Complementing studies on vulnerable youths with reddit data

Mauri, Andrea; Psyllidis, Achilleas; Bozzon, Alessandro; Lee, Ju Sung; Pridmore, Jason; Van Zoonen, Liesbet; Giest, Sarah

DOI
10.1145/3464385.3464703

Publication date
2021

Document Version
Final published version

Published in
CHItaly 2021 - Frontiers of HCI

Complementing Studies on Vulnerable Youths with Reddit Data

Andrea Mauri
Delft University of Technology
Delft, The Netherlands
a.mauri@tudelft.nl

Achilleas Psyllidis
Delft University of Technology
Delft, The Netherlands
a.psyllidis@tudelft.nl

Alessandro Bozzon
Delft University of Technology
Delft, The Netherlands
a.bozzon@tudelft.nl

Ju-Sung Lee
Erasmus University
Rotterdam, The Netherlands
lee@eshcc.eur.nl

Jason Pridmore
Erasmus University
Rotterdam, The Netherlands
pridmore@eshcc.eur.nl

Liesbet van Zoonen
Erasmus University
Rotterdam, The Netherlands
vanzoonen@essb.eur.nl

Sarah Giest
Leiden University
Leiden, The Netherlands
s.n.giest@fgga.leidenuniv.nl

ABSTRACT

Social web data increasingly complement studies of various social phenomena, especially when the availability of traditional data is limited. One such case is that of vulnerable young populations that are disengaged from employment, education, or training; usually referred to as NEETs. This paper explores the extent to which social media data and discussion websites could complement conventional sources in the study of NEETs. We focus on user-generated content posted to the dedicated r/NEET subreddit, which gathers subscribers who self-identify as NEETs. We develop and implement a data processing pipeline for the analysis of the behavioral patterns and main concerns of this social group. Our analysis of Reddit data reaches similar conclusions to official reports from governmental institutions in Europe. The paper also provides insights into health-related issues and latent interests of NEETs, not recorded in official reports and related literature.

CCS CONCEPTS

• **Human-centered computing** → **Ethnographic studies**; • **Information systems** → **Social networks**.

KEYWORDS

NEETs, Reddit, social media analysis, text mining, topic analysis, vulnerable youth, NLP

ACM Reference Format:
Andrea Mauri, Achilleas Psyllidis, Alessandro Bozzon, Ju-Sung Lee, Jason Pridmore, Liesbet van Zoonen, and Sarah Giest. 2021. Complementing Studies on Vulnerable Youths with Reddit Data. In *CHItaly 2021: 14th Biannual Conference of the Italian SIGCHI Chapter (CHItaly '21), July 11–13, 2021, Bolzano, Italy*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3464385.3464703

CHItaly '21, July 11–13, 2021, Bolzano, Italy

ACM ISBN 978-1-4503-8977-8/21/06.
https://doi.org/10.1145/3464385.3464703

1 INTRODUCTION

In the domain of social sciences, surveys and interviews continue to be the primary sources of information when it comes to the study of social phenomena. Often supported by demographic records on population-related attributes (e.g., income, education, marital status, and employment), they are integral in developing evidence-based policies to tackle pressing societal issues. Despite their high level of reliability, these data sources also suffer from several drawbacks: infrequent updates (e.g., on a yearly or, on some occasions, a decennial basis); high deployment costs in terms of both human and material resources, and sampling bias, to name a few. The recent proliferation of social media, blogs, and discussion websites has opened up new avenues in social studies. A wealth of existing literature in both Web and HCI communities has used various social web data as alternative sources to study personality issues [34], socio-economic behavior [1, 23], health [9], and indications of psychological problems such as depression [2, 11, 36], among other topics. Compared to conventional sources, social web data are characterized by frequent updates and high granularity, while they can be extracted at low or no costs. However, issues of representativeness, veracity, and bias continue to challenge the fidelity of research that uses such data [17].

Especially with regard to representativeness, it has been shown that younger populations are the predominant contributors of social media data [17]. The content posted in the form of text, pictures, and videos reflects their activity and general attitude towards everyday life at a fine-grained level that is difficult to be recorded in administrative data or in the infrequent dedicated surveys. The lack of this extent of granularity in traditional data sources, to date, has been the main reason why the behavior and characteristics of young people, especially those in vulnerable positions (e.g., unemployed, poorly educated), has been a challenging topic in sociology and related fields [5, 25]. Drawing on this, we assert—and consequently show in this paper—that by observing the interactions of young vulnerable people on social media and analyzing the data

their produce we could complement and substantially improve their understanding. Hence, this paper investigates the extent to which new forms of social data could complement conventional sources in the study of vulnerable young groups, focusing on the analysis of content posted to related groups on Reddit. Reddit is an online media aggregation and discussion platform, where users share personal stories and offer commentary and opinions on groups that reflect topics they are interested in. A large number of these groups gathers users around sensitive topics, such as depression, alcoholism, loneliness, and others. We focus on a specific group of vulnerable youths, namely NEETs (not in education, employment, or training) [24]. The study of NEETs with conventional data sources has been particularly challenging because the majority of people belonging to this group are often difficult to find in, and often being excluded, from official records [5, 25]. Currently, the use of social web data in the study of NEETs' behavior and activity patterns remains largely untapped.

We extract *public* posts and comments from the dedicated `r/NEET` subreddit, which gathers about 4,800 subscribers[1], who self-identify as NEETs. Driven by existing literature in sociology and policy sciences around this topic, we design and implement a data processing pipeline for the analysis of the behavioral patterns and main concerns (e.g. isolation from society, depression, mental health issues, etc.) of NEETs. We compare our findings to related reports from the European Union and OECD[2] to identify the ways they corroborate and or diverge from official data and statistics. Our analysis reaches similar conclusions to those reported in the official records, with some differences likely due to the focus of the Reddit population. As this kind of research raises considerable ethical issues, we engaged in various reflexive and analytic practices to make sure the purpose and process of the study did not violate individual privacy nor compromised the integrity of these vulnerable youth as a social group. We expand on this in the last section of the paper. The added value of our study, concerning the NEET issue, comes from uncovering prominent latent topics emerging from the discussions (e.g., health issues and general interests of the group), which are usually untapped in the existing literature. From a HCI perspective, this work contributes to the growing literature devoted to understanding people's behavior and issues through social media and to bridge the gap between HCI and policy making [35]. Our findings suggest that Reddit is a valuable source of complementary information about NEETs that can be used as a starting point for the design of more targeted and detailed studies.

2 RELATED WORK

Literature about NEETs. The NEET term appeared for the first time in the late-80s in the UK [14] as an alternative way to classify unemployed people under 18 years old of age, as youth unemployment was officially under-recognized. Since then, this condition has been extensively studied by both research and government institutions. Robson [33] studied the main factors that could predict a youth as being a NEET and found they are were characterized by low income, education, and social capital; moreover, females were more at risk than men. A report by Eurofound [12], released in 2011, investigates the risk factors and the economic and social consequences of NEETs' disengagement from education and the job market. Historically women have been more at risk of unemployment than men, but at the time of this report, in Europe, the rates had started to converge. Risk factors include: having a disability, having an immigrant background, low education, living in a remote area, living in a household with low income, having parents with low education, and/or who are divorced. NEETs showed distrust in institutions, low participation and interest in politics, low presence in bonding organizations (e.g., religious, trade unions, professional organizations) and bridging organizations (e.g, welfare organizations, local community, human rights), and exhibit mental health issues at a higher rate than non-NEETs. This is also confirmed by more recent national studies performed in Italy [32], UK [15, 41], and Romania [7]. Also, several additional studies [5, 14, 25] have shown how NEETs are heterogeneous and comprise many types of people, ranging from individuals disengaged from the labor market due to illness or difficult family background, to ones that chose to become NEET to pursue personal goals or to wait for a specific job. It is clear that these different groups have different needs and can be aided through different policies.

Social Media and Vulnerable Groups. A number of studies have explored the use of social media by various vulnerable groups, including NEETs. For instance, Twitter textual and network features have been used to detect depression symptoms [10], predict recovery and relapse from Alcohol Use Disorder [40], and detect extremism and radicalization [13]. Also, social media usage reveals socio-economic characteristics: correlation have been found between employment rate and correct use of the language, diverse mobility patterns, and diurnal rhythm [1, 23]. Similarly, works in HCI domain studied how data from social media can be used to better design online therapy for mental health problems [22], also highlighting how the peer-to-peer interaction helps to emerge deeper insights of people's issues [30]. They also studied how social media data can complement studies about eating disorder [8], depression [2, 36], personality traits [16] and general people characteristics [19]. Closer to the NEET case, recently Facebook data has been used to get a better understanding of Italian NEETs by tracking their digital behavior and comparing it with employed people [37]. Other works highlighted the importance of online communities for the NEETs, as they help them to stay connected while struggling with problems related to mental health and social anxiety [21, 38]. Reddit, in particular, has served as a venue and resource for those affected by mental health issues. Studies have examined how Reddit forums are used by and aid such individuals, in the context suicide prevention [20], automatic diagnosis of mental health issues [3] and emotional support [31] .

3 METHOD: REDDIT ANALYSIS

Drawing on the literature summarized in the Section 2, we focus on the the commonly studied factors that contribute to the NEET condition: gender, age, education, health, and living context. In order to infer these personal characteristics from Reddit content, we develop a pattern-matching approach that extracts direct mentions of the aforementioned attributes in the user-generated text. A pattern is a linguistic triple $\langle subject, predicate, object \rangle$. Since we are looking

[1] as of 15th January 2019
[2] Organization for Economic Co-operation and Development

for sentences referring to information related to the person, the *subject* is always the first singular person "I". The sub-patterns for *predicate* and *object*, then, reflect the attributes we want to extract. The *predicate* is a verb (plus eventual preposition), while the *object* is a sequence of nouns, adjectives and/or adverbs. The object can be either a single value coded in the pattern or a list of allowed values or Part-of-Speech tags.

Table 1 shows some examples. The first one refers to the gender attribute and extracts mentions about the posting redditor's being a male; it matches all the sentences starting with *I am*, followed by an article ($\langle DT \rangle$) and a list of terms that may identify a male (e.g., man, guy, male, etc.). The second refers to the age attribute and extract mentions about being of a particular age; it matches sentences starting with *I am* followed by a digit ($\langle CD \rangle$).

In addition, we manually created a subreddits mapping for the attributes `gender` and `health`. That is, for assigning the gender, we also considered subreddits where the author of a post is more likely to belong to a specific gender (e.g., `r/GirlGamers`, `r/trans`, `r/malefashionadvice`, etc...). For the health attribute, we considered subreddits where people post to ask for support and seek help for a specific health issue (e.g., `r/depression`, `r/ADHD`, `r/cancer`, etc...).

While this method may not be completely error-proof, upon examining a sample of user comments in the subreddits captured in our data – which we detail later, we infer that most users "belong" to the community represented by the subreddit. That is, the users in our study are largely not trolls or bots or advertisers.

We extract three basic and evaluative measurements from our pattern-matching approach. First, we consider the *precision* of the pattern matching approach, by computing the ratio between the number of mentions that correctly refer to the attribute and the total number of mentions extracted. This is done by looking manually at the meaning of the mention in the text it was extracted from. Secondly, to evaluate the *quantity* of information shared by redditors, we look at the ratio between the number of users that disclose a specific attribute to all users. We call this metric *Disclosure*. Finally, to measure the *reliability* of the information, we compute one minus the ratio between the number of users mentioning contradicting pieces of information and the total number of users sharing it. We call this metric *Coherence*. Given the limitations in investigating NEETs through sociodemographic attributes alone – as suggested by previous studies [5] - and in order to form a more complete depiction of NEETs, we try to acquire a better understanding of the different facets of the NEET group by utilizing a bottom-up approach, where we analyze hidden behavioral patterns that emerge from the posts and comments. We perform a topic modeling analysis using Latent Dirichlet Allocation (LDA) [6] and, based on this, we further perform clustering analysis on the users, using a spectral clustering algorithm [27], to distinguish the different groups of users.

3.1 Data Collection and pre-processing

Before describing the collection and analysis processes, we define the terms we will use thorough the paper: a *post* is the main element of discussion that a *user* - or *redditor* - can create in a subreddit; while a *comment* is an answer to a *post* or to another *comment*. We will use *submission* as an umbrella term to refer to both. We crawled all the public posts created on the `r/NEET` subreddit using the Python Reddit API wrapper[3]. For each user who submitted at least one post in `r/NEET`, we retrieved their most recent submissions, both *comments* and *posts*, sent in other subreddits. We retrieved a total of 264,308 submissions: 31,674 posts and 232,634 comments from 734 users. The considered time ranges from May 28, 2011, to December 5, 2018[4]. Before the text analysis step, post and comments are preprocessed; special characters are removed, abbreviations, acronyms, and slang are expanded (e.g., *kinda* becomes *kind of*, *gonna* becomes *going to*, etc.). The text is then POS (part-of-speech) tagged and tokenized (i.e. sentences are separated into their constituent words).

3.2 Gender

For identifying the gender of a user, we collate their mentions of being male, female, and trans (e.g, "I am a man", "I am a trans girl..."). We also complement this identification with the information about the subreddit on which the user submitted posts. The gender is then defined by selecting the one that was mentioned the most. We extracted gender mentions with a precision of 96%. A total of 198 (27%) redditors disclose their gender, 118 through text and 80 by posting to gender-related subreddits. 152 (77%) redditors explicitly reveal themselves to be male, 36 female, and 10 trans. The coherence reaches a value of 95% for the text and 98% for the subreddits. The incoherent users can be classified as follow: a) 5 users mentioned to be male and female at a different point in time; b) 2 users posted in both male and female related subreddits; c) 10 users posted on male subreddits while mentioning to be female and vice versa.

3.3 Age

We look for sentences where the users state their age (e.g., "I am 22 years old", "As a 30 years old.."). After removing the incorrect mentions, in the case of multiple mentions, we consider the age of a redditor the value occurring the most, except consecutive values, where we consider the highest one. For example, if a user mentions two times to be 19 and one times to be 20, we consider the redditor to be 20 years old. Given the 7 years time range of our data, reports of increasing age are possible. Our approach has a precision of 97% in extracting mentions regarding age and infers the age of 264 (35%) users. As shown in Figure 1, the age ranges between 16 and 43 years, with a peak in the mid-20s' and early 30s'. Dismissing increasing consecutive ages by a given user as accurate, we find age coherence to be 91%.

3.4 Education

For inferring the education levels of users, we look for statements mentioning the possession of a diploma or degree (e.g., "I got a degree in..", "I have a diploma"), having graduated (e.g., "I graduated from high school", "I graduated from hs") and being a drop out (e.g, "I dropped out last semester", "I drop out from high school"). Our approach achieves a precision of 94% and 90% respectively in extracting information about the education level attained and the

[3] https://praw.readthedocs.io/en/latest/index.html

[4] All the data collected and produced in this work is published here https://doi.org/10.5281/zenodo.3565430

Table 1: Example of patterns.

Pattern	Sentence	Mention
$I\,am \wedge \langle DT \rangle \wedge (guy\|man)$	hey i am a guy who watches wentworth too lol	i am a guy
$I\,am\ \langle CD \rangle$	i am 23 and beer no longer has any great effect on me	i am 23

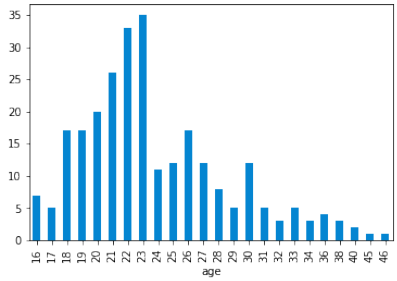
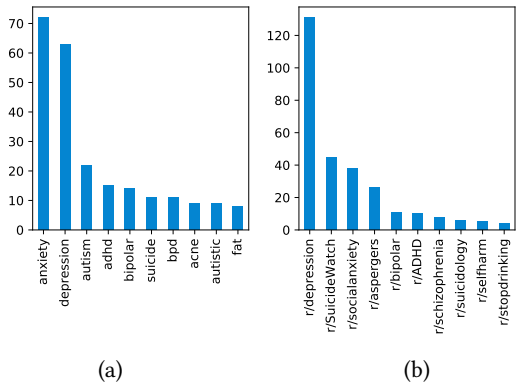


Figure 1: Age distribution

dropout. 127 (17%) users disclose information about their education. Interestingly, 53% of them declare to have graduated from college and 45% mention to have dropped out. 94% of people are coherent in their all post history. 7 users make contradicting statements, saying they both dropped out and graduated from college. Upon further inspection, we discover 5 of them graduated after having dropped out.

3.5 Health

We look in the text for mentions of being diagnosed or suffering from health issues (e.g., "I was diagnosed with..", "I'm suffering from ...", "I have ..."). For identifying relevant object sub-patterns, we use a custom made dictionary built crawling DBPedia for all the terms referring to the class *Disease*[5].



(a) (b)

Figure 2: Top 10 mentioned health issues (a) and top 10 health related subreddits where the posts were submitted to (b)

[5]http://dbpedia.org/ontology/Disease

Our approach achieves a precision of 88% in extracting mentions regarding health issues. 245 (34%) users claim to suffer from some kind of illness, and an additional 62 redditors who post on `r/NEET`, also post on health related subreddits, for a total of 317 (43%). In this case, contradictory mentions do not exist; so we do not report the coherence value for this attribute. Figure 2(a) and 2(b) show respectively the most mentioned health issues and the contributed health-related subreddits. In both cases, the majority are related to mental health issues such as depression, anxiety, and bipolar disorder.

3.6 Living Context

Here we are interested in mentions that refer to where and how a user lives. We seek to identify the *location* - i.e., city or country - the *residential status* - i.e., does he/she live alone or with the family? - and *isolation* of the living place (e.g., rural vs urban area). Our approach achieves a precision of 83% in extracting mentions regarding the living context. 189 (26%) users write about their living situation. 127 (17%) of them disclose their location at either country or city level. 45 (6%) talk about their residential status, and finally, 46 (6%) mention the isolation of their location. For the geographic location and residential status, the coherence is 98%, while it is 97% for the isolation level. The data show users tend to live with their family (58%) and in rural areas (59%). Most of the redditors seem to come from the United States, followed by Europe and Canada.

3.7 Topical Analysis

Topic analysis typically requires the number of topics to be set a priori. To determine the optimal number, we assessed the coherence for topic models for a range of topic numbers and selected the optimal number when the coherence measure's gain (i.e. first derivative) sharply diminished, similar to the elbow method employed in standard factor analysis. Then, we manually inspected a sample of submissions affiliated with each topic in order to devise a descriptive label, characterized in Table 2 by distinct bags of words, in which the words are sorted by its prominence the topic.

NEETs on Reddit discuss a wide range of topics, as shown in Table 2. They talk about leisure activities such as gaming (T0 and T16), hobbies—that include reading and writings (T7), watching videos and listening to music (T18), but also work (T3). They share their problems with health (T1 and T9), social interactions (T5), and sex (T12), and solicit suggestions to improve their lifestyle (T6 and T13). They tell stories about their personal and daily lives (T10 and T14). They discuss politics (T4), society (T17), and religion (T11), and some of them express anger and hate (T15). Figure 3 shows the overall distribution of topics in posts and comments, where each post and comment is mapped to its most representative topic. In both cases, the most occurring topic is *Gaming* (T0), but, for posts

Table 2: List of topics extracted.

Topic	Keywords	Label
T0	game, play, good, fun, player, team, win, pretty, level, card	Gaming
T1	feel, experience, problem, thing, anxiety, depression, feeling, mind, issue, thought	Mental Health
T2	lol, nice, dude, joke, funny, lmao, op, cool, wait, bro	Web Slang
T3	work, job, year, money, neet, pay, school, college, live, class	Jobs
T4	country, state, world, american, city, law, place, live, war, government	Politics
T5	people, friend, talk, family, person, kid, social, life, parent, love	Social Interaction
T6	eat, day, food, high, drug, week, low, make, good, drink	Nutrition
T7	read, learn, write, book, word, find, type, make, art, good	Hobbies
T8	buy, money, free, high, number, sell, order, pay, make, cost	Buying and selling
T9	face, hand, head, side, big, eye, make, body, small, back	Physical health
T10	time, point, change, lot, happen, part, bad, put, sound, reason	Personal stories
T11	human, world, exist, god, make, true, good, animal, religion, fact	Religion and philosophy
T12	man, woman, guy, girl, sex, date, incel, female, male, ugly	Sex and dating
T13	find, give, pretty, ill, bad, hope, great, time, place, hard	Self improvement
T14	day, time, start, back, leave, year, ago, sleep, long, night	Daily life
T15	fuck, shit, give, bad, hate, make, stop, call, literally, suck	Hate and Anger
T16	fight, kill, level, hit, make, damage, attack, good, power, shoot	Roleplaying
T17	people, problem, call, care, person, white, agree, hate, black, opinion	Society
T18	watch, show, youtube, video, love, movie, music, sound, character, listen	Video and Music
T19	post, question, reddit, comment, action, rule, concern, link, moderator, subreddit	Reddit

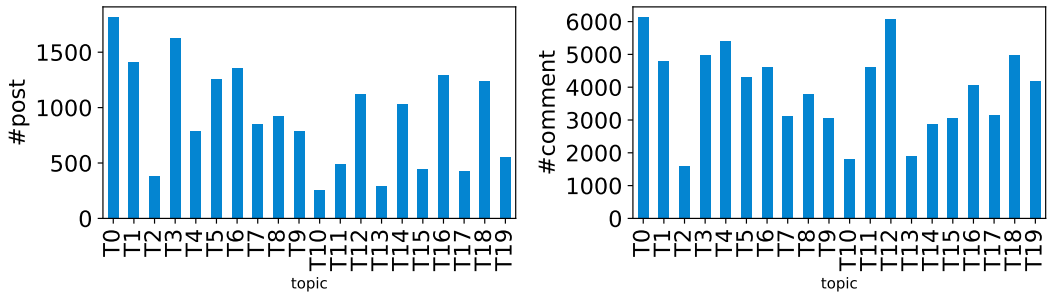


Figure 3: Overall distribution of topics in posts and comments. For the enumeration of the topics, we refer the reader to Table 2.

only, the second most discussed topic is *Job* (T3). *Mental health* (T1) is an important topic, appearing at the top of the topic rankings for both posts and comments. A Chi-Square test on the two distributions revealed there is a statistically significant difference between the frequencies of occurrence of topics in posts and comments[6]. Interestingly, *Sex* (T12) is the most discussed topic in the comments, but less so (7th) in posts. Further analysis shows the majority of comments pertaining this topic are posted in subreddits where the 'incel' (involuntary celibates) community is predominant (e.g., `r/braincels`, `r/ForeverAlone` and, to some extent, `r/MGTOW`).

3.8 Redditor Cluster Analysis

We perform a cluster analysis on the users through a Spectral Clustering algorithm using as features the number of times a topic occurs in the posts of a user. Hence, these clusters characterize and distinguish users based on the topical profile or distribution of their submissions. We computed the affinity matrix using a Gaussian kernel and choose the number of clusters by looking at the index of the largest gap between its eigenvalues (n=5) [39]. Figure 4 shows the distribution of topics in the five clusters.

Cluster 1. This is the largest cluster and include the 68% of our data's redditors. Users belonging to this cluster submit mainly posts related to jobs. However, these discussions appear to occur mainly within the `r/NEET` subreddit, and not in other, more specialized subreddits (e.g., `r/jobs`). They also discuss hobbies (T0 and T7) and other interests.

Cluster 2. People belonging to this group discuss how to socialize (T5) and roleplaying games (T16). They tend to discuss these topics on general purpose subreddits as `r/AskReddit`, but also on more specific ones like `r/socialskill`.

[6] $X2 = 1374.08$ $DoF = 19$ $p < 0.001$

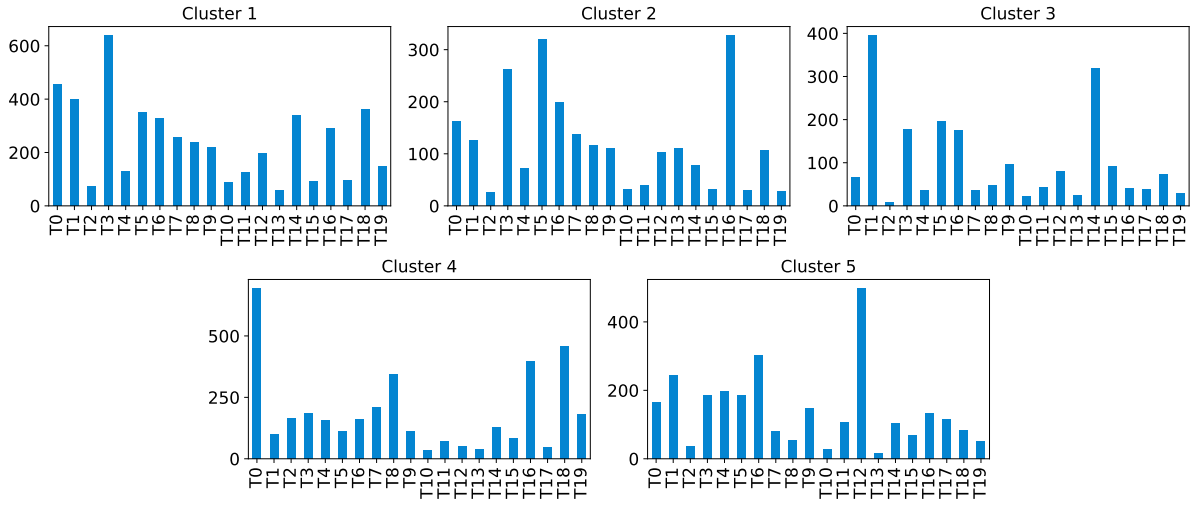


Figure 4: Distribution of topics in the different clusters. For the enumeration of the topics, we refer the reader to Table 2

Cluster 3. Users belonging to this cluster use reddit to share their struggles with depression, anxiety and other mental health issues. Generally, they post both in "peer-support" (e.g. `r/depression` and `r/suicidewatch`) and medical advice (e.g., `r/AskDocs`) subreddits.

Cluster 4. Redditors in this cluster use Reddit mainly to talk about leisure activities such as gaming (T0), watching videos, and listening to music (T18).

Cluster 5. People belonging to this group mainly discuss sex and relationships. Moreover, they frequent subreddits dedicated to the incel commmunity (e.g., `r/braincels` and `r/ForeverAlone`).

4 DISCUSSION

Studies from both Eurofound [12, 25] and OECD [29] report females being more at risk to become NEETs, with 12.3% females compared to 11.7% males in the former, and 16.7% females versus 11.8% males for the second. In our analysis, instead, the NEET group seems to largely comprise males (77%). However, we need to consider that the Reddit demographic is skewed toward a young male population [4]. The OECD [29] report only covers age 15 to 29, while Eurofound [12] considers an even more constrained age range from 15 to 24. Conversely, our analysis shows that, even though the majority of users disclosing the age fit in those ranges, the ages of those who are either NEETs or involved with NEETs extends beyond 29, reaching in some cases the 40s. This shift in age was recently reported also in [26]. In the literature [29], education is seen as an important protection against becoming NEET, with 48% of them having attained upper secondary level education (i.e., high school diploma), and only the 8% the tertiary level (i.e., college). Nevertheless, in our analysis 53% of the users - who disclose information about the education - declare having obtained a degree and 48% at least a high school diploma. In agreement with official reports [12] - even though we were able to extract geographic location or area from only few redditors - 60% of them complain about living in *rural* areas. As many traditional [5, 12, 15] and novel [21, 38] studies have shown, our analysis reveals that 43% of redditors either declare themselves to be suffering from health issues (depression and anxiety mainly) or post content on health-related subreddits. The *Mental Health* topic extracted by the topic modeling analysis is discussed by 74% of users. However, in our case, it is not possible to say if poor health is the cause of the NEET condition or the other way around. One of the main consequences of being NEET is the isolation from society and radicalization [12, 28]. In our study, we find elements of extremist behavior with the presence of the incel community. Related to this, 64% of our sample of redditors discussed the *Sex* topic in either posts or comments at least once. It is not clear, though, if they approach the incel ideology once they become NEET, or if this belief is already present. The incel phenomenon is very complex, which includes isolation, far-right movements, and health issues [18]. A detailed study of this movement is out of the scope of this paper and shall be addressed in future work.

Our topical analysis shows NEETs are interested in *Politics* and *Society* even though official studies [12] report low participation and interest in politics, suggesting NEETs are engaged in the discussion about those themes, but lack of trust in existing institutions, as also reported by [21]. As previously mentioned, literature [5, 14, 25] asserts the NEET group is composed of very diverse types of people. In particular [25] distinguish seven groups: re-entrants - people that will soon re-enter in employment - short-term unemployed, long-term unemployed, NEET for illness or disabilities, NEET for family responsibilities, discouraged workers - people who stopped looking for a job because they believe there are no job opportunities for them - volunteer NEETs - people who are not looking for a job because they are pursuing alternative career paths. The topics of discussions we uncovered appears to exclude the group of *Re-entrants* and *Volunteer NEETs*, as most probably people belonging to those groups do not identify as NEETs. According to the age range, they seem to belong either to the *Short-term* or *Long-term* unemployed group. Our cluster analysis found a group of NEET

mainly talking about mental health issues - (cluster 3). This may suggest the presence of people identifying themselves as NEET due illness or disabilities group. Also, the presence of *NEETs for family responsibilities* is hinted by the users who claim to be still living with their family. The *discouraged workers* are difficult to detect as discouragement and disillusion can be linked to depression - an ever-present topic of discussion, making it difficult to ascertain which is the main cause. People belonging to clusters 3,4, and 5 are not actively talking about jobs on Reddit, but this does not mean they are discouraged.

Ethical issues. We retrieved Reddit content using the platform API, as it is allowed by its User Agreement[7] and Privacy Policy[8]. Beyond the legal framework, there are ethical considerations for this type of digital ethnography. As researchers, we are weighing the relative anonymity of the subreddit and the opportunity for some open discussions regarding the experiences of being a NEET against the fact that we are not engaging with participants directly. We have not identified any of the participants on the NEETs subreddit or further identified their participation within Reddit more broadly; the data collection was limited and driven by the current literature on NEETs. There are concerns that this type of 'distant' digital ethnography, results in research about NEETs rather than exploring the topic with them. This is addressed in several ways. First, the chosen approach allows a way of data collection that is not time-consuming for participants and gives them a voice without burdening them by asking for participation in an interview or survey. Further, the goal of the paper is to raise awareness around the issues that NEETs face – as described by them - and to add to known literature on NEETs drawing directly from persons self-identifying as such. The method of this paper reflects that and tries to highlight common issues among a group of NEETs. Still some ethical concerns remain regarding how to deal with the implications of our analyses when designing data-driven interventions (e.g., confidentiality, data misuse) as they can have serious consequences (e.g. profiling, discrimination, perpetuation of problematic stereotypes). Any actors willing to implement such intervention needs to engage all the relevant stakeholders (e.g., researchers, designers, civil servants, clinicians, etc..) in order to both provide help and respect the rights of the vulnerable people involved.

5 CONCLUSION

In this paper we presented a study of the Reddit platform as a complementary source of information about vulnerable youth, focusing on the NEET community. Our findings exhibit some similarities to the ones described in official reports. Differences have been found in the gender declared by the redditors, due to the skewness of the Reddit population, in the age range, and in the education levels. NEETs on Reddit share the daily struggle to deal with health and emotional issues such as depression and anxiety. We were able to highlight the presence of extremist behavior, and a large variety of interests such as society, politics, videogames, music, arts, reading, and writing. While the NEET members of Reddit may not be representative of the entire population, our work exposes substantial concerns and conditions of this subpopulation. We cannot consider local factors, as Reddit data do not come with the user's location. In this work, we use an retrieval approach based on manually built patterns; an unknown part of data may remain untapped. Nevertheless, we believe this work contributes to understanding how Reddit and social media can be used to complement information about NEETs and design more targeted and detailed studies. Future work will focus on improving the automatic approach to achieve a higher recall, engaging directly with the individuals, including other vulnerable youth groups, and applying the method on other social media and online platforms.

[7] https://www.redditinc.com/policies/user-agreement
[8] https://www.redditinc.com/policies/privacy-policy

ACKNOWLEDGMENTS

This work was supported by the NWO project "JOIN Jongeren in een veerkrachtige samenleving. Naar nieuwe arrangementen voor inclusiviteit en participatie" (grant number 400.17.603) and it was carried out on the Dutch national e-infrastructure with the of SURF Cooperative.

REFERENCES

[1] Jacobo Levy Abitbol, Márton Karsai, Jean-Philippe Magué, Jean-Pierre Chevrot, and Eric Fleury. 2018. Socioeconomic Dependencies of Linguistic Patterns in Twitter: A Multivariate Analysis. (apr 2018). https://doi.org/10.1145/3178876.3186011 arXiv:1804.01155

[2] Nazanin Andalibi, Pinar Ozturk, and Andrea Forte. 2015. Depression-Related Imagery on Instagram. In *Proceedings of the 18th ACM Conference Companion on Computer Supported Cooperative Work and Social Computing* (Vancouver, BC, Canada) *(CSCW'15 Companion)*. Association for Computing Machinery, New York, NY, USA, 231–234. https://doi.org/10.1145/2685553.2699014

[3] American Psychiatric Association et al. 2013. *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub.

[4] B Y Michael Barthel, Galen Stocking, Jesse Holcomb, and Amy Mitchell. 2016. Nearly Eight-in-Ten Reddit Users Get News on the Site. 25 (2016). www.pewresearch.org.

[5] Federico Batini, Vanessa Corallino, Giulia Toti, and Marco Bartolucci. 2017. NEET: A Phenomenom Yet to Be Explored. *Interchange* 48, 1 (01 Feb 2017), 19–37. https://doi.org/10.1007/s10780-016-9290-x

[6] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research* 3 (2003), 993–1022.

[7] Mariana Bălan. 2015. Methods to Estimate the Structure and Size of the "Neet" Youth. *Procedia Economics and Finance* 32 (2015), 119–124. https://doi.org/10.1016/S2212-5671(15)01372-6

[8] Stevie Chancellor, Zhiyuan Lin, Erica L. Goodman, Stephanie Zerwas, and Munmun De Choudhury. 2016. Quantifying and Predicting Mental Illness Severity in Online Pro-Eating Disorder Communities. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work and Social Computing* (San Francisco, California, USA) *(CSCW '16)*. Association for Computing Machinery, New York, NY, USA, 1171–1184. https://doi.org/10.1145/2818048.2819973

[9] Aron Culotta. 2014. Estimating County Health Statistics with Twitter. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 1335–1344. https://doi.org/10.1145/2556288.2557139

[10] Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013. Social media as a measurement tool of depression in populations. In *Proceedings of the 5th Annual ACM Web Science Conference on - WebSci '13*. ACM Press, New York, New York, USA, 47–56. https://doi.org/10.1145/2464464.2464480

[11] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting Depression via Social Media. *ICWSM* 2 (2013), 128–137. https://doi.org/10.1109/IRI.2012.6302998 arXiv:1511.02556

[12] Eurofound. 2011. Young people and NEETs in Europe : First findings. *Europe* (2011), 1–8. https://doi.org/10.2806/3177

[13] Miriam Fernandez, Moizzah Asif, and Harith Alani. 2018. Understanding the Roots of Radicalisation on Twitter. In *Proceedings of the 10th ACM Conference on Web Science* (Amsterdam, Netherlands) *(WebSci '18)*. ACM, New York, NY, USA, 1–10. https://doi.org/10.1145/3201064.3201082

[14] A. Furlong. 2007. The zone of precarity and discourses of vulnerability: NEET in the UK. *Journal of Social Sciences and Humanities* 381 (2007), 101–121. http://eprints.gla.ac.uk/36831/

[15] Sidra Goldman-Mellor, Avshalom Caspi, Louise Arseneault, Nifemi Ajala, Antony Ambler, Andrea Danese, Helen Fisher, Abigail Hucker, Candice Odgers, Teresa Williams, Chloe Wong, and Terrie E Moffitt. 2016. Committed to work but

vulnerable: self-perceptions and mental health in NEET 18-year olds from a contemporary British cohort. *Journal of Child Psychology and Psychiatry and Allied Disciplines* 57, 2 (feb 2016), 196–203. https://doi.org/10.1111/jcpp.12459

[16] Liang Gou, Michelle X. Zhou, and Huahai Yang. 2014. KnowMe and ShareMe: Understanding Automatically Discovered Personality Traits from Social Media and User Sharing Preferences. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 955–964. https://doi.org/10.1145/2556288.2557398

[17] Eszter Hargittai. 2018. Potential Biases in Big Data: Omitted Voices on Social Media. *Social Science Computer Review* 0, 0 (2018), 0894439318788322. https://doi.org/10.1177/0894439318788322 arXiv:https://doi.org/10.1177/0894439318788322

[18] Sylvia Jaki, Tom De Smedt, Maja Gwó, Rudresh Panchal, Alexander Rossa, and Guy De Pauw. 2018. Online Hatred of Women in the Incels . me Forum : Linguistic Analysis and Automatic Detection. (2018), 1–30. https://organisms.be/downloads/incels.pdf

[19] Kyriaki Kalimeri, Mariano G. Beiró, Matteo Delfino, Robert Raleigh, and Ciro Cattuto. 2019. Predicting demographics, moral foundations, and human values from digital behaviours. *Computers in Human Behavior* 92 (2019), 428–445. https://doi.org/10.1016/j.chb.2018.11.024

[20] Ramakanth Kavuluru, María Ramos-Morales, Tara Holaday, Amanda G Williams, Laura Haye, and Julie Cerel. 2016. Classification of Helpful Comments on Online Suicide Watch Forums. In *Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics - BCB '16*, Vol. 2016. NIH Public Access, 32–40. https://doi.org/10.1145/2975167.2975170

[21] Hanna Kirjavainen and Harri Jalonen. 2020. The Many Faces of Social Withdrawal in Hikikomori. In *Well-Being in the Information Society. Fruits of Respect*, Mirella Cacace, Raija Halonen, Hongxiu Li, Thao Phuong Orrensalo, Chenglong Li, Gunilla Widén, and Reima Suomi (Eds.). Springer International Publishing, Cham, 156–168.

[22] Reeva Lederman, Greg Wadley, John Gleeson, Sarah Bendall, and Mario Álvarez-Jiménez. 2014. Moderated online social therapy: Designing and evaluating technology for mental health. *ACM Transactions on Computer-Human Interaction (TOCHI)* 21, 1 (2014), 1–26.

[23] Alejandro Llorente, Manuel Garcia-Herranz, Manuel Cebrian, and Esteban Moro. 2015. Social Media Fingerprints of Unemployment. *PLoS ONE* 10, 5 (may 2015), e0128692. https://doi.org/10.1371/journal.pone.0128692 arXiv:1411.3140

[24] Massimiliano Mascherini. 2012. Young people and NEETs in Europe: first findings. https://doi.org/10.2806/3177

[25] Massimiliano Mascherini, Stefanie Ledermaier, and European Foundation for the Improvement of Living and Working Conditions. 2016. *Exploring the diversity of NEETs*.

[26] Walter Matli. 2021. The Changing Nature and Use of the Concept NEET in Contemporary Society: Normalising the NEET Age Cohort. *Handbook of Research on Institutional, Economic, and Social Impacts of Globalization and Liberalization* (2021), 394–411. https://doi.org/978-1-7998-4459-4.ch022

[27] Andrew Y Ng, Michael I Jordan, Yair Weiss, et al. 2002. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems* 2 (2002), 849–856.

[28] OECD. 2014. Education at a Glance 2014. http://www.oecd.org/education/Education-at-a-Glance-2014.pdf

[29] OECD. 2019. Youth not in employment, education or training (NEET) (indicator). https://doi.org/10.1787/72d1033a-en

[30] Kathleen O'Leary, Stephen M. Schueller, Jacob O. Wobbrock, and Wanda Pratt. 2018. "Suddenly, We Got to Become Therapists for Each Other": Designing Peer Support Chats for Mental Health. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3173905

[31] Albert Park, Mike Conway, and Annie T. Chen. 2018. Examining thematic similarity, difference, and membership in three online mental health communities from reddit: A text mining and visualization approach. *Computers in Human Behavior* 78 (2018), 98 – 112. https://doi.org/10.1016/j.chb.2017.09.001

[32] Claudio Quintano, Paolo Mazzocchi, and Antonella Rocca. 2018. The determinants of Italian NEETs and the effects of the economic crisis. *Genus* 74, 1 (dec 2018), 5. https://doi.org/10.1186/s41118-018-0031-0

[33] Karen Robson. 2008. Becoming NEET in Europe: A Comparison of Predictors and Later-Life Outcomes. In *Global Network on Inequality Mini-Conference 22 February 2008*. www.youth-inequalities.org

[34] H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Stephanie M. Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin E. P. Seligman, and Lyle H. Ungar. 2013. Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. *PLoS ONE* 8, 9 (sep 2013), e73791. https://doi.org/10.1371/journal.pone.0073791

[35] Anne Spaa, Abigail Durrant, Chris Elsden, and John Vines. 2019. Understanding the Boundaries between Policymaking and HCI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3290605.3300314

[36] Sho Tsugawa, Yusuke Kikuchi, Fumio Kishino, Kosuke Nakajima, Yuichi Itoh, and Hiroyuki Ohsaki. 2015. Recognizing Depression from Twitter Activity. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 3187–3196. https://doi.org/10.1145/2702123.2702280

[37] Alessandra Urbinati, Kyriaki Kalimeri, Andrea Bonanomi, Alessandro Rosina, Ciro Cattuto, and Daniela Paolotti. 2020. Young Adult Unemployment Through the Lens of Social Media: Italy as a Case Study. In *Social Informatics*, Samin Aref, Kalina Bontcheva, Marco Braghieri, Frank Dignum, Fosca Giannotti, Francesco Grisolia, and Dino Pedreschi (Eds.). Springer International Publishing, Cham, 380–396.

[38] Mark Wong. 2020. Hidden youth? A new perspective on the sociality of young people 'withdrawn' in the bedroom in a digital age. *New Media & Society* 22, 7 (2020), 1227–1244. https://doi.org/10.1177/1461444820912530 arXiv:https://doi.org/10.1177/1461444820912530

[39] Lihi Zelnik-Manor and Pietro Perona. 2004. Self-tuning Spectral Clustering. In *Proceedings of the 17th International Conference on Neural Information Processing Systems* (Vancouver, British Columbia, Canada) *(NIPS'04)*. MIT Press, Cambridge, MA, USA, 1601–1608. http://dl.acm.org/citation.cfm?id=2976040.2976241

[40] Yue Zhang, Arti Ramesh, Jennifer Golbeck, Dhanya Sridhar, and Lise Getoor. 2018. A Structured Approach to Understanding Recovery and Relapse in AA. In *Proceedings of the 2018 World Wide Web Conference* (Lyon, France) *(WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 1205–1214. https://doi.org/10.1145/3178876.3186019

[41] Carolina V Zuccotti and Jacqueline O'Reilly. 0. Ethnicity, Gender and Household Effects on Becoming NEET: An Intersectional Analysis. *Work, Employment and Society* 0, 0 (0), 0950017017738945. https://doi.org/10.1177/0950017017738945 arXiv:https://doi.org/10.1177/0950017017738945