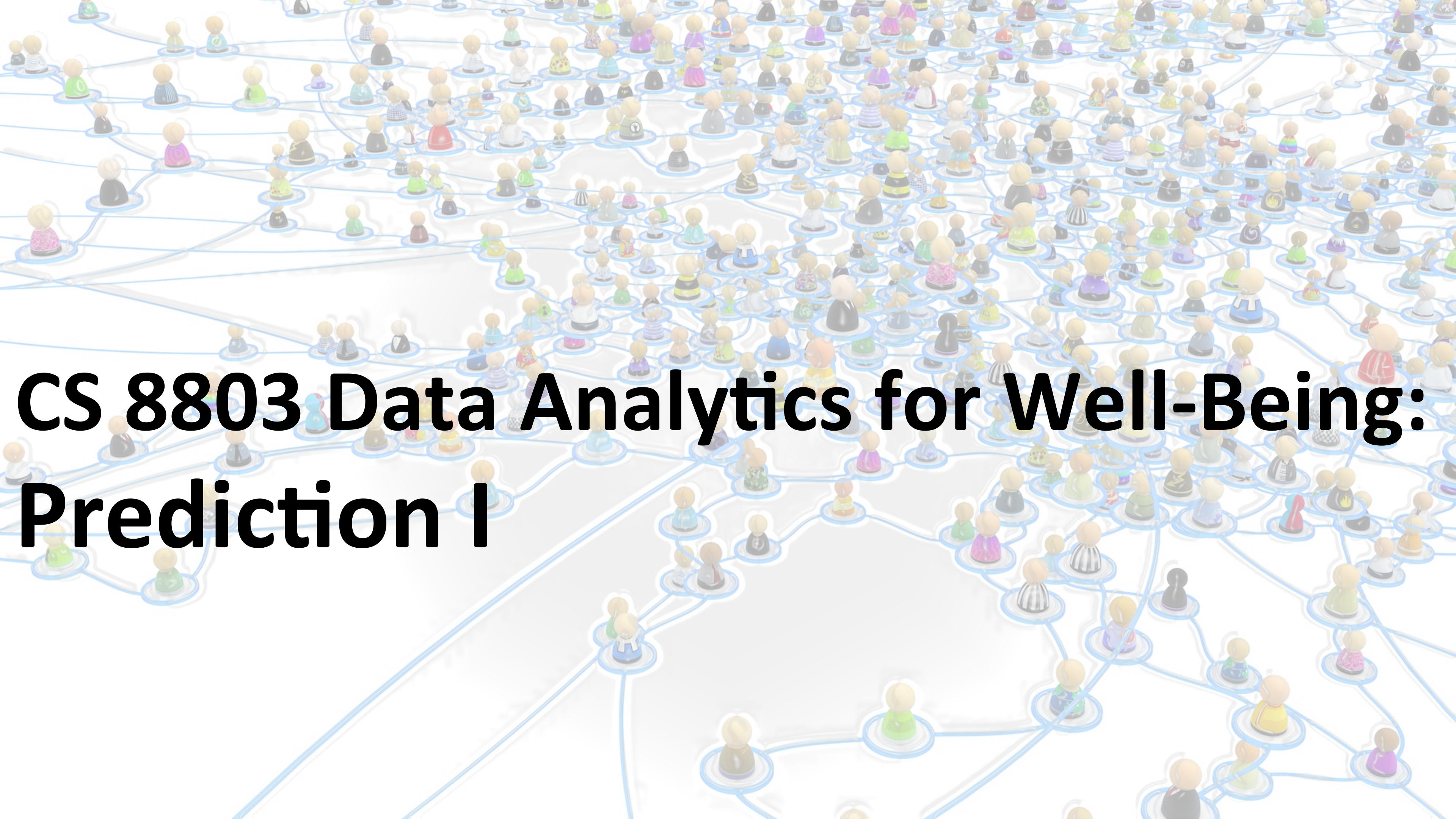


CS 8803 Data Analytics for Well-Being: Prediction I

Munmun De Choudhury

munmund@gatech.edu

Week 9 | March 7, 2016

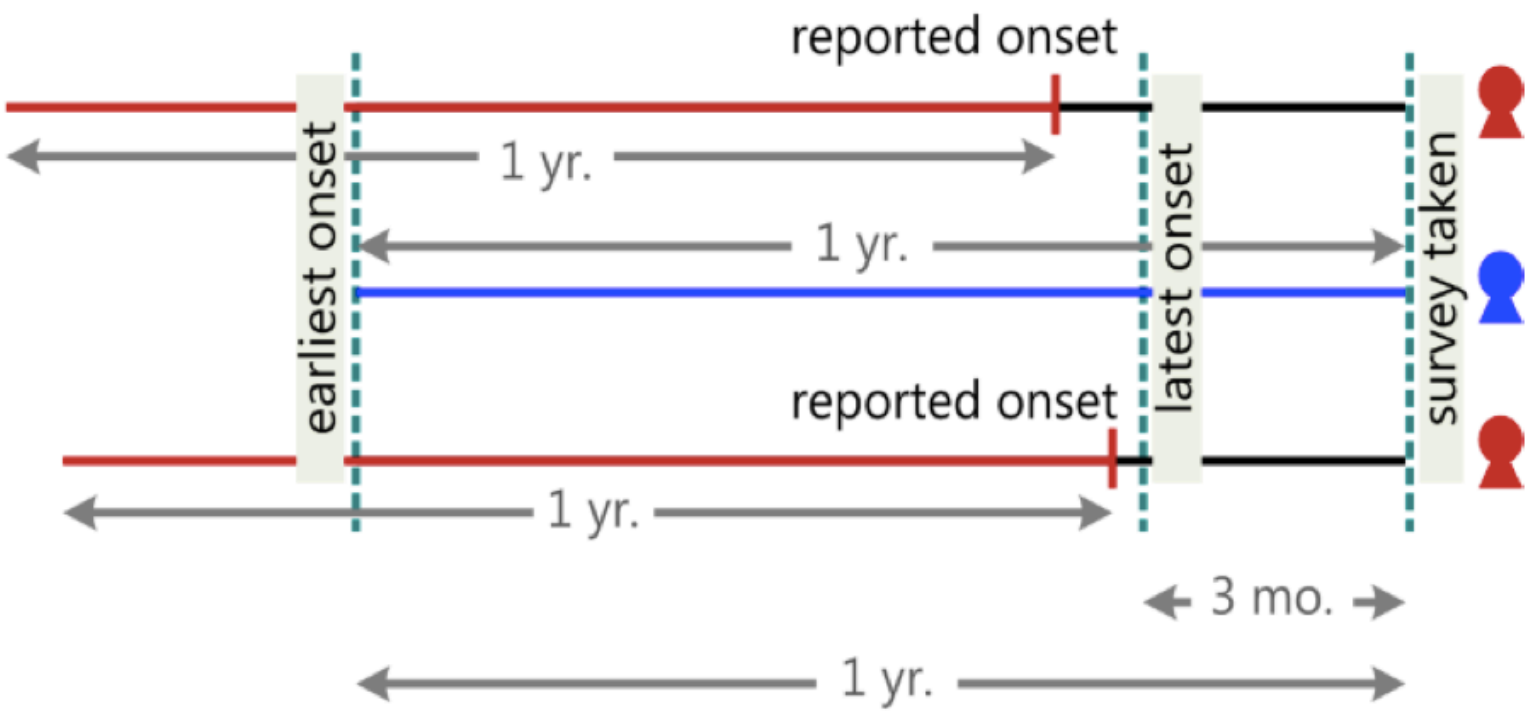
Please sign up for the class presentations!!!!!

Mid-term progress presentations due on Mar 16

Predicting Depression via Social Media

Summary

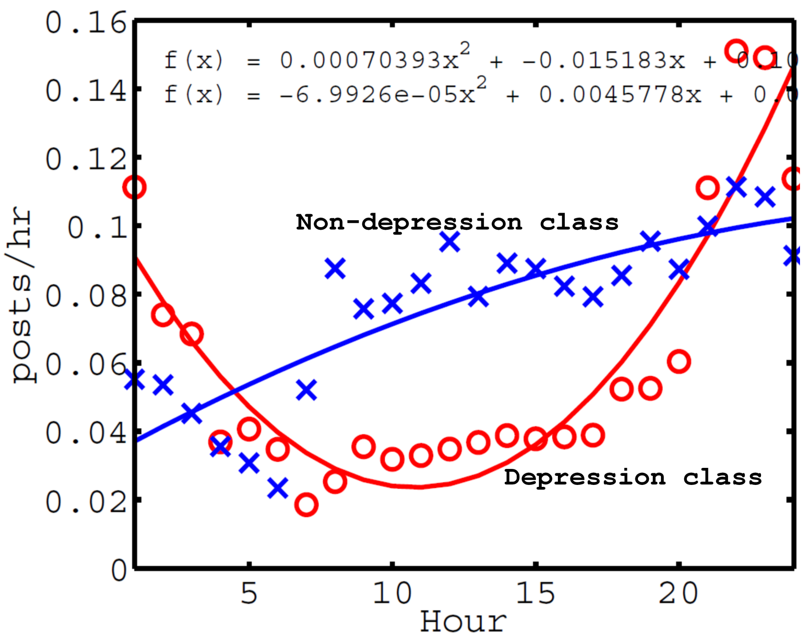
- Can social media activities and connectedness predict risk to major depressive disorder?
- Recruitment of a sample of Twitter users through a survey methodology over Amazon's Mechanical Turk
  - ~40% provided access to Twitter data



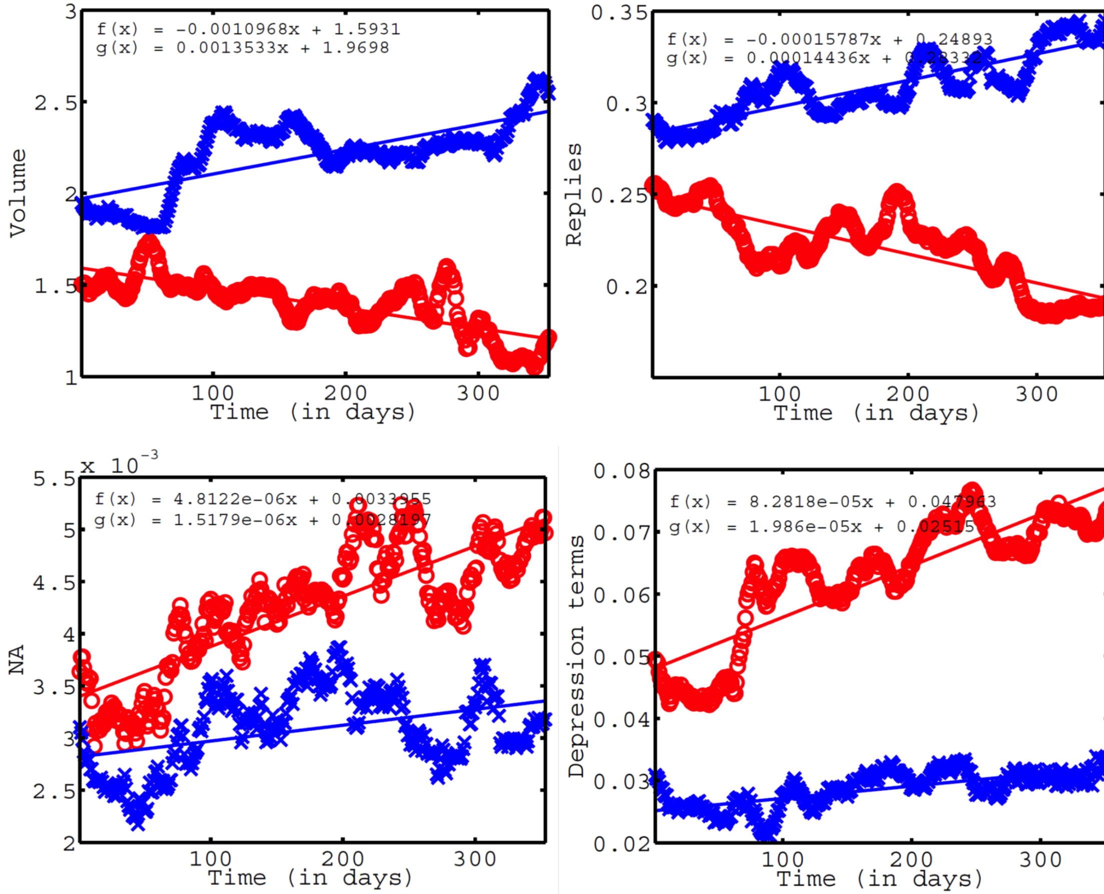
Summary

- Social engagement
- "Insomnia index" – mean *z*-score of an individual's volume of Twitter activity per hour
- Ego-centric social graph – nodal properties (*inlinks, outlinks*); dyadic properties (*reciprocity, interpersonal exchange*); neighborhood properties (*density, clustering coefficient, two-hop neighborhood, embeddedness, number of ego components*)
- Language
  - Depression lexicon – top uni- and bigrams compiled from Yahoo! Answers category on mental health
  - Linguistic style

Summary



f(x) = 0.00070393x2 + -0.015183x + 0.10
f(x) = -6.9926e-05x2 + 0.0045778x + 0.0
Non-depression class
Depression class
posts/hr
Hour
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
5
10
15
20



f(x) = -0.0010968x + 1.5931
g(x) = 0.0013533x + 1.9698
Volume
Time (in days)
f(x) = -0.00015787x + 0.24893
g(x) = 0.00014436x + 0.28332
Replies
Time (in days)
x 10-3
f(x) = 4.8122e-06x + 0.0033955
g(x) = 1.5179e-06x + 0.0028197
NA
Time (in days)
f(x) = 8.2818e-05x + 0.047963
g(x) = 1.986e-05x + 0.02515
Depression terms
Time (in days)
100
200
300

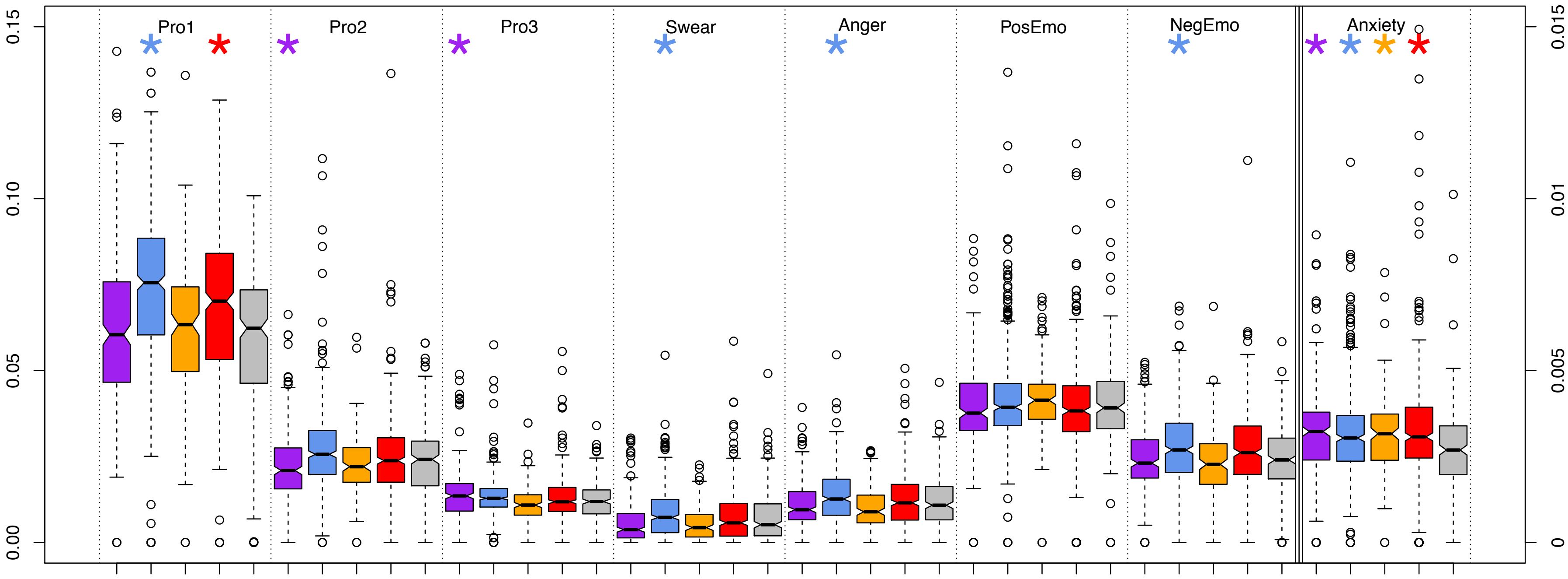
Summary

Egonetwork measures	Depres. class	Non-depres. class
#followers/inlinks	26.9 (σ=78.3)	45.32 (σ=90.74)
#followees/outlinks	19.2 (σ=52.4)	40.06 (σ=63.25)
Reciprocity	0.77 (σ=0.09)	1.364 (σ=0.186)
Prestige ratio	0.98 (σ=0.13)	0.613 (σ=0.277)
Graph density	0.01 (σ=0.03)	0.019 (σ=0.051)
Clustering coefficient	0.02 (σ=0.05)	0.011 (σ=0.072)
2-hop neighborhood	104 (σ=82.42)	198.4 (σ=110.3)
Embeddedness	0.38 (σ=0.14)	0.226 (σ=0.192)
#ego components	15.3 (σ=3.25)	7.851 (σ=6.294)

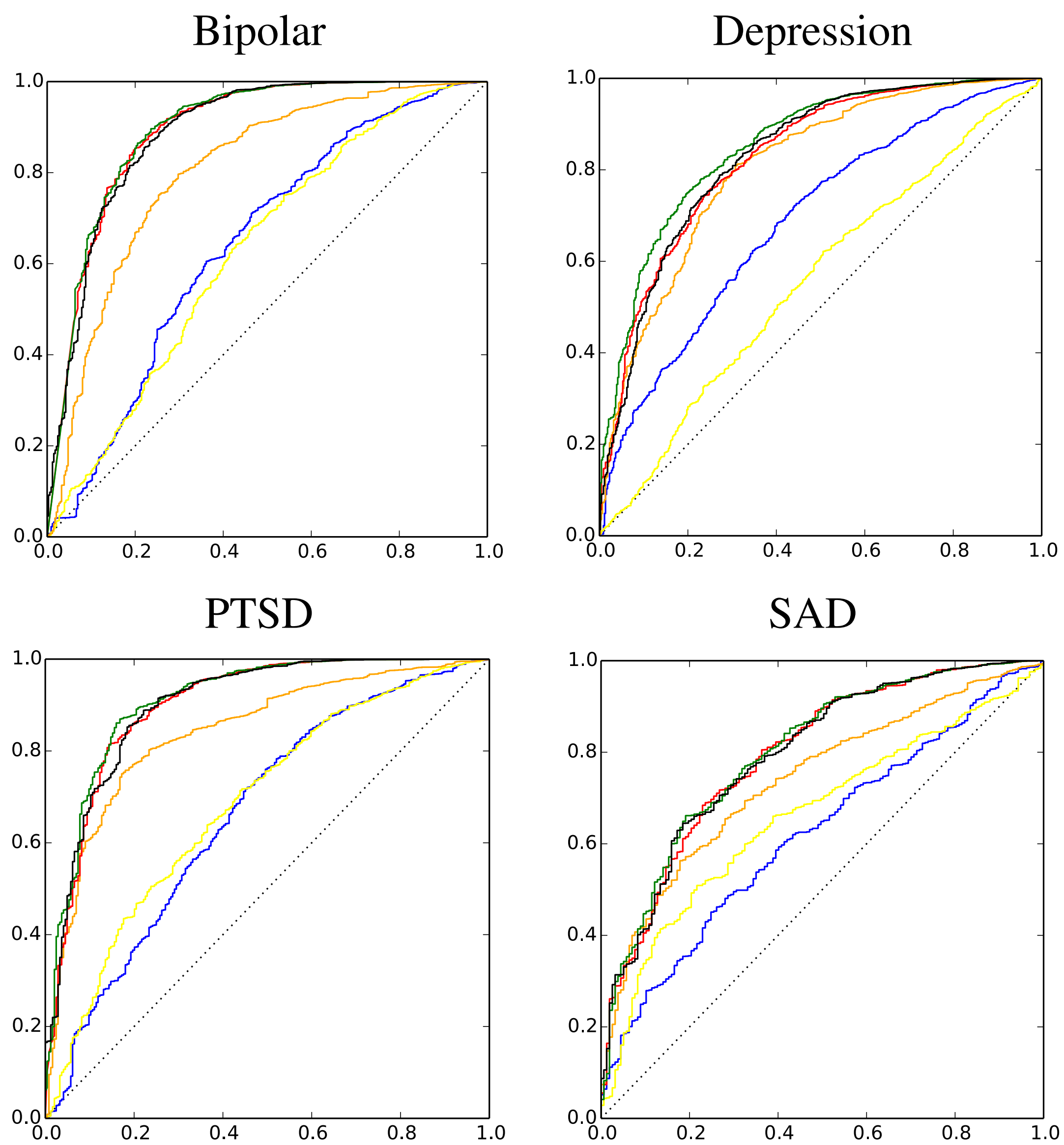
Quantifying Mental Health Signals on Twitter

Summary

- Overcome challenges of solicited survey data – use self-identified mental health diagnoses on Twitter for obtaining ground truth.



Summary



What are the differences you observed between the two studies? What are the strengths and limitations of those differences?

A consistent challenge in many prediction tasks like these, is gathering gold standard information (or ground truth). What could be different ways to get at this problem?

In many ways, predictive models are never 100% perfect. How can social media platforms leverage the predictive methodologies outlined in the two papers?

Both papers used Twitter as the data source of study – in ways it is a nice platforms where an individual can make their profile however they wish it to be. Would the same findings hold on a platform that enforces real identities, like Facebook?

Depression is not an online condition, but one that spans both the online and the offline life. The papers do not take offline attributes into their models.

Is there a way to that into account? What would be the most significant offline attributes to consider?

Anurag brings up the concern of “awareness contamination”. To what extent do you think this would impact a predictive model like the one proposed in the papers going forward? How would you combat that?

The analyses in Coppersmith et al. indicate that depression, PTSD, and bipolar disorder have high mutual correlation, but correlation is low between SAD and the others. Why do you think that is the case – is there anything in the model or the data that account for it?

In the first paper, the ground truth was obtained from Amazon mechanical turk workers. Anything odd or specific about this population that may have affected the findings? What would be alternative ways of recruiting people?

Coppersmith et al shows how *n*-gram language models can provide better performance over LIWC features. What is the downside of using language models though?