Source:

<https://www.google.com/search?q=resnick+maryland&oq=resnick+maryland&gs_lcrp=EgZjaHJvbWUyBggAEEUYOTIICAEQABgWGB4yCAgCEAAYFhgeMggIAxAAGBYYHjIICAQQABgWGB7SAQg5NDY5ajBqMagCALACAA&sourceid=chrome&ie=UTF-8>

<https://www.google.com/search?q=munmun+de+choudhury+google+scholar&oq=munmun+de+&gs_lcrp=EgZjaHJvbWUqBwgCEAAYgAQyBggAEEUYOTIHCAEQLhiABDIHCAIQABiABDIICAMQABgWGB4yBggEEAUYLNIBCDg1NDZqMGoxqAIAsAIA&sourceid=chrome&ie=UTF-8>

<https://www.google.com/search?q=mark+dredze&oq=mark+dredze&gs_lcrp=EgZjaHJvbWUyBggAEEUYOTIHCAEQABiABDIHCAIQABiABDIICAMQABgWGB4yCggEEAAYDxgWGB4yCggFEAAYDxgWGB7SAQkxNTY2NmoxajGoAgCwAgA&sourceid=chrome&ie=UTF-8>

**Mark Dreze**

**Mental Health Surveillance over social media with digital media**

chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/<https://aclanthology.org/W19-3013.pdf>

**MunMun**

**Mental health discourse on reddit: self-disclosure, social support, anonymity**

<https://ojs.aaai.org/index.php/ICWSM/article/view/14526/14375>

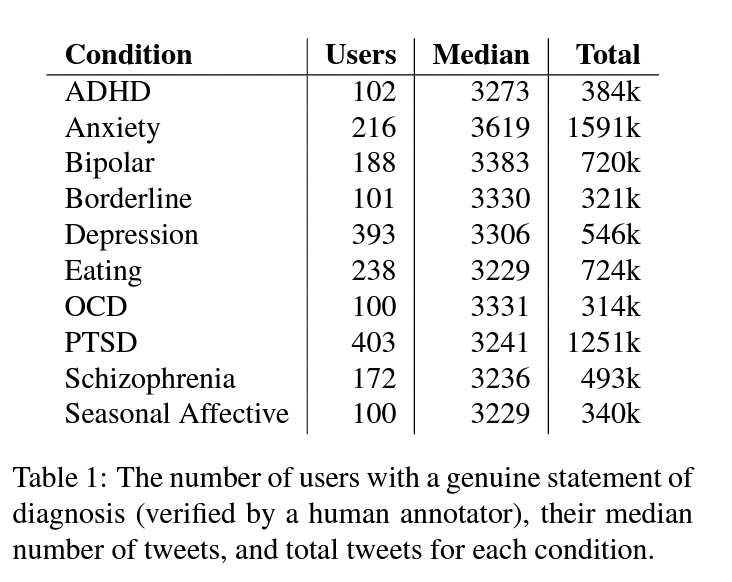
* cumulative density distribution (CDF) of number Of redditors authoring posts and authoring comments. X axis— number of posts/comments, Y axis—number of users

**Mark Dredze**

**From ADHD to SAD: Analyzing the Language of Mental Health on Twitter through Self-Reported Diagnoses** - page 5

chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/<https://aclanthology.org/W15-1201.pdf>

* Using LIWC - quantifiable signals relevant to mental health in the path
* Character n-gram language models - CLM



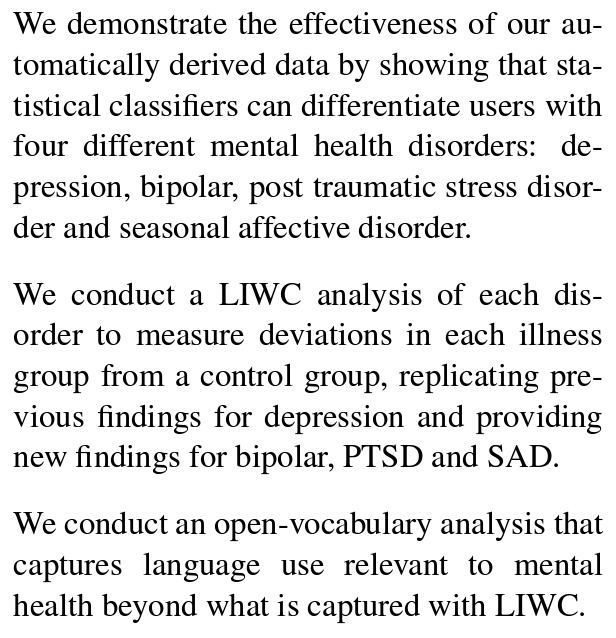
**Discovering health topics in social media using topic models** - model description

<https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0103408>

* Topic modeling: LDA - probability distribution over topics, associated with distribution over words, clustering words into topics

**Quantifying mental health signal in twitter** - LIWC, LM,

chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/<https://aclanthology.org/W14-3207.pdf>

* 

text document has some probability distribution over “topics,” and each such topic is associated with a distribution over words. Topics are not observed as input, rather they are inferred. Topic models are unsupervised models; they can be thought of as automatically clustering words into topics and associating documents with those topics.

**Description of the Dataset**

What does the dataset contain?

* Title
* Post text
* Post URL
* User id
* Score
* Upvoted ratio
* Total comments
* New features - NLP

We have 9 different subreddit groups regarding mental illness, mental health, depression, bipolar, adhd, anxiety, suicidewatch, healthAnxiety, ptsd - total 7 groups of mental health illness

And combined all these dataset into one meta dataset with new columns added of name of subreddit group

Modeling that we are going to use

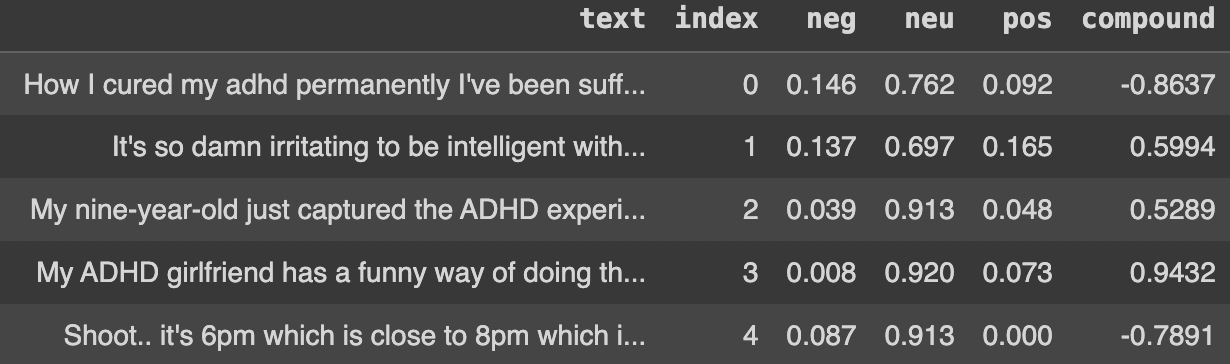
* Sentiment analysis - pos, neg, nat
* LIWC (basically sentiment analysis) <https://medium.com/bright-ml/nlp-feature-extraction-from-liwc-in-python-ded333bc97aa>
* For mental health subreddit - use topic modeling (which mental illness they talk about the most?)

Clarify the posts -

not users

Construct/Features

* Frequency of reddit usage : score , user id
  + Instead using “Frequency” we are using
* Readability of posts: title, post text, total comments, upvote ratio
* Emotionality of posts: compound sentiment score
* Prediction of mental illness: construct 1 to 3 (no assigned feature yet)



1. Clarify unit of analysis
2. For separated subreddit -anxiety : Using Liwc and CLM for posts - emotions for sentiment analysis
3. (Topic modeling for mental health subreddit - subreddit posts?? contains all mental illness, so we can see which topic/word (mental illness) has most frequency) – keep?
4. Should we get rid of frequency, and readability for constructs, modify emotionality (to behavior) and keep mental illness prediction
5. Hypothesis: by examining users’ behavior on social media, we can detect their potential susceptibility to experiencing mental health
6. Temporal precedence + covariance ? Is there some changes in the state of the reddit data/trends

Use the label: is also a concern

Hypothesis -

1. Selection bias? - is it in internal validity? – problem when we have groups we're comparing and the process of selection of groups may drive your observe behavior

are people in depression submission more or less likely to read the news

before vs after: not selection bias (mortality)

find out what your causal theory

1. Ask question 3 and 4

What is the intervention/what happened that might have affected the changes in our group

1. What does a combination of quasi-experimental designs mean? How to we build a combination > how to demonstrate it on pilot data, if we only have one observation (for each group) in our data

Differences between dp vs anxiety major capital, covid vaccines, mental health around covid.. What are the things that to evaluate? Quasi experiment around the thing different reddit

2021-2023 middle covid - mental health awareness , sep - suicide awareness month …

Anxiety vs dep

Comparison that is reasonable

Change the questions → use multiple measures

**New Research Question and Hypothesis**

**From chatgpt..**

* Active participation and engagement in Reddit mental health subreddits - measured by frequency and depth of interactions such as comments, upvoted ratios, frequency of posting collected by user id - is associated with improved mental health outcomes among users
* Text analysis of user-generated content within reddit mental health subreddits can identify linguistic markers associated with changes in mental health status, and the presence of these markers can predict shifts in well-being and mental health outcomes
  + Topic modeling: LDA
  + Sentiment analysis: nltk, vader, liwc, clm
  + Language style analysis: analyze changes in linguistic style - reflect shifts in mental health status - LIWC linguistic inquiry and word count
* By sentiment of post text, the upvoted ratio, and post score in reddit within mental health related subreddits are associated with observable changes in users’ linguistic markers or participation patterns and these features can be used to predict mental health indicators

**What i came up - using multiple measurements (title, post text, upvoted ratio, score, user id, date)**

Comparative analysis of linguistic markers and post engagement within nine mental illness subreddit groups can reveal distinct mental health behaviors of users

* + Regression analysis: post popularity measured by score - dependent - and upvoted ratio/post text sentiment - independent -. (multiple linear regression, logistic regression, poisson regression)
  + Text analysis: topic modeling - LDA, Sentiment analysis - liwc, clm, nltk-sid
  + Clustering and classification: cluster user id and posts based on characteristics - k mean clustering for user id & classify posts into categories based on text
  + Descriptive statistics: distribution of upvoted ratios, scores and user engagement

Adding to the hypothesis, we wanted to provide you with further details regarding the measurement of mental health behaviors

We plan to employ two different measurement approaches:

1. self-disclosure: using self-disclosure with if users used specific keywords "diagnosed" or "feel"

2. text analysis: we will analyze the text data from posts within each subreddit groups by frequency of words most commonly used and sentiment of the posts (polarity score of negative, positive or neutral) to compare each mental illness

<https://medium.com/bright-ml/nlp-feature-extraction-from-liwc-in-python-ded333bc97aa>

<https://medium.com/decision-automation/using-machine-learning-models-in-making-decisions-751ca1f4808c>

"Hypothesis: Comparative analysis of post engagement and linguistic markers within nine distinct mental illness subreddit groups can reveal distinct mental health patterns of users.

This alternative hypothesis focuses on the comparative analysis of post engagement and linguistic markers without assuming prior knowledge of mental health status. It suggests that the examination of user behavior and language use in these subreddits will reveal distinct patterns that can be used to differentiate between the subreddits and identify unique characteristics of each online community.