Understanding who uses Reddit: Profiling individuals with a self-reported bipolar disorder diagnosis

Glorianna Jagfeld[⋆], **Fiona Lobban**[⋆], **Paul Rayson**[▽], **Steven H. Jones**[⋆]
[⋆]Spectrum Centre for Mental Health Research
[▽]School of Computing and Communications
Lancaster University, United Kingdom
{g.jagfeld, f.lobban, p.rayson, s.jones7}@lancaster.ac.uk

Abstract

Recently, research on mental health conditions using public online data, including Reddit, has surged in NLP and health research but has not reported user characteristics, which are important to judge generalisability of findings. This paper shows how existing NLP methods can yield information on clinical, demographic, and identity characteristics of almost 20K Reddit users who self-report a bipolar disorder diagnosis. This population consists of slightly more feminine- than masculine-gendered mainly young or middle-aged US-based adults who often report additional mental health diagnoses, which is compared with general Reddit statistics and epidemiological studies. Additionally, this paper carefully evaluates all methods and discusses ethical issues.

1 Introduction and related work

People who experience extreme mood states that interfere with their functioning, meet the criteria for bipolar disorder (BD) according to the diagnostic manuals Diagnostic and Statistical Manual of Mental Disorders (DSM) (American Psychiatric Association, 2013) and International Classification of Diseases (ICD) (World Health Organisation, 2018). DSM and ICD operationalise extreme mood states in terms of major depressive episodes, 'almost daily depressed mood or diminished interest in activities with additional symptoms for at least 14 days' (World Health Organisation, 2018) and (hypo-)manic episodes, 'a distinct period of abnormally and persistently elevated, expansive, or irritable mood and abnormally and persistently increased goal-directed activity or energy' that lasts at least seven (four) days (American Psychiatric Association, 2013, p. 124).

DSM and ICD distinguish several BD subtypes based on the lifetime frequency and intensity of (hypo-)manic and depressed episodes. The only requirement for a diagnosis of bipolar I disorder (BD-I) is at least one lifetime manic episode, whereas bipolar II disorder (BD-II) requires at least one hypomanic and one major depressive episode (American Psychiatric Association, 2013, pp. 126, 132). Cyclothymic disorder applies to numerous periods of hypomanic and depressive symptoms during at least two years that do not meet criteria for hypomanic or major depressive episodes (American Psychiatric Association, 2013, p. 139).

Bipolar mood episodes are often recurring (Treuer and Tohen, 2010; Gignac et al., 2015), so many individuals living with BD require life-long treatment (Goodwin et al., 2016) and have a heightened suicide risk (Novick et al., 2010). However, characteristics and outcomes of people meeting BD criteria are diverse, with some living well, (e.g., Warwick et al., 2019) and even functioning on a high level (Akers et al., 2019).

1.1 Online forums as research data source

Online forums have become an increasingly attractive source for research data, enabling non-reactive data collection, where researchers do not influence data creation, at large scale (Fielding et al., 2016). Natural language processing (NLP) research in this area has focused on predicting people at risk of BD (Coppersmith et al., 2014; Cohan et al., 2018; Sekulić et al., 2018). Health researchers have explored the lived experience of BD with qualitative analyses of online posts (Mandla et al., 2017; Sahota and Sankar, 2019). Unlike in clinical studies, usually little or no demographic information is available for online forum users, so it is unclear to what populations these results generalise (Ruths and Pfeffer, 2014). For example, language differences between Twitter users with self-reported Major depressive disorder (MDD) or Post-traumatic stress disorder (PTSD) correlated highly with their personality and demographic characteristics (Preoţiuc-Pietro et al., 2015). So it is unclear whether these findings really indicate mental health (MH) diagnoses or other user characteristics.

Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology, pages 1–14
June 11, 2021. ©2021 Association for Computational Linguistics

1.2 The online discussion forum Reddit

Besides MH-specific platforms (Kramer et al., 2004; Vayreda and Antaki, 2009; Bauer et al., 2013; Latalova et al., 2014; Poole et al., 2015; McDonald and Woodward-Kron, 2016; Campbell and Campbell, 2019), blogs (Mandla et al., 2017), and Twitter (Coppersmith et al., 2014; Ji et al., 2015; Saravia et al., 2016; Budenz et al., 2019; Huang et al., 2019), much recent research of user-generated online content in BD has focused on the international online discussion forum Reddit[1] (Gkotsis et al., 2016, 2017; Cohan et al., 2018; Sekulić et al., 2018; Sahota and Sankar, 2019; Yoo et al., 2019).

The platform Reddit is among the most visited internet sites worldwide (Alexa Internet, 2020), hosting a number of subforums ('subreddits') for general topics as well as interest groups. There is a vast and growing amount of BD-related content on Reddit, with more than 50K new posts per month in the four largest BD-related subreddits[2]. Anyone can view posts without registration and the Reddit API offers free access to all historic posts. Reddit profiles do not provide any user characteristics besides the username and sign-up date in a structured format or comparable to a Twitter bio. While some surveys provide general information on Reddit users, none of the BD-specific studies looked at particular user characteristics of their sample, which is important (Amaya et al., 2019).

1.3 Research questions and contributions

The above considerations motivate our research questions: What characteristics of Reddit users who disclose a BD diagnosis can be automatically inferred from their public Reddit information and how do they compare to general Reddit users and clinical populations? What are the ethical considerations around determining users' characteristics and ways to minimise potential negative impacts?

This work has two main contributions, both of which may be relevant to different parts of the CLPsych community. Crucially, the authors are an interdisciplinary team of NLP and clinical psychology researchers, as well as practising clinical psychologists, who regularly consult with people with lived experience of BD in an advisory panel.

First, this paper estimates and discusses clinical, demographic and identity characteristics of Reddit users who self-report a BD diagnosis (see Figure 3 for a visual results summary). This has implications for future BD-focused research on Reddit and helps to contextualise previous work. Moreover, this information is relevant for clinicians who may want to recommend certain online forums to clients and to clinical researchers interested in recruiting via Reddit. Second, this work shows how simple rule-based and off-the-shelf state-of-the-art NLP methods can estimate Reddit user characteristics, and carefully discusses ethical considerations and harm-mitigating ways of doing so. These findings and discussions apply to other, also non-clinical, subgroups of Reddit users. The evaluation with manual annotations evaluates published NLP methods in an applied setting.

2 Methods

2.1 User identification

In this work, the identification of Reddit users with lived experience of BD adapts previous approaches based on self-reported diagnosis statements, e.g., 'I was diagnosed with BD today' (Coppersmith et al., 2015; Cohan et al., 2018; Sekulić et al., 2018). Importantly, this captures self-*reported* diagnoses by a professional and not *self-diagnoses*, which were excluded. Contrary to existing datasets of Reddit posts by people with a self-reported BD diagnosis, all posts of identified people were retained and not only those unrelated to MH concerns. This enables subsequent research on the lived experience of people with BD. All available Reddit posts (January 05 - March 19) that mentioned 'diagnosis' and a BD term (see below) were downloaded from Google BigQuery. User account meta-data (id, username, UTC timestamp of sign-up) for all matching posts was retrieved via the Reddit python API praw[3] to remove posts by users who had deleted their profile after creation of the BigQuery tables. Each of the 170K posts was classified as self-reported diagnosis post after automatically removing quoted content if it met the following criteria adapted from Cohan et al. (2018) (see Table 1 for examples):

- Contains at least one condition term for BD.
- Matches at least one inclusion pattern, i.e., BD diagnosis of any type by a professional.
- Does not match any exclusion pattern, e.g., self-diagnosis.

[1] https://www.reddit.com/

[2] r/bipolar, r/BipolarReddit, r/bipolar2, r/bipolarSOs

[3] https://github.com/praw-dev/praw

Component	Number	Examples
Inclusion patterns	145	As someone with a diagnos*, my recent CONDITION diagnos*, I went to a DOCTOR and got diagnos*
CONDITION terms	92	Bipolar, manic depression, BD-I, BD-II, cyclothymia
DOCTOR terms	18	Doctor, pdoc, shrink
Exclusion patterns	74	Not formally diagnos*, self diagnos*, she's diagnos*

Table 1: Components of patterns to identify English self-reported diagnosis statements; *: wildcard

- The distance between at least one condition term and the beginning or end of an inclusion phrase is less than the experimentally determined threshold of 55 characters.

Subsequently, all posts (id, submissions title, text, subreddit, user id, UTC timestamp of time posted) of the 21K user accounts with at least one self-reported diagnosis post were downloaded via praw. The first author checked the self-reported diagnosis statements of all accounts with more than 1.5K submissions or 200K comments or whose name included 'bot' or 'auto', removing 30 automated user accounts (bots). Finally, 960 user accounts with a self-reported psychotic disorder diagnosis were removed because this constitutes an exclusion criterion for BD (American Psychiatric Association, 2013, pp. 126, 134).

2.2 User characteristics extraction/inference

Several NLP methods were applied and compared to extract or infer clinical (MH comorbidities = diagnoses additional to BD), demographic (age, country of residence), and identity (gender) characteristics of Reddit users with a self-reported BD diagnosis. See Appendix A for more details on the age, country, and gender methods and their previously published performance. The first and third author manually annotated self-reported BD diagnoses, age, country, and gender for random included users for evaluation.

2.2.1 Mental health comorbidities

Frequencies for other self-reported MH diagnoses were obtained by matching all dataset posts against inclusion patterns for other diagnoses, in the same way as for identifying self-reported BD diagnoses. Condition terms for nine major DSM-5 and ICD-11 diagnoses were extended from Cohan et al. (2018): Anxiety disorder (Generalised/Social anxiety disorder, Panic disorder), Attention deficit hyperactivity disorder (ADHD), Borderline personality disorder (BPD), MDD, PTSD, Psychotic disorder (Schizophrenia/Schizoaffective disorder), Obsessive compulsive disorder (OCD), Autism spectrum disorder (ASD), and Eating disorder (ED).

2.2.2 Age

Two methods to recognise a user's age relative to one of their posts were compared. An approximate date of birth was calculated from the post timestamp to then calculate the user's age when posting for the first time and their mean age over all posts.

- **Self-reported**: Reddit users sometimes self-report their age and gender in a bracketed format, e.g. 'I [17f] just broke up with bf [18m]'. Regular expressions extracted age and gender from such self-reports in submission titles.

- **Language use**: Tigunova's (2019) neural network model predicts the age group of users with at least ten posts from their contents and language style. Training data for this model came from Tigunova et al. (2020) who automatically labelled Reddit users with their self-reported age (see Appendix A.1).

- **Hybrid**: The Hybrid method assigns the extracted age from the Self-reported method if available, and otherwise the predicted age from the Language use method because evaluation revealed that the Self-reported method had higher accuracy but lower coverage than the Language use method (see Section 4.2).

2.2.3 Country of residence

The only published method for Reddit user localisation to date (Harrigian, 2018) infers a user's country of residence via a dirichlet process mixture model[4]. It uses the distribution of words, posts per subreddit, and posts per hour of the day (timezone proxy) of a user's up to 250 most recent comments.

[4] https://github.com/kharrigian/smgeo

2.2.4 Binary gender

Three methods to recognise binary gender (feminine (f)/masculine (m)) leveraging different types of information were compared. All three methods pertain to a performative gender view, which posits that people understand their and others' gender identity by certain behaviours (including language) and appearances that society stipulates for bodies of a particular sex (Larson, 2017). Non-binary gender identities were not included due to a lack of NLP methods to detect them.

- **Username**: The character-based neural network model of Wang and Jurgens (2018) predicts whether a username strongly performs f or m gender, otherwise it assigns no label.
- **Self-reported**: See Section 2.2.2.
- **Language use**: The neural network model by Tigunova et al. (2019) predicts gender for Reddit users with at least ten posts from the post texts. It was trained on data automatically labelled with self-reported gender provided by Tigunova et al. (2020) (see Appendix A.1).
- **Hybrid**: Evaluation revealed an accuracy ranking of Username > Self-reported > Language use and the inverse for coverage (Section 4.2). The Hybrid method assigns a binary gender identity in a sequential approach, disregarding possible disagreements between methods: If the Username method found the username to perform f or m gender, it takes this prediction, otherwise assumes the self-reported gender if available, and else resorts to the predictions of the Language use method.

3 Ethical considerations

At least four main ethical considerations arise for the work presented here: Concerns around (1) consent and (2) anonymity of Reddit users, around the (3) selection, category labels, and assignment of user characteristics (MH diagnoses, age, country, gender), and (4) potentially harmful uses of the presented dataset and methods. The Lancaster University Faculty of Health and Medicine research ethics committee reviewed and approved this study in May 2019 (reference number FHMREC18066).

3.1 Consent

If and how research on social media data needs to obtain informed consent is debated (Eysenbach and Till, 2001; Beninger et al., 2014; Paul and Dredze, 2017), mainly because it is not straightforward to determine if posts pertain to a public or private context. Legally, the Reddit privacy policy[5] explicitly allows copying of user contents by third parties via the Reddit API, but it is unclear to what extends users are aware of this (Ahmed et al., 2017). In practice it is often infeasible to seek retrospective consent from hundreds or thousands of social media users. Current ethical guidelines for social media research (Benton et al., 2017; Williams et al., 2017) and practice in comparable research projects (O'Dea et al., 2015; Ahmed et al., 2017), regard it as acceptable to waive explicit consent if users' anonymity is protected. Therefore, Reddit users in this work were not asked for consent.

3.2 Anonymity

In line with guidelines for ethical social media health research (Benton et al., 2017), this research only shares anonymised and paraphrased excerpts from posts in publications. Otherwise, it is often possible to recover usernames via a web search with the verbatim post text (see also Section 3.5).

3.3 Rationales for user characteristics

As stated in the introduction, user characteristics are important to determine about which populations research on this dataset may generalise. The NLP community increasingly expects data statements for datasets (Bender and Friedman, 2018), which include speaker age and gender specifications. As Section 4.3 shows, characteristics of Reddit users with a self-reported BD diagnosis deviate from both general Reddit user statistics and epidemiological studies, which therefore do not constitute useful proxies. Relying entirely on self-reported information introduces selection biases because not all user groups may be equally inclined to explicitly share certain characteristics. This motivates using statistical methods to infer Reddit users' age, country, and gender here.

The user characteristics comorbid MH issues, age, country, and gender were chosen because they impact peoples' lived experience in BD as discussed in the following. This work identifies users with a self-reported BD diagnosis because collecting posts from BD-specific subreddits does not suffice as carers and people who are unsure if

[5] `https://www.redditinc.com/policies/privacy-policy`

they meet diagnostic criteria also post there. Other self-reported MH diagnoses were extracted because people with BD diagnoses frequently experience additional MH issues (Merikangas et al., 2011). Self-reported diagnoses capture only users who explicitly and publicly share their diagnosis. This research does not infer any users' MH state.

Depp and Jeste (2004), among others, provide evidence for age-related differences in BD symptoms and experiences, also through increasing importance of physical health comorbidities with ageing. Age estimates were grouped in the same way as in a US survey of Reddit users for comparison.

Healthcare systems, including provision of MH care, vastly differ between countries, even within Western countries such as the US, UK, and Germany. The MH services people can access may influence their experience of BD, motivating estimation of their country of residence. While Harrigian (2018) predicts longitude/latitude coordinates in 0.5 steps, these are mapped to countries because more fine-grained user localisations are not needed.

Using a gender variable in NLP deserves special consideration because it concerns people's identity (Larson, 2017). Biological sex can impact on the experience of BD, primarily through issues around childbirth and menopause, also related to mood-impacting hormonal changes (Diflorio and Jones, 2010); Sajatovic et al. (2011) found effects of gender identity on treatment adherence in BD. This work only uses binary m/f gender labels since no NLP method with more diverse categories was available. The gender recognition methods could cause harm to individual users if they were misgendered and then incorrectly addressed or referred to. This project minimises such harm because the labels only serve to estimate the gender distribution and not to target individual users.

3.4 Dual use

This research aims to learn more about Reddit users who share their experiences with BD to yield findings that will ultimately lead to new or improved interventions that support living well with BD. However, most research, even when conducted with the best intentions, suffers from the dual-use problem (Jonas, 1984), in that it can be misused or have consequences that affect people's life negatively. Adverse consequences of this study could arise for the Reddit users included in the dataset if they are sought out based on their self-reported BD diagnosis to be targeted with, e.g. medication advertisements. The large number of Reddit posts in this dataset can serve as training data for machine learning systems that assign a likelihood to other Reddit/social media users for meeting BD criteria (e.g., Cohan et al., 2018; Sekulić et al., 2018). For example, health insurance companies could misuse this, using applicants' social media profiles in risk assessments.

3.5 Transparency: Dataset and code release

Based on all above considerations, the dataset will only be shared with other researchers upon request and under a data usage agreement that specifies ethical usage of the dataset as detailed in this section. The dataset release necessarily contains the original post texts but with replaced post and user ids. This requires verbatim web searches with the post texts to seek out individual Reddit users and thus complicates automatisation and scaling. User characteristics, including the manually annotated subsets, will only be shared separately with researchers who justify a specific need for them. To aid transparency, the code and patterns to identify self-reported MH diagnoses, age, and gender are released[6].

Variable	Users	Agreement (%)	Labels (%)
Self-rep. BD diag.	100	97.0	Yes: 97.0 No: 3.0
Date of birth	116	99.1	Date: 90.5 ?: 19.5
Country	100	90.0	US: 46.0 CA: 9.0 GB: 8.0 Other: 25.0 ?: 12.0
Gender	116	95.7	F: 51.7 M: 34.5 Trans: 0.9 ?: 13.8

Table 2: Number of users in manual annotation, raw annotator agreement, and label distributions after resolving disagreements in discussion (?: no label assigned due to lack of user-provided information on Reddit)

[6] https://github.com/glorisonne/reddit_bd_user_characteristics

Variable	Users^{test}	Method	Accuracy^{test}	Coverage^{test}	Coverage^{all}
Age group	105	Self-reported	100.0%	98.1%	11.5%
		Language use	60.6%	94.3%	66.0%
		Hybrid	99.0%	100%	68.3%
Country	88	Words, subreddits, timing	78.4%	100%	100%
Gender	100	Username	100%	12.0%	10.9%
		Self-reported	97.9%	94.0%	11.9%
		Language use	84.2%	95.0%	66.0%
		Hybrid	97.0%	100%	71.5%

Table 3: Accuracy ($\frac{\text{correct}}{\text{total}}$) for user metadata extraction and inference methods (see Section 2.2) for manually annotated users (test), coverage ($\frac{\text{predicted}}{\text{total}}$) for manually annotated (test) and all (all, n=19,685) users

4 Results and discussion

The self-reported BD diagnosis matching method identified 19,685 Reddit users who together had 21,407,595 public Reddit posts between March 2006 and March 2019. Compared to 9K unique user accounts who posted in the four largest BD-related subreddits in May 2020, this likely only constitutes a small fraction of Reddit users with a BD diagnosis that could be reliably automatically identified (see following subsection).

4.1 Manual annotation

Two authors manually annotated random subsets of users to evaluate all automatically extracted or inferred information according to the annotation guidelines[7]. As shown in Table 2 agreement for all annotations was above 90%, demonstrating feasibility and high reliability.

The annotators checked all extracted self-reported bipolar disorder diagnosis statements of 100 random included users, disagreeing only for three users (see first line of Table 2)[8]. The pattern matching approach for self-reported diagnosis statements mistakenly identified only three users (subsequently removed from the dataset) based on reports of other MH diagnoses where the word bipolar occurred close to the diagnosis term as well[9].

To facilitate manual age and gender annotation, 116 users where randomly selected from the 2854 (14%) of users where the Self-reported age or gender extraction method matched. This explains the discrepancy between the coverage of the Self-reported method in Table 3 for the test set and full dataset. The annotators only checked whether date of birthor gender could be unambiguously extracted from all of a users' posts that matched a self-reported age and gender pattern. The test set for the gender evaluation results in Table 3 comprises only users labelled as m/f and excludes one manually identified transgender person.

4.2 Evaluation of NLP methods

Table 3 shows accuracy and coverage for the user characteristics extraction and inference methods described in Section 2.2 against the manually labelled users for which the annotators could determine a label. For age, the Self-reported method outperforms the Language use method for accuracy but not coverage[10]. The Hybrid method, subsequently used in Section 4.3.2, achieves 99% test set accuracy and 68% coverage on the full dataset. Harrigian's (2018) method assigns a country estimate to every user with 78% test set accuracy. For gender, accuracy decreases from the Username, Self-reported, and Language use method, while coverage increases [11]. The Hybrid gender identification method, used in Section 4.3.2, achieves 97% test set accuracy, gender-labelling 72% of users.

[7] `https://github.com/glorisonne/reddit_bd_user_characteristics/blob/master/ManualAnnotationGuidelines.pdf`

[8] No attempt was made to evaluate recall of user identification. Given an international prevalence of meeting BD criteria of about 2% (Merikangas et al., 2011) and expecting numbers of posts per account close to the average of 1,224 in the collected dataset, it was deemed infeasible to manually check all posts of randomly selected user accounts for self-reported bipolar disorder diagnosis statements.

[9] Paraphrased excerpts of incorrectly identified self-reported BD diagnoses: 'clinical depression with bipolar tendencies', 'diagnosed with BPD today, thought it was BD for years', 'diagnosed with depression, but sure I've got bipolar'.

[10] The Language use method for age/gender does not have full coverage because it requires at least ten posts per user. The methods agree for 62.6% of the 1,788 users where both assign an age group.

[11] For 195 users where all three methods assign a gender identity, they agree on 73.8% (90.8% agreement between the Username and Self-reported method, 80% between the Language use and Username or Self-reported method).

Diagnosis	Dataset n=19,685 (%)	SMHD n=6,434 (%)	Epidemiological studies (%)
MDD	30.2	27.4	N/A
Anxiety disorder	15.8	12.8	13.3-16.8*, n=921-1,537
ADHD	12.9	9.6	17.6†, n=399
BPD	8.4	N/A	16$, n=1,255
PTSD	6.5	5.1	10.8*, n=1,185
OCD	3.9	3.4	10.7*, n=808
ASD	2.2	2.0	Unknown
ED	1.0	0.8	5.3-31⊙, n=51-1,710

Table 4: Self-reported comorbid diagnoses with BD in this work, the SMHD dataset, and epidemiological studies: *Nabavi et al. (2015), †(McIntyre et al., 2010), $(Zimmerman and Morgan, 2013), ⊙(Álvarez Ruiz and Gutiérrez-Rojas, 2015)

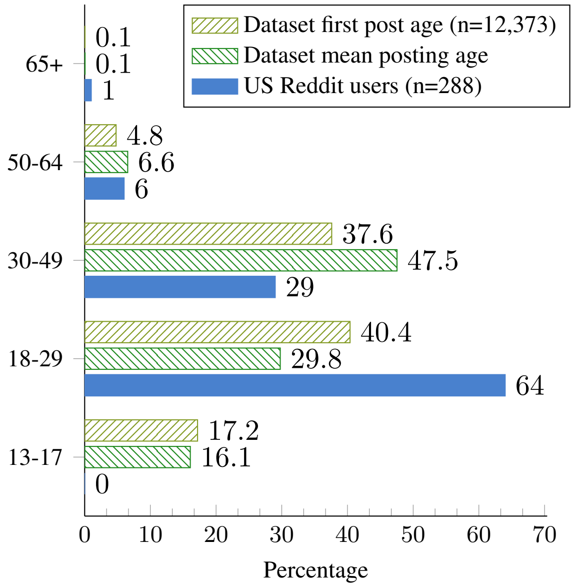

Figure 1: Age of Reddit users

4.3 Reddit users' characteristics

The following subsections compare characteristics of Reddit users with a self-reported BD diagnosis to general Reddit users and epidemiological statistics.

4.3.1 Mental health comorbidities

Table 4 shows how many users disclosed other concurrent or lifetime MH diagnoses besides BD. Rates for self-reported MH diagnoses in addition to BD are sightly higher in our dataset compared to the Self-reported MH diagnoses (SMHD) dataset (Cohan et al., 2018), potentially because our dataset covers 27 more months of posts.

Like psychotic disorder (5.2% of users prior to exclusion), a MDD diagnosis is mutually exclusive with BD according to the DSM (American Psychiatric Association, 2013, pp. 126, 134)[12]. A large part of identified self-reported MDD diagnoses were false positives where 'depression' occurred near to a BD diagnosis statement. More conservatively only considering self-reported MDD diagnosis posts that do not also match BD patterns, results in 8.7% users reporting both diagnoses. MDD and Psychotic disorder diagnoses jointly with BD might indicate subsequently changed (mis-)diagnoses or disagreement of professionals. Surveys in Germany (Pfennig et al., 2011) and the US (Hirschfeld et al., 2003) have shown that often more than ten years pass between onset of BD symptoms and receiving the diagnosis, with two thirds of people being misdiagnosed, most frequently with MDD. Moreover, field trials for BD diagnoses with DSM-V criteria only showed moderate clinician agreement (Freedman et al., 2013).

Comorbidity rates for anxiety disorders, BPD and PTSD align with results from epidemiological studies. Rates for comorbid ADHD, OCD, and ED are lower in the Reddit dataset population, which might in part be due to incomplete coverage of the patterns to capture diagnosis self-reports. Additionally, epidemiological studies can be expected to yield higher comorbidity rates because they determine if participants meet criteria for various diagnoses with clinical interviews, whereas Reddit users may not have (or report) diagnoses for every condition they meet the criteria of. Overall, 50.7% of users reported at least one additional MH diagnosis, slightly less than three quarters of surveyed people in the World Mental Health Survey Initiative who met criteria for at least one other DSM-IV disorder besides BD (Merikangas et al., 2011).

More than 2% of users reported an ASD diagnosis in addition to BD, with no epidemiological studies on ASD prevalence with BD yet. Dell'Osso et al. (2019) found significant levels of autistic traits among 43% of people with a BD diagnosis.

4.3.2 Age

As shown in Figure 1, less Reddit users with a self-reported BD diagnosis are 18-29 but more

[12] The dataset includes users with self-reported MDD but not psychotic disorder because depression but not psychosis is a core aspect of extreme mood, our focus of future research.

Country	Dataset (%)	Reddit.com traffic (%)	12-months prev. (%)
US	81.9	49.69	0.68
UK	5.6	7.93	1.11
Canada	4.9	7.85	0.75
Australia	1.7	4.32	1.15
Germany	1.4	3.17	0.83

Table 5: Top 5 estimated countries of residence of Reddit users with a self-reported BD diagnosis, location of reddit.com site visitors (Statista.com, 2020) and 12-months prevalence of BD (Global Burden of Disease Collaborative Network, 2018)

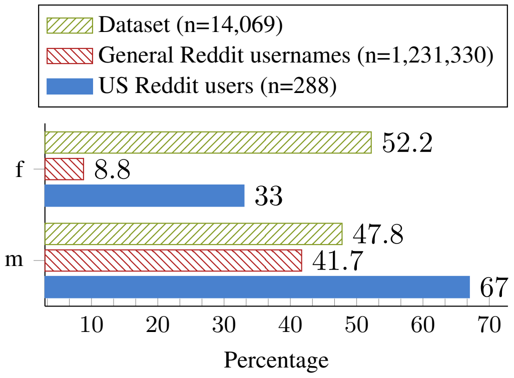

Figure 2: Binary gender of Reddit users

30-49 years old compared to average US Reddit users (Barthel et al., 2016, p. 7)[13]. The age of onset of BD symptoms is most frequently in late adolescence and early adulthood (Pini et al., 2005; Merikangas et al., 2011, p. 6). In line with this, the majority of Reddit users who disclose a BD diagnosis are between 13-29 years old at their first post. In the Global Burden of Disease study 2013, BD 12-months prevalence rates were significantly elated for 20-54 year olds Ferrari et al. (2016, p. 447). In our dataset, almost 80% of the Reddit users are 18-49 years old at their first post.

4.3.3 Country of residence

As shown in Table 5, more than 80% of the Reddit users with a self-reported BD diagnosis are estimated to live in the US, and 95% in one of the English-speaking countries US, UK, Canada, Australia. This ranking aligns with site visitors of the Reddit desktop version (Statista.com, 2020), although US users are even more prevalent in the BD dataset. All of the top-5 countries in the dataset have a 12-months prevalence of BD diagnoses above the global average of 0.62% according to the 2017 Global Burden of Disease Study (Global Burden of Disease Collaborative Network, 2018).

4.3.4 Binary gender

Figure 2 shows that the Hybrid method assigned feminine gender to slightly more than half of the Reddit users for which it ascribed a gender identity. This sharply contrasts with only 9% feminine vs. 41% masculine gender-performing usernames among Reddit users who posted in the top 10K subreddits with most posts (Wang and Jurgens, 2018). A survey of adult US Reddit users (Barthel et al., 2016) found that two thirds were men.

In epidemiological studies, biological men and women are equally likely to meet criteria for BD overall (Pini et al., 2005, American Psychiatric Association, 2013, p. 124) although there is evidence that BD-II is more frequently diagnosed among women (Diflorio and Jones, 2010). Sajatovic et al. (2011) found that biological men with a BD diagnosis scored significantly lower on masculine gender identity than the general male population, while there were no gender identity differences for biological women. Considering a majority of male Reddit users and sex-equal prevalence of the diagnosis, feminine-gender-identifying people with a BD diagnosis seem to be more likely to use Reddit and/or to disclose their diagnosis. The increased rates of female-gender identifying Reddit users with a self-reported BD diagnosis might also point towards a higher relative frequency of BD-II diagnoses (compared to BD-I) in this population.

5 Limitations and implications

5.1 Limitations

First, unlike in clinical studies with face-to-face interactions, we cannot assume that every Reddit user in the dataset corresponds to one person. Additionally, self-reported diagnoses cannot be confirmed with diagnostic interviews as in clinical research.

Furthermore, there are several limitations to the NLP methods to infer user characteristics. The method to extract self-reported MH diagnoses does not distinguish between actual comorbidities and misdiagnoses or previous diagnoses, for which symptoms may have resolved. Manual evaluation of ten users with BPD comorbidity showed that seven reported concurrent diagnoses, one a BD to BPD change, one a BPD misdiagnosis, and one re-

[13] The Barthel et al. (2016) survey only targeted adults, therefore there are no 13-17-year-old users.

ferred to BD by 'BPD'. Harrigian's (2018) method indicates the predominantly reflected country in a user's most recent posts, disregarding relocations.

The Self-reported age and gender extraction method is fallible to users providing incorrect information, for example disguising themselves as younger than they really are on dating subreddits. Finally, none of the gender inference methods allow us to estimate how many users identify as transgender or non-binary. Such indications were also too diverse to be captured in the regular expressions for self-reported age and gender. Still, four of the subreddits with more than 10K posts by users with a self-reported BD diagnosis target transgender people, indicating that a proportion of the users in this research may not identify with their born sex.

5.2 Health research implications

Most importantly this work provides the first large-scale characterisation of Reddit users with a self-reported BD diagnosis, who are on average 27.7 years old at their first post, seem to overwhelmingly live in the US, and are more likely to identify with the feminine gender. Insofar they deviate from general Reddit as well as epidemiological statistics and also from participants in clinical studies.

A large meta-analysis of psychological interventions for BD (Oud et al., 2016) showed that in 55 trials conducted across twelve countries (35% in the US) comprising 6,060 adults with BD, 89% had recruited participants with a mean age higher than the 30 year-average of adult Reddit users with a self-reported BD diagnosis. 67% of the trials recruited a higher percentage of females than the 52% figure in the Reddit dataset (Oud et al., 2016, Table DS2). This cautions against generalising findings from Reddit data to all people with a BD diagnosis, but stresses its complementary role to clinical studies with different selection biases.

Another important implication is that NLP analysis of Reddit social media users largely confirmed high prevalence rates for comorbid MH conditions with BD from epidemiological studies. Besides clinically established comorbidities with, e.g., Anxiety disorder and ADHD, the present analysis also revealed substantial prevalence of ASD, for which there is little clinical research to date. Reddit may constitute a useful platform to learn about the experiences of people with BD with such currently under-researched comorbidities and may be a way to target them for recruitment to clinical studies.

5.3 NLP research implications

This work evaluated state-of-the-art methods to infer Reddit user characteristics (Harrigian, 2018; Wang and Jurgens, 2018; Tigunova et al., 2019) and demonstraed their utility in applied research. A hybrid method achieved the best accuracy and coverage for age and gender identity by using high-accuracy information from self-reports (or a gender-performing username) when available, filling in information for more users with less accurate predictions from a neural network language use-based method (Tigunova et al., 2019).

Importantly, gender-inference methods so far are limited to detecting binary gender, although, e.g., 0.4% of the US population identify as transgender (Meerwijk and Sevelius, 2017). Off-the-shelf NLP tools supporting a wider range of gender identities may be more inclusive and give more visibility to these groups of people in research. However, important ethical considerations arise around identifying people with transgender and non-binary gender identities, which are often stigmatised.

6 Conclusion

This paper set out to automatically profile Reddit users under consideration of ethical aspects. A combination of pattern-based and previously published NLP methods served to estimate clinical, demographic, and identity characteristics of nearly 20K Reddit users with a self-reported BD diagnosis. Half of the Reddit users disclosed MH diagnoses besides BD and 80% were located in the US. From the users for which age or gender could be estimated, 80% were between 18-49 years old and 52% performed or identified with feminine gender.

These findings indicate about which populations BD-focused research on Reddit may generalise. Additionally, this work may serve as a model for how to provide more information on other specific Reddit populations as requested by recent transparency and accountability movements in NLP.

Acknowledgements

We would like to thank Anna Tigunova and Keith Harrigian for their assistance in applying their Reddit user profiling NLP tools. We would also like to express our heartfelt thanks to Daisy Harvey, Stephen Mander, and the anonymous reviewers for helpful comments on a draft version of this article, to Andrew Moore for testing the code release, and to Alistair Baron for the initial idea for this work.

References

Wasim Ahmed, Peter A. Bath, and Gianluca Demartini. 2017. Using Twitter as a data source: an overview of ethical, legal and methodological challenges. In Kandy Woodfield, editor, *The Ethics of Online Research*, pages 79–107. Emerald Books.

Nadia Akers, Fiona Lobban, Claire Hilton, Katerina Panagaki, and Steven H. Jones. 2019. Measuring social and occupational functioning of people with bipolar disorder: A systematic review. *Clinical Psychology Review*, 74.

Alexa Internet. 2020. reddit.com.

Eva M. Álvarez Ruiz and Luis Gutiérrez-Rojas. 2015. Comorbidity of bipolar disorder and eating disorders. *Revista de Psiquiatria y Salud Mental*, 8(4):232–241.

Ashley Amaya, Ruben Bach, Florian Keusch, and Frauke Kreuter. 2019. New Data Sources in Social Science Research: Things to Know Before Working With Reddit Data. *Social Science Computer Review*, pages 1–18.

American Psychiatric Association. 2013. *DSM-5*. Washington, DC.

Michael Barthel, Galen Stocking, Jesse Holcomb, and Amy Mitchell. 2016. Nearly Eight-in-Ten Reddit Users Get News on the Site. Technical report.

Rita Bauer, Michael Bauer, Hermann Spiessl, and Tanja Kagerbauer. 2013. Cyber-support: An analysis of online self-help forums (online self-help forums in bipolar disorder). *Nordic Journal of Psychiatry*, 67(3):185–190.

Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Kelsey Beninger, Alexandra Fry, Natalie Jago, Hayley Lepps, Laura Nass, and Hannah Silvester. 2014. Research using Social Media; Users' Views.

Adrian Benton, Glen Coppersmith, and Mark Dredze. 2017. Ethical Research Protocols for Social Media Health Research. In *Proceedings of the First Workshop on Ethics in Natural Language Processing*, pages 94–102.

Alexandra Budenz, Ann Klassen, Jonathan Purtle, Elad Yom Tov, Michael Yudell, and Philip Massey. 2019. Mental illness and bipolar disorder on Twitter: implications for stigma and social support. *Journal of Mental Health*, 29(2):191–199.

Iain H. Campbell and Harry Campbell. 2019. Ketosis and bipolar disorder: controlled analytic study of online reports. *BJPsych Open*, 5(4):1–6.

Arman Cohan, Bart Desmet, Sean Macavaney, Andrew Yates, Luca Soldaini, Sean Macavaney, and Nazli Goharian. 2018. SMHD: A Large-Scale Resource for Exploring Online Language Usage for Multiple Mental Health Conditions. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, pages 1485–1497.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying Mental Health Signals in Twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 51–60.

Glen Coppersmith, Mark Dredze, Craig Harman, and Kristy Hollingshead. 2015. From ADHD to SAD: Analyzing the Language of Mental Health on Twitter through Self-Reported Diagnoses. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*.

Liliana Dell'Osso, Barbara Carpita, Carlo Antonio Bertelloni, Elisa Diadema, Filippo Maria Barberi, Camilla Gesi, and Claudia Carmassi. 2019. Subthreshold autism spectrum in bipolar disorder: Prevalence and clinical correlates. *Psychiatry Research*, 281(October):112605.

Colin A. Depp and Dilip V. Jeste. 2004. Bipolar disorder in older adults: A critical review. *Bipolar Disorders*, 6(5):343–367.

Arianna Diflorio and Ian Jones. 2010. Is sex important Gender differences in bipolar disorder. *International Review of Psychiatry*, 22(5):437–452.

Gunther Eysenbach and James E. Till. 2001. Ethical issues in qualitative research on internet communities. *BMJ*, 323(7055):1103–1105.

Alize J. Ferrari, Emily Stockings, Jon Paul Khoo, Holly E. Erskine, Louisa Degenhardt, Theo Vos, and Harvey A. Whiteford. 2016. The prevalence and burden of bipolar disorder: findings from the Global Burden of Disease Study 2013. *Bipolar Disorders*, 18(5):440–450.

Nigel G. Fielding, Raymond M. Lee, Grant Blank, and Dietmar Janetzko. 2016. Nonreactive Data Collection Online. *The SAGE Handbook of Online Research Methods*, pages 76–91.

Robert Freedman, David A. Lewis, Robert Michels, Daniel S. Pine, Susan K. Schultz, Carol A. Tamminga, Glen O. Gabbard, Susan Shur-Fen Gau, Daniel C. Javitt, Maria A. Oquendo, Patrick E. Shrout, Eduard Vieta, and Joel Yager. 2013. The Initial Field Trials of DSM-5: New Blooms and Old Thorns. *American Journal of Psychiatry*, 170(1):1–5.

Andréanne Gignac, Alexander McGirr, Raymond W Lam, and Lakshmi N Yatham. 2015. Recovery and recurrence following a first episode of mania: a systematic review and meta-analysis of prospectively characterized cohorts. *The Journal of clinical psychiatry*, 76(9):1241–1248.

George Gkotsis, Anika Oellrich, Tim Hubbard, Richard Dobson, Maria Liakata, Sumithra Velupillai, and Rina Dutta. 2016. The language of mental health problems in social media. In *Proceedings of the Third Workshop on Computational Lingusitics and Clinical Psychology*, pages 63–73.

George Gkotsis, Anika Oellrich, Sumithra Velupillai, Maria Liakata, Tim J.P. Hubbard, Richard J.B. Dobson, and Rina Dutta. 2017. Characterisation of mental health conditions in social media using Informed Deep Learning. *Scientific Reports*, 7:1–11.

Global Burden of Disease Collaborative Network. 2018. Global Burden of Disease Study 2017 (GBD 2017) Results. Technical report, Institute for Health Metrics and Evaluation (IHME), Seattle, United States.

G. M. Goodwin, P. M. Haddad, I. N. Ferrier, J. K. Aronson, T. R.H. Barnes, A. Cipriani, D. R. Coghill, S. Fazel, J. R. Geddes, H. Grunze, E. A. Holmes, O. Howes, S. Hudson, N. Hunt, I. Jones, I. C. MacMillan, H. McAllister-Williams, D. R. Miklowitz, R. Morriss, M. Munafò, C. Paton, B. J. Saharkian, K. E.A. Saunders, J. M.A. Sinclair, D. Taylor, E. Vieta, and A. H. Young. 2016. Evidence-based guidelines for treating bipolar disorder: Revised third edition recommendations from the British Association for Psychopharmacology. *Journal of Psychopharmacology*, 30(6):495–553.

Keith Harrigian. 2018. Geocoding without geotags: a text-based approach for reddit. In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*, pages 17–27.

Robert M. A. Hirschfeld, Lydia Lewis, and Lana A. Vornik. 2003. Perceptions and impact of bipolar disorder: How far have we really come? Results of the National Depressive and Manic-Depressive Association 2000 survey of individuals with bipolar disorder. *The Journal Of Clinical Psychiatry*, 64(2):161–174.

Yen-Hao Huang, Yi-Hsin Chen, Fernando H. Calderon, Ssu-Rui Lee, Shu-I Wu, Yuwen Lai, and Yi-Shin Chen. 2019. Leveraging Linguistic Characteristics for Bipolar Disorder Recognition with Gender Differences. In *Proceedings of the 2019 KDD Workshop on Applied Data Science for Healthcare (DSHealth '19)*.

Xiang Ji, Soon Ae Chun, Zhi Wei, and James Geller. 2015. Twitter sentiment classification for measuring public health concerns. *Social Network Analysis and Mining*, 5(1):1–25.

Hans Jonas. 1984. *The Imperative of Responsibility: Foundations of an Ethics for the Technological Age*. University of Chicago Press, Chicago.

Adam D.I. Kramer, Susan R. Fussell, and Leslie D. Setlock. 2004. Text analysis as a tool for analyzing conversation in online support groups. *Conference on Human Factors in Computing Systems - Proceedings*, pages 1485–1488.

Brian Larson. 2017. Gender as a Variable in Natural-Language Processing: Ethical Considerations. In *Proceedings of the First Workshop on Ethics in Natural Language Processing*, pages 1–11.

Klara Latalova, Jan Prasko, Dana Kamaradova, Katerina Ivanova, Lubica Jurickova, Latalova K., Prasko J., Kamaradova D., and Ivanova K. 2014. Bad on the net, or bipolars' lives on the web: Analyzing discussion web pages for individuals with bipolar affective disorder. *Neuro Endocrinology Letters*, 35(3):206–212.

Anika Mandla, Jo Billings, and Joanna Moncrieff. 2017. "Being Bipolar": A Qualitative Analysis of the Experience of Bipolar Disorder as Described in Internet Blogs. *Issues in Mental Health Nursing*, 38(10):858–864.

D McDonald and R Woodward-Kron. 2016. Member roles and identities in online support groups: Perspectives from corpus and systemic functional linguistics. *Discourse and Communication*, 10(2):157–175.

Roger S McIntyre, Sidney H Kennedy, Joanna K Soczynska, Ha T T Nguyen, Timothy S Bilkey, Hanna O Woldeyohannes, Jay A Nathanson, Shikha Joshi, Jenny S H Cheng, Kathleen M Benson, and David J Muzina. 2010. Attention-deficit/hyperactivity disorder in adults with bipolar disorder or major depressive disorder: results from the international mood disorders collaborative project. *Primary Care Companion To The Journal Of Clinical Psychiatry*, 12(3).

Esther L. Meerwijk and Jae M. Sevelius. 2017. Transgender population size in the United States: A meta-regression of population-based probability samples. *American Journal of Public Health*, 107(2):e1–e8.

Kathleen R. Merikangas, Robert Jin, Jian-ping He, Ronald C. Kessler, Sing Lee, Nancy A. Sampson, Maria Carmen Viana, Laura Helena Andrade, Chiyi Hu, Elie G. Karam, Maria Ladea, Maria Elena Medina Mora, Mark Oakley Browne, Yutaka Ono, Jose Posada-Villa, Rajesh Sagar, and Zahari Zarkov. 2011. Prevalence and correlates of bipolar spectrum disorder in the world mental health survey initiative. *Archives of general psychiatry*, 68(3):241–251.

Behrouz Nabavi, Alex J. Mitchell, and David Nutt. 2015. A Lifetime Prevalence of Comorbidity Between Bipolar Affective Disorder and Anxiety Disorders: A Meta-analysis of 52 Interview-based Studies of Psychiatric Population. *EBioMedicine*, 2(10):1405–1419.

Danielle M. Novick, Holly A. Swartz, and Ellen Frank. 2010. Suicide attempts in bipolar I and bipolar II disorder: A review and meta-analysis of the evidence. *Bipolar Disorders*, 12(1):1–9.

Bridianne O'Dea, Stephen Wan, Philip J. Batterham, Alison L. Calear, Cecile Paris, and Helen Christensen. 2015. Detecting suicidality on Twitter. *Internet Interventions*, 2(2):183–188.

Matthijs Oud, Evan Mayo-Wilson, Ruth Braidwood, Peter Schulte, Steven H. Jones, Richard Morriss, Ralph Kupka, Pim Cuijpers, and Tim Kendall. 2016. Psychological interventions for adults with bipolar disorder: Systematic review and meta-analysis. *British Journal of Psychiatry*, 208(3):213–222.

Michael J. Paul and Mark Dredze. 2017. Social Monitoring for Public Health. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 9(5):1–183.

Andrea Pfennig, B. Jabs, S. Pfeiffer, B. Weikert, K. Leopold, and M. Bauer. 2011. Health care service experiences of bipolar patients in Germany - Survey prior to the introduction of the S3 Guideline for diagnostics and treatment of bipolar disorders. *Nervenheilkunde*, 30(5):333–340.

Stefano Pini, Valéria De Queiroz, Daniel Pagnin, Lukas Pezawas, Jules Angst, Giovanni B. Cassano, and Hans Ulrich Wittchen. 2005. Prevalence and burden of bipolar disorders in European countries. *European Neuropsychopharmacology*, 15(4):425–434.

Ria Poole, Daniel Smith, and Sharon Simpson. 2015. How Patients Contribute to an Online Psychoeducation Forum for Bipolar Disorder: A Virtual Participant Observation Study. *JMIR Mental Health*, 2(3:e21).

Daniel Preoţiuc-Pietro, Johannes Eichstaedt, Gregory Park, Maarten Sap, Laura Smith, Victoria Tobolsky, H. Andrew Schwartz, and Lyle Ungar. 2015. The role of personality, age, and gender in tweeting about mental illness. In *Proceedings ofthe 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 21–30.

Derek Ruths and Jürgen Pfeffer. 2014. Social media for large studies of behavior. *Science*, 346(6213):1063–1064.

Puneet K.C. Sahota and Pamela L. Sankar. 2019. Bipolar Disorder, Genetic Risk, and Reproductive Decision-Making: A Qualitative Study of Social Media Discussion Boards. *Qualitative Health Research*.

Martha Sajatovic, Weronika Micula-Gondek, Curtis Tatsuoka, and Christopher Bialko. 2011. The relationship of gender and gender identity to treatment adherence among individuals with bipolar disorder. *Gender Medicine*, 8(4):261–268.

Elvis Saravia, Chun Hao Chang, Renaud Jollet De Lorenzo, and Yi Shin Chen. 2016. MIDAS: Mental illness detection and analysis via social media. *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2016*, pages 1418–1421.

Ivan Sekulić, Matej Gjurković, and Jan Šnajder. 2018. Not Just Depressed: Bipolar Disorder Prediction on Reddit. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment Social Media Analysis (WASSA)*, pages 72–78.

Statista.com. 2020. Regional distribution of desktop traffic to Reddit.com as of September 2020, by country.

Anna Tigunova, Paramita Mirza, Andrew Yates, and Gerhard Weikum. 2019. Listening between the lines: Learning personal attributes from conversations. *The Web Conference 2019 - Proceedings of the World Wide Web Conference, WWW 2019*, 2:1818–1828.

Anna Tigunova, Andrew Yates, Paramita Mirza, and Gerhard Weikum. 2020. RedDust: a Large Reusable Dataset of Reddit User Traits. In *Proceedings ofthe 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 6118–6126.

T. Treuer and M. Tohen. 2010. Predicting the course and outcome of bipolar disorder: A review. *European Psychiatry*, 25(6):328–333.

Agnès Vayreda and Charles Antaki. 2009. Social Support and Unsolicited Advice in a Bipolar Disorder Online Forum. *Qualitative Health Research*, 19(7):931–942.

Zijian Wang and David Jurgens. 2018. It's going to be okay: Measuring Access to Support in Online Communities. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 33–45.

Helen Warwick, Sara Tai, and Warren Mansell. 2019. Living the life you want following a diagnosis of bipolar disorder: A grounded theory approach. *Clinical Psychology Psychotherapy*.

Matthew L. Williams, Pete Burnap, and Luke Sloan. 2017. Towards an Ethical Framework for Publishing Twitter Data in Social Research: Taking into Account Users' Views, Online Context and Algorithmic Estimation. *Sociology*, 51(6):1149–1168.

World Health Organisation. 2018. *International classification of diseases for mortality and morbidity statistics (11th Revision)*.

Minjoo Yoo, Sangwon Lee, and Taehyun Ha. 2019. Semantic network analysis for understanding user experiences of bipolar and depressive disorders on Reddit. *Information Processing and Management*, 56(4):1565–1575.

Mark Zimmerman and Theresa A. Morgan. 2013. Problematic Boundaries in the Diagnosis of Bipolar Disorder: The Interface with Borderline Personality Disorder. *Current Psychiatry Reports*, 15(422).

A Further method details

A.1 Age and gender: Language use

Tigunova et al.'s (2019) HAM$_{\text{CNN-attn}}$ model predicts an age group[14] and gender for Reddit users with at least ten posts based on their up to 100 most recent posts. Separate HAM$_{\text{CNN-attn}}$ models were trained on the RedDust dataset (Tigunova et al., 2020) with the HAM open-source implementation[15] with the hyper-parameters specified by Tigunova et al. (2020) (128 CNN filters of size 2, attention layer with 150 units, 70 training epochs). Likely due to random seed variation, our trained age model had an area under the curve (AUROC) score of 0.80 compared to 0.88 in Tigunova et al. (2020). Our trained gender model had 84.9% accuracy on the RedDust test set compared to 86.0% reported by Tigunova et al. (2020).

A.2 Age: Hybrid method

Two corrections were applied prior to the Hybrid method: The first author checked all users with a self-reported average posting age below 16 or above 60. Age at account creation predictions younger than 13 by the Language use approach were discarded as Reddit requires an age of at least 13 when signing up.

A.3 Country

The Reddit country inference method (Harrigian, 2018) initially was a proprietary project but later the first author, Keith Harrigian, rebuilt it for the public release[16] used in this work. Therefore, the training data and model performance, provided by Keith Harrigian in personal email communication on 5th March 2021, slightly differ from the original publication. The training data consists of 56,853 automatically location-labelled users (top 5: 68.8% US, 9.4% Canada, 7.0% UK, 3.3% Australia, 1.0% Germany), of which 8.2% were identified based on self-reported locations in r/AmateurRoomPorn and the remainder by self-reported locations in reply to 'Where are you from?' questions (Harrigian, 2018). Label precision was 97.6% in a manual evaluation of 500 users[17].

The 'Global' (as opposed to US only) model was used to predict user locations, which achieves 35.6% accuracy at 100 miles in 5-fold cross validation, equal to the originally reported performance in Harrigian (2018). Overall country-level accuracy is 81.9% and is generally higher for users with more training data (95.1% US, 65.1% Canada, 82.8% UK, 44.1% Australia, 41.1% Germany).

A.4 Gender: Username method

Wang and Jurgens (2018, p. 38) trained their long short-term memory (LSTM) gender inference model on 80% of 4,900,250 Twitter and 367,495 Reddit usernames, automatically labelled with self-reported m or f gender identity. Following them, the present work assumes usernames to perform masculine (m) gender for model predictions of 0.1 or lower, and feminine (f) for 0.9 or higher. This model and setting achieved 0.92 precision with 0.18 recall in 10% held-out Twitter and Reddit username test data (Wang and Jurgens, 2018, Figure 5 in supplementary material).

[14] younger than 14, 14-23, 24-45, 46-65, 66+, relative to the user's most recent post

[15] https://github.com/Anna146/HiddenAttributeModels

[16] https://github.com/kharrigian/smgeo

[17] https://github.com/kharrigian/smgeo#dataset-noise

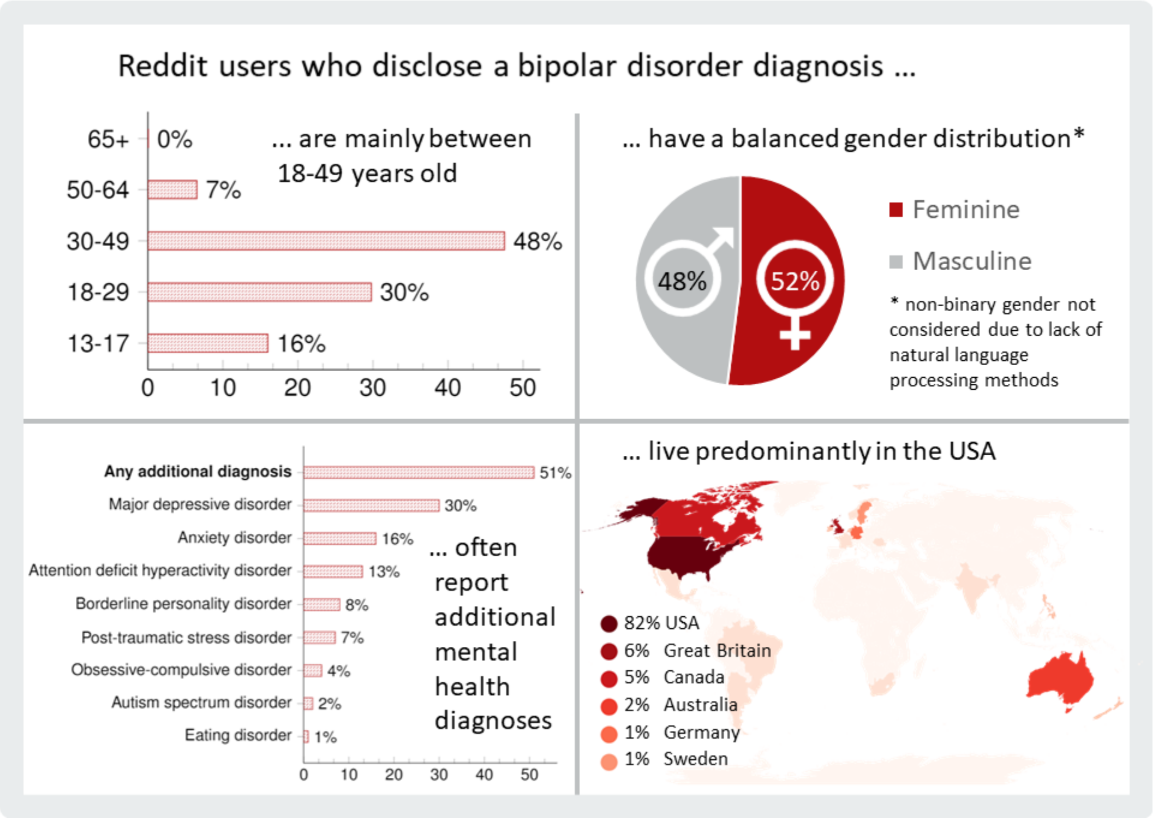

Figure 3: Visual summary of the characteristics of Reddit users who self-report a bipolar disorder diagnosis