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

Lancaster University

THE SPECTRUM CENTRE

UCREL

What and why?
Background & motivation

Extreme mood experiences, bipolar disorder & recovery

- Mood is a continuum[1]:

- Mania or hypomania + depression → bipolar disorder diagnosis

What constitutes recovery?

- Clinical: no symptoms for >= 8 weeks
- Personal: [aiming for] satisfying, hopeful, contributing life even with limitations caused by symptoms[2]

1. Jones, Lobban, Cook (2010) 2. Anthony (1993)

Previous research on personal recovery in bipolar disorder

Few quantitative[1-2] & qualitative[3-11] studies

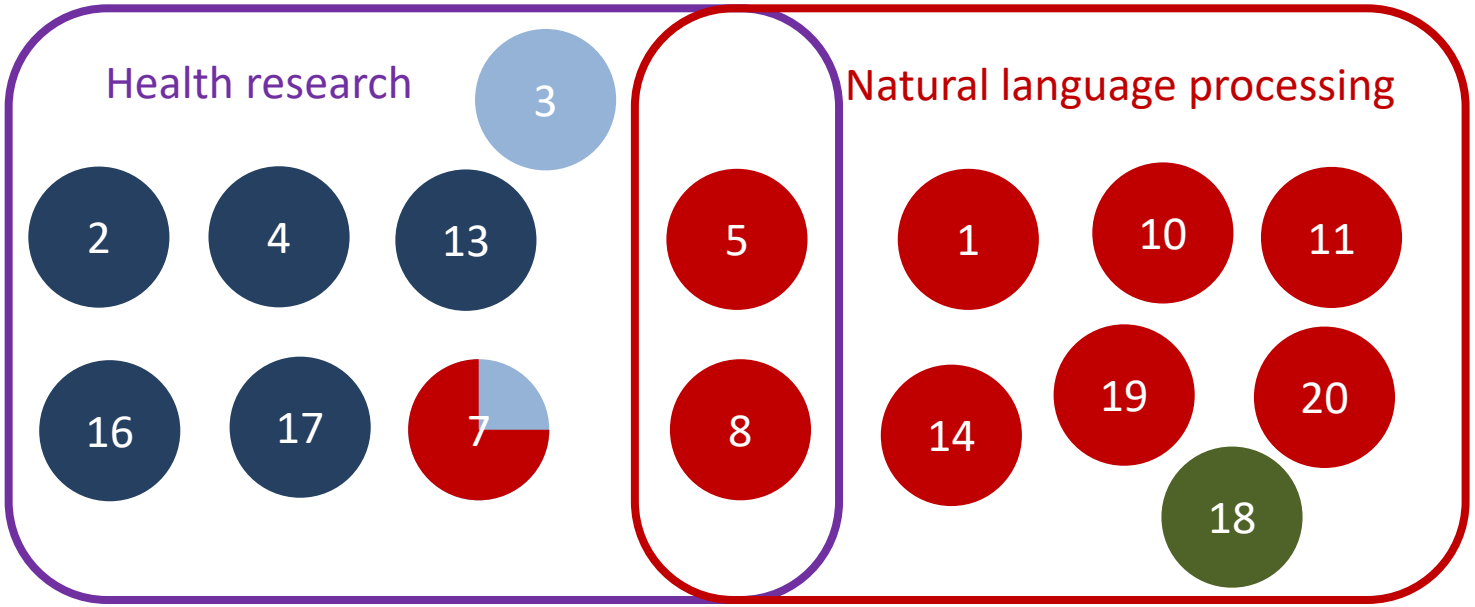
- Small samples
- Researcher-guided data production: interviews, questionnaires
- Mainly attracted (or only recruited) people who regard themselves in personal recovery

Potential of mixed-methods analysis of social media posts

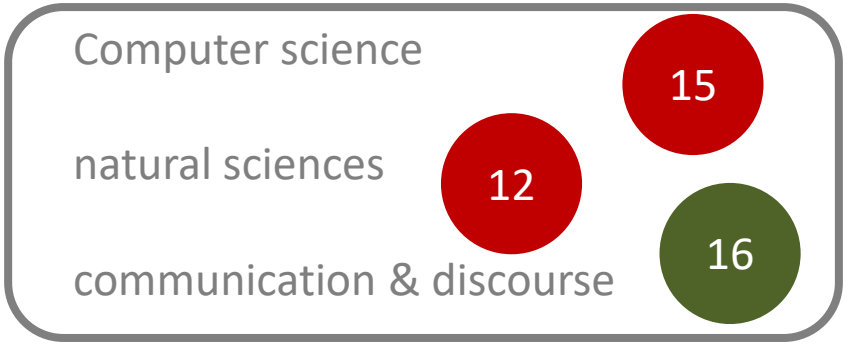
- Larger sample
- More diverse participants
- Data production not researcher-influenced

1. Jones et al. (2013) 2. Jones et al. (2015) 3. Mansell et al. (2010) 4. Michalak et al. (2012) 5. Todd et al. (2012) 6. Tjoflat & Ramvi (2013) 7. Fernandez et al. (2014) 8. Veseth et al. (2014) 9. Borg et al. (2015) 10. Morrison et al. (2016) 11. Warwick et al. (2019)

Bipolar disorder studies with social media posts by field & method



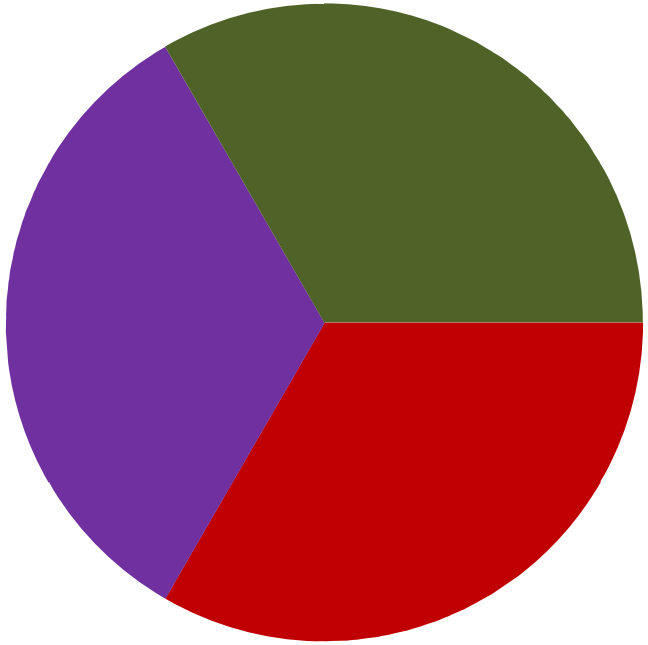
Main analysis methods:
Manual qualitative
Manual mixed/quantitative
Natural language processing
Corpus linguistics



PhD project aim & approach

Better understanding of personal recovery in bipolar disorder to make recommendations for mental health care improvements

Health research:
Personal recovery, implications



Corpus linguistics:
Analysis

Computational linguistics:
Data collection + processing

Study overview[1]

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

Lancaster University

THE SPECTRUM CENTRE

UCREL

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

1. Coppersmith et al. (2015) 2. Cohan et al. (2018)
3. bigquery.cloud.google.com/dataset/fh-bigquery:reddit_posts, /fh-bigquery:reddit_comments

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 2. Lui & Baldwin (2012)

Corpus processing

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

1. Garside et al. (1997), C7 tagset: http://ucrel.lancs.ac.uk/claws7tags.html 2. Rayson et al. (2004) 3. Rayson & Garside (2000) 4. Hardie (2014) 5. Cohan et al. (2018) 6. 2. https://www.sketchengine.eu/

UCREL Semantic Analysis System (USAS)[1]

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

1. Rayson et al. (2004)

Lancaster University

THE SPECTRUM CENTRE

UCREL

Results

Lancaster University

THE SPECTRUM CENTRE

UCREL

Dataset user demographics

Country of residence

Dataset rank	Country	Reddit dataset (N=19,816)	reddit.com traffic[2]	Systematic review (N=88, NA=14)[3]
1	USA	82%	50%	
2	Great Britain	6%	8%	39%
3	Canada	5%	7%	15%
4	Australia	2%	4%	11%
5	Germany	1%	3%	
6	Sweden	1%		
…				
14	Norway	0.15%		22%

- Dataset prediction based on posts' texts, subreddits posted in, posting time of day distribution[1]

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)

Gender	Reddit dataset (N= 2,344)[1]	Adult US Reddit users[2] (N=288)	Systematic review (N=88)
Female	65%	33%	66%
Male	35%	67%	34%

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

Age	Reddit dataset (N= 2,271) [1]	Adult US Reddit users[2] (N=288)	Systematic review (N=88, NA=34) [3]
13-17	6%	N/A	0%
18-29	78%	64%	7%
30-49	15%	29%	52%
50-64	1%	6%	41%
65+	0%	1%	0%

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 criterium; N=88: 24-68, mean 46

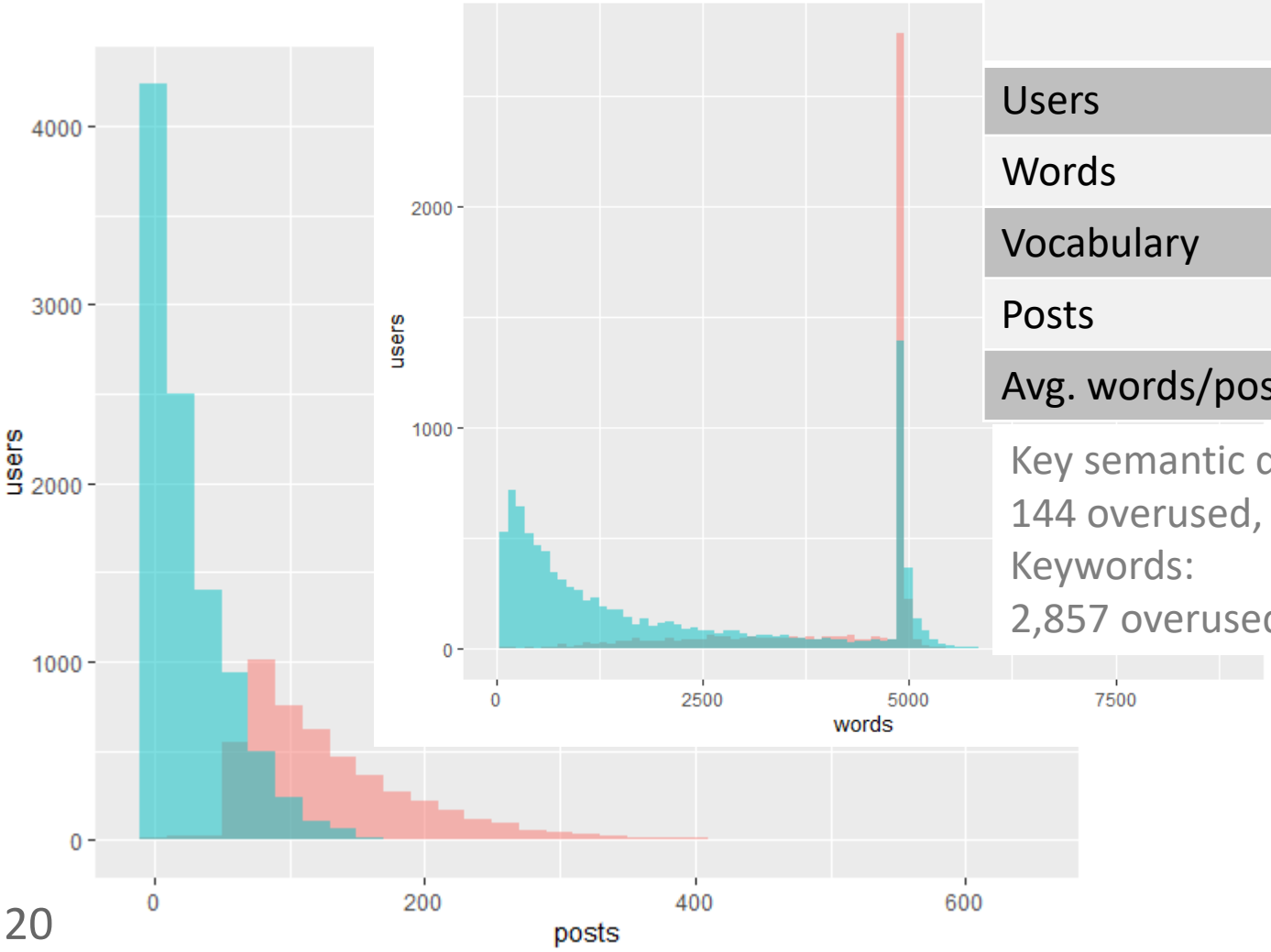
Lancaster University

THE SPECTRUM CENTRE

UCREL

Corpus & key semantic domains

Bipolar subreddits & reference corpus statistics



	Corpus	Control
Users	10K	5K
Words	20.6M	20.6M
Vocabulary	156K	405K
Posts	268K	661K
Avg. words/posts	77	31

Key semantic domains:
144 overused, 230 underused
Keywords:
2,857 overused, 9,764 underused

Key semantic domains in the bipolar subreddits corpus

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

1. Rayson and Garside (2000)

Domain-specific grouping of key semantic domains[1] with keywords[2]

- Mental health symptoms (~400K 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 (~200K): diagnosed, doctor, medication
- Recovery & self-management (~44K): 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

term[1]	frequency
diagnosed	20740
doctor	17524
medication	17498
psychiatrist	12949
diagnosis	11955
therapist	9910
therapy	8953

terms	freq.
medication(s)/meds	57,981
(psycho)therapy/therapies	9,246

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

→ Medication dominant treatment, therapy mainly in addition

1. keywords (log likelihood, $p<0.0001$) overused in corpus of 7 most frequent words for each semantic sub-domain

Key semantic domains
B2: Health and disease

B2- (~4x overused)	Freq.[1]	B2 (~4x overused)	Freq.	B2+ (~2x overused)	Freq.
mania	14110	mental health	5674	healthy	3620
disorder	11915	health	2647	recovery	908
symptoms	10542	Medicaid	160	recover	502
side effects	7338	wellness	158	wellbeing	453
crazy	5836	asymptomatic	29	recovering	342
illness	5354			recovered	184
mental illness	4700			snap out of it	124
total	154833	**total**	9114	**total**	7753

1. Keywords overused in corpus of up to 7 most frequent words for each semantic sub-domain

Lancaster University

THE SPECTRUM CENTRE

UCREL

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 semantics[1]
- Relate to systematic review findings
  - 9 personal recovery processes: Connectedness, Hope & optimism, Identity, Meaning & purpose, Empowerment[2], 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

bipolar UK

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)

- Anthony WA. Recovery from mental illness: the guiding vision of the mental health system in the 1990s. Vol. 16, Psychosocial Rehabilitation Journal. 1993.16(4):11–23.
- Barthel, M. *et al.* (2016) 'Nearly Eight-in-Ten Reddit Users Get News on the Site', Available at: www.pewresearch.org.
- Borg M, Veseth M, Binder P-E, Topor A. The role of work in recovery from bipolar disorders. Vol. 12, Qualitative Social Work. 2013.12(3):323–39.
- Coole, M., Rayson, P. and Mariani, J. (2016) 'LexiDB: A scalable corpus database management system', *Proceedings - 2016 IEEE International Conference on Big Data, Big Data 2016*. IEEE, pp. 3880–3884. doi: 10.1109/BigData.2016.7841062.
- Cohan A, Desmet B, Yates A, Soldaini L, MacAvaney S, Goharian N. 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). 2018. p. 1485–97.
- Coppersmith, G. *et al.* (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*
- Fernandez ME, Breen LJ, Simpson TA. Renegotiating identities: Experiences of loss and recovery for women with bipolar disorder. Vol. 24, Qualitative Health Research. 2014.24(7):890–900.

Full references (II)

- Garside R, Smith N. A Hybrid Grammatical Tagger: CLAWS4. In: Garside R, Leech G, McEnery A, editors. Corpus Annotation: Linguistic Information from Computer Text Corpora. London: Longman; 1997. p. 102–21.
- Hardie A. Statistical identification of keywords, lockwords and collocations as a two-step procedure. In: Proceedings of the Annual Conference of the International Computer Archive for Modern and Medieval English (ICAME). 2014. p. 49. http://cass.lancs.ac.uk/log-ratio-an-informal-introduction/
- Harrigian, K. (2019) '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*, pp. 17–27. https://github.com/kharrigian/smgeo
- Harvey, K. (2012) 'Disclosures of depression', *International Journal of Corpus Linguistics*, 17(3), pp. 349–379. doi: 10.1075/ijcl.17.3.03har.
- Hunt, D. and Brookes, G. (2020) *Corpus, discourse and mental health*. *Discourse and Mental Health*. Bloomsbury Academic. doi: 10.4324/9780203701928.
- Jones S, Lobban F, Cook A. Understanding Bipolar Disorder - Why some people experience extreme mood states and what can help. British Psychological Society; 2010. 85 p.
- Jones S, Mulligan LD, Higginson S, Dunn G, Morrison AP. The bipolar recovery questionnaire: Psychometric properties of a quantitative measure of recovery experiences in bipolar disorder. Vol. 147, Journal of Affective Disorders. 2013.147(1–3):34–43.

Full references (III)

- Jones SH, Smith G, Mulligan LD, Lobban F, Law H, Dunn G, et al. Recovery-focused cognitive-behavioural therapy for recent-onset bipolar disorder: Randomized controlled pilot trial. Vol. 206, British Journal of Psychiatry. 2015.206(1):58–66.
- Leamy, M. *et al.* (2011) 'Conceptual framework for personal recovery in mental health: Systematic review and narrative synthesis', *British Journal of Psychiatry*, 199(6), pp. 445–452. doi: 10.1192/bjp.bp.110.083733.
- Lui M, Baldwin T. Langid.py: An Off-the-shelf Language Identification Tool. Aclweb.Org. 2012.(July):25–30. https://github.com/saffsd/langid.py
- Mansell W, Powell S, Pedley R, Thomas N, Jones SA. The process of recovery from bipolar i disorder: A qualitative analysis of personal accounts in relation to an integrative cognitive model. Vol. 49, British Journal of Clinical Psychology. 2010.49(2):193–215.
- McDonald, D. and Woodward-Kron, R. (2016) 'Member roles and identities in online support groups: Perspectives from corpus and systemic functional linguistics', *Discourse and Communication*, 10(2), pp. 157–175. doi: 10.1177/1750481315615985.
- McIntyre RS, Kennedy SH, Soczynska JK, Nguyen HTT, Bilkey TS, Woldeyohannes HO, et al. Attention-Deficit/Hyperactivity Disorder in Adults With Bipolar Disorder or Major Depressive Disorder: Results From the International Mood Disorders Collaborative Project. Vol. 12, Primary Care Companion to the Journal of Clinical Psychiatry. 2010.12(3):321–42.

Full references (IV)

- Michalak EE, Hole R, Holmes C, Velyvis V, Austin J, Pesut B, et al. Implication for psychiatric care of the word “recovery” in people with bipolar disorder. Vol. 42, Psychiatric Annals. 2012.42(5):173–8.
- Morrison AP, Law H, Barrowclough C, Bentall RP, Haddock G, Jones SH, et al. Psychological approaches to understanding and promoting recovery in psychosis and bipolar disorder: a mixed-methods approach. 2016.
- Nabavi B, Mitchell AJ, Nutt D. A Lifetime Prevalence of Comorbidity Between Bipolar Affective Disorder and Anxiety Disorders: A Meta-analysis of 52 Interview-based Studies of Psychiatric Population. Vol. 2, EBioMedicine. 2015.2(10):1405–19.
- Rayson P, Archer D, Piao S, McEnery T. The UCREL semantic analysis system. In: Proceedings of the beyond named entity recognition semantic labelling for NLP tasks workshop. 2004. p. 7–12.
- Rayson P, Garside R. Comparing corpora using frequency profiling. In: Proceedings of the workshop on Comparing Corpora,
held in conjunction with the 38th annual meeting of the Association for Computational Linguistics (ACL). 2000. p. 1–6.
- Sekulić I, Gjurković M, Šnajder J. Not Just Depressed: Bipolar Disorder Prediction on Reddit. In: Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis (WASSA). 2018. p. 72–8.

Full references (V)

- Staiano, K. V. (1986) *Interpreting Signs of Illness*. Berlin, Boston: De Gruyter Mouton.
- Tjoflåt M, Ramvi E. I am Me! Experiencing Parenting While Dealing With One's Own Bipolar Disorder. Vol. 11, Social Work in Mental Health. 2013.11(1):75–97.
- Todd NJ, Jones SH, Lobban FA. "Recovery" in bipolar disorder: How can service users be supported through a self-management intervention? A qualitative focus group study. Vol. 21, Journal of Mental Health. 2012.21(2):114–26.
- Veseth M, Binder P-E, Borg M, Davidson L. Toward caring for oneself in a life of intense ups and downs: A reflexive-collaborative exploration of recovery in bipolar disorder. Vol. 22, Qualitative Health Research. 2012.22(1):119–33.
- Warwick H, Tai S, Mansell W. Living the life you want following a diagnosis of bipolar disorder: A grounded theory approach. Clinical Psychology & Psychotherapy. 2019.

References for bipolar disorder studies with social media posts (I)

1. Text analysis as a tool for analyzing conversation in online support groups Kramer at al. (2004)
2. Social Support and Unsolicited Advice in a BD Online Forum Vayreda and Antaki (2009)
3. Cyber-support: An analysis of online self-help forums (online self-help forums in BD) Bauer et al. (2013)
4. Bad on the net, or bipolars' lives on the web: Analyzing discussion web pages for individuals with bipolar affective disorder Latalova et al. (2014)
5. Quantifying Mental Health Signals in Twitter Coppersmith et al. (2014)
6. How Patients Contribute to an Online Psychoeducation Forum for BD: A Virtual Participant Observation Study Poole et al. (2015)
7. Mental illness and bipolar disorder on Twitter: implications for stigma and social support Budenz et al. (2015)
8. From ADHD to SAD: Analysing the Language of Mental Health on Twitter through Self-Reported Diagnoses Coppersmith et al. (2015)
9. Member roles and identities in online support groups: Perspectives from corpus and systemic functional linguistics McDonald andWoodward-Kron (2016)
10. The language of mental health problems in social media Gkotsis et al. (2016)

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)