Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Behavioral Sciences

Detecting depression and mental illness on social media: an integrative review

Sharath Chandra Guntuku[1], David B Yaden[1], Margaret L Kern[2], Lyle H Ungar[1] and Johannes C Eichstaedt[1]

Although rates of diagnosing mental illness have improved over the past few decades, many cases remain undetected. Symptoms associated with mental illness are observable on Twitter, Facebook, and web forums, and automated methods are increasingly able to detect depression and other mental illnesses. In this paper, recent studies that aimed to predict mental illness using social media are reviewed. Mentally ill users have been identified using screening surveys, their public sharing of a diagnosis on Twitter, or by their membership in an online forum, and they were distinguishable from control users by patterns in their language and online activity. Automated detection methods may help to identify depressed or otherwise at-risk individuals through the large-scale passive monitoring of social media, and in the future may complement existing screening procedures.

Addresses
[1] University of Pennsylvania, Philadelphia, PA, United States
[2] The University of Melbourne, Melbourne, Australia

Corresponding author: Eichstaedt, Johannes C (Johannes.penn@gmail.com)

Current Opinion in Behavioral Sciences 2017, **18**:43–49

This review comes from a themed issue on **Big data in the behavioural sciences**

Edited by **Michal Kosinski** and **Tara Behrend**

http://dx.doi.org/10.1016/j.cobeha.2017.07.005

Introduction

The widespread use of social media may provide opportunities to help reduce undiagnosed mental illness. A growing number of studies examine mental health within social media contexts, linking social media use and behavioral patterns with stress, anxiety, depression, suicidality, and other mental illnesses. The greatest number of studies of this kind focus on depression. Depression continues to be under-diagnosed, with roughly half the cases detected by primary care physicians [1] and only 13–49% receiving minimally adequate treatment [2]. Automated analysis of social media potentially provides methods for early detection. If an automated process could detect elevated depression scores in a user, that individual could be targeted for a more thorough assessment, and provided with further resources, support, and treatment. Studies to date have either examined how the use of social media sites correlates with mental illness in users [3] or attempted to detect mental illness through analysis of the content created by users. This review focuses on the latter: studies aimed at predicting mental illness using social media. We first consider methods used to predict depression, and then consider four approaches that have been used in the literature. We compare the different approaches, provide direction for future studies, and consider ethical issues.

Prediction methods

Automated analysis of social media is accomplished by building predictive models, which use 'features,' or variables that have been extracted from social media data. For example, commonly used features include users' language encoded as frequencies of each word, time of posts, and other variables (see Figure 2). Features are then treated as independent variables in an algorithm (e.g. Linear Regression [4] with built in variable selection [5], or Support Vector Machines (SVM)) [6] to predict the dependent variable of an outcome of interest (e.g. users' mental health). Predictive models are trained, using an algorithm, on part of the data (the training set) and then are evaluated on the other part (the test set) to avoid overfitting — a process called cross-validation. The prediction performances are then reported as one of several possible metrics (see Table 1).

Assessment criteria

Several approaches have been studied for collecting social media data with associated information about the users' mental health. Participants are either recruited to take a depression survey and share their Facebook or Twitter data (section A below), or data is collected from existing public online sources (sections B, C, and D below; see Figure 1). These sources include searching public Tweets for keywords to identify (and obtain all Tweets from) users who have shared their mental health diagnosis (section B), user language on mental illness related forums (section C), or through collecting public Tweets that mention mental illness keywords for annotation (section D). The approaches using public data (sections B, C, D) have the advantage that much larger samples can, in

Table 1

Prediction performances achieved by different mental illness studies reviewed in this paper. The relevant dataset, features, and prediction settings are provided.

Ref.	Year	Dataset			Section	Mental Illness Criteria	Features (predictors)						Outcome Type	Model	Metric	Performance
		Platform	N (users)	Cases (conditions; base rate [BR])			n-grams	LIWC	Sentiment	Topics	Metadata	Others				
[8]	2013	Twitter	476	Depression = 171 (BR = 36%)	A	survey (CESD + BDI)		Y	Y		Y	Social Network	Binary	PCA, SVM w/ RBF Kernel	Accuracy	.72
[13]	2014	Facebook	165	Post-partum Depression = 28 (BR = 17%)	A	survey (PHQ-9)		Y	Y		Y	User Activity, Social Capital	Binary	Logistic Regression	pseudo-R2[b]	.36
[14]	2014	Facebook	28,749	(continous Depression score)	A	survey (Personality)	Y	Y		Y			Continuous	Ridge Regression	Correlation	.38
[12]	2015	Twitter	209	Depression = 81 (BR = 39%)	A	survey (CESD)	Y	Y	Y	Y	Y	User Activity	Binary	SVM	Accuracy	.69
[11]	2016	Twitter	378	Depression = 105 (BR = 28%) PTSD = 63 (BR = 17%)	A	survey (CESD)		Y	Y		Y	Time-Series, LabMT	Binary	Random Forests	AUC	Depression = .87 PTSD = .89
[40]	2014	Twitter	5,972	PTSD = 244 (BR = 4%)	B	self-declared	Y	Y					Binary	(not reported)	ROC	(AUC not reported)
[42]	2014	Twitter	21,866	11,866 (across 4 Conditions, BR = 54%)	B	self-declared	Y	Y	Y		Y	User Activity	Binary	Log linear classifier	Precision[a]	Depression = .48 Bipolar = .64 PTSD = .67 SAD = .42
[17]	2015	Twitter	1,957	Depression = 483 (BR = 25%) PTSD = 370 (BR = 19%)	B	self-declared	Y	Y	Y	Y		Age, Gender, Personality	Binary	Logistic Regression	AUC	Depression = .85 PTSD = .91
[21]	2015	Twitter	4,026	2,013 (across 10 Conditions, BR = 50%)	B	self-declared	Y	Y					Binary	(not reported)	Precision[a]	Depression = .48 Bipolar = .63 Anxiety = .85 Eating Dis. = .76
[41]	2016	Twitter	250	Suicide Attempt = 125 (BR = 50%)	B	self-declared	Y		Y		Y	User Activity	Binary	(not reported)	Precision[a]	.70
[43]	2016	Twitter	900	Depression = 326 (BR = 36%)	B	self-declared	Y						Binary	Naive Bayes	AUC	.70
[19]	2017	Twitter	9,611	4820 (across 8 Conditions, BR = 50%)	B	self-declared	Y					Gender	Multi-Task	Neural Network	AUC	Depression = .76 Bipolar = .75 Depression = .76 Suicide Attempt = .83

AUC: Area Under the Receiver Operating Characteristic (ROC) Curve; Precision: fraction of cases ruled positive that are truly positive; Accuracy: fraction of cases that are correctly labeled by the model; SVM: Support Vector Machines; PCA: Principal Component Analysis; RBF — Radial Basis Function.

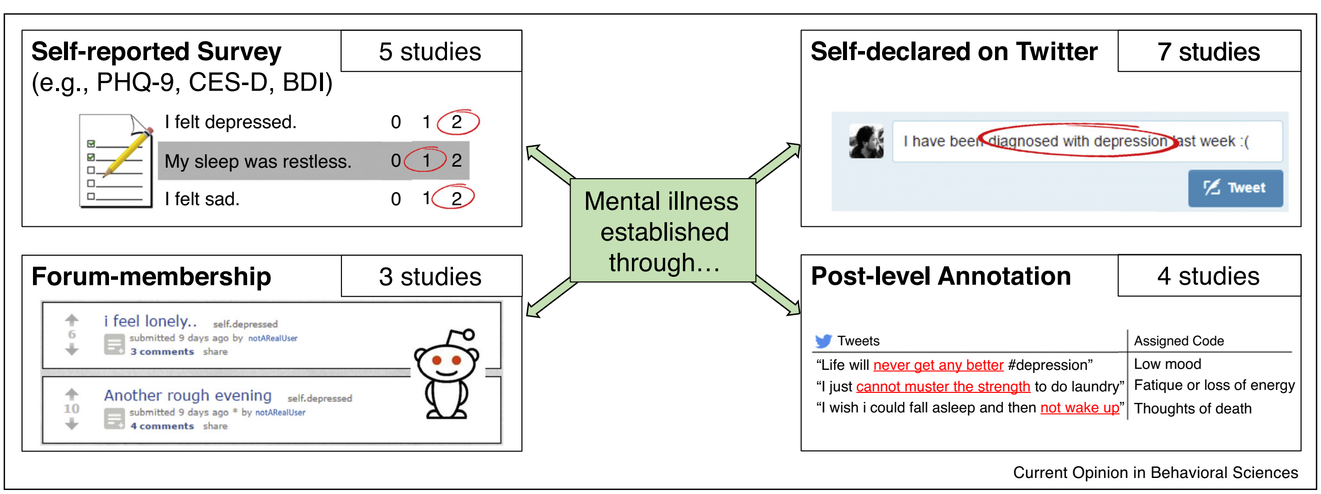
[a]Precision with 10% False Alarms.

[b]Within-sample (not cross-validated).

[c]Using the Depression facet of the Neuroticism factor measured by the International Personality Item Pool (IPIP) proxy to the NEO-PI-R Personality Inventory [38].

Studies highlighted in green report AUCs; AUCs are not base rate dependent and can be compared across studies.

Figure 1



Data sources used in studies as assessment criteria to establish mental illness status. The number of studies selected for review in the present article is provided. The most commonly used self-reported screening surveys for depression include the PHQ-9 = Patient Health Questionnaire [7], CES-D = Centers for Epidemiological Studies Depression Scale Revised [9], BDI = Beck Depression Inventory [10].