be a useful exercise in the training of machine learning models for text analysis or, as undertaken by [Burnap et al., 2017], to further identify content to analyse with other methods.

Text as a Proxy for Interaction(s)

Beyond explicit processing of text to derive knowledge or insights from social media data, many researchers do not focus on just the content itself, but rather recognise that text is produced as a side effect of some communicative act between users. Thus, text can also be used to model underlying activities, community structure and more generally interaction(s). Whilst this is common in more general social media studies, it is not that common in studies of marginal phenomena where researchers (often social scientists) are limited by available tools, which are both expensive and often clumsy. There is, however, a wealth of open-source software libraries available for the study of networks; with the Python-based NetworkX [Hagberg et al., 2008] and Gephi [Bastian et al., 2009]

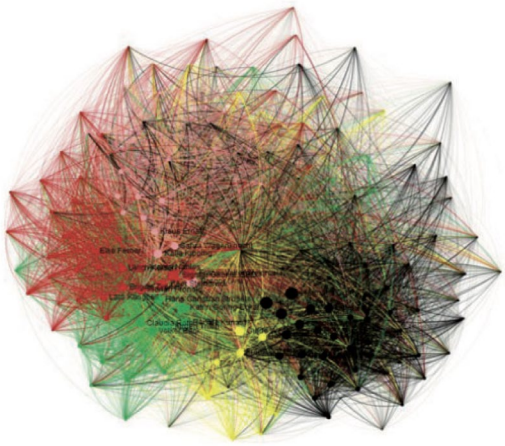
[17] A burst here refers to a sudden, yet significant, increase in a term or topic. It typically signifies a rapid change in discussion on a platform.

[18] https://radimrehurek.com/gensim/models/phrases.html - last accessed 24/03/2022

being very popular. As the network analysis literature is vast, the discussion here will emphasise on how (at a high-level) text is converted into a network or other forms of data for analysis.

By mapping each interaction (here a mention, retweet, comment, upvote / like, cross post etc.) between all elements (here users, subreddits, hashtags etc.) it is possible to compile a graph for analysis. The simplest case is the transformation of text elements (e.g. tweet, comment, or post), metadata (time of day, whether a retweet or not, author etc.) or content (tags, mentions, hashtags, URLs etc) into metrics which can be used to measure activity intensity, intent, direction or interactions in general (e.g. [Caton et al., 2015, Murthy et al., 2016, Burnap and Williams, 2015]). As summarised by [Burnap et al., 2014] these entail social features (tweet / post, comment, and post metadata, such as retweets, upvotes etc.), temporal features (e.g. lagged response rates) and content features. A social graph can be used for tasks like information flow and understanding directions of information, identifying key members of a community, community or (sub)group detection, community interaction(s), etc. by applying different (social) network analysis methods. To give a concrete example, Figure 2 shows a colour coded (coloured by political party) network where nodes are political Facebook pages, and edges are interactions between these pages, represented by users having posted on both Facebook pages. Here the interaction was derived only by collecting and comparing lists of active users on each Facebook page.

Figure 2: Example social network graph representing interactions between different political parties. This figure is reproduced from [Caton et al., 2015], the colours represent different political parties, with each node representing a Politician's Facebook page, and the edges links between them, represented by Facebook users commenting on both pages.



Illustrative Research Scenarios

Scenario 1: What are (recent) trends or topics discussed in the context of marginal work(ers)?

This scenario captures the idea of "emerging trends" and topics. This is an exceptionally well researched area that is often applied to specific and focused studies, e.g. who will win the election, what do viewers of the FA cup final talk about on social media, what are emerging product ideas, etc. Key in this scenario is to illustrate that similar techniques could be used to identify "themes" of discussion around specific marginal phenomena. This scenario would also support research designs that seek to explore whether topics, themes and phrase use have changed over time. It captures methodologies for the analysis of trends and topics on social media.

In terms of the utility of a research question like this, it could be imagined as a means to generate new insights into a marginal phenomenon, that may then be further explored in other research designs: deriving new survey questions. Similarly, it could also be used in scenario two as a means to generate a general corpus of data that can be transformed into a social graph for community-based analyses. In terms of complexity, this is the easiest of the three scenarios, and the most likely use of social media to generate new actionable insights in the study of marginal phenomena. At minimum, some programming (likely Python) is needed to extract, transform and process text content and to then store the processed data for final analysis.

Scenario 2: Who are members of marginal work(er) population, and what does the online community look like? Are there "key" users in this community?

Here, the results of scenario 1 could be applied to detect a community of users, and then construct a sample population for a specific marginal phenomenon. Such a community might support various research designs (beyond surveys), e.g. content analysis studies, and could also act as a bootstrap mechanism for other research designs, by for example discovering new participants for surveys or key users to aid in the advertisement of a survey. It captures methodologies and technical approaches to social network analysis.

In terms of the utility of a research question like this, it could be imagined as a means to identify potential respondents in a more traditional design, i.e. survey. Similarly, it could also be used to better understand key aspects of a community in augmentation of scenario 1, i.e. having identified the following themes, topics etc. how are these themes communicated through different marginalised communities. In terms of complexity, this really depends upon the ease of generating useful data and correspondingly network structure from the data as well as the actual social network analysis methods applied. At minimum, some programming is required as well as some fluency with graph algorithms, their application(s) and implementation details.

Scenario 3: Can different "localisations" of marginal work be uncovered?

This scenario captures more complex and ultimately realistic research questions that attempt to drill down into social media data. This results in a highly complex research design especially if multiple platforms are needed. The goal is to differentiate the findings of a study similar to scenario 1 by slicing the data across specific dimensions when considering a pan-European context. Where specific examples include, but are not limited to, language used, region, gender, age and other macro non-

invasive sociodemographic or sociocultural factors. Possible research studies here could seek to understand if marginal phenomena are represented differently within different sub-groups across these dimensions capturing and potentially reconciling differences in the manifestation of marginal phenomena across platforms, (spoken) languages, countries or regions etc.

In terms of the utility of a research question like this, it could be imagined as a means to identify diversification of opinions / content aligned to sociodemographic or sociocultural properties of interest, e.g. location, native language etc. This scenario has a very high complexity that is unlikely to be realised without significant resources. It is hampered by a general lack of tools that can reliably process languages other than English, that mainstream social media platforms have a significant English bias, and also that context related domain and local expertise is needed to refine the process of data gathering; i.e. the identification of subreddits, key terms or users etc.

Scenario 1: How We Talk About Marginal Work

It is important to consider the implications noted earlier in this section for this scenario. For simplicity, we will assume that the Twitter scenario will operate on a (near) real time basis (and thus leverage the streaming API[19]) whereas Reddit will work more historically (and thus leverage the Push-Shift API[20]). We'll also assume that we will only seek and process text content in English (and consequently, part of the pre-processing and filtering of content will be to remove non-English posts and comments). This assumption will be revisited in scenario 3.

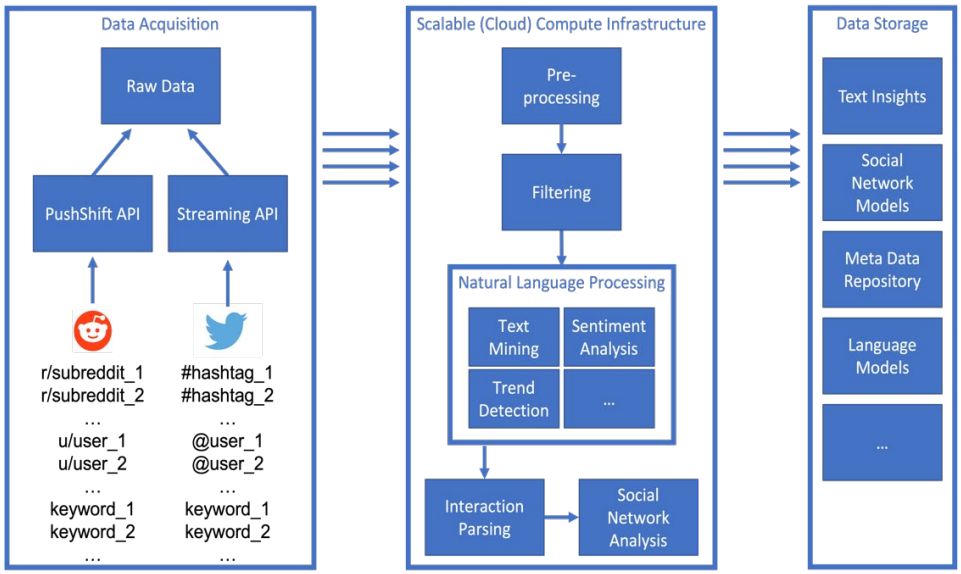
Next, we need to identify key entry points into the platforms to draw a sample of text data. As previously noted, this can involve many steps. Yet, as the working paper seeks to provide a general overview, the following would be recommended: source and elicit key domain-specific terminology (including appropriate sociolinguistic terms) from key stakeholders, as well as parse and process relevant (micro)blogs to develop a lexicon of initial key search terms. Key terms could be derived either with the aid of human annotators to label terms as (ir)relevant. Alternatively, a set of reference documents (e.g. context-neutral newspaper articles of various categories in the target language) could be used in addition to standard approaches (like removing stop words, specific part of speech tagged words: articles, pronouns etc.) to aid the removal of "common" but frequent words that are not domain specific. Finally, this process could also be augmented using Google trends to potentially identify additional search terms under various contexts of employment.

Using standard search tools on Twitter would also reveal an initial population of tweets (making suggestions of hashtags to include in the study) and potential users to more explicitly include in the sampling process. This would result in a set of candidate search terms for use with the Twitter Streaming API. Similarly, for Reddit, the same set of terms could be used to search for subreddits which make use of these terms or simply to search for posts / comments. Thus, tailoring the data curation process. With a set of initial keywords identified (and note that these can be updated as the study progresses, and more domain knowledge acquired) the extraction and processing architecture can be considered and is shown in Figure 3. Note that this architecture will be similar across all scenarios. Often a scalable cloud infrastructure serves a good basis for an architecture like this, as it can scale with the data and processing requirements. Note that the key computational aspects are

[19] The Twitter streaming API enables real-time extraction of tweets as they are tweeted. It uses a set of filters to identify relevant tweets (e.g. presence of keywords), and returns these in accordance to Twitter's rate limitations.
[20] The PushShift API enables historical access to Reddit data see: [Baumgartner et al., 2020]

captured in Figure 3: extracting the raw text data, which often creates a stream of data to be processed (as raw data should not be placed in offline storage in alignment with many platforms' terms and conditions), and thus motivates a scalable computational backend to process data as extracted from the platform. Note that some approaches will require a data staging area where extracted data is held in memory (or similar) until it can be batch-processed. Batch-processing here is often needed by methods that compute statistical measures between documents within a corpus. Similarly, depending on the granularity of the study, content may need to be assembled into documents prior to processing.

Figure 3: General analytical architecture for social media analysis



For indicative purposes only, we will assume that using Twitter data, we would like to observe "trending" terms over time, i.e. highlight the emergence of "unusual" or new terms, as a proxy for potentially interesting new phenomena in marginal work. The time-scale of a study of this type is likely months, and as already noted, we have to be quite careful how we sample Twitter data over time. Noted here is the methodology of [Kilroy et al., 2020] where it was shown that using a large(r) number of small(er) time windows that are later combined is useful in the identification of new and emerging trends over time with Twitter data. [Kilroy et al., 2020] also derive a "term importance" score for the identification of trends and this would be suggested here too. Thus, in each time window, a ranked list of "important" but "bursty" terms could be derived and presented for human interpretation to derive insights from (this would usually be between 10 and 100 terms per time window, as is often the case in these types of study, e.g. [Kilroy et al., 2020]. By tracking such terms over time their sustained social media use could be tracked to later inform other research designs and also potentially to highlight new (sociolinguistic) domain specific terminology. Similarly, sentiment-based tools could highlight the tone of discussions surrounding keywords of interest, although, this may be better suited to platforms like Reddit, where content tends to be longer.

Using Reddit, the same text processing approach could in fact be used. However, the timescale would be quite different. Operating historically, and not in (near) real-time such a study could be

operationalised in a potentially shorter space of time; bounded by the time needed to implement the project and computational resources needed to process the data. Unlike Twitter, Reddit data is more historically accessible and based on API rate limitations, a well sized corpus of data (assuming one exists) could be curated relatively quickly; days or weeks to collect months or even years of social media content depending on the target timeline of the study. Of note is the work of [Hada et al., 2021] that investigated the semi-automated curation (semi-automated here because there was an element of human annotation in the process) of a Reddit corpus that would closely align with this scenario in design and operationalisation.

Scenario Summary: Realistically a (simple) study of emerging themes and topics pertaining to one or more marginal phenomena should be possible. Study success will depend on the ability to find a suitable initial lexicon of terms to enable keyword-based search with Twitter or Reddit to uncover Tweets, Hashtags, users, subreddits etc. There is a wealth of natural language processing options available to support different research designs once a sufficient text corpus is curated. The key remark at this stage really is the lack of structured guidance (as noted by [Arastoopour Irgens, 2022]) on study design beyond the literature discussed in this working paper. As such, it would be suggested to commence with a small-scale feasibility study in an area where the initial lexicon of data is well known and where the research team is confident in its use on social media platforms.

Scenario 2: Discovering Marginalised Workers and/or Communities

There are two main use cases that this scenario can address for studying marginal phenomenon: 1) identifying communities of users (e.g. those that discuss marginal work) and 2) identifying key users; e.g. those that may “influence” others or who have large structural roles in information dissemination. Similar to scenario 1, a key aspect for community detection or to detect individual(s) of “interest” is to generate a suitable corpus of social media content, thus a similar approach to corpus construction as in scenario 1 would be applied. However, here the emphasis is less on text processing and more concerned with interaction effects within the corpus. Yet, that notwithstanding, tone or sentiment analysis methods could identify if individuals or communities generally communicate positively or negatively, if that is of interest. For this scenario, we are going to use a simple mechanism of community construction from tweets (Twitter) and posts / comments (Reddit). Note however, that because we are less interested in the text content, we would not necessarily need to remove non-English content (and could if desired, even use language detection methods to tag users with the language(s) they use).

To construct a community or network from Twitter data, the easiest approach is to view interactions (retweets, use of specific hashtags, tags via handle indicating a reply or other communicative intent, and being a follower of a user). Here, each action indicates some level of interest or interactivity with content elements. Thus, we build a dataset (graph) of nodes (users) and edges (representing one or more interactions). Note we could weight edges according to the number of interactions between two nodes. Typically, we would represent this graph using an adjacency matrix where each cell in the matrix represents the strength of interaction (0: no interaction or 1 to *n* interactions), this can also encode the directionality of interaction too, as each pair of users have 2 cells in the matrix. In fact, this is a similar approach, in part, to [Lutkenhaus et al., 2019].

To construct a community from Reddit, it would be tempting to look at subscriptions to specific subreddits, however, as noted by [Datta and Adar, 2019] users may not necessarily be actively involved in subreddits they subscribe to, and similarly, many social media users are simply "lurkers" [Liu et al., 2014]. [21] Instead, it may be better to follow an interaction-based approach similar to the Twitter dataset. Here key interactions would be up/down votes (to note the acceptance of a post / comment), post-based interactions, i.e. comments between users, and mentions. This too would be used to derive an adjacency matrix.

With our adjacency matrices as the basic data structure, we can apply specific social network analysis techniques to identify "interesting" aspects of our community. Network analysis is the study of the networks that emerge when the links within our matrices are graphed [Lim et al., 2013]. Here, even just the presence or absence of homophily can be quite informative: it provides a view of the dynamics of the (sub)community. However, for our scenario, we are going to leverage centrality analysis to find: 1) key nodes (users) in the networks; 2) "Popular" users who are perhaps influential leaders, hubs or experts in a network; 3) key users that bridge different communities; and 4) users with an ability to communicate with large numbers of other users. Specifically, we are going to use in/out degree centrality, and betweenness centrality[22] (two of the most basic social network analysis metrics for a node), and we want users with "high" centrality values. High in-degree centrality would indicate that a user receives a lot of interactions, high out-degree centrality that they reach a lot of users, high betweenness centrality that they bridge subnetworks, i.e. that they have "good" information dissemination potential. This is an approach often leveraged to identify key users within communities of users for product placement (to ensure the "right" users are given early access and then provide reviews to larger audiences, see for example [Chau and Xu, 2012]).

For example, Figure 4 illustrates the ideas of centrality and why these measures might be useful for studying marginal phenomenon: users with "high" centrality are "interesting" for several reasons: 1) those with high out-centrality can reach many other users which can be extremely useful for recruiting users for surveys. Similarly, users with high betweenness centrality can aid in maximising the reach of messages (like recruitment drives for surveys) further because of how they are positioned in the social structure of multiple communities. There are of course no guarantees here, and users may need specific incentives to cooperate.

To identify (sub)groupings (i.e. colour coding in Figure 4: red, green and purple clusters of users), there are many possible techniques. The easiest is to (automatically) assign labels to nodes. For

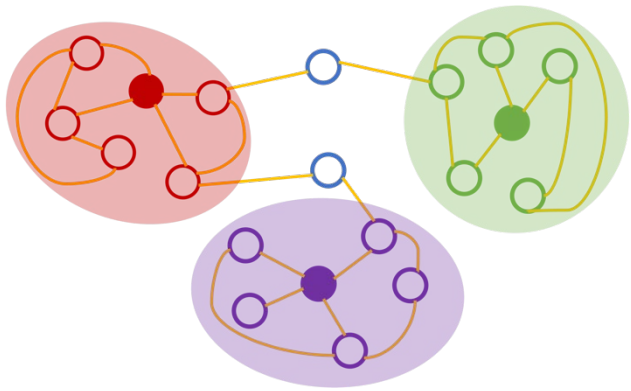
[21] Passive consumers of content without actively engaging in discussion. The kinds of information that can be gathered on lurkers depends on the extent of their passive consumption. If they do nothing and never engage with content (e.g. like, share, etc.) it can be quite difficult to gather much information on them. Beyond their presence in lists of followers (e.g. Twitter, and Instagram) or subscribers (e.g. Reddit, and YouTube). However, as soon as a user engages with content, the portions of their profile which are publicly visible are accessible. This occurs because when a user engages with content, this engagement is stored, and accessible through the social media platform's API(s); the ID of the user becomes associated with the post, and is the user profile queried. Note that this level of user spotlighting can be seen as privacy invasion, and that typically ethical review processes will need a reasonable justification for the need to spotlight specific users.

[22] Betweenness centrality is a way of detecting the amount of influence a node has over the flow of information, see https://neo4j.com/docs/graph-data-science/current/algorithms/betweenness-centrality/ for a practical example and more details.

example, this could be based on vocabulary, topic, hashtag, or subreddit "similarity", or any other comparable set of measures that can be quantitatively compared in terms of their (dis)similarity. Users that are "similar" are placed "near" to each other in the graph to form clusters. Of course, forms of clustering can be used as well, if appropriate. Yet the challenge here would be to derive a meaningful notion of cluster "quality". To return to the idea of community tone, we could also cluster users based on the tone of their social media posts / comments.

Once some mechanism for forming communities has been devised, we can conduct various analyses of the content corresponding to a specific community, e.g. the purple community illustrated in Figure 4. Here, we could filter the original social media corpus corresponding to only members of the purple community and apply specific studies to this subset. For example, we could look at the tone of posts / comments / tweets from this community when a specific term, such as "marginal worker",[23] is mentioned to estimate their acceptance of this type of worker. Alternatively, we could attempt to garner a larger set of the content produced by these users. However, doing so is not so straightforward, Twitter has quite severe limitations on user history extraction via its APIs. For Reddit, this is quite doable from a technical standpoint. This additional form of community study could further reveal whether this community should be included in other analyses. Similarly, we could repeat the analysis outlined for scenario one for just one community. However, there might be reservations in "spotlighting" users to this extent depending on the exact research question(s) being posed.

Figure 4: Sample simple network illustrating node of "high" centrality. Here, the circles represent users (nodes) the yellow lines are edges, and we have 3 communities: green, purple and red. The solid fill nodes have high in/out centrality, and the blue nodes have high betweenness centrality as they connect otherwise unrelated communities.



Scenario Summary: Similarly, to scenario 1, the success of scenario 2 depends on the ability to find a suitable set of Twitter or Reddit content to construct a social graph to analyse. There are no guarantees that the social media platform has suitable representation of the phenomena being

[23] Note that "marginal worker" is unlikely to be used as a term in online discourse, and serves here only as an example.

studied. Yet, should it be possible to curate an appropriate dataset, there is a wealth of social network analysis methods available to support different research designs and a large number of options to represent interactions within the data to construct the social graph. In terms of possible use cases, the most likely would be to: 1) identify social media users to either promote traditional surveys, or potentially recruit directly as participants, or 2) build and study community dynamics. Note that we can relax the only English content assumption for constructing the social graph, but any forms of "deeper" analysis that makes use of text content requires further consideration. This will be discussed in scenario 3.

Scenario 3: Localisation Studies: Comparing languages or regions

In scenarios 1 and 2, there was an assumption that only English content will be processed. Similarly, there was also an assumption that the social media platform represents one collection of users dividable into different subsets. These assumptions could be somewhat relaxed for scenario 2, this was only to construct a social graph from general notions of interaction(s), which can be performed in a language agnostic manner up to an extent and can segment users into different groups. In this last scenario, we want to explore a specific challenge for Eurofound: research that spans countries, languages and cultures and seeking to unravel differentiated manifestations of marginal phenomena. Or more specifically how are different marginalised phenomena manifested across the pan-European landscape; when and how do they change? It's worth noting already here that this is not easy to accomplish, especially as the number of differentiating factors (e.g. language, region, (sub)population(s)) increases. Thus, realisations of this scenario will require significant resources to undertake.

There is one overarching key challenge for this scenario: improving language representation and diversity. The proportion of content in English is extremely high on "mainstream" social media platforms:[24] there is a distinct English bias. This does not mean that other languages are absent, in fact Twitter has a large user community outside the English-speaking countries,[25] but rather that for some social media platforms languages other than English might be significantly less prevalent for specific topics, which might foster biases in the research design, and by extension results. It may also limit what can be achieved with social media platforms in some countries.

There is, however, a rapidly growing set of tools for processing languages other than English. With for examples, studies on Twitter for French [Mazoyer et al., 2020], German [Cieliebak et al., 2017], Spanish [Gonzalez et al., 2021], Portuguese [de Melo and Figueiredo, 2020], Dutch [Lutkenhaus et al., 2019], and there are numerous examples of other languages as well. However, it is noteworthy that in addition to the observation that there are already few resources and guidelines for studying marginal populations (and by extension marginal phenomena) on social media [Arastoopour Irgens, 2022], many of the non-English studies are on quite mainstream topics: elections, vaccination acceptance, etc. Thus, these topics are "easy" to define already and generally discussed on social

[24] Reportedly as high as 97% for Reddit, see: https://towardsdatascience.com/the-most-popular-languages-on-reddit-analyzed-with-snowflake-and-a-java-udtf-4e58c8ba473c – last accessed March 24, 2022

[25] See: https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries – last accessed March 24, 2022

media platforms: they are newsworthy. It is also noteworthy that many also rely on leveraging more complex methods and (pre-trained) deep learning model architectures that are then adapted to specific tasks based on the content to hand. This requires a significant amount of data, which as mentioned, might not be too easily available, and significant resources too in terms of computation and analyst skill.

This notwithstanding, it doesn't mean such studies are infeasible from a technical perspective, but that care is needed; they are more resource intensive, and have a higher degree of complexity. Simply put, a study seeking to understand localised (i.e. national) differences in perception towards marginal work(ers) would need to repeat scenario 1 for each country (most likely via language) of interest, yet note that countries like Germany, Austria and to some extent Luxembourg, Switzerland and parts of Italy all use German, and thus disambiguation between certain regions may be a separate challenge. A similar challenge would occur with Spanish (e.g. Latin American countries), Portuguese (e.g. Brazil), French (e.g. North African countries, Belgium, Switzerland etc.). So it might be easier to note that general language-based tendencies are discernible, but focusing the social media lens onto very specific countries in the absence of significant data and/or localised vernacular might require some novel methods to be developed. It is difficult to say as there are not many studies of this type in the literature; at least beyond perceptions of English vs. one other language (often French, German, Arabic etc.), and even then these studies usually rely on comparing the tone (sentiment) of discussions at a macro level as opposed to a topic level. While on Twitter, such a challenge may make the study too prohibitive, the structured nature of Reddit (via subreddits) may appease some of the data collection issues if the subject study is sufficiently topical, as subreddits exist to discuss contemporary topics in most countries. This inherent structure could act as a convenient starting place for small(er)-scale feasibility studies.

It might be tempting to consider "normalising" text content through automated translation tools. So for example, because English is a better supported language for natural language processing tasks, one might consider using machine translation to translate all content into English, and thus removing the challenges of handling multiple languages. On the surface, this would seem like a sensible approach, as all content would be in one language appearing to significantly reduce effort. Obviously assuming good translation tools and some additional pre-processing to handle both synonyms and cognates. Yet, this assumption is quite precarious. There has been significant work that shows that there are significant biases in either machine translation, non-native-level language use (such as from a poor translation), and that other facets of language (often stemming from different pronouns, genders, races, sexual orientations etc.) can significantly affect machine learning based text processing systems e.g. [Dixon et al., 2018, Díaz et al., 2018, Kiritchenko and Mohammad, 2018, Zhiltsova et al., 2019, Blodgett et al., 2020, Saunders and Byrne, 2020, Saunders et al., 2020]. Returning to the definition of a marginal population, many of these sociocultural constructs are likely present in the text content. Thus, automated translation is not recommended.

In cases where multiple platforms are needed due to insufficient data for one or more localised viewpoints there are specific challenges to consider. For example, collecting data from Russian social media platform VK and/or the German Xing (assuming appropriate APIs to do so). In fact, there have been explicit calls in the literature to better develop research practices for multi-platform studies (e.g. [Hall et al., 2018]). Briefly summarised, the main technical challenges of note for multi-platform studies are: 1) potentially different or additional instantiations of the research design considerations

noted at the start of this section; 2) additional mechanisms, (re)sellers, or limits for extracting data; 3) potentially different terms and conditions as well as legislation to navigate; 4) new biases (as discussed in [Hall et al., 2018]) may arise depending on the research design. None of these completely rule out such a research design, but they may significantly increase the level of technical expertise needed and by extension project resources.

Scenario Summary: All of the noted requirements for scenarios 1 and 2 are repeated in scenario 3, yet amplified due to its multifaceted nature. Yet, this multifaceted nature adds new challenges, which are potentially prohibitive: they significantly increase the magnitude of project resourcing required. Whilst scenario 1 and 2 can be likely conducted by a team with Masters level technical education, scenario 3 requires the project team to have significant research and technical expertise likely to be found in more senior data scientists and/or PhD and post-doctoral researchers. This aside, there is a strong recommendation that any study falling under a remit similar to scenario 3 first commence with a set small(er) scale feasibility studies capturing a select few target localisations. Such feasibility studies should serve a few purposes: 1) highlight any technical gaps in the research implementation; 2) give an indication of the likelihood to collect reasonable data; 3) assist the team in scoping the project; and 4) generally highlighting the potential return on investment. Technically, there is nothing outrageous about a scenario 3 study given sufficient resources and technical skill, but this is largely uncharted ground in the literature for marginal phenomenon where there is already a general lack of research guidelines when adopting social media as research resource [Arastoopour Irgens, 2022].

Overview and General Observations

Returning to the original question, could technical analyses of social media (either via text analytics or social network analysis) complement exercises like the EWCS, or EQLS? Yes, but it is unlikely that there is an “out-of-the-box” solution. This occurs as approaches in the literature to use social media to study marginal phenomena (or rather marginal populations) are either bespoke technical implementations or leverage qualitative research methods. This section has explored three hypothetical research scenarios to try and structure the research landscape for technical social media analysis and introduced a small selection of key methods as content anchors for these scenarios: 1) trend and topic discovery; 2) community detection (potentially as an additional form of recruitment), and 3) localised studies that would seek to explore different cultural perceptions of specific marginal phenomena.

The single biggest challenge in leveraging social media to study marginal phenomena, in isolation or in addition to other research methods, is accessing “good” data. This is not a problem more generally for studying social phenomena (e.g. elections, riots, product reviews, and other macro-economic events) as the general use case of social media is to allow users to discuss events of the “now”, and such events are “significant enough” that millions of content units are generated. Yet for marginal phenomena, this is unlikely to be the case. Thus, significant work is needed to ensure that the social media data sample is representative and not just an echo-chamber of a few select topics that were already obvious and well known. This is probably the most critical part of the research process, as there is a wealth of methods that can be applied once there is a dataset to analyse. This, metaphorically, is like first finding a haystack in a field of haystacks within which to search for a

needle. The success of social media-based studies will hinge on the domain knowledge of the research team to curate a meaningful set of search and filtering terms to allow a representative sample of data to be extracted. Yet, we must also not forget associated ethical concerns in the research design and consider them early in the research design process.

Another important aspect not mentioned, but important nonetheless, is reproducibility. Common practice in social media research to foster reproducibility is to record the identifiers of key content (posts and comments usually). Using this information, provided the user who generated this content hasn't subsequently removed it from the platform, it is possible to rebuild the initial data using the platform's APIs. A key comment here is that the limitations on historic access via identifiers tends to be more prohibitive, i.e. it takes at least as long to reconstruct the dataset, and sometimes significantly longer. Storing and even making publicly available the list of identifiers used also does not violate most platforms' terms and conditions as the source content is not being (re)distributed. In this sense, researchers can somewhat avoid biases arising from drawing a different sample of the platform (as outlined by [González-Bailón et al., 2014]) and maintain a certain amount of reproducibility. Variation in results would instead arise due to any stochastic behaviour in the modelling stage(s), which can be controlled for methodologically, for example using techniques like cross-fold validation, seeding random number generators etc. and ensuring the dataset is suitably large.

The skill set needed for this nature of research is inherently interdisciplinary. Substantial domain knowledge of the marginal phenomena is needed, as too is local and language knowledge. Familiarity with social media platforms (as a user) is also a must, it is often underestimated how important it is to know the quirks, netiquette, and common practices of the platforms that will be used to gather data. At a technical level, there are multiple different skill sets or roles that can be defined, which can be mapped back to Figure 3. Firstly, is the role to extract, clean, transform and preprocess data. This is comparable to the role of a data engineer: someone that establishes a stable data pipeline from social media platform(s) to offline storage.[26] Such a role requires expertise in web APIs, programming (likely python), databases, and a solid foundation data science from the perspective of data preparation. Secondly, is the role similar to that of a data architect: someone that can design flexible data models to store partially processed data. This is important because of the volume of social media data that can be extracted, it's typically quite difficult to extract, prepare and load the data into some model in one continuous process, and thus the idea of data staging areas (where data is partially processed) is useful. This is especially the case if the team lacks experience in analysing social media data, as partially processed data allow for restarting the analysis pipeline at various stages. Key skills for this role are: data modelling, databases and their administration. Thirdly, is a cloud or scalable computing role. Once various data processing pipelines are in place, these will need to be scaled up in terms of computational infrastructure. Often this will leverage public cloud offerings and platforms, thus suggesting a role in this area. Key skills for this role are: a familiarity with scaling up data pipelines, experience in deploying data models and pipelines to cloud platforms, and experience in the parallelisation of computational workflows. Finally, is a set or roles related to the methods of data analysis. This is most likely in the areas of

[26] Note that most social media platforms prohibit raw data being stored offline.

natural language processing and social network analysis (thus these are the key skills along with machine or statistical learning experience). It is not that common to see extensive training in natural language processing or text analytics on master's level programmes, but depending on the level of difficulty of the task (e.g. scenario 1 is much easier than 3) masters level may be more than sufficient, as too is a candidate that previously held a data science role (even without prior experience of natural language process and social network analysis). Here, experience from other projects of a similar nature is key. Whilst it certainly would be beneficial to have qualified researchers, i.e. members of the team with PhDs, this might not be needed for feasibility studies or instrumentations of tasks like scenario 1. However, as the complexity of the research task increases, the need for experienced researchers with the appropriate technical skill set also increases. Note that multiple roles mentioned here could be filled by one individual.

Summary and Concluding Remarks

This working paper has sought to review whether social media could be used to augment or complement existing approaches for studying marginal phenomena (surveys), within the running example of marginal work(ers). Yet this has been just an example scenario to provide context. In this section, a summary of the main points discussed throughout the paper will be used to explicitly highlight how social media could be used to complement existing survey-based research methods currently employed by Eurofound.

The first possible use case for social media discussed how the literature has used social media as a recruitment platform for surveys and other methods (e.g. interviews). This is in line with Eurofound's recent endeavours in the Living, working and COVID eSurvey. Here targeted social media adverts are be used to attract specific target respondents to improve their representation in studies. In the example of marginal work(ers), this could include migrant or marginalised workers as well as those who feel they have been subjected to forms of exploitation. Key to using social media in this manner is to leverage domain-specific knowledge and awareness in the definition of target user profiles, and careful consideration of how this population is curated, and what incentives they have to participate, i.e. targeted ads are not enough. The research has outlined additional layers of control and curation that can be employed to further improve sample representation, where an approach of specific merit is to curate a seed set of respondents via targeted ads that fulfil some set of characteristics, and then snowball sample from this seed set via recommendations leveraging expected homophily within the networks of these individuals. Other aspects of note are additional resources alongside the ads (videos, local champions, stakeholder involvement etc.) to better justify why respondents should participate. This use of social media, however, is not without some risk. The literature reports the possibility of toxic or defamatory comments towards the marginal population and researchers themselves being major considerations in associated ethical risk assessments.

The second main use case for social media to leverage it as a source of data with the goal of informing research on specific marginal phenomena. To position this several hypothetical research scenarios were discussed in the context of potential research designs that could support them using either Twitter or Reddit as sources of data. These scenarios explored trend and topic analysis, community detection, and aspects localisation and differentiation across a pan-European context.

Trend and topic modelling could be used to uncover views of or towards specific marginal phenomena, discover vernacular or language use in its discourse, and identify trends and topics in online discourse. Here, social media could be used to: 1) better position surveys by highlighting specific areas to ask questions within identified sociolinguistic context(s); 2) key topics discussed with respect to the marginal phenomena; 3) illustrate how discourse has changed or evolved over time as different topics or content trends emerge (potentially making survey questions redundant); 4) aspects of tonality (e.g. positive vs. negative) content and views among others. With these use cases, as with all social media studies, it is critical to initialise the data sampling process as effectively as possible, and some initial structure in this regard was provided: leverage domain knowledge, expand this with other sources of lexical detail (blogs, newspaper articles, employment stakeholders etc.), and consider key entry points (significant users, thematically relevant hashtags, subreddits etc.).

Building on macro studies of this type, forms of community detection and analysis can be performed to segment online discourse. This can be based on many aspects, but a simple example could be discussion around pro-migrant vs. anti-migrant workers based on forms of sentiment analysis. Leveraging users and the interactions provides other forms of structures (graphs) that can be analysed. This was the context of scenario 2: discovering marginal communities. Here, using communities as a structure can afford other forms of analysis, such as further identifying key topics, themes, terms, or events within an online (sub)community related to one or more marginal phenomena that could be used to inform new surveys, or survey questions. Similarly, any identified "key" users within a community could be approached or otherwise recruited as study participants, asked to share invitations to participate, or approached to champion specific community participation. Yet here some caution is needed. Users have not given explicit informed consent for their social media content to be processed. Permission is given by the platform to process the data and often captured in the platform's terms and conditions, which few users actually read. Thus, this can create ethical and privacy concerns. On the legal front, there are also very precise terms and conditions on how data can be used and for what kinds of research purposes. This is usually a differentiation between academic research, and research for revenue generation and is often quite expansively and explicitly covered in the API terms and conditions of use. This would need closer consultation prior to any study commencing as not all platforms treat research uses of their data in the same way. However, once appropriately reviewed, the nature of these communities can reveal detail to inform representation statistics and goals in Eurofound surveys. Similarly, when a community is discovered, content-based studies can more precisely target this community to better curate content data samples. As such, example means to compliment surveys with community detection are: 1) discovering "key" community users as recruitment champions, or respondents; 2) refining content based studies (preceding use case) using discovered community structure; 3) approximating relative size(s) between communities as a heuristic for representation targets; 4) engage with "key" communities with methods like focus groups or interviews to help inform new questions, or better understand barriers to participation, or go beyond the limitations of standardised questions. This could also be undertaken asynchronously. Specifically for platforms like Reddit, this could also involve interactions with subreddit moderators.

Finally, the scenario of localisation was discussed. This covers the possibility of generating dimensionalised (where dimensions could include any ethically discernible sociodemographic or sociocultural characteristic of a user) views of marginal phenomena. For example, how do people in different European countries, using different languages view or discuss topics related to marginal work(ers). Key complimentary potential in studies of this nature is in creating drill down potential in social media studies across different user contexts. As discussed, the methods used here (topic modelling, trend analysis and community detection) are all standard use cases for using social media to study social phenomena in general, and likely only need to be refined for studying marginal phenomena, and the majority of effort is expected to reside in data curation. Thus, the technical skill set (discussed in the previous section) and corresponding project resources are not prohibitive. However, this cannot be said for localisation studies. These are vastly more complicated, and potentially large-scale research projects demanding experienced research staff to a robust methodology which may need to be customised and tailored to different categories of feature. For

example, if the key feature of interest is country, each country may need a bespoke instrumentation of the main methodology employed. Whilst localisation studies have significant research potential, it may not be in the interest of Eurofound to develop these methodologies alone, but rather do so in the context of large(r) scale research and innovation collaborations.

To conclude this working paper, it can be stated that social media has the potential to compliment surveys in the study of hard to reach, marginalised (sub)populations, or more generally marginalised phenomena. There are a significant number of studies in the literature to illustrate this. Yet, this may not be in the exact form of question answering that is more commonplace in surveys. Instead, it is likely to assist in participant recruitment, and in the discovery of key defining attributes of specific phenomena manifested as artefacts of content and discourse, or online social community structures. However, not all the methods and techniques discussed here need to be applied to marginal phenomena. Many of the same techniques can be applied to study "small" populations, and in fact not much changes in the approach to do so; depending on how small the population is. The main challenge is to curate a sufficiently sized corpus of social media content aligned to the population or corresponding research question(s). Here, structured social media platforms like Reddit may lend themselves better to such studies if subreddits aligned to the population or characteristics of the population exist. If not, the data curation approach should be the same as that discussed for marginal phenomena. Once a corpus of content has been curated, the methods discussed throughout this paper and the literature can be applied.

Key takeaways of this working paper surrounding the use of social media data to complement survey-based research methods are:

1. This is an interdisciplinary research environment with a strong reliance on domain knowledge.
2. Feasibility studies are critical: we cannot know ex-ante if there is sufficient discourse on a platform to answer research questions pertaining to marginal phenomena.
3. There are many potential biases in social media research that can significantly affect research results. These need to be carefully navigated.
4. Using social media data well requires an experienced team with a wide range of technical competences.
5. There is a wealth of potentially useful data on social media platforms for studying marginal phenomena: the main challenge is to find it.
6. A good mechanism to execute simple feasibility studies is to host competitions and hackathons around specific areas of interest for Eurofound.

References

[Ahmed et al., 2013] Ahmed, N., Jayasinghe, Y., Wark, J. D., Fenner, Y., Moore, E. E., Tabrizi, S. N., Fletcher, A., and Garland, S. M. (2013). Attitudes to chlamydia screening elicited using the social networking site facebook for subject recruitment. *Sexual health*, 10(3):224–228.

[Aiello et al., 2013] Aiello, L. M., Petkos, G., et al. (2013). Sensing trending topics in Twitter. *IEEE Transactions on Multimedia*, 15(6):1268–1282.

[Allan et al., 1998] Allan, J. et al. (1998). Topic detection and tracking pilot study final report.

[Altshuler et al., 2015] Altshuler, A. L., Storey, H. L. G., and Prager, S. W. (2015). Exploring abortion attitudes of us adolescents and young adults using social media. *Contraception*, 91(3):226–233.

[Amir et al., 2016] Amir, S., Wallace, B. C., Lyu, H., Carvalho, P., and Silva, M. J. (2016). Modelling context with user embeddings for sarcasm detection in social media. In Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning, pages 167–177.

[Amon et al., 2016] Amon, K. L., Paxton, K., Klineberg, E., Riley, L., Hawke, C., and Steinbeck, K. (2016). Insights into facebook pages: an early adolescent health research study page targeted at parents. *International journal of adolescent medicine and health*, 28(1):69–77.

[Andrews, 2012] Andrews, C. (2012). Social media recruitment. *Applied Clinical Trials*, 21(11).

[Arastoopour Irgens, 2022] Arastoopour Irgens, G. (2022). Using knowledgeable agents of the digital and data feminism to uncover social identities in the #blackgirlmagic twitter community. *Learning, Media and Technology*, 47(1):79–94.

[Arcia, 2014] Arcia, A. (2014). Facebook advertisements for inexpensive participant recruitment among women in early pregnancy. *Health Education & Behavior*, 41(3):237–241.

[Bastian et al., 2009] Bastian, M., Heymann, S., and Jacomy, M. (2009). Gephi: an open source software for exploring and manipulating networks. In *Proceedings of the international AAAI conference on web and social media*, volume 3, pages 361–362.

[Bauermeister et al., 2012] Bauermeister, J. A., Zimmerman, M. A., Johns, M. M., Glowacki, P., Stoddard, S., and Volz, E. (2012). Innovative recruitment using online networks: lessons learned from an online study of alcohol and other drug use utilizing a web-based, respondent-driven sampling (webrds) strategy. *Journal of studies on alcohol and drugs*, 73(5):834–838.

[Baumgartner et al., 2020] Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., and Blackburn, J. (2020). The pushshift reddit dataset. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 830–839.

[Bender et al., 2014] Bender, K., Begun, S., DePrince, A., Haffejee, B., and Kaufmann, S. (2014). Utilizing technology for longitudinal communication with homeless youth. *Social Work in Health Care*, 53(9):865–882.

[Berinsky et al., 2012] Berinsky, A. J., Huber, G., and Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon.com's mechanical turk. *Political Analysis*, 20:351–368.

[Blei et al., 2003] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

[Blodgett et al., 2020] Blodgett, S. L., Barocas, S., Daumé III, H., and Wallach, H. (2020). Language (technology) is power: A critical survey of "bias" in nlp. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476.

[Bollen et al., 2011] Bollen, J., Gonçalves, B., Ruan, G., and Mao, H. (2011). Happiness is assortative in online social networks. *Artificial life*, 17(3):237–251.

[Bolukbasi et al., 2016] Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, pages 4349–4357.

[Boyd and Crawford, 2011] Boyd, D. and Crawford, K. (2011). Six provocations for big data. In *A decade in internet time: Symposium on the dynamics of the internet and society*.

[Boyd and Crawford, 2012] Boyd, D. and Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information,*

communication & society, 15(5):662–679.

[Buchanan, 2012] Buchanan, E. (2012). Ethical decision-making and internet research. *Association of Internet Researchers*.

[Bunting et al., 2021] Bunting, A. M., Frank, D., Arshonsky, J., Bragg, M. A., Friedman, S. R., and Krawczyk, N. (2021). Socially-supportive norms and mutual aid of people who use opioids: An analysis of reddit during the initial covid-19 pandemic. *Drug and alcohol dependence*, 222:108672.

[Burnap et al., 2017] Burnap, P., Colombo, G., Amery, R., Hodorog, A., and Scourfield, J. (2017). Multi-class machine classification of suicide-related communication on twitter. *Online social networks and media*, 2:32–44.

[Burnap et al., 2015] Burnap, P., Rana, O. F., Avis, N., Williams, M., Housley, W., Edwards, A., Morgan, J., and Sloan, L. (2015). Detecting tension in online communities with computational twitter analysis. *Technological Forecasting and Social Change*, 95:96–108.

[Burnap and Williams, 2015] Burnap, P. and Williams, M. L. (2015). Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & internet*, 7(2):223–242.

[Burnap et al., 2014] Burnap, P., Williams, M. L., Sloan, L., Rana, O., Housley, W., Edwards, A., Knight, V., Procter, R., and Voss, A. (2014). Tweeting the terror: modelling the social media reaction to the woolwich terrorist attack. *Social Network Analysis and Mining*, 4(1):1–14.

[Cacioppo et al., 2009] Cacioppo, J. T., Fowler, J. H., and Christakis, N. A. (2009). Alone in the crowd: the structure and spread of loneliness in a large social network. *Journal of personality and social psychology*, 97(6):977.

[Carlini et al., 2015] Carlini, B. H., Safioti, L., Rue, T. C., and Miles, L. (2015). Using internet to recruit immigrants with language and culture barriers for tobacco and alcohol use screening: a study among brazilians. *Journal of immigrant and minority health*, 17(2):553–560.

[Carr and Hayes, 2015] Carr, C. T. and Hayes, R. A. (2015). Social media: Defining, developing, and divining. *Atlantic journal of communication*, 23(1):46–65.

[Carter-Harris et al., 2016] Carter-Harris, L., Ellis, R. B., Warrick, A., Rawl, S., et al. (2016). Beyond traditional newspaper advertisement: leveraging facebook-targeted advertisement to recruit long-term smokers for research. *Journal of medical Internet research*, 18(6):e5502.

[Cassa et al., 2013] Cassa, C. A., Chunara, R., Mandl, K., and Brownstein, J. S. (2013). Twitter as a sentinel in emergency situations: lessons from the boston marathon explosions. *PLoS currents*, 5.

[Caton et al., 2012] Caton, S., Dukat, C., Grenz, T., Haas, C., Pfadenhauer, M., and Weinhardt, C. (2012). Foundations of trust: Contextualising trust in social clouds. In *2012 Second International Conference on Cloud and Green Computing*, pages 424–429. IEEE.

[Caton et al., 2015] Caton, S., Hall, M., and Weinhardt, C. (2015). How do politicians use facebook? an applied social observatory. *Big Data & Society*, 2(2):2053951715612822.

[Chau and Xu, 2012] Chau, M. and Xu, J. (2012). Business intelligence in blogs: Understanding consumer interactions and communities. *MIS quarterly*, pages 1189–1216.

[Chen et al., 2011] Chen, J. J., Menezes, N. J., Bradley, A. D., and North, T. A. (2011). Opportunities for crowdsourcing research on amazon mechanical turk. *Interfaces*, 5.

[Cheng et al., 2014] Cheng, X., Yan, X., et al. (2014). Btm: Topic modeling over short texts. *IEEE Transactions on Knowledge and Data Engineering*, 26(12):2928–2941.

[Choi et al., 2020] Choi, J., Oh, S., Yoon, J., Lee, J.-M., and Coh, B.-Y. (2020). Identification of time-evolving product opportunities via social media mining. *Technological Forecasting and Social Change*, 156:120045.

[Chu and Snider, 2013] Chu, J. L. and Snider, C. E. (2013). Use of a social networking web site for recruiting canadian youth for medical research. *Journal of Adolescent Health*, 52(6):792–794.

[Cieliebak et al., 2017] Cieliebak, M., Deriu, J. M., Egger, D., and Uzdilli, F. (2017). A twitter corpus and benchmark resources for german sentiment analysis. In *5th International Workshop on Natural Language Processing for Social Media, Boston MA, USA, 11 December 2017*, pages 45–51. Association for Computational Linguistics.

[Coe and Griffin, 2020] Coe, K. and Griffin, R. A. (2020). Marginalized identity invocation online: The case of president donald trump on twitter. *Social media+ society*, 6(1):2056305120913979.

[Comito et al., 2019] Comito, C., Forestiero, A., and Pizzuti, C. (2019). Bursty event detection in twitter streams. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 13(4):1–28.

[Cresci et al., 2018] Cresci, S., Lillo, F., Regoli, D., Tardelli, S., and Tesconi, M. (2018). $ fake: Evidence of spam and bot activity in stock microblogs on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

[Datta and Adar, 2019] Datta, S. and Adar, E. (2019). Extracting inter-community conflicts in reddit. In *Proceedings of the international AAAI conference on Web and Social Media*, volume 13, pages 146–157.

[Davidson et al., 2019] Davidson, T., Bhattacharya, D., and Weber, I. (2019). Racial bias in hate speech and abusive language detection datasets. *arXiv preprint arXiv:1905.12516*.

[de Melo and Figueiredo, 2020] de Melo, T. and Figueiredo, C. M. (2020). A first public dataset from brazilian twitter and news on covid-19 in portuguese. *Data in brief*, 32:106179.

[Deerwester et al., 1990] Deerwester, S., Dumais, S. T., et al. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407.

[Devito et al., 2019] Devito, M. A., Walker, A. M., Birnholtz, J., Ringland, K., Macapagal, K., Kraus, A., Munson, S., Liang, C., and Saksono, H. (2019). Social technologies for digital wellbeing among marginalized communities. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*, pages 449–454.

[Díaz et al., 2018] Díaz, M., Johnson, I., Lazar, A., Piper, A. M., and Gergle, D. (2018). Addressing age-related bias in sentiment analysis. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 412. ACM.

[Dixon et al., 2018] Dixon, L., Li, J., Sorensen, J., Thain, N., and Vasserman, L. (2018). Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73. ACM.

[Dosono and Semaan, 2020] Dosono, B. and Semaan, B. (2020). Decolonizing tactics as collective resilience: Identity work of aapi communities on reddit. *Proceedings of the ACM*

on Human-Computer interaction, 4(CSCW1):1–20.

[Fayyad et al., 1996] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). The kdd process for extracting useful knowledge from volumes of data. *Communications of the ACM*, 39(11):27–34.

[Ferrara et al., 2016] Ferrara, E., Varol, O., Davis, C., Menczer, F., and Flammini, A. (2016). The rise of social bots. *Communications of the ACM*, 59(7):96–104.

[Fox and Ralston, 2016] Fox, J. and Ralston, R. (2016). Queer identity online: Informal learning and teaching experiences of lgbtq individuals on social media. *Computers in Human Behavior*, 65:635–642.

[Freelon, 2018] Freelon, D. (2018). Computational research in the post-api age. *Political Communication*, 35(4):665–668.

[Friess and Eilders, 2015] Friess, D. and Eilders, C. (2015). A systematic review of online deliberation research. *Policy & Internet*, 7:319–339.

[Gadiraju et al., 2017] Gadiraju, U., Yang, J., and Bozzon, A. (2017). Clarity is a worthwhile quality. pages 5–14.

[Gaffney and Matias, 2018] Gaffney, D. and Matias, J. N. (2018). Caveat emptor, computational social science: Large-scale missing data in a widely-published reddit corpus. *PloS one*, 13(7):e0200162.

[Geiger et al., 2011] Geiger, D., Seedorf, S., Nickerson, R., and Schader, M. (2011). Managing the crowd : Towards a taxonomy of crowdsourcing processes. pages 1–11.

[Gerber and Krzywdzinski, 2019] Gerber, C. and Krzywdzinski, M. (2019). Brave new digital work? new forms of performance control in crowdwork.

[Gonzales, 2017] Gonzales, A. L. (2017). Disadvantaged minorities' use of the internet to expand their social networks. *Communication Research*, 44(4):467–486.

[Gonzalez et al., 2021] Gonzalez, J. A., Hurtado, L.-F., and Pla, F. (2021). Twilbert: Pretrained deep bidirectional transformers for spanish twitter. *Neurocomputing*, 426:58–69.

[González-Bailón et al., 2014] González-Bailón, S., Wang, N., Rivero, A., Borge-Holthoefer, J., and Moreno, Y. (2014). Assessing the bias in samples of large online networks. *Social Networks*, 38:16–27.

[Gorwa and Guilbeault, 2020] Gorwa, R. and Guilbeault, D. (2020). Unpacking the social media bot: A typology to guide research and policy. *Policy & Internet*, 12(2):225–248.

[Gupta et al., 2013] Gupta, A., Lamba, H., and Kumaraguru, P. (2013). $1.00 per rt #bostonmarathon #prayforboston: Analyzing fake content on twitter. In *2013 APWG eCrime researchers summit*, pages 1–12. IEEE.

[Hada et al., 2021] Hada, R., Sudhir, S., Mishra, P., Yannakoudakis, H., Mohammad, S., and Shutova, E. (2021). Ruddit: Norms of offensiveness for english reddit comments. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2700–2717.

[Hagberg et al., 2008] Hagberg, A., Swart, P., and S Chult, D. (2008). Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States).

[Hall and Caton, 2017] Hall, M. and Caton, S. (2017). Am i who i say i am? unobtrusive self-representation and personality recognition on facebook. *PloS one*, 12(9):e0184417.

[Hall et al., 2018] Hall, M., Mazarakis, A., Chorley, M., and Caton, S. (2018). Editorial of the special issue on following user pathways: Key contributions and future directions in crossplatform social media research. *International Journal of Human–Computer Interaction*, 34(10):895–912.

[Hart et al., 2013] Hart, R. P., Childers, J. P., and Lind, C. J. (2013). *Political tone: How leaders talk and why*. University of Chicago Press.

[Hasyim, 2019] Hasyim, M. (2019). Linguistic functions of emoji in social media communication. Opcion, 35.

[Hernandez-Romieu et al., 2014] Hernandez-Romieu, A. C., Sullivan, P. S., Sanchez, T. H., Kelley, C. F., Peterson, J. L., Del Rio, C., Salazar, L. F., Frew, P. M., and Rosenberg, E. S. (2014). The comparability of men who have sex with men recruited from venue-time-space sampling and facebook: a cohort study. *JMIR research protocols*, 3(3):e3342.

[Hughes et al., 2012] Hughes, D. J., Rowe, M., Batey, M., and Lee, A. (2012). A tale of two sites: Twitter vs. facebook and the personality predictors of social media usage. *Computers in human behavior*, 28(2):561–569.

[Hurtado et al., 2019] Hurtado, S., Ray, P., and Marculescu, R. (2019). Bot detection in reddit political discussion. In *Proceedings of the Fourth International Workshop on Social Sensing*, pages 30–35.

[Ifrim et al., 2014] Ifrim, G., Shi, B., and Brigadir, I. (2014). Event detection in twitter using aggressive filtering and hierarchical tweet clustering. In *Second Workshop on Social News on the Web (SNOW), Seoul, Korea, 8 April 2014*. ACM.

[Ipeirotis et al., 2010] Ipeirotis, P. G., Provost, F., and Wang, J. (2010). Quality management on amazon mechanical turk. pages 64–67. ACM Press.

[Isaak and Hanna, 2018] Isaak, J. and Hanna, M. J. (2018). User data privacy: Facebook, cambridge analytica, and privacy protection. *Computer*, 51(8):56–59.

[Jeong et al., 2019] Jeong, B., Yoon, J., and Lee, J.-M. (2019). Social media mining for product planning: A product opportunity mining approach based on topic modeling and sentiment analysis. *International Journal of Information Management*, 48:280–290.

[Kahneman et al., 1999] Kahneman, D. et al. (1999). Objective happiness. *Well-being: The foundations of hedonic psychology*, 3(25):1–23.

[Kazai et al., 2011] Kazai, G., Kamps, J., and Milic-Frayling, N. (2011). Worker types and personality traits in crowdsourcing relevance labels. page 1941.

[Kazai et al., 2012] Kazai, G., Kamps, J., and Milic-Frayling, N. (2012). The face of quality in crowdsourcing relevance labels: Demographics, personality and labeling accuracy. pages 2583–2586.

[Kilroy et al., 2020] Kilroy, D., Caton, S., and Healy, G. (2020). Finding short lived events on social media. In *28th AIAI Irish Conference on Artificial Intelligence and Cognitive Science (AICS)*.

[Kilroy et al., 2022] Kilroy, D., Healy, G., and Caton, S. (2022). Using machine learning to improve lead times in the identification of emerging customer needs. IEEE Access.

[Kim et al., 2019] Kim, W., Ko, T., Rhiu, I., and Yun, M. H. (2019). Mining affective experience for a kansei design study on a recliner. *Applied ergonomics*, 74:145–153.

[Kiritchenko and Mohammad, 2018] Kiritchenko, S. and Mohammad, S. (2018). Examining gender and race bias in two hundred sentiment analysis systems. pages 43–53.

[Kleinberg, 2003] Kleinberg, J. (2003). Bursty and hierarchical structure in streams. *Data mining and knowledge discovery*, 7(4):373–397.

[Ko et al., 2020] Ko, T., Rhiu, I., Yun, M. H., and Cho, S. (2020). A novel framework for identifying customers' unmet needs on online social media using context tree. *Applied Sciences*, 10(23):8473.

[Kralj Novak et al., 2015] Kralj Novak, P., Smailović, J., Sluban, B., and Mozetič, I. (2015). Sentiment of emojis. PloS one, 10(12):e0144296.

[Kramer et al., 2014] Kramer, J., Rubin, A., Coster, W., Helmuth, E., Hermos, J., Rosenbloom, D., Moed, R., Dooley, M., Kao, Y.-C., Liljenquist, K., et al. (2014). Strategies to address participant misrepresentation for eligibility in web-based research. *International journal of methods in psychiatric research*, 23(1):120–129.

[Krause et al., 2019] Krause, M., Afzali, F. M., Caton, S., and Hall, M. (2019). Is quality control pointless? In *Proceedings of the 52nd Hawaii International Conference on System Sciences*.

[Krishnamoorthy, 2015] Krishnamoorthy, S. (2015). Linguistic features for review helpfulness prediction. *Expert Systems with Applications*, 42(7):3751–3759.

[Krosnick, 1999] Krosnick, J. A. (1999). Survey research. *Annual review of psychology*, 50(1):537–567.

[Law et al., 2017] Law, E., Gajos, K. Z., Wiggins, A., Gray, M. L., and Williams, A. (2017). Crowdsourcing as a tool for research: Implications of uncertainty. pages 1544–1561. Association for Computing Machinery.

[Lee and Seung, 1999] Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791.

[Lenhart et al., 2015] Lenhart, A., Duggan, M., Perrin, A., Stepler, R., Rainie, H., Parker, K., et al. (2015). Teens, social media & technology overview 2015.

[Li et al., 2012] Li, R., Lei, K. H., Khadiwala, R., and Chang, K. C.-C. (2012). Tedas: A twitter-based event detection and analysis system. In *2012 IEEE 28th International Conference on Data Engineering*, pages 1273–1276. IEEE.

[Lim et al., 2013] Lim, E.-P., Chen, H., and Chen, G. (2013). Business intelligence and analytics: Research directions. *ACM Transactions on Management Information Systems (TMIS)*, 3(4):1–10.

[Liu et al., 2014] Liu, Y., Kliman-Silver, C., and Mislove, A. (2014). The tweets they are achangin': Evolution of twitter users and behavior. In *Eighth International AAAI Conference on Weblogs and Social Media*.

[Lohse et al., 2013] Lohse, B., Wamboldt, P., et al. (2013). Purposive facebook recruitment endows cost-effective nutrition education program evaluation. *JMIR research protocols*, 2(2):e2713.

[Lord et al., 2011] Lord, S., Brevard, J., and Budman, S. (2011). Connecting to young adults: an online social network survey of beliefs and attitudes associated with prescription opioid misuse among college students. *Substance use & misuse*, 46(1):66–76.

[Lundmark and LeDrew, 2019] Lundmark, E. and LeDrew, S. (2019). Unorganized atheism and the secular movement: Reddit as a site for studying 'lived atheism'. *Social Compass*, 66(1):112–129.

[Lutkenhaus et al., 2019] Lutkenhaus, R. O., Jansz, J., and Bouman, M. P. (2019). Mapping the dutch vaccination debate on twitter: Identifying communities, narratives, and interactions. *Vaccine: X*, 1:100019.

[Madahali and Hall, 2020] Madahali, L. and Hall, M. (2020). Application of the benford's law to social bots and information operations activities. In *2020 International Conference on Cyber Situational Awareness, Data Analytics and Assessment (CyberSA)*, pages 1–8. IEEE.

[Markey, 1926] Markey, J. F. (1926). A redefinition of social phenomena: giving a basis for comparative sociology. *American Journal of Sociology*, 31(6):733–743.

[Mazoyer et al., 2020] Mazoyer, B., Cagé, J., Hervé, N., and Hudelot, C. (2020). A french corpus for event detection on twitter. In *Proceedings of the 12th language resources and evaluation conference*, pages 6220–6227.

[McInroy, 2016] McInroy, L. B. (2016). Pitfalls, potentials, and ethics of online survey research: Lgbtq and other marginalized and hard-to-access youths. *Social Work Research*, 40(2):83–94.

[Metaxas and Mustafaraj, 2012] Metaxas, P. T. and Mustafaraj, E. (2012). Social media and the elections. *Science*, 338(6106):472–473.

[Mitchell et al., 2016] Mitchell, J., Lee, J.-Y., and Stephenson, R. (2016). How best to obtain valid, verifiable data online from male couples? lessons learned from an ehealth hiv prevention intervention for hiv-negative male couples. *JMIR public health and surveillance*, 2(2):e6392.

[Morgan et al., 2013] Morgan, A. J., Jorm, A. F., and Mackinnon, A. J. (2013). Internet-based recruitment to a depression prevention intervention: lessons from the mood memos study. *Journal of medical Internet research*, 15(2):e2262.

[Muresan et al., 2016] Muresan, S., Gonzalez-Ibanez, R., Ghosh, D., and Wacholder, N. (2016). Identification of nonliteral language in social media: A case study on sarcasm. Journal of the Association for Information Science and Technology, 67(11):2725–2737.

[Murthy et al., 2016] Murthy, D., Gross, A., and Pensavalle, A. (2016). Urban social media demographics: An exploration of twitter use in major american cities. *Journal of ComputerMediated Communication*, 21(1):33–49.

[Myers et al., 2017] Myers, Z. R., Swearer, S. M., Martin, M. J., and Palacios, R. (2017). Cyberbullying and traditional bullying: the experiences of poly-victimization among diverse youth. *International Journal of Technoethics (IJT)*, 8(2):42–60.

[Nartey, 2021] Nartey, M. (2021). Centering marginalized voices: a discourse analytic study of the black lives matter movement on twitter. *Critical Discourse Studies*, pages 1–16.

[Nemer, 2016] Nemer, D. (2016). Online favela: The use of social media by the marginalized in brazil. *Information technology for development*, 22(3):364–379.

[Niemeyer et al., 2018] Niemeyer, C., Teubner, T., Hall, M., and Weinhardt, C. (2018). The impact of dynamic feedback and personal budgets on arousal and funding behaviour in participatory budgeting. *Group Decision and Negotiation*, 27:611–636.

[Nosek et al., 2002] Nosek, B. A., Banaji, M. R., and Greenwald, A. G. (2002). E-research: Ethics, security, design, and control in psychological research on the internet. *Journal of Social Issues*, 58(1):161–176.

[Oleson et al., 2011] Oleson, D., Sorokin, A., Laughlin, G., Hester, V., Le, J., and Biewald, L. (2011). Programmatic gold: Targeted and scalable quality assurance in crowdsourcing. In *Workshops at the Twenty-Fifth AAAI conference on artificial intelligence*.

[Oppenlaender et al., 2020] Oppenlaender, J., Milland, K., Visuri, A., Ipeirotis, P., and Hosio, S. (2020). Creativity on paid crowdsourcing platforms. Association for Computing Machinery.

[O'Donnell et al., 2016] O'Donnell, P., Tierney, E., O'Carroll, A., Nurse, D., and MacFarlane, A. (2016). Exploring levers and barriers to accessing primary care for marginalised groups and identifying their priorities for primary care provision: a participatory learning and action research study. *International journal for equity in health*, 15(1):1–16.

[Patton et al., 2017] Patton, D. U., Lane, J., Leonard, P., Macbeth, J., and Smith Lee, J. R. (2017). Gang violence on the digital street: Case study of a south side chicago gang member's twitter communication. *new media & society*, 19(7):1000–1018.

[Pedersen et al., 2015] Pedersen, E. R., Helmuth, E. D., Marshall, G. N., Schell, T. L., PunKay, M., and Kurz, J. (2015). Using facebook to recruit young adult veterans: online mental health research. *JMIR research protocols*, 4(2):e3996.

[Pennebaker et al., 2001] Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001). Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

[Petrović et al., 2010] Petrović, S., Osborne, M., and Lavrenko, V. (2010). Streaming first story detection with application to twitter. In *Human language technologies: The 2010 annual conference of the north american chapter of the association for computational linguistics*, pages 181–189. Association for Computational Linguistics.

[Pew Research Centre, 2017] Pew Research Centre (2017). Social media fact sheet. *Pew Research Center: Washington, DC, USA*.

[Podsakoff et al., 2003] Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., and Podsakoff, N. P. (2003). Common method biases in behavioral research: a critical review of the literature and recommended remedies. *Journal of applied psychology*, 88(5):879.

[Powell et al., 2020] Powell, A., Scott, A. J., and Henry, N. (2020). Digital harassment and abuse: Experiences of sexuality and gender minority adults. *European Journal of Criminology*, 17(2):199–223.

[Prpić et al., 2015] Prpić, J., Taeihagh, A., and Melton, J. (2015). The fundamentals of policy crowdsourcing. *Policy and Internet*, 7:340–361.

[Qiu et al., 2012] Qiu, L., Lin, H., Leung, A. K., and Tov, W. (2012). Putting their best foot forward: Emotional disclosure on facebook. *Cyberpsychology, Behavior, and Social Networking*, 15(10):569–572.

[Ramo and Prochaska, 2012] Ramo, D. E. and Prochaska, J. J. (2012). Broad reach and targeted recruitment using facebook for an online survey of young adult substance use. *Journal of medical Internet research*, 14(1):e1878.

[Ratkiewicz et al., 2011] Ratkiewicz, J., Conover, M., Meiss, M., Gonçalves, B., Flammini, A., and Menczer, F. (2011). Detecting and tracking political abuse in social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 5, pages 297–304.

[Reyes et al., 2012] Reyes, A., Rosso, P., and Buscaldi, D. (2012). From humor recognition to irony detection: The figurative language of social media. Data & Knowledge Engineering, 74:1–12.

[Reyes et al., 2013] Reyes, A., Rosso, P., and Veale, T. (2013). A multidimensional approach for detecting irony in twitter. Language resources and evaluation, 47(1):239–268.

[Romo-Fernández et al., 2013] Romo-Fernández, L. M., Guerrero-Bote, V. P., and MoyaAnegón, F. (2013). Co-word based thematic analysis of renewable energy (1990–2010). *Scientometrics*, 97(3):743–765.

[Rost et al., 2013] Rost, M., Barkhuus, L., Cramer, H., and Brown, B. (2013). Representation and communication: Challenges in interpreting large social media datasets. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 357–362.

[Russomanno et al., 2019] Russomanno, J., Patterson, J. G., Tree, J. M. J., et al. (2019). Social media recruitment of marginalized, hard-to-reach populations: development of recruitment and monitoring guidelines. *JMIR Public Health and Surveillance*, 5(4):e14886.

[Saeed et al., 2019] Saeed, Z., Abbasi, R. A., Razzak, I., Maqbool, O., Sadaf, A., et al. (2019). Enhanced heartbeat graph for emerging event detection on Twitter using time series networks. *Expert Systems with Applications*, 136:115–132.

[Saunders and Byrne, 2020] Saunders, D. and Byrne, B. (2020). Reducing gender bias in neural machine translation as a domain adaptation problem. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7724–7736.

[Saunders et al., 2020] Saunders, D., Sallis, R., and Byrne, B. (2020). Neural machine translation doesn't translate gender coreference right unless you make it. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 35–43.

[Schifanella et al., 2016] Schifanella, R., de Juan, P., Tetreault, J., and Cao, L. (2016). Detecting sarcasm in multimodal social platforms. In Proceedings of the 24th ACM international conference on Multimedia, pages 1136–1145.

[Schiffer and Schatz, 2008] Schiffer, K. and Schatz, E. (2008). Marginalisation, social inclusion and health. experiences based on the work of correlation–european network social inclusion & health. *Amsterdam: Foundation Regenboog AMOC*.

[Schwarz et al., 1985] Schwarz, N., Hippler, H.-J., Deutsch, B., and Strack, F. (1985). Response scales: Effects of category range on reported behavior and comparative judgments. *Public Opinion Quarterly*, 49(3):388–395.

[Sheng et al., 2008] Sheng, V. S., Provost, F., and Ipeirotis, P. G. (2008). Get another label? improving data quality and data mining using multiple, noisy labelers. pages 614–622.

[Siddiqui, 2014] Siddiqui, F. R. (2014). Annotated bibliography on participatory consultations to help aid the inclusion of marginalized perspectives in setting policy agendas. *International Journal for Equity in Health*, 13(1):1–16.

[Singh and Tucker, 2015] Singh, A. S. and Tucker, C. S. (2015). Investigating the heterogeneity of product feature preferences mined using online product data streams. In *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, volume 57083, page V02BT03A020. American Society of Mechanical Engineers.

[Smith et al., 2015] Smith, G., Richards, R. C., and Gastil, J. (2015). The potential of participedia as a crowdsourcing tool for comparative analysis of democratic innovations. *Policy and Internet*, 7:243–262.

[Smith et al., 1996] Smith, H. J., Milberg, S. J., and Burke, S. J. (1996). Information privacy: Measuring individuals' concerns about organizational practices. *MIS quarterly*, pages 167–196.

[Subramanian et al., 2019] Subramanian, J., Sridharan, V., Shu, K., and Liu, H. (2019). Exploiting emojis for sarcasm detection. In International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation, pages 70–80. Springer.

[Trinquart and Galea, 2015] Trinquart, L. and Galea, S. (2015). Mapping epidemiology's past to inform its future: metaknowledge analysis of epidemiologic topics in leading journals, 1974–2013. *American journal of epidemiology*, 182(2):93–104.

[Tuarob and Tucker, 2015] Tuarob, S. and Tucker, C. S. (2015). Quantifying product favorability and extracting notable product features using large scale social media data. *Journal of Computing and Information Science in Engineering*, 15(3).

[Uchitpe et al., 2016] Uchitpe, M., Uddin, S., and Lynn, C. (2016). Predicting the future of project management research. *Procedia-Social and Behavioral Sciences*, 226:27–34.

[Webster, 2020] Webster, K. (2020). *A Textual Analysis of Online Asexual Representation and Visibility on Reddit*. PhD thesis, Syracuse University.

[Weeden et al., 2013] Weeden, S., Cooke, B., and McVey, M. (2013). Underage children and social networking. *Journal of research on technology in education*, 45(3):249–262.

[Xie et al., 2016] Xie, W. et al. (2016). Topicsketch: Real-time bursty topic detection from twitter. *IEEE Transactions on Knowledge and Data Engineering*, 28(8):2216–2229.

[Yancey et al., 2006] Yancey, A. K., Ortega, A. N., and Kumanyika, S. K. (2006). Effective recruitment and retention of minority research participants. *Annu. Rev. Public Health*, 27:1–28.

[Yang et al., 2016] Yang, J., Redi, J., Demartini, G., and Bozzon, A. (2016). Modeling task complexity in crowdsourcing. *The Fourth AAAI Conference on Human Computation and Crowdsourcing*, pages 249–258.

[Zhang and Qu, 2015] Zhang, Y. and Qu, Z. (2015). A novel method for online bursty event detection on Twitter. In *2015 6th IEEE International Conference on Software Engineering and Service Science (ICSESS)*, pages 284–288. IEEE.

[Zhiltsova et al., 2019] Zhiltsova, A., Caton, S., and Mulway, C. (2019). Mitigation of unintended biases against non-native english texts in sentiment analysis. In *AICS*, pages 317–328.

[Zhou et al., 2020] Zhou, F., Ayoub, J., Xu, Q., and Jessie Yang, X. (2020). A machine learning approach to customer needs analysis for product ecosystems. *Journal of Mechanical Design*, 142(1).

[Zimmer, 2020] Zimmer, M. (2020). "but the data is already public": on the ethics of research in facebook. In *The Ethics of Information Technologies*, pages 229–241. Routledge.

[Zwitter, 2014] Zwitter, A.(2014).Bigdataethics. *Big Data & Society*, 1(2):2053951714559253.

WPEF21052

The European Foundation for the Improvement of Living and Working Conditions (Eurofound) is a tripartite European Union Agency established in 1975. Its role is to provide knowledge in the area of social, employment and work-related policies according to Regulation (EU) 2019/127.