SMHD: A Large-Scale Resource for Exploring Online Language Usage for Multiple Mental Health Conditions

Arman Cohan, Bart Desmet, Andrew Yates,
Luca Soldaini, **Sean MacAvaney**, Nazli Goharian

COLING 2018 “Area Chair Favorite Paper”

4.2% of Americans suffer from Serious Mental Illnesses (SMI).

Only 65% of those with SMI received any treatment in past year.

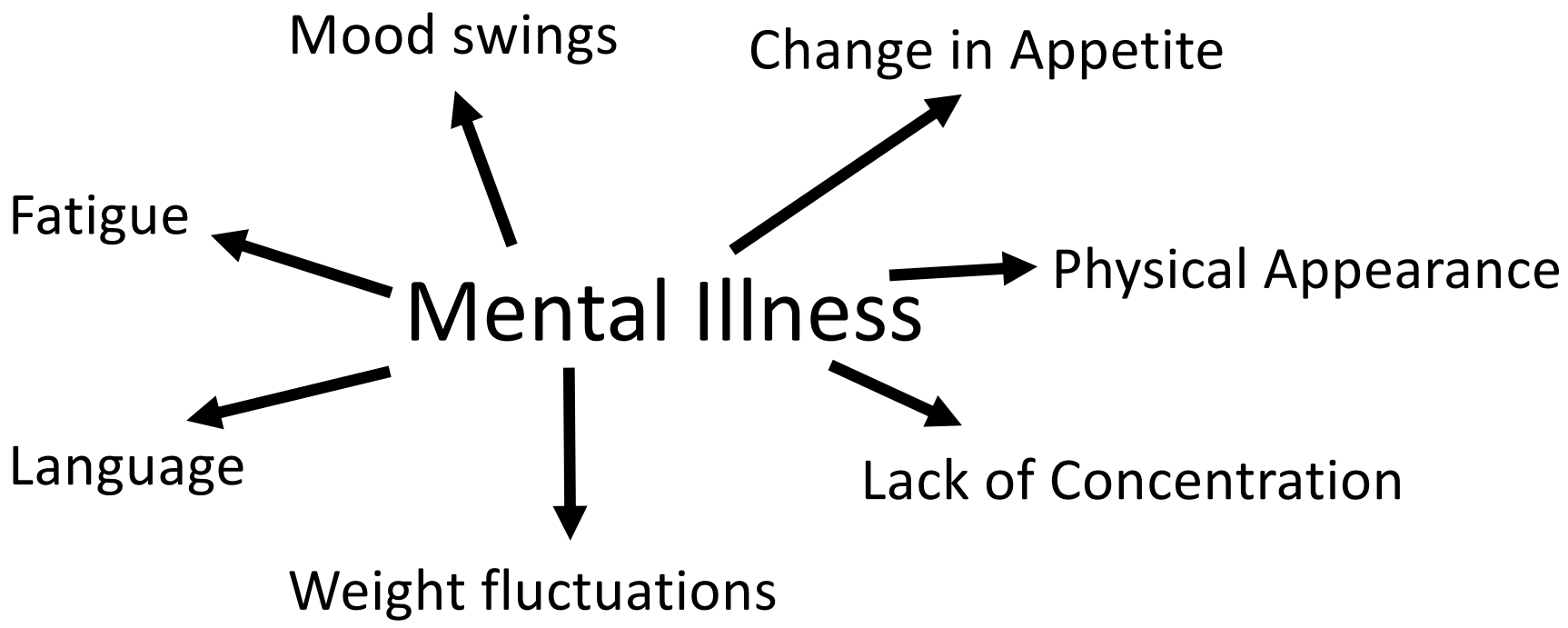
US Health and Human Services, 2016

Suicide is the 10th leading cause of death in the US.

45,000 Americans take their own life each year.

CDC Vital Statistics Report, 2016

Doctors and mental health providers evaluate many possible symptoms to make a diagnosis.

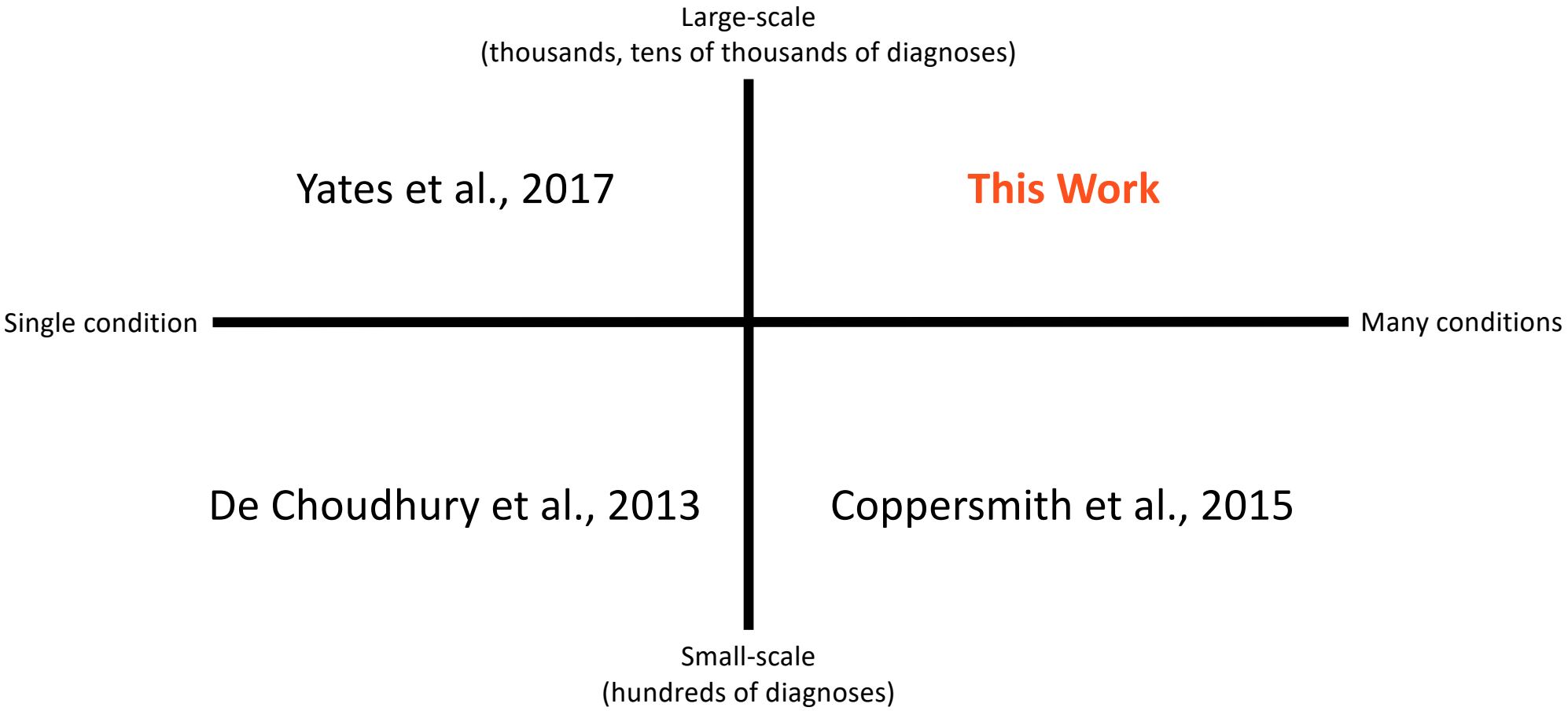


Let’s try to better understand how people’s language is affected by mental illness.

Goals:

- **Large-scale** – people show symptoms in a variety of ways; the more data the better
- **High quality** – a low false-positive rate for diagnoses is important
- **Multiple conditions** – conditions do not exist in isolation; the interaction between conditions is valuable to explore

Previous work has not explored this space.



We present:

SMHD: a dataset of Self-reported Mental Health Diagnoses

- Large-scale labeled data from Reddit
- 37k diagnosed users from high-precision diagnosis statements
- Carefully-selected “control” users for each
- 9 mental health conditions
  - ADHD, anxiety, autism spectrum disorder, bipolar disorder, depression, eating disorder, OCD, PTSD,
- Available to researches with signed DUA

Source of Relevance:
Self-reported diagnosis statements

was officially diagnosed with ADHD last year.

I have a diagnosed history of PTSD.

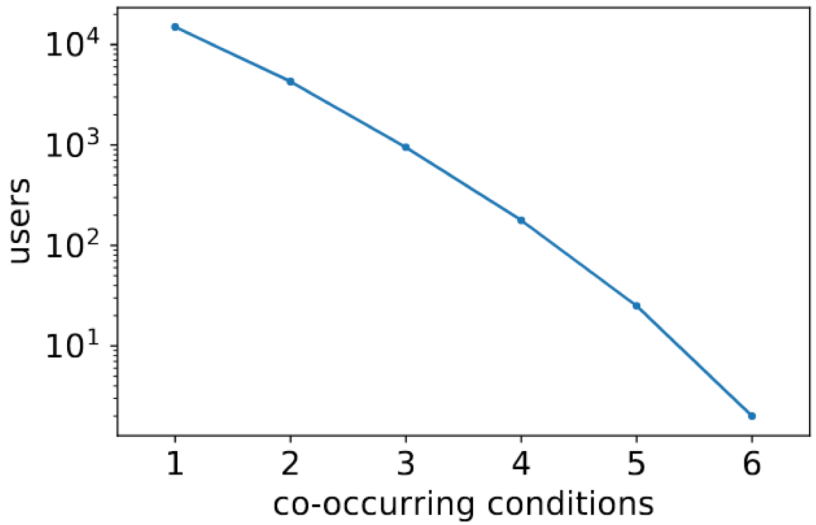
my dr just diagnosed me as schizo.

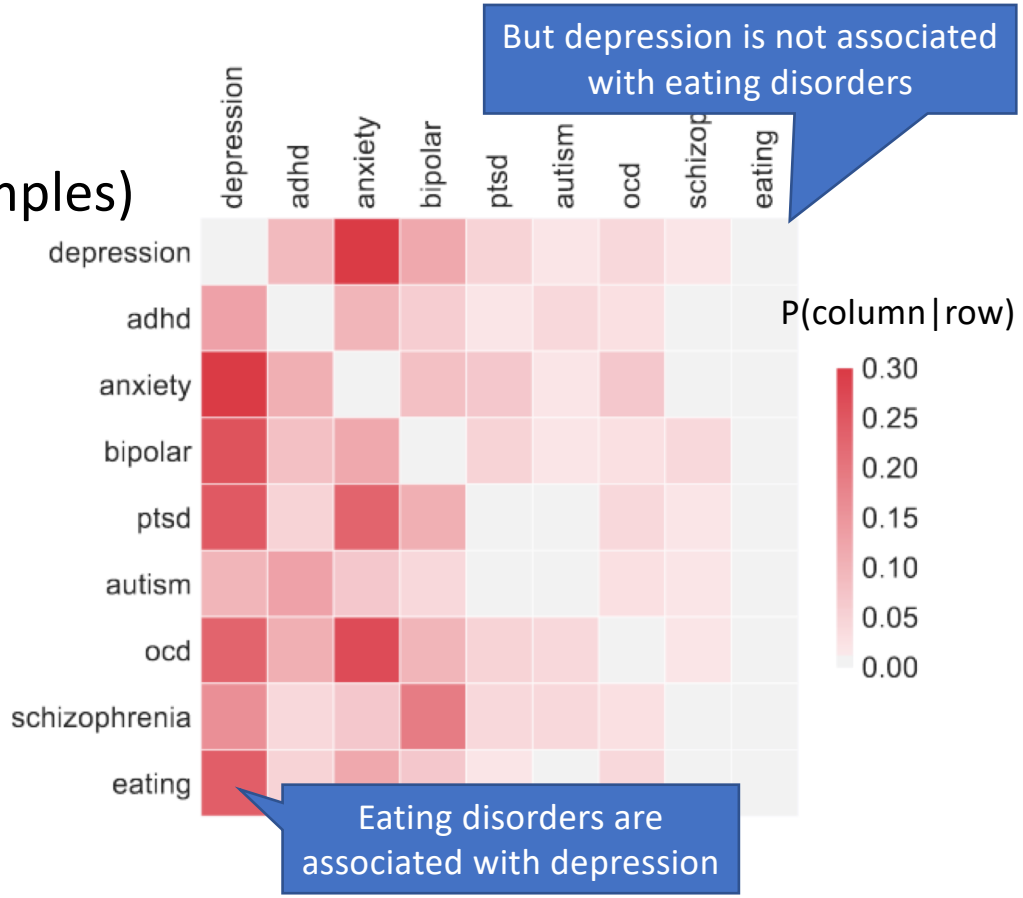
Self-reported diagnosis rules

1. Diagnosis pattern patterns, e.g. “i am diagnosed with”, “they recently diagnosed me”, “have [condition] diagnosed”…
2. Condition names matching terms found in dictionaries (MedSyn, Behaviorial), manually adjusted to remove hypernyms or general terms (e.g., add), and add terms found when manually inspecting the data
3. At least 50 posts unrelated to mental health (no mental health terminology, and not in mental health sub-forums)
   - To remove bias, we only analyze these posts

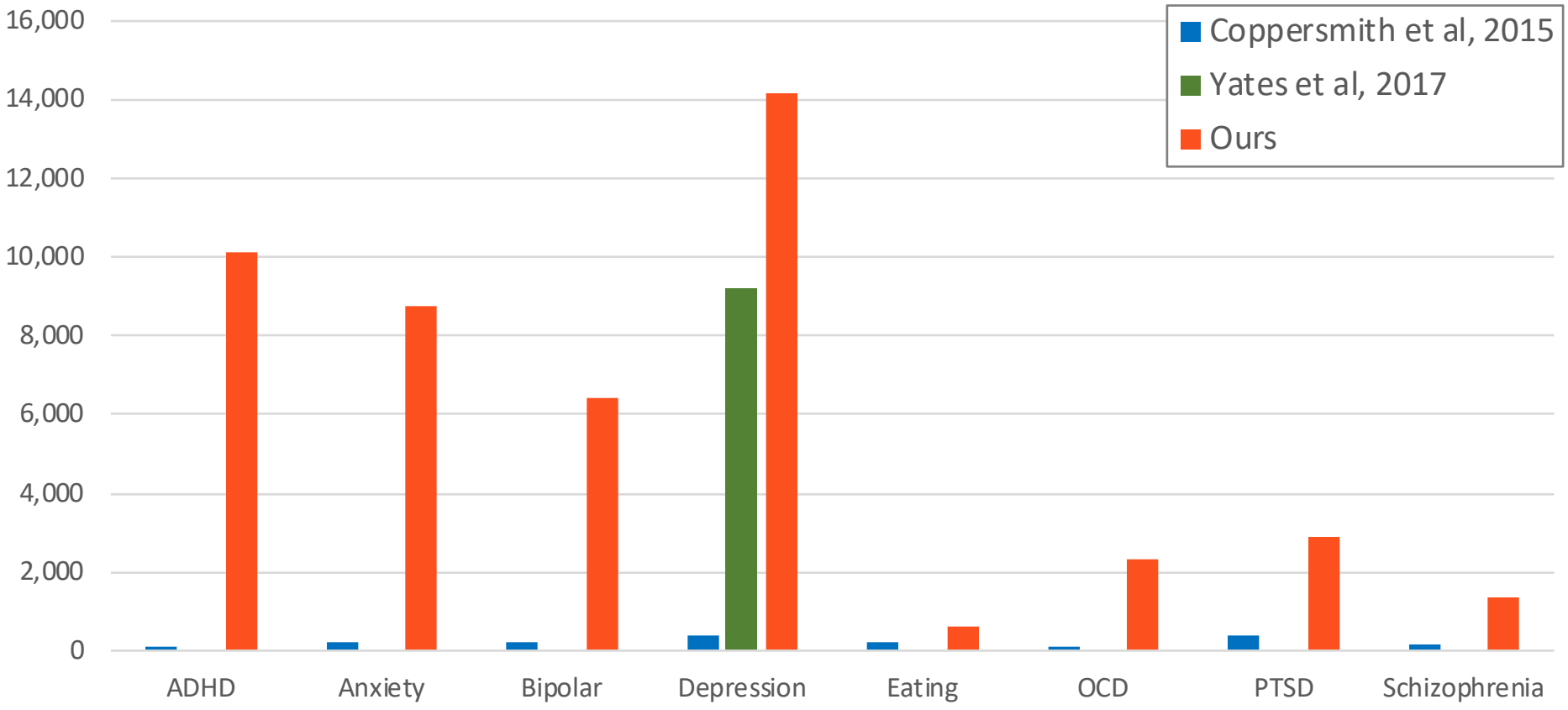
Self-reported diagnosis results

- 36,948 users
- 95.8% precision (900 held-out samples)





Ours much larger than previous datasets.



Does not include autism (ours=2,911) borderline personality (Coppersmith=101) or Seasonal Affective (Coppersmith=100).

Control Users

- Choose 9 control users per diagnosed user
- Stringent criteria:
  - Must not have any posts that contain mental health language or be in a mental health sub-forum.
  - Must have at least 50 posts.
  - Most have posted in the same sub-forum

Dataset Characteristics

condition	posts		tokens		characters
	per user	total	per post	total	per post
control	310.0 (157.8)	115,669k	26.2 (48.3)	3,031.6M	133.9 (252.9)
depression	162.2 (84.2)	1,272k	45.1 (80.0)	57.4M	227.5 (406.9)
adhd	164.7 (83.6)	872k	46.5 (82.7)	40.5M	237.5 (433.5)
anxiety	159.7 (83.0)	795k	46.4 (83.0)	36.9M	233.9 (422.8)
bipolar	157.6 (82.4)	575k	45.5 (86.5)	26.2M	230.6 (447.0)
ptsd	160.7 (84.7)	258k	53.1 (114.0)	13.7M	267.8 (581.7)
autism	168.3 (84.5)	248k	46.5 (82.3)	11.6M	237.9 (434.0)
ocd	158.8 (81.4)	203k	46.4 (90.1)	9.4M	234.2 (459.5)
schizophrenia	157.3 (80.5)	123k	49.2 (105.6)	6.1M	253.8 (566.6)
eating	161.4 (81.0)	53k	46.3 (73.7)	2.5M	232.6 (372.8)

Textual Analysis

- Linguistic Inquiry and Word Count (LIWC) categories
- Compare scores in categories using Welch's t-test ($p<0.001$)
- P-values adjusted with Bonferroni correction

LIWC category	depression	adhd	anxiety	bipolar	ptsd	autism	ocd	schizophrenia	eating
Summary Language Variables									
Clout	$-0.06^{\ddagger}$	–	$-0.1^{\ddagger}$	–	–	–	–	–	–
Authentic	$0.2^{\ddagger}$	$0.15^{\ddagger}$	$0.22^{\ddagger}$	$0.18^{\ddagger}$	$0.21^{\ddagger}$	$0.14^{\ddagger}$	$0.23^{\ddagger}$	–	$0.24^{\star}$
WPS	$0.08^{\ddagger}$	$0.1^{\ddagger}$	$0.08^{\ddagger}$	$0.06^{\ddagger}$	$0.1^{\ddagger}$	$0.12^{\ddagger}$	$0.08^{\dagger}$	–	–
Dictionary words	$0.27^{\ddagger}$	$0.22^{\ddagger}$	$0.28^{\ddagger}$	$0.24^{\ddagger}$	$0.31^{\ddagger}$	$0.2^{\ddagger}$	$0.28^{\ddagger}$	$0.22^{\ddagger}$	$0.3^{\ddagger}$
Total function words	$0.27^{\ddagger}$	$0.21^{\ddagger}$	$0.28^{\ddagger}$	$0.24^{\ddagger}$	$0.3^{\ddagger}$	$0.26^{\ddagger}$	$0.28^{\ddagger}$	$0.23^{\ddagger}$	$0.27^{\ddagger}$
Total pronouns	$0.22^{\ddagger}$	$0.14^{\ddagger}$	$0.24^{\ddagger}$	$0.2^{\ddagger}$	$0.25^{\ddagger}$	$0.17^{\ddagger}$	$0.26^{\ddagger}$	$0.18^{\ddagger}$	$0.26^{\dagger}$
Personal pronouns	$0.23^{\ddagger}$	$0.14^{\ddagger}$	$0.26^{\ddagger}$	$0.21^{\ddagger}$	$0.22^{\ddagger}$	$0.14^{\ddagger}$	$0.23^{\ddagger}$	$0.2^{\ddagger}$	$0.27^{\dagger}$
1st pers singular	$0.23^{\ddagger}$	$0.16^{\ddagger}$	$0.28^{\ddagger}$	$0.22^{\ddagger}$	$0.23^{\ddagger}$	$0.17^{\ddagger}$	$0.26^{\ddagger}$	$0.17^{\dagger}$	$0.28^{\dagger}$
3rd pers singular	$0.09^{\ddagger}$	–	$0.1^{\ddagger}$	$0.08^{\star}$	$0.17^{\star}$	–	–	–	–
Impersonal pronouns	$0.06^{\ddagger}$	$0.05^{\star}$	$0.07^{\dagger}$	–	$0.11^{\dagger}$	$0.09^{\star}$	$0.13^{\dagger}$	–	–
Prepositions	$0.12^{\ddagger}$	$0.12^{\ddagger}$	$0.12^{\ddagger}$	$0.12^{\ddagger}$	$0.16^{\ddagger}$	$0.12^{\ddagger}$	$0.11^{\dagger}$	–	–
Auxiliary verbs	$0.12^{\ddagger}$	$0.1^{\ddagger}$	$0.14^{\ddagger}$	$0.11^{\ddagger}$	$0.14^{\ddagger}$	$0.15^{\ddagger}$	$0.13^{\ddagger}$	–	–
Common Adverbs	$0.09^{\ddagger}$	$0.07^{\ddagger}$	$0.08^{\ddagger}$	$0.06^{\star}$	–	–	–	–	–
Conjunctions	$0.17^{\ddagger}$	$0.14^{\ddagger}$	$0.18^{\ddagger}$	$0.16^{\ddagger}$	$0.17^{\ddagger}$	$0.13^{\ddagger}$	$0.11^{\dagger}$	–	$0.3^{\ddagger}$
Other Grammar									
Common verbs	$0.15^{\ddagger}$	$0.11^{\ddagger}$	$0.17^{\ddagger}$	$0.13^{\ddagger}$	$0.15^{\ddagger}$	$0.13^{\ddagger}$	$0.19^{\ddagger}$	–	–
Numbers	$-0.1^{\ddagger}$	$-0.09^{\ddagger}$	$-0.11^{\ddagger}$	$-0.09^{\ddagger}$	$-0.11^{\ddagger}$	–	$-0.1^{\ddagger}$	–	$-0.13^{\dagger}$
Psychological Variables									
Positive emotion	–	–	–	–	–	$-0.08^{\star}$	–	–	–
Anxiety	$0.07^{\dagger}$	–	$0.07^{\star}$	–	–	–	–	–	–
Social processes	$0.11^{\ddagger}$	$0.07^{\ddagger}$	$0.11^{\ddagger}$	$0.1^{\ddagger}$	$0.15^{\ddagger}$	–	–	–	–
Family	$0.06^{\ddagger}$	–	$0.06^{\star}$	–	$0.12^{\dagger}$	–	–	–	–
Female references	$0.13^{\ddagger}$	$0.07^{\ddagger}$	$0.1^{\ddagger}$	$0.13^{\ddagger}$	$0.22^{\dagger}$	–	–	–	–
Cognitive processes	$0.12^{\ddagger}$	$0.13^{\ddagger}$	$0.14^{\ddagger}$	$0.09^{\ddagger}$	$0.12^{\dagger}$	$0.16^{\ddagger}$	$0.13^{\dagger}$	–	–
Insight	$0.09^{\ddagger}$	$0.07^{\ddagger}$	$0.1^{\ddagger}$	$0.08^{\ddagger}$	–	$0.1^{\star}$	$0.17^{\dagger}$	–	–
Discrepancy	–	$0.06^{\star}$	–	–	–	–	–	–	–
Tentative	$0.07^{\ddagger}$	$0.08^{\ddagger}$	$0.08^{\ddagger}$	$0.07^{\star}$	–	–	–	–	–
Differentiation	$0.08^{\ddagger}$	$0.08^{\ddagger}$	$0.1^{\ddagger}$	–	–	$0.09^{\star}$	–	–	–
Biological processes	$0.06^{\ddagger}$	–	–	–	–	–	–	–	–
Health	$0.08^{\ddagger}$	$0.07^{\dagger}$	$0.08^{\ddagger}$	$0.11^{\ddagger}$	–	–	–	–	–
Time orientation									
Past focus	$0.08^{\ddagger}$	$0.05^{\star}$	$0.09^{\ddagger}$	$0.08^{\dagger}$	$0.09^{\star}$	–	$0.11^{\star}$	–	–
Present focus	$0.09^{\ddagger}$	$0.06^{\ddagger}$	$0.1^{\ddagger}$	$0.07^{\star}$	–	–	–	–	–
Personal concerns									
Relativity	$0.05^{\dagger}$	$0.06^{\dagger}$	–	–	–	–	–	–	–
Time	$0.06^{\ddagger}$	–	$0.06^{\star}$	–	–	–	–	–	–
Work	–	$0.06^{\star}$	–	–	–	–	–	–	–
Leisure	$-0.07^{\ddagger}$	$-0.07^{\ddagger}$	$-0.07^{\ddagger}$	$-0.09^{\ddagger}$	$-0.12^{\ddagger}$	$-0.09^{\dagger}$	–	–	–
Money	$-0.06^{\ddagger}$	–	$-0.05^{\ddagger}$	$-0.06^{\ddagger}$	–	$-0.1^{\ddagger}$	–	–	$-0.12^{\ddagger}$
Informal language	$-0.07^{\ddagger}$	$-0.07^{\ddagger}$	$-0.06^{\ddagger}$	–	$-0.12^{\ddagger}$	–	$-0.09^{\star}$	–	–
Netspeak	$-0.07^{\ddagger}$	$-0.06^{\ddagger}$	$-0.06^{\ddagger}$	$-0.06^{\star}$	$-0.1^{\ddagger}$	$-0.08^{\star}$	–	–	$-0.15^{\ddagger}$
Assent	–	–	–	–	$-0.06^{\star}$	–	–	–	–

Control users more likely to exhibit “clout” than users with depression or anxiety.

“Authentic” language is more common among most conditions (e.g., using more personal pronouns).

Topics such as leisure and money are more common among control users.

Can we predict the condition based on posts?

FastText model generally performs better than others (F1)

Method	Logistic Regression	XGBoost	SVM	FastText	CNN
Depression	42.9	43.3	52.8	**53.6**	50.5
ADHD	31.6	34.1	44.9	**47.0**	35.1
Anxiety	41.1	45.7	53.6	**53.7**	48.0
Bipolar	34.6	44.7	**51.6**	50.8	38.3
PTSD	32.5	53.4	57.1	**57.6**	54.9
Autism	14.6	34.2	42.9	**49.8**	38.3
OCD	9.3	39.6	41.6	**44.5**	32.9
Schizophrenia	4.4	38.5	39.5	**45.3**	37.0
Eating	0.0	24.8	23.6	41.2	**44.5**

SMHD: A Large-Scale Resource for Exploring Online Language Usage for Multiple Mental Health Conditions

- Large dataset of mental health language, spanning multiple conditions http://ir.cs.georgetown.edu/resources/
- Observations of significant differences between diagnosed users and control users on a variety of categories
- Baseline classification results show prediction is a difficult task