Talking about lived experience in bipolar disorder: a corpus linguistic analysis of Reddit social media posts

Glorianna Jagfeld ∙ g.jagfeld@lancaster.ac.uk ∙ @glorisonne

Glorianna Jagfeld
Steven Jones, Fiona Lobban
Spectrum Centre for Mental Health Research

Paul Rayson
University Centre for Computer Corpus Research on Language
Outline

1. Background & project overview
2. Reddit data collection & corpus construction
3. Results
   a) User demographics
   b) Key semantic domains in bipolar subreddit posts
4. Next steps, open questions, limitations
What and why?
Background & motivation
Extreme mood experiences, bipolar disorder & recovery

• Mood is a continuum\(^1\):

\[ \text{mania} \quad \text{hypomania} \quad \text{depression} \]

\[ \text{elated} \quad \text{normal} \quad \text{low} \]

• Mania or hypomania + depression $\rightarrow$ bipolar disorder diagnosis

What constitutes recovery?

• Clinical: no symptoms for $\geq$ 8 weeks
• Personal: [aiming for] satisfying, hopeful, contributing life even with limitations caused by symptoms\(^2\)

---

Previous research on personal recovery in bipolar disorder

Few quantitative\textsuperscript{1-2} & qualitative\textsuperscript{3-11} studies

- Small samples
- Researcher-guided data production: interviews, questionnaires
- Mainly attracted (or only recruited) people who regard themselves in personal recovery

Potential of mixed-methods analysis of social media posts

- Larger sample
- More diverse participants
- Data production not researcher-influenced

Bipolar disorder studies with social media posts by field & method

Main analysis methods:
- Manual qualitative
- Manual mixed/quantitative
- Natural language processing
- Corpus linguistics

Health research
- 2
- 4
- 16
- 17
- 3

Natural language processing
- 1
- 5
- 8
- 14
- 19
- 11
- 10
- 18

Computer science
- 12

Natural sciences
- 15

Communication & discourse
- 16
Better understanding of personal recovery in bipolar disorder to make recommendations for mental health care improvements

Corpus linguistics:
Analysis

Computational linguistics:
Data collection + processing

Health research:
Personal recovery, implications
Study overview

1) Structured setting: Systematic review of qualitative research on personal recovery in bipolar disorder

2) Unstructured setting: Corpus linguistic analysis of Reddit social media posts

3) Combination: Comparison of social media posts with interview transcripts

1. Project proposal approved by faculty research ethics committee
How?
Data & methods
Reddit user identification

- Identify posts with self-reported diagnosis statements\(^1\):
  - 90 term variants for bipolar disorder\(^2\): manic depression, BD-I, ...
  - 145 diagnosis phrases with placeholders\(^2\):
    my <doctor> diagnosed me with <bipolar term>, ...
  - 74 exclusion phrases\(^2\): self diagnos*, ...
- Checked all Reddit posts from 06/2005-05/2019 (1.9 TB of data) via Google BigQuery\(^3\)
- 97% precision: manually verified matched diagnosis statements for 100 random included users

Dataset & corpus construction

- Download all posts of 20K identified users via Reddit API\(^1\) → 24M posts, 1.1 billion words
- Automatic language identification\(^2\) → keep only English (97%)
- Build corpus from posts in 7 bipolar subreddits (total 628K posts/60M words)
  - Contains 44% of all posts mentioning bipolar disorder
  - Balance #words per user to curtail long tail:
    - Remove 5% of users with fewest words (< 76)
    - Cap number of words (> 4855) for 20% most prolific users

1. [https://github.com/praw-dev/praw](https://github.com/praw-dev/praw)
2. Lui & Baldwin (2012)
Corpus processing

- POS tagging with CLAWS\textsuperscript{1} (137 tags)
- Semantic domain tagging with USAS\textsuperscript{2}
- Keywords & key semantic domains
  - Reference corpus: posts by control users from SMHD dataset\textsuperscript{5} selected to match size of bipolar corpus
  - Keyness: log likelihood\textsuperscript{3} (p<0.0001, Bonferroni correction)
  - Effect size: binary log of relative risk (‘log ratio’)\textsuperscript{4}
- Keyword context exploration with SketchEngine\textsuperscript{6}

# UCREL Semantic Analysis System (USAS)\(^1\)

<table>
<thead>
<tr>
<th>A: general and abstract terms</th>
<th>B: the body and the individual</th>
<th>C: arts and crafts</th>
<th>E: emotion</th>
</tr>
</thead>
<tbody>
<tr>
<td>F: food and farming</td>
<td>G: government and public</td>
<td>H: architecture, housing and the home</td>
<td>I: money and commerce in industry</td>
</tr>
<tr>
<td>K: entertainment, sports and games</td>
<td>L: life and living things</td>
<td>M: movement, location, travel and transport</td>
<td>N: numbers and measurement</td>
</tr>
<tr>
<td>O: substances, materials, objects and equipment</td>
<td>P: education</td>
<td>Q: language and communication</td>
<td>S: social actions, states and processes</td>
</tr>
<tr>
<td>T: time</td>
<td>W: world and environment</td>
<td>X: psychological actions, states and processes</td>
<td>Y: science and technology</td>
</tr>
<tr>
<td>Z: names and grammar</td>
<td>21 general domains, 232 subdomains</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1. Rayson et al. (2004)
Results
Dataset user demographics
# Country of residence

<table>
<thead>
<tr>
<th>Dataset rank</th>
<th>Country</th>
<th>Reddit dataset (N=19,816)</th>
<th>reddit.com traffic</th>
<th>Systematic review (N=88, NA=14)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>USA</td>
<td>82%</td>
<td>50%</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Great Britain</td>
<td>6%</td>
<td>8%</td>
<td>39%</td>
</tr>
<tr>
<td>3</td>
<td>Canada</td>
<td>5%</td>
<td>7%</td>
<td>15%</td>
</tr>
<tr>
<td>4</td>
<td>Australia</td>
<td>2%</td>
<td>4%</td>
<td>11%</td>
</tr>
<tr>
<td>5</td>
<td>Germany</td>
<td>1%</td>
<td>3%</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Sweden</td>
<td>1%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Norway</td>
<td>0.15%</td>
<td></td>
<td>22%</td>
</tr>
</tbody>
</table>

- Dataset prediction based on posts’ texts, subreddits posted in, posting time of day distribution

1 Harrigian (2018)
2 https://www.statista.com/statistics/325144/reddit-global-active-user-distribution/ (only data for top 5 countries available)
3 8 studies conducted 2010-2019, country of residence inferred from study country and recruitment strategy
Gender

- Self-reported age + gender in submission titles: 'Me [17f] just broke up with partner [17m]'
- Extract with regular expressions (gender: male/female only)

<table>
<thead>
<tr>
<th>Gender</th>
<th>Reddit dataset (N=2,344)(^1)</th>
<th>Adult US Reddit users(^2) (N=288)</th>
<th>Systematic review (N=88)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>65%</td>
<td>33%</td>
<td>66%</td>
</tr>
<tr>
<td>Male</td>
<td>35%</td>
<td>67%</td>
<td>34%</td>
</tr>
</tbody>
</table>

1 92% accuracy in manual evaluation of extracted gender for 86 random users 2 Barthel et al. (2016)
### Age

- Calculate date of birth with post timestamp + extracted age
- Calculate age for all posts with timestamp + date of birth

<table>
<thead>
<tr>
<th>Age</th>
<th>Reddit dataset (N= 2,271)</th>
<th>Adult US Reddit users (N=288)</th>
<th>Systematic review (N=88, NA=34)</th>
</tr>
</thead>
<tbody>
<tr>
<td>13-17</td>
<td>6%</td>
<td>N/A</td>
<td>0%</td>
</tr>
<tr>
<td>18-29</td>
<td>78%</td>
<td>64%</td>
<td>7%</td>
</tr>
<tr>
<td>30-49</td>
<td>15%</td>
<td>29%</td>
<td>52%</td>
</tr>
<tr>
<td>50-64</td>
<td>1%</td>
<td>6%</td>
<td>41%</td>
</tr>
<tr>
<td>65+</td>
<td>0%</td>
<td>1%</td>
<td>0%</td>
</tr>
</tbody>
</table>

1 User age in dataset = average over their post ages, 91% accuracy in manual evaluation of extracted date of birth for 100 users (13-16 years, mean 25)  
2 Barthel et al. (2016)  
3 Average participant age >= 16 study inclusion criterion; N=88: 24-68, mean 46
Corpus & key semantic domains
Bipolar subreddits & reference corpus statistics

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>Users</td>
<td>10K</td>
</tr>
<tr>
<td>Words</td>
<td>20.6M</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>156K</td>
</tr>
<tr>
<td>Posts</td>
<td>268K</td>
</tr>
<tr>
<td>Avg. words/posts</td>
<td>77</td>
</tr>
</tbody>
</table>

Key semantic domains:
144 overused, 230 underused

Keywords:
2,857 overused, 9,764 underused
## Key semantic domains in the bipolar subreddits corpus

<table>
<thead>
<tr>
<th>A: general and abstract terms</th>
<th>B: the body and the individual</th>
<th>C: arts and crafts</th>
<th>E: emotion</th>
</tr>
</thead>
<tbody>
<tr>
<td>F: food and farming</td>
<td>G: government and public</td>
<td>H: architecture, housing and the home</td>
<td>I: money and commerce in industry</td>
</tr>
<tr>
<td>K: entertainment, sports and games</td>
<td>L: life and living things</td>
<td>M: movement, location, travel and transport</td>
<td>N: numbers and measurement</td>
</tr>
<tr>
<td>O: substances, materials, objects and equipment</td>
<td>P: education</td>
<td>Q: language and communication</td>
<td>S: social actions, states and processes</td>
</tr>
<tr>
<td>T: time</td>
<td>W: world and environment</td>
<td>X: psychological actions, states and processes</td>
<td>Y: science and technology</td>
</tr>
</tbody>
</table>

20 statistically significant (log likelihood\(^1\), \(p < 0.0001\)) sub-domains overused least 2 times compared to control with frequency > 1K, covers 4% of all terms

---

Domain-specific grouping of key semantic domains\(^1\) with keywords\(^2\)

- Mental health symptoms (\(~400\)K mentions)
  - Bipolar disorder-specific
    - Mania: manic, mania, impulsive, hypo
    - Depression: depression, depressed, depressive, sad, cry
    - Extreme mood: mixed, extreme, mood swings
  - Other: anxiety, PTSD, ADHD, concentration, psychotic, addict
- Professional treatment (\(~200\)K): diagnosed, doctor, medication
- Recovery & self-management (\(~44\)K): recovery, wellbeing, stable, mindfulness, vigilant
- Misc: life: personal narrative; time: rate of recurrence, routines

---

1. Of 20 key semantic domains as defined on previous slide
2. Selected from 20 most frequent keywords per semantic category
Key semantic domains
B3: Medicines/medical treatment

<table>
<thead>
<tr>
<th>term</th>
<th>frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>diagnosed</td>
<td>20740</td>
</tr>
<tr>
<td>doctor</td>
<td>17524</td>
</tr>
<tr>
<td>medication</td>
<td>17498</td>
</tr>
<tr>
<td>psychiatrist</td>
<td>12949</td>
</tr>
<tr>
<td>diagnosis</td>
<td>11955</td>
</tr>
<tr>
<td>therapist</td>
<td>9910</td>
</tr>
<tr>
<td>therapy</td>
<td>8953</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>terms</th>
<th>freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>medication(s)/meds</td>
<td>57,981</td>
</tr>
<tr>
<td>(psycho)therapy/therapies</td>
<td>9,246</td>
</tr>
</tbody>
</table>

- Most frequent trigrams with therapy: go going to therapy (541), meds/medication and therapy (323)

→ Medication dominant treatment, therapy mainly in addition
## Key semantic domains

### B2: Health and disease

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>mania</td>
<td>14110</td>
<td>mental health</td>
<td>5674</td>
<td>healthy</td>
<td>3620</td>
</tr>
<tr>
<td>disorder</td>
<td>11915</td>
<td>health</td>
<td>2647</td>
<td>recovery</td>
<td>908</td>
</tr>
<tr>
<td>symptoms</td>
<td>10542</td>
<td>Medicaid</td>
<td>160</td>
<td>recover</td>
<td>502</td>
</tr>
<tr>
<td>side effects</td>
<td>7338</td>
<td>wellness</td>
<td>158</td>
<td>wellbeing</td>
<td>453</td>
</tr>
<tr>
<td>crazy</td>
<td>5836</td>
<td>asymptomatic</td>
<td>29</td>
<td>recovering</td>
<td>342</td>
</tr>
<tr>
<td>illness</td>
<td>5354</td>
<td></td>
<td></td>
<td>recovered</td>
<td>184</td>
</tr>
<tr>
<td>mental illness</td>
<td>4700</td>
<td></td>
<td></td>
<td>snap out of it</td>
<td>124</td>
</tr>
<tr>
<td><strong>total</strong></td>
<td><strong>154833</strong></td>
<td><strong>total</strong></td>
<td><strong>9114</strong></td>
<td><strong>total</strong></td>
<td><strong>7753</strong></td>
</tr>
</tbody>
</table>

1. Keywords overused in corpus of up to 7 most frequent words for each semantic sub-domain
So what?
Next steps, open questions, limitations
Next steps & open questions

• More detailed analysis of ‘recovery’ posts
  – How is recovery described: outcome/process, seen as possible?
  – Only 7% of users in 845 posts use term ‘recovery’
  – Find related terms for recovery via distributional semantics¹

• Relate to systematic review findings
  – 9 personal recovery processes: Connectedness, Hope & optimism,
    Identity, Meaning & purpose, Empowerment², coping with losses,
    balancing acceptance & ambitions, ...

• Posts outside bipolar subreddits (mental health/general)?

¹ https://code.google.com/archive/p/word2vec/
² Leamy et al. (2011)
Limitations

- Biased population: Reddit users disclosing diagnosis
- Imperfect text processing tools
  - Lower term counts ≠ lower concept mentions?
    - Symptom-related terms list more exhaustive?
    - Personal recovery expressed more indirectly?
- (So far) focus on largest phenomena, not individual experiences
Wrap-up

• Interdisciplinary project between clinical psychology & corpus/computational linguistics
  – Involves people with lived experience of bipolar disorder via consultation panel
• Main topics in Reddit posts by people with bipolar disorder diagnosis: symptoms, professional treatment, less discussion of recovery & self-management

Information and support addresses

bipolarUK

Mind
For better mental health
Thanks for your attention!

Looking forward to your questions and comments

Glorianna Jagfeld · Spectrum Centre for Mental Health Research · Lancaster University
g.jagfeld@lancaster.ac.uk · @glorisonne
Full references (I)


Full references (III)


Full references (V)

<table>
<thead>
<tr>
<th>Reference</th>
<th>Authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Text analysis as a tool for analyzing conversation in online support</td>
<td>Kramer et al. (2004)</td>
</tr>
<tr>
<td>groups</td>
<td></td>
</tr>
<tr>
<td>2. Social Support and Unsolicited Advice in a BD Online Forum</td>
<td>Vayreda and Antaki (2009)</td>
</tr>
<tr>
<td>forums in BD)</td>
<td></td>
</tr>
<tr>
<td>4. Bad on the net, or bipolar’s lives on the web: Analyzing discussion</td>
<td>Latalova et al. (2014)</td>
</tr>
<tr>
<td>web pages for individuals with bipolar affective disorder</td>
<td></td>
</tr>
<tr>
<td>5. Quantifying Mental Health Signals in Twitter</td>
<td>Coppersmith et al. (2014)</td>
</tr>
<tr>
<td>6. How Patients Contribute to an Online Psychoeducation Forum for BD: A</td>
<td>Poole et al. (2015)</td>
</tr>
<tr>
<td>Virtual Participant Observation Study</td>
<td></td>
</tr>
<tr>
<td>7. Mental illness and bipolar disorder on Twitter: implications for</td>
<td>Budenz et al. (2015)</td>
</tr>
<tr>
<td>stigma and social support</td>
<td></td>
</tr>
<tr>
<td>8. From ADHD to SAD: Analysing the Language of Mental Health on Twitter</td>
<td>Coppersmith et al. (2015)</td>
</tr>
<tr>
<td>through Self-Reported Diagnoses</td>
<td></td>
</tr>
<tr>
<td>9. Member roles and identities in online support groups: Perspectives</td>
<td>McDonald and Woodward-Kron (2016)</td>
</tr>
<tr>
<td>from corpus and systemic functional linguistics</td>
<td></td>
</tr>
<tr>
<td>10. The language of mental health problems in social media</td>
<td>Gkotsis et al. (2016)</td>
</tr>
</tbody>
</table>
## References for bipolar disorder studies with social media posts (II)

11. Multitask Learning for Mental Health Conditions with Limited Social Media Data  
   Benton et al. (2017)

12. Characterisation of mental health conditions in social media using Informed Deep Learning  
   Gkotsis et al. (2017)

13. Being bipolar*: A qualitative analysis of the experience of BD as described in internet blogs  
   Mandla et al. (2018)

14. SMHD: A Large-Scale Resource for Exploring Online Language Usage for Multiple Mental  
    Health Conditions  
   Cohan et al. (2018)

15. Semantic network analysis for understanding user experiences of bipolar and depressive  
    disorders on Reddit  
   Yoo et al. (2019)

16. BD, Genetic Risk, and Reproductive Decision-Making: A Qualitative Study of Social Media  
    Discussion Boards  
   Sahota and Sankar (2019)

17. Ketosis and BD: controlled analytic study of online reports  
   Campbell and Campbell (2019)

18. Analyzing Judgment in Bipolar Depression Patients’ Narratives Using Syntactic Patterns: A  
    Corpus-Based Study  
   Abdo et al. (2019)

19. Not Just Depressed: BD Prediction on Reddit  
   Sekulic et al (2018)

20. Adapting Deep Learning Methods for Mental Health Prediction on Social Media  
    Sekulic and Strube (2019)