SCIENTIFIC REPORTS

OPEN

Forecasting the onset and course of mental illness with Twitter data

Andrew G. Reece[1], Andrew J. Reagan[2,3], Katharina L. M. Lix[4], Peter Sheridan Dodds[2,3], Christopher M. Danforth[2,3] & Ellen J. Langer[1]

Received: 8 September 2016
Accepted: 18 September 2017
Published online: 11 October 2017

We developed computational models to predict the emergence of depression and Post-Traumatic Stress Disorder in Twitter users. Twitter data and details of depression history were collected from 204 individuals (105 depressed, 99 healthy). We extracted predictive features measuring affect, linguistic style, and context from participant tweets (N = 279,951) and built models using these features with supervised learning algorithms. Resulting models successfully discriminated between depressed and healthy content, and compared favorably to general practitioners' average success rates in diagnosing depression, albeit in a separate population. Results held even when the analysis was restricted to content posted before first depression diagnosis. State-space temporal analysis suggests that onset of depression may be detectable from Twitter data several months prior to diagnosis. Predictive results were replicated with a separate sample of individuals diagnosed with PTSD (N_{users} = 174, N_{tweets} = 243,775). A state-space time series model revealed indicators of PTSD almost immediately post-trauma, often many months prior to clinical diagnosis. These methods suggest a data-driven, predictive approach for early screening and detection of mental illness.

Social media data provide valuable clues about physical and mental health conditions. This holds true even in cases where social media users are not yet aware that their health has changed. For example, searching for information on certain health symptoms has been shown to provide accurate early-warning indicators for hard-to-detect cancers[1]. Social media networks have been used to plot the trajectory of disease outbreaks[2–4], and to track regional dietary health[5]. In addition to physical ailments, predictive screening methods have successfully identified markers in social media data for a number of mental health issues, including addiction[6], depression[7–12], Post-Traumatic Stress Disorder (PTSD)[13,14], and suicidal ideation[15]. The field of predictive health screening with social media data is still in its infancy, however, and considerable refinements are needed to develop methodologies that can effectively augment health care. In this report, we present a set of improved methods and novel contributions for predicting and tracking depression and PTSD on Twitter.

Depression has emerged as the leading mental health condition of interest among computational social scientists[7–12], as it is a relatively common mental disorder[16] and influences a range of behaviors and patterns of communication[17]. Underdiagnosis of depression remains a persistent problem; a recent survey of a major metropolitan area found nearly half (45%) of all cases of major depression were undiagnosed[18]. PTSD, while less common[19], is frequently comorbid with major depression[20]. Studies have found that PTSD is underdiagnosed or under-treated by a majority of primary-care physicians[21,22]. The costs of underdiagnosis of these conditions, both to human quality of life and health care systems, are considerable. Computational methods for early screening and diagnosis of depression and PTSD have the potential to make a positive impact on a major public health issue, with minimal associated costs and labor intensity.

Improvements and novel contributions. Early efforts to detect depression and PTSD signals in Twitter data have been promising. Park *et al.*[11] established that Twitter users suffering from depression tended to post tweets containing more negative emotional sentiment compared to healthy users. De Choudhury *et al.*[7] successfully identified new mothers suffering from postpartum depression, based on changes in Twitter usage and tweet content. In a separate analysis, De Choudhury *et al.*[8] found that depressive signals were observable in tweets made

[1]Department of Psychology, Harvard University, Cambridge, MA, 02138, USA. [2]Computational Story Lab, Vermont Advanced Computing Core, and the Department of Mathematics and Statistics, University of Vermont, Burlington, VT, 05401, USA. [3]Vermont Complex Systems Center, University of Vermont, Burlington, VT, 05401, USA. [4]Department of Management Science and Engineering, Stanford University, Palo Alto, CA, 94305, USA. Correspondence and requests for materials should be addressed to A.G.R. (email: andrew.reece@post.harvard.edu) or C.M.D. (email: chris.danforth@uvm.edu)

by individuals with Major Depressive Disorder. In addition, De Choudhury *et al.*[23] found that increased social isolation, measured via Facebook data, was predictive of postpartum depression in mothers. A small number of studies have attempted to identify PTSD markers in Twitter data[13,14].

This growing literature has employed progressively more sophisticated methods for making intelligent inferences about Twitter users' mental health based on their online activity. Despite these advances, we identified a number of methodological issues in recent reports which we have improved upon in the present work. A brief review of these modifications provide motivation for the results that follow.

De Choudhury *et al.*[8] built a predictive model using tweets from depressed individuals posted within a year prior to their self-reported onset of a recent depressive episode. Subjects were included for analysis if they had experienced at least two depressive episodes within that one year period, meaning that the data used to make predictions contained tweets posted after the first onset of depression. The date of first depression diagnosis for each individual was not explicitly accounted for in their model. As a result, model training data may have contained both tweets posted during a previous depressive episode, as well as tweets posted after subjects had already received a formal diagnosis. Both of these possibilities seem especially likely, as depression-related terms such as *diagnosis*, *antidepressants*, *psychotherapy*, and *hospitalization* were significant predictors in their model, along with the names of specific antidepressant medications (e.g. *serotonin*, *maprotiline*, and *nefazodone*).

We chose to use only tweets posted prior to the date of subjects' first depression diagnosis, rather than focus on recent depressive episodes, for three reasons. First, self-reported information about depressive symptoms is often inaccurate[24]. By contrast, a clinical diagnosis is an explicit event that does not rely on subjective impressions, as may be the case with self-reported onset dates. Second, individuals diagnosed with depression often come to identify with their diagnosis[25,26], and subsequent choices, including how to portray oneself on social media, may be influenced by this identification. It is possible that the predictive signals indicated in De Choudhury *et al.*[8] were not tracking depressive symptoms, per se, but rather identified purposeful communication choices on the part of depressed Twitter users. Third, and most important, if we are able to accurately discriminate between depressed and healthy participants using only tweets posted prior to first diagnosis, this would support a stronger claim than has been made previously - namely, that Twitter data are capable not only of detecting depression, but can do so before the first diagnosis has been made.

While date of first diagnosis provides a more reliable temporal marker than self-reported onset of symptoms, onset timing is also valuable to researchers and health care professionals looking to better understand depression. This is especially true regarding the onset of an individual's *first* depressive episode. Winokur[27] found that over 50% of depression patients experienced first onset at least 6 months prior to diagnosis. The months during which individuals suffering from depression are undiagnosed and untreated pose a significant health risk. Given that the changes that occur with the first onset of depression may be reflected in social media data, we hypothesized that a computational approach could model the progression of depression without any explicit estimates of onset. Using only the content of participants' tweets, we generated a time series model which charts the course of illness in depressed individuals, and compared this with healthy participants' data. To our knowledge, De Choudhury *et al.*[8] represent the state-of-the-art in depression screening on Twitter, and our own work was informed by their innovative methods. Accordingly, we report model accuracy scores from De Choudhury *et al.*[8] along with our results as a point of comparison.

All of the above methodological improvements were also applied to our PTSD analysis, which used a separate cohort of study participants. Extant literature on PTSD detection in Twitter data[13,14] differ from our analysis in important ways. Previous research used bulk collections of public tweets, and assigned PTSD labels to users based on tweets which mentioned a PTSD diagnosis. By comparison, we communicated directly with participants, and excluded any who could not report a specific date on which they received a professional clinical diagnosis. Our analytical approach incorporated a wide array of metadata features and semantic measures, which were limited[14] or missing entirely[13] from earlier research. Most importantly, previous research focused only on differentiating PTSD users from healthy users, without any consideration of timing with respect to the dates of traumatic events or diagnoses. Our models focused specifically on identifying predictive markers of PTSD prior to diagnosis date, as well as tracking the course of this disorder over time.

Comparison with trained healthcare professionals. Mitchell, Vaze, and Rao[28] evaluated general practitioners' abilities to correctly diagnose depression in their patients, without assistance from scales, questionnaires, or other measurement instruments. Out of 50,371 patient outcomes culled from 118 studies, 21.9% of patients were actually depressed. General practitioners were able to correctly rule out depression in 81% of non-depressed patients, but only correctly diagnosed depressed patients 42% of the time. Taubman-Ben-Ari *et al.*[22] tested primary-care physicians' abilities to detect PTSD. PTSD prevalence was 7.5% for men and 10.5% for women in the observed sample (N = 683). Physicians correctly identified 2.5% of PTSD cases, and out of all PTSD diagnoses made, only 43% were accurate. We refer to general practitioner accuracy rates[22,28] as an informal benchmark for the quality of our computational models.

Method

The methods used in recruitment, data collection, and analysis are adopted from Reece and Danforth[12]. The present study was reviewed and approved by the Harvard University Institutional Review Board, approval #15-2529, as well as the University of Vermont Institutional Review Board, approval #CHRMS-16-135. All experimental procedures were performed in accordance with Institutional Review Board guidelines. All study participants provided informed consent and acknowledged all of the study goals, expectations, and procedures, including data privacy, prior to any data collection. Surveys were built using the Qualtrics survey platform, and analyses were performed using Python and R. Twitter data collection apps were written in Python, using the Twitter developer's Application Programming Interface (API).

Data Collection. Participants were recruited using Amazon's Mechanical Turk (MTurk) crowdwork platform, and we collected user data from both the survey on MTurk and participants' Twitter history. Recruitment and data collection procedures were identical for depression and PTSD samples, with the exception of the condition-specific questionnaire used for screening. Separate surveys were created for affected and healthy samples. The term "affected" here refers to participants affected by either depression or PTSD, respectively. In the affected sample surveys, participants were invited to complete a questionnaire that involved passing a series of inclusion criteria, responding to a standardized clinical assessment survey, answering questions related to demographics and mental health history, and sharing social media history. We used the CES-D (Center for Epidemiologic Studies Depression Scale) questionnaire to screen participant depression levels[29]. CES-D assessment quality has been demonstrated as on-par with other depression inventories, including the Beck Depression Inventory and the Kellner Symptom Questionnaire[30,31]. The Trauma Screening Questionnaire (TSQ) was used to screen for PTSD[32]. A comparison cohort of healthy participants were screened to ensure no history of depression or PTSD, respectively, and for active Twitter use.

Qualified participants were asked to share their Twitter usernames and history. An app embedded in the survey allowed participants to securely log into their Twitter accounts and agree to share their data. Upon securing consent, we made a one-time collection of participants' Twitter posting history, up to the most recent 3,200 tweets (this limit is imposed by the Twitter API). In total we collected 279,951 tweets from 204 Twitter users for the depression analysis, and 243,775 tweets from 174 Twitter users for the PTSD analysis. Details on participant data protection measures are outlined below.

Inclusion criteria. The surveys for affected samples collected age data from participants, and asked qualified participants questions related to their first clinical diagnosis of either depression or PTSD, as well as questions about social media usage at the time of diagnosis. These questions were given in addition to the CES-D or TSQ scales. The purpose of these questions was to determine:

- The date of first clinical diagnosis of the condition,
- Whether or not the individual suspected having the condition before diagnosis, and,
- If so, the number of days prior to diagnosis that this suspicion began

In the case that participants could not recall exact dates, they were instructed to approximate the actual date.

The survey for healthy participants collected age and gender data from participants. It also asked four questions regarding personal health history, which were used as inclusion criteria for this and three other studies. These questions were as follows:

- Have you ever been pregnant?
- Have you ever been clinically diagnosed with depression?
- Have you ever been clinically diagnosed with Post-Traumatic Stress Disorder?
- Have you ever been diagnosed with cancer?

Participants' responses to these questions were not used in analysis, and only served to include qualified respondents in each of the various studies, including the depression- and PTSD-related studies reported here.

Participant safety and privacy. This study design raised two important issues regarding ethical research practices, as it concerned both individuals with mental illness and potentially personally identifiable information. We were unable to guarantee strict anonymity to participants, given that usernames and personal information posted to Twitter are often inherently specific to participants' identities (such as usernames and tweets containing real names). As we potentially had the capacity to link study participants' identities to sensitive health information, study participants were informed of the risks of being personally identified from their social media data. Participants were assured that no personal identifiers, including usernames, would ever be made public or published in any format. We used Turk Prime, an interface for conducting MTurk studies, to mask participants' MTurk worker IDs from our records. We made it clear that any links between social media data and private personal health data would be available only to our team of researchers, and participants were able to request to have their data removed at any time.

Improving data quality. In an effort to minimize noisy and unreliable data, we applied several quality assurance measures in our data collection process. MTurk workers who have completed at least 100 tasks, with a minimum 95% approval rating, have been found to provide reliable, valid survey responses[33]. We restricted survey visibility only to workers with these qualifications. Survey access was also restricted to U.S. IP addresses, as MTurk data collected from outside the United States are generally of poorer quality[34]. All participants were only permitted to take the survey once.

We excluded participants with a total of fewer than five Twitter posts. We also excluded participants with CES-D scores of 21 or lower (depression), or TSQ scores of 5 or lower (PTSD). Studies have indicated that a CES-D score of 22 represents an optimal cutoff for identifying clinically relevant depression[35,36]; an equivalent TSQ cutoff of 6 has been found to be optimal in the case of PTSD[32]. We note here that in the study that inspired the present work, De Choudhury *et al.*[8] used two depression scales (CES-D and BDI), and filtered individuals whose depression score did not correlate across the both scales. This additional criteria is a methodological strength of De Choudhury *et al.*[8] with respect to the present work.

Depression	Users	Posts	Posts $\mu(\sigma)$	Posts (median)
Total	204	279,951	1373 (1282)	862
Depressed	105	164,218	1564 (1332)	1127
Healthy	99	115,733	1169 (1200)	574
PTSD	**Users**	**Posts**	**Posts μ (σ)**	**Posts (median)**
Total	174	243,775	1401 (1284)	946.5
Has PTSD	63	91,589	1564 (1332)	1058
Healthy	111	152,186	1371 (1268)	893

Table 1. Summary statistics for depression and PTSD tweet collection ($N_{depr} = 279{,}951$, $N_{ptsd} = 243{,}775$).

Summary statistics. All data collection took place between February 1, 2016 and June 10, 2016. Across both depressed and healthy groups, we collected data from 204 Twitter users, totaling 279,951 tweets. This number includes up to 3,200 tweets from each participant's Twitter history; the analyses in this report focus only on tweets from depressed users created before the date of first depression diagnosis. The mean number of posts per user was 1372.71 (SD = 1281.74). This distribution was skewed by a smaller number of frequent posters, as evidenced by a median value of just 861 posts per user. See Table 1 for summary statistics.

In the depressed group, 147 crowdworkers successfully completed participation and provided access to their Twitter data. Imposing the CES-D cutoff reduced the number of viable participants to 105. The mean age for viable participants was 30.3 years (SD = 8.34), with a range of 18 to 64 years. Dates of participants' first depression diagnoses ranged from March 2010 to February 2016, with nearly all diagnosis dates (92%) occurring in the period 2013–2015. In the healthy group, 99 participants completed participation and provided access to their Twitter data. The mean age for this group was 33.9 years, with a range of 19 to 63 years, and 42% of respondents were female. (Gender data were not collected for affected sample surveys).

For the PTSD analysis, we collected data from 174 Twitter users, totaling 243,775. The mean number of posts per user was 1372.71 (SD = 1281.74). This distribution was skewed by a smaller number of frequent posters, as evidenced by a median value of just 862 posts per user.

In the PTSD sample, 73 crowdworkers successfully completed participation and provided access to their Twitter data. Imposing the TSQ cutoff reduced the number of viable participants to 63. The mean age for viable participants was 30.64 years (SD = 7.57), with a range of 21 to 54 years. Dates of participants' first PTSD diagnoses ranged from April 2010 to December 2015, with nearly all diagnosis dates (94%) occurring in the period 2013–2015. In the healthy group, 111 participants completed participation and provided access to their Twitter data. The mean age for this group was 33.25 years, with a range of 19 to 63 years, and 51% of respondents were female.

Feature extraction. We extracted several categories of predictors from the Twitter posts collected. Predictor selection for both depression and PTSD was based on prior machine learning models of depression in Twitter data[7,8], as the two conditions' high comorbidity rates suggest their predictive signals may exhibit considerable overlap[20]. Depressed Twitter users have been observed to tweet less frequently than non-depressed users[8], so we used total tweets per user, per day, as a measure of user activity. Tweet metadata was analyzed to assess average word count per tweet (here, a word is defined as a set of characters surrounded by whitespace), whether or not the tweet was a retweet, and whether or not the tweet was a reply to someone else's tweet. The labMT, LIWC 2007, and ANEW unigram sentiment instruments were used to quantify the happiness of tweet language[37–40]. The use of labMT, which has shown strong prior performance in analyzing happiness on Twitter[41,42], is novel with respect to depression screening; ANEW and LIWC have been successfully applied in previous studies on depression and Twitter[7,8,14]. LIWC was also used to compile frequency counts of various parts of speech (e.g., pronouns, verbs, adjectives) and semantic categories (e.g., food words, familial terms, profanity) as additional predictors[37].

Units of observation. Determining the best time span for analysis raises a difficult question: When and for how long does mental illness occur? Receiving a clinical diagnosis of depression or PTSD does not imply that an individual remains in a persistent state of illness, and so to conduct analysis with an individual's entire posting history as a single unit of observation is a dubious proposition. At the other extreme, to take one tweet as a unit of observation runs the risk of being too granular.

De Choudhury *et al.*[8] looked at all of a given user's tweets in a single day, and aggregated those data into per-person, per-day units of observation. In this report we have followed the convention of aggregated "user-days" as a primary unit of analysis, rather than try to categorize a person's entire history, or analyze each individual tweet. In our own previous research, however, we have found that many Twitter users do not generate enough daily content to make for robust unigram sentiment analysis[43]. For completeness, we conducted analyses using both daily and weekly units of observation. Both analyses yielded predictive models of similar strengths, with the weekly model showing a slight, but consistent, edge in performance. We report accuracy metrics from both analyses, but restrict other results to the daily-unit analysis to allow for more direct comparison with previous research. Details of weekly-unit analytical results are available in Supplementary Information, section II. When reporting results we occasionally refer to observations or tweets as "depressed", e.g., "depressed tweets received fewer likes". It would be more correct to use the phrase "tweet data from depressed participants, aggregated by user-days" instead of "depressed tweets", but we chose to sacrifice a degree of technical correctness for the sake of clarity.

Statistical framework. *Machine learning models.* We trained supervised machine learning classifiers to discriminate between affected and healthy sample members' observations. Classifiers were trained on a randomly-selected 70% of total observations, and tested on the remaining 30%. Out of several candidate algorithms, a 1200-tree Random Forests classifier demonstrated best performance. Stratified five-fold cross-validation was used to optimize Random Forests hyperparameters, and final accuracy scores were averaged over five separate randomized runs. See Supplementary Information, section III for optimization details. Precision, recall, specificity, negative predictive value, and F1 accuracy scores are reported, and general practitioners' unassisted diagnostic accuracy rates as reported in Mitchell, Vaze, and Rao[28] (MVR) and Taubman-Ben-Ari *et al.*[22] (TBA) are used as informal benchmarks for depression and PTSD, respectively. In addition to the fact that our results are drawn from different samples, using different observational units, than our chosen comparisons, comparing point estimates of accuracy metrics is not a statistically formal means of model comparison. We felt it was more meaningful to frame our findings in a realistic context, however informal, rather than to benchmark against a naive statistical model that simply predicted the majority class for all observations.

Time series analysis. Predictive screening methods use indirect indicators, such as language use on social media, to infer health status. We are not actually interested in the average word count of depressed individuals' tweets, for example, but rather we hope that this measure will allow us some access to the underlying variable we truly care about: depression. Accordingly, state-space models, which use observable data to estimate the status of a latent, or hidden, variable over time, may provide useful insights. We trained a two-state Hidden Markov Model (HMM) to detect differential changes between affected and healthy groups over time. We used the hmmlearn Python module[44] to fit emission and transition matrices (using expectation-maximization) and hidden state sequence (using the Viterbi path algorithm); this module also provided mean parameter estimates for all predictors, for each latent state.

The use of HMM presents an interpretability challenge: how to know whether resulting latent states have any relationship to the clinical condition of interest? Consider the case of depression: Finding evidence that HMM had, in fact, recovered two states from our data that closely resembled the depressed and healthy classes was prerequisite to making any inferences based on HMM output. We addressed this issue by comparing HMM output with mean differences between depressed and healthy observations in the raw data. If the directions of the differences between HMM mean parameter estimates generally agree with the true differences in the data, this provides evidence that the two sample groups in our data (depressed and healthy) are well-characterized by HMM latent states. For example, if depressed observations contained more sad words on average than healthy observations (variable name: "LIWC_sad"), then the HMM state with the higher LIWC_sad estimate is more likely to be the depressed one, given that HMM does track depression (i.e., the latent states generated by HMM map onto "depressed" and "healthy"). If, on the other hand, HMM-generated states are weakly or not at all related to depression, there should be no clear alignment between HMM means and means in the raw data. The same procedure was applied when fitting an HMM to the PTSD data.

Word shift graphs. Machine learning algorithms provide powerful predictive capability, but most algorithms offer little in the way of context and interpretation. Word shift graphs use the labMT happiness scores, the most important predictor for both depression and PTSD analyses, to show qualitatively how inter-group differences may be driven by the usage of specific words in tweets[39]. We present word shift graphs comparing the way tweet language adjusted happiness scores in affected and healthy samples. The visualization ranks words by their contribution to the happiness difference between the two groups (for more explanation of word shift rankings, see [45,46]). Word shifts are generated from a different statistical method than the machine learning algorithm we used to make predictions, and so should be treated as exploratory analyses independent of our main findings.

Results

For the depression study, we analyzed 74,990 daily observations (23,541 depressed) from 204 individuals (105 depressed). For the PTSD study, we analyzed 54,197 daily observations (13,008 PTSD) from 174 individuals (63 PTSD). Observations from affected sample members accounted for 31.4% and 24% of the entire datasets for depression and PTSD, respectively.

Machine learning classifier. Results are reported for both daily and weekly units of observation (see Table 2 and Fig. 1).

Our best depression classifier, averaged over cross-validation iterations, improved over both Mitchell *et al.*[28] and De Choudhury *et al.*[8] on several metrics. Our depression model's precision rate was considerably higher, with just over 1 false positive for every 10 depression diagnoses. By comparison, general practitioners from Mitchell *et al.*[28] incorrectly diagnosed patients as having depression in more than half of all diagnoses.

Our best PTSD classifier improved considerably over the primary-care physicians from Taubman-Ben-Ari *et al.*[22] (TBA). Whereas more than half of all PTSD diagnoses made by TBA physicians were incorrect, our model was correct in roughly 9 out of every 10 (88.2%) of its PTSD predictions. Model recall rate was strong, with 68.3% discovery of actual PTSD sample observations.

The labMT happiness score was the strongest predictor of both depression and PTSD. Notably, average labMT average happiness over user days showed only modest correlation with ANEW ($r_{depr} = 0.37$, $r_{ptsd} = 0.36$) and LIWC ($r_{depr} = 0.36$, $r_{ptsd} = 0.37$), suggesting that labMT identifies relevant prediction signals not fully captured by other sentiment instruments. The additional benefit offered by labMT in this context may be a reflection of its inclusion of the 5000 most frequently used words on Twitter, including slang[39]. The second most important variable was word count, which represented the average number of words per tweet. Sentiment-related variables from ANEW and LIWC accounted for most of the remaining top predictors (see Fig. 1).

Depression	MVR μ	DC μ	Daily $\mu(\sigma)$	Weekly $\mu(\sigma)$
Recall	0.510	0.614	0.518 (0.000)	0.521 (0.000)
Specificity	0.813	N/A	0.958 (0.000)	0.969 (0.000)
Precision	0.42	0.742	0.852 (0.000)	0.866 (0.000)
NPV	0.858	N/A	0.812 (0.000)	0.841 (0.000)
F1	0.461	0.672	0.644 (0.000)	0.651 (0.000)
PTSD	**TBA μ**	**NHC μ**	**Daily $\mu(\sigma)$**	**Weekly $\mu(\sigma)$**
Recall	0.249	0.82	0.683 (0.000)	0.658 (0.000)
Specificity	0.979	N/A	0.988 (0.000)	0.994 (0.000)
Precision	0.429	0.86	0.882 (0.000)	0.934 (0.000)
NPV	0.602	N/A	0.959 (0.000)	0.954 (0.000)
F1	0.315	0.84	0.769 (0.000)	0.772 (0.000)

Table 2. Classification accuracy metrics for daily and weekly models ($N_{depr} = 74{,}990$, $N_{ptsd} = 54{,}197$). Accuracy scores from Mitchell *et al.*[28] (MVR), De Choudhury *et al.*[8] (DC), Taubman-Ben-Ari *et al.*[22] (TBA), and Nadeem, Horn, & Coppersmith[13] (NHC) are included for comparison to depression (MVR, DC) and PTSD (TBA, NHC) results. Table cells marked N/A indicate unavailable metrics from previous studies.

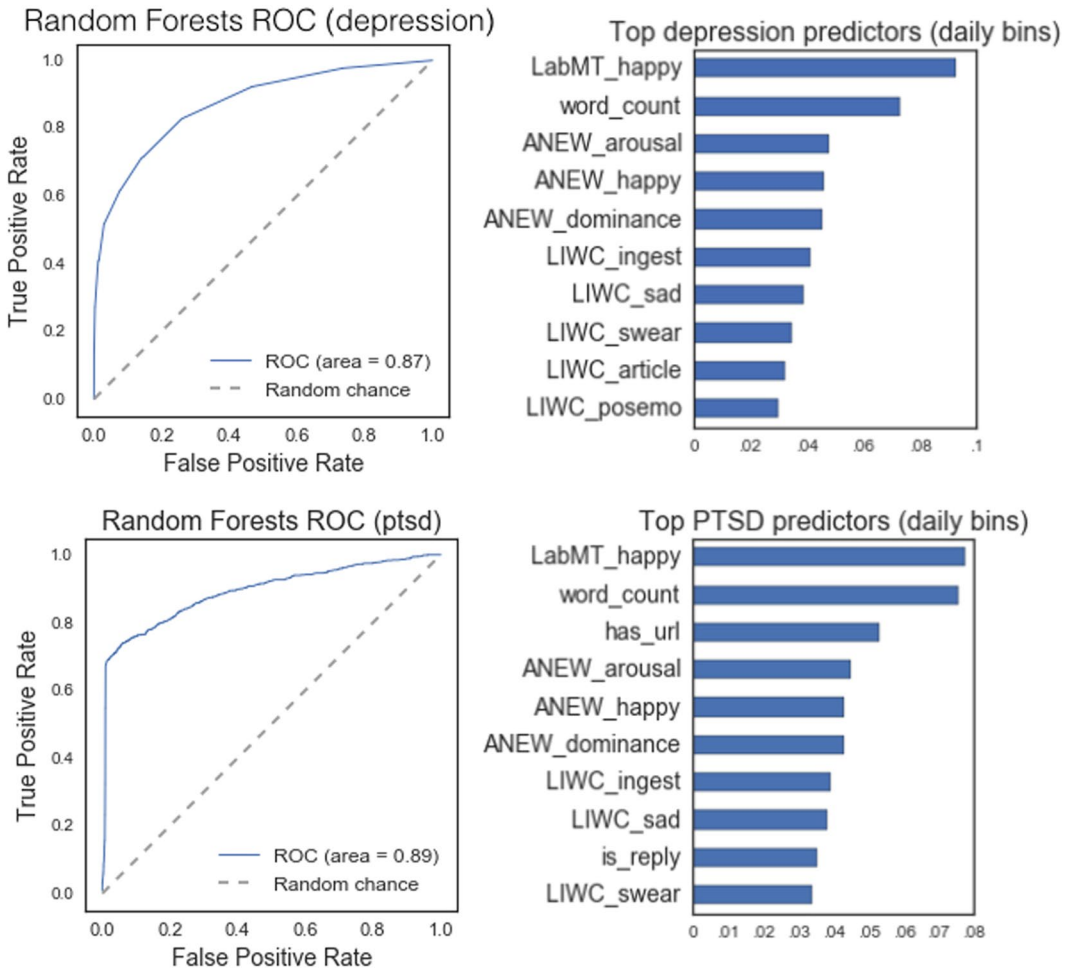

Figure 1. ROC curve and top predictors for Random Forests algorithm, for depression and PTSD samples ($N_{depr} = 74,990$, $N_{ptsd} = 54{,}197$). Predictor names ending in "_happy" are happiness measures; LIWC predictors[37] refer to the occurrence of semantic categories (eg. LIWC_ingest refers to food and eating words, LIWC_swear refers to profanity).

As with most decision-tree classifiers, the Random Forests algorithm provides information on the relevance, but not the directionality, of predictors. In other words, we can know how important a variable was to the algorithm, but not if it was positively or negatively associated with the response variable. Word shift graphs, reported

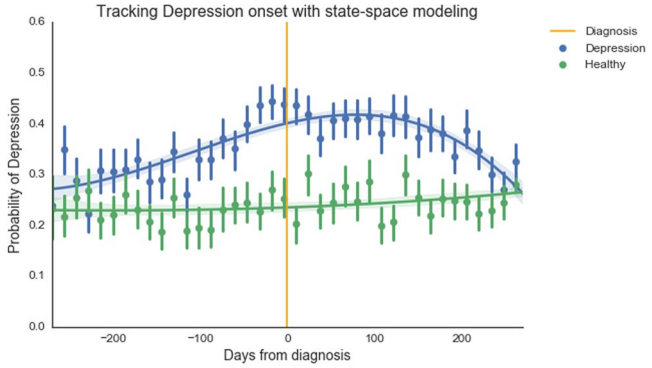

Figure 2. Hidden Markov Model showing probability of depression (N = 74,990). X-axis represents days from diagnosis. Healthy data are plotted from a consecutive time span of equivalent length. Trend lines represent cubic polynomial regression fits with 95% CI bands, points are aggregations of 14 day periods, with error bars indicating 95% CI on central tendency of daily values.

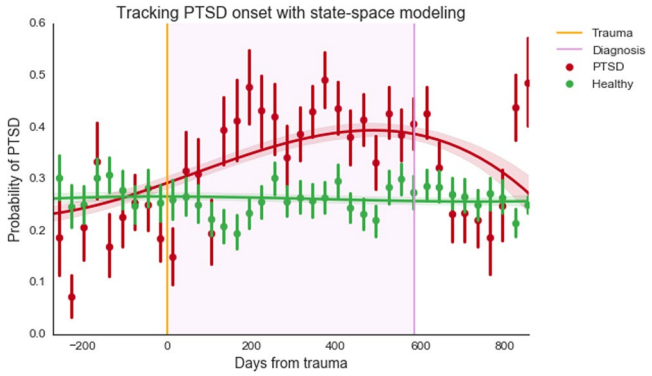

Figure 3. Hidden Markov Model showing probability of PTSD (N = 54,197). X-axis represents days from trauma event. Healthy data are plotted from a consecutive time span of equivalent length. The purple vertical line indicates mean number of days to PTSD diagnosis, post-trauma, and the purple shaded region shows the average period between trauma and diagnosis. Trend lines represent cubic polynomial regression fits with 95% CI bands, points are aggregations of 30 day periods, with error bars indicating 95% CI on central tendency of daily values.

below, offer some indication of directionality but are computed differently than Random Forests and should not be used to directly interpret Random Forests output. However, as a means of informing early detection methods, predictor directionality becomes less important, as the goal is identification, rather than explanation, of possible mental health issues.

Time course description. A Hidden Markov Model simulated affected and healthy states. HMM states were determined to accurately track with affected and healthy groups by comparing differences in mean parameter estimates between the model fit and original data. Across all 40 predictors, HMM means were in agreement with true means for the depression sample in 38 cases (95% agreement), and were in 100% agreement with true means for the PTSD sample. This evidence strongly suggested that the two states identified by HMM were closely aligned with the affected and healthy classes in our data, and we have reported HMM results based on this assumption. See Figs 2 and 3.

Depressed individuals showed a slightly higher probability of depression even in the period nine months prior to diagnosis, and gradually diverged from healthy data points. Healthy individuals showed a steady, lower probability of depression, which did not change noticeably over an 18-month period. By three months prior to diagnosis, depressed subjects showed a marked rise in probability of being in a depressed state, whereas healthy individuals showed little or no change over the same time period. Post-diagnosis, probability of depression began to decrease after 3–4 months (90–120 days). This trajectory matches closely with average improvement time frames observed in therapeutic programs[47]. Given that HMM constructed latent states from unlabeled data, it is striking that HMM not only reconstructed the division between depressed and healthy groups, but also generated a plausible timeline for depression onset and recovery. Similarly, tweets from individuals with PTSD deviated from healthy tweets within months after the date of the traumatic event that caused PTSD (indicated by the orange line in Fig. 3), and well over a year before the average time to diagnosis (the mean time period from trauma to diagnosis was 586 days). A decrease in PTSD probability can be observed shortly after diagnosis, indicating possible improvement due to treatment.

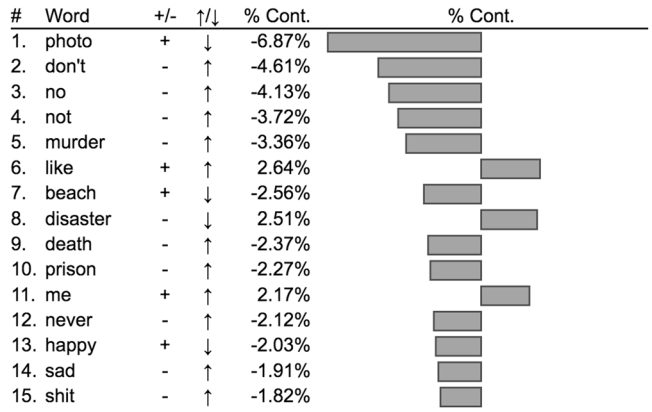

Figure 4. Depression word-shift graph revealing contributions to difference in Twitter happiness observed between depressed (5.98) and healthy (6.11) participants. In column 3, (−) indicates a relatively negative word, and (+) indicates a relatively positive word, both with respect to the average happiness of all healthy tweets. An up (down) arrow indicates that word was used more (less) by the depressed class. Words on the left (right) contribute to a decrease (increase) in happiness in the depressed class. See Appendix III for PTSD word-shift graph.

While these results suggest promise, further analysis of the nature of aggregate time courses, including investigation of more sophisticated methods of time-series analysis[44,48,49] will be a subject of future work. In particular, the health trajectories associated with individuals (not in aggregate), other forms of text-based communication, and applications to other mental illnesses should be explored in greater detail.

Word shift graphs. We averaged labMT happiness scores across observations in each class, after the removal of common neutral words and re-tweeted promotional material[39]. Neutral words were words with labMT happiness scores between 4 and 6, on a 1–9 scale. This included many common parts of speech, including articles and pronouns, which contributed little to understanding inter-group differences in valenced language. Some of the positive language observed more frequently among healthy individuals came from re-tweets of promotional or other advertising material (e.g., "win", "free", "gift"). We removed obvious promotional retweets when generating word shift graphs, as their removal did not significantly change mean tweet-happiness differences between groups, and the resulting graphs gave better impressions of what participants personally tweeted about. We observed that tweets authored by the depressed class were sadder ($h_{avg} = 6.01$) than the healthy class ($h_{avg} = 6.15$). In Fig. 4, we rank order individual words with respect to their contribution to this observed difference, and display the top contributing words. PTSD word shift graphs are included in Appendix III.

The dominant contributor to the difference between depressed and healthy classes was an increase in usage of negative words by the depressed class, including "no", "never", "prison", "murder", and "death". The second largest contributor was a decrease in positive language by the depressed class, relative to the healthy class, including fewer appearances of "happy", "beach", and "photo". The increased usage of negatively valenced language by depressed individuals is congruent with previous research[50].

Discussion

The aim of the present study was to identify predictive markers of depression and PTSD based on users' Twitter data using computational methods. Our findings strongly support the claim that computational methods can effectively screen Twitter data for indicators of depression and PTSD. Our method identified these mental health conditions earlier and more accurately than the performance of trained health professionals, and was more precise than previous computational approaches. Our state-space models portrayed a timeline for depression which is impressively realistic, given that it was generated analyzing only the text of 140-character messages. In addition, HMM identified a rise in probability of PTSD within six months, post-trauma, compared to the average 19 month delay between trauma event and diagnosis experienced by the individuals in our sample. Word shifts provided context for the specific differences in language that shifted happiness scores between samples. These advances make improvements on existing predictive screening technology, as well as contribute novel methods for the identification and tracking of mental illness.

The HMM depression timeline is an intriguing finding, and should be treated with both optimism and caution. HMM assigned each data point a probability of belonging to two latent state-spaces while "blind" to our actual states of interest, as affected/healthy labels were removed from HMM training data. Considering that the model could have used criteria completely unrelated to mental health to delineate between the two latent classes, it is noteworthy that the resulting states' mean estimates for each variable in both PTSD and depression analyses closely resembled the mean estimates for affected and healthy participants' data, respectively. This adds support for the claim that affected-condition and healthy Twitter users' data are objectively different, in addition to providing justification for the use of HMM assignments as indicators of depression/PTSD signals at a given point in time. Despite this evidence in favor of applying HMM to analyze mental health trajectories, HMM is an unsupervised learning procedure and so conclusive HMM-based inferences should be approached cautiously and with close attention on validation procedures. The diverging trajectories observed in our HMM time series suggest

that, with careful attention to model validity, state-space modeling may be used to identify and track the onset of certain mental illnesses over time, using only Twitter data.

The labMT happiness measure proved to be the most important predictor in our model, and was considerably stronger than ANEW or LIWC happiness indicators. This is in line with a series of previous findings, which have found labMT measures to be a superior for tracking happiness in Twitter data[40], and suggests that future research in this field should incorporate this instrument for more accurate measurements. That average tweet word count was the second most important predictor is intriguing, especially as increases in word count were positively associated with depression and PTSD. If anything, depression is often characterized by reduced communication[17], although word count is distinct from posting frequency, which was not a significant predictor in our models. The current depression and PTSD literatures are largely devoid of studies relating verbosity to these conditions, and so this finding may motivate new inquiries into behavioral traits of mental health disorders as observed on social media.

From a practical standpoint, our model showed considerable improvement over the ability of unassisted general practitioners to correctly diagnose depression and PTSD. Despite the imprecise nature of this comparison, given the paucity of data currently available to serve as benchmarks for the type of analysis performed in the present study, our model's relative success seems encouraging. Health care providers may be able to improve quality of care and better identify individuals in need of treatment based on the simple, low-cost methods outlined in this report. Especially given that mental health services are unavailable or underfunded in many countries[51], this computational approach, which only requires patients' digital consent to share their social media history, may open avenues to care which are currently difficult or impossible to provide. Future investigations would use the same participant pool to collect both health professionals' assessments, as well as computational models of participant social media behavior, to allow for more precise comparison.

The present findings may be limited by the non-specific use of the term "depression" in participant surveys. While earlier research identified depression predictors in Twitter data for Major Depressive Disorder and postpartum depression[7,8], we used a more general category in our recruitment and data collection to build a predictive model capable of screening for common depressive signals. We acknowledge that depression diagnoses exist across a clinical spectrum. It is possible that participants with a specific type of depression were responsible for the observed results. Future research might examine other specific depression classes, including manic depression and dysthymia, to determine whether predictive screening models should be segmented per diagnosis type.

It is also possible that inferences from our results are limited specifically to Twitter users who have been diagnosed with depression or PTSD, and who are willing to share their social media history with researchers. Current literature on depression treatment suggests that people who seek out mental health services are usually "well-informed and psychologically minded, experience typical symptoms of depression and little stigma, and have confidence in the effectiveness of treatment, few concerns about side effects, adequate social support, and high self-efficacy"[52]. Since it is possible only a subset of Twitter users will fit this description, we recommend making conservative inferences about depression, as well as PTSD, based on our findings.

Considering the frequent comorbidity of depression and PTSD[20], together with the similarity in predictor importance observed across our analyses of these two conditions, the signals driving our predictive models may share considerable overlap. While our results do not offer strict segregation between these two conditions, this gives little cause for concern when considered from a mental health screening perspective. If the desired outcome is to identify individuals who may be in need of mental health services, whether an individual is flagged for evaluation for depression with possible associated PTSD, or vice-versa, becomes an academic distinction. If anything, the issue of comorbidity may serve as a useful reminder that this computational method should not be regarded as a standalone diagnostic tool, but rather as a technology for early identification of potential mental health issues.

As the methods employed in the present study aim to infer health related information about individuals, some additional cautionary considerations are in order. Data privacy and ethical research practices are of particular concern, given recent admissions that individuals' Facebook and dating profile data were experimentally manipulated or exposed without permission[53,54]. Indeed, we observed a response rate reflecting a seemingly reluctant population. Of the 2,261 individuals who began our survey, 790 (35%) refused to share their Twitter username and history, even after we identified ourselves as an "academic, not-for-profit research team" and provided the above-mentioned guarantees about data privacy. Future research should prioritize establishing confidence among experimental participants that their data will remain secure and private. Complicating efforts to build socio-technical tools such as the models presented in this study, data trends often change over time, degrading model performance without frequent calibration[55]. As such, our results should be considered a methodological proof-of-concept upon which to build and refine subsequent models.

This report provides an outline for an accessible, accurate, and inexpensive means of improving depression and PTSD screening, especially in contexts where in-person assessments are difficult or costly. In concert with robust data privacy and ethical analytics practices, future models based on our work may serve to augment traditional mental health care procedures. More generally, our results support the idea that computational analysis of social media can be used to identify major changes in individual psychology.

References

1. Paparrizos, J., White, R. W. & Horvitz, E. Screening for Pancreatic Adenocarcinoma Using Signals From Web Search Logs: Feasibility Study and Results. *J Oncol Pract*: JOPR010504 (2016).
2. Christakis, N. A. & Fowler, J. H. Social network sensors for early detection of contagious outbreaks. *PLoS ONE* **5**(9), e12948 (2010).
3. Li, J. & Cardie, C. Early Stage Influenza Detection from Twitter. *arXiv*:1309.7340 [cs] (2013).
4. Schmidt, C. W. Trending now: Using social media to predict and track disease outbreaks. *Environ Health Perspect* **120**(1), a30–a33 (2012).
5. Alajajian, S. E. *et al.* The Lexicocalorimeter: Gauging public health through caloric input and output on social media. *arXiv* **1507**, 05098 (2015).

6. Moreno, M., Christakis, D., Egan, K., Brockman, L. & Becker, T. Associations between displayed alcohol references on Facebook and problem drinking among college students. *Arch Pediatr Adolesc Med* **166**(2), 157–163 (2012).
7. De Choudhury, M., Counts, S. & Horvitz, E. Predicting postpartum changes in emotion and behavior via social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (ACM: New York), pp. 3267–3276 (2013).
8. De Choudhury, M., Gamon, M., Counts, S. & Horvitz, E. Predicting depression via social media. In *Seventh International AAAI Conference on Weblogs and Social Media* (2013).
9. Katikalapudi, R., Chellappan, S., Montgomery, F., Wunsch, D. & Lutzen, K. Associating internet usage with depressive behavior among college students. *IEEE Tech Soc Magazine* **31**(4), 73–80 (2012).
10. Moreno, M. A. *et al*. Feeling bad on Facebook: Depression disclosures by college students on a social networking site. *Depress Anxiety* **28**(6), 447–455 (2011).
11. Park, M., Cha, C. & Cha, M. Depressive moods of users portrayed in Twitter. In *Proceedings of the ACM SIGKDD Workshop on healthcare informatics (HI-KDD)* (pp. 1–8) (2012).
12. Reece, A. G. & Danforth, C. M. Instagram photos reveal predictive markers of depression. *EPJ Data Science* **6**(1), 15 (2017).
13. Nadeem, M., Horn, M. & Coppersmith, G. Identifying depression on Twitter. *arXiv*:1607.07384 (2016).
14. Coppersmith, G., Harman, C. & Dredze, M. Measuring Post-Traumatic Stress Disorder in Twitter. In *Eighth International AAAI Conference on Weblogs and Social Media* (2014).
15. De Choudhury, M., Kiciman, E., Dredze, M., Coppersmith, G. & Kumar, M. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (ACM: New York), pp. 2098–2110 (2016).
16. Ferrari, A. *et al*. Global variation in the prevalence and incidence of major depressive disorder: a systematic review of the epidemiological literature. *Psychol Med* **43**(3), 471–481 (2013).
17. American Psychiatric Association. Diagnostic and statistical manual of mental disorders (4th ed., text rev.) (2000).
18. Gwynn, R. C. *et al*. Prevalence, diagnosis, and treatment of depression and generalized anxiety disorder in a diverse urban community. *Psychiatr Serv* **59**(6), 641–647 (2008).
19. Stein, M. B., McQuaid, J. R., Pedrelli, P., Lenox, R. & McCahill, M. E. Posttraumatic stress disorder in the primary care medical setting. *Gen Hosp Psychiatry* **22**(4), 261–269 (2000).
20. Campbell, D. G. *et al*. Prevalence of Depression–PTSD comorbidity: Implications for clinical practice guidelines and primary care-based interventions. *J Gen Intern Med* **22**(6), 711–718 (2007).
21. Munro, C. G., Freeman, C. P. & Law, R. General practitioners' knowledge of post-traumatic stress disorder: a controlled study. *Br J Gen Pract* **54**(508), 843–847 (2004).
22. Taubman-Ben-Ari, O., Rabinowitz, J., Feldman, D. & Vaturi, R. Post-traumatic stress disorder in primary-care settings: prevalence and physicians' detection. *Psychol Med* **31**(03), 555–560 (2001).
23. Choudhury, M. D., Counts, S., Horvitz, E. J. & Hoff, A. Characterizing and predicting postpartum depression from shared facebook data. In Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing (CSCW'14). ACM, New York, NY, USA, 626–638, https://doi.org/10.1145/2531602.2531675 (2014).
24. Eaton, W. W., Neufeld, K., Chen, L. & Cai, G. A comparison of self-report and clinical diagnostic interviews for depression: Diagnostic interview schedule and schedules for clinical assessment in neuropsychiatry in the baltimore epidemiologic catchment area follow-up. *Arch Gen Psychiatry* **57**(3), 217–222 (2000).
25. Cornford, C. S., Hill, A. & Reilly, J. How patients with depressive symptoms view their condition: A qualitative study. *Fam Pract* **24**(4), 358–364 (2007).
26. Karp, D. A. Living with depression: Illness and identity turning points. *Qual Health Res* **4**(1), 6–30 (1994).
27. Winokur, G. Duration of Illness prior to Hospitalization (Onset) in the Affective Disorders. *Neuropsychobiology* **2**(2–3), 87–93 (1976).
28. Mitchell, A. J., Vaze, A. & Rao, S. Clinical diagnosis of depression in primary care: a meta-analysis. *Lancet* **374**(9690), 609–619 (2009).
29. Radloff, L. S. The CES-D scale: A self-report depression scale for research in the general population. *Appl Psych Manage* **1**(3), 385–401 (1977).
30. Fountoulakis, K. N. *et al*. Comparison of depressive indices: Reliability, validity, relationship to anxiety and personality and the role of age and life events. *J Affect Disord* **97**(1–3), 187–195 (2007).
31. Zich, J. M., Attkisson, C. C. & Greenfield, T. K. Screening for depression in primary care clinics: The CES-D and the BDI. *Int J Psychiatry Med* **20**(3), 259–277 (1990).
32. Brewin, C. R. *et al*. Brief screening instrument for post-traumatic stress disorder. *Br J Psychiatry* **181**(2), 158–162 (2002).
33. Peer, E., Vosgerau, J. & Acquisti, A. Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behav Res Methods* **46**(4), 1023–1031 (2013).
34. Litman, L., Robinson, J. & Rosenzweig, C. The relationship between motivation, monetary compensation, and data quality among US- and India-based workers on Mechanical Turk. *Behav Res Methods* **47**(2), 519–528 (2014).
35. Cuijpers, P., Boluijt, P. & van Straten, A. Screening of depression in adolescents through the Internet. *Eur Child Adolesc Psychiatry* **17**(1), 32–38 (2007).
36. Haringsma, R., Engels, G. I., Beekman, A. T. F. & Spinhoven, P. The criterion validity of the Center for Epidemiological Studies Depression Scale (CES-D) in a sample of self-referred elders with depressive symptomatology. *Int J Geriatr Psychiatry* **19**(6), 558–563 (2004).
37. Pennebaker, J. W., Boyd, R. L., Jordan, K. & Blackburn, K. The development and psychometric properties of LIWC2015. *UT Faculty/Researcher Works* (2015).
38. Bradley, M. M. & Lang. P. J. Affective norms for English words (ANEW): Instruction manual and affective ratings. Technical report C-1, the center for research in psychophysiology, University of Florida (1999).
39. Dodds, P. S., Harris, K. D., Kloumann, I. M., Bliss, C. A. & Danforth, C. M. Temporal Patterns of Happiness and Information in a Global Social Network: Hedonometrics and Twitter. *PLoS ONE* **6**(12), e26752 (2011).
40. Reagan, A. J., Tivnan, B. F., Williams, J. R., Danforth, C. M. & Dodds, P. S. Benchmarking sentiment analysis methods for large-scale texts: A case for using continuum-scored words and word shift graphs (2016).
41. Cody, E. M., Reagan, A. J., Mitchell, L., Dodds, P. S. & Danforth, C. M. Climate change sentiment on Twitter: An unsolicited public opinion poll. *PloS one* **10**(8), e0136092 (2015).
42. Frank, M. R., Mitchell, L., Dodds, P. S. & Danforth, C. M. Happiness and the patterns of life: A study of geolocated tweets. *Scientific Reports* 3(2625) (2013).
43. Bliss, C. A., Kloumann, I. M., Harris, K. D., Danforth, C. M. & Dodds, P. S. Twitter reciprocal reply networks exhibit assortativity with respect to happiness. *Journal of Computational Science 3(5)*: pp. 388–397 (2012).
44. Gao, Z., Small, M., Kurths, J. Complex network analysis of time series, *Europhysics Letters*, Vol 116, 5 (2016).
45. Dodds, P. S. *et al*. Human language reveals a universal positivity bias. *Proceedings of the National Academy of Sciences* **112**(8), 2389–2394 (2015).
46. Storylab. Hedonometer 2.0: Measuring happiness and using word shifts. http://goo.gl/oM9W4Z (2014).
47. Schulberg, H. C., Katon, W., Simon, G. E. & Rush, A. Treating major depression in primary care practice: An update of the agency for health care policy and research practice guidelines. *Arch Gen Psychiatry* **55**(12), 1121–1127 (1998).

48. Gao, Z., Cai, Q., Yang, Y., Dang, W., Zhang, S. Multiscale limited penetrable horizontal visibility graph for analyzing nonlinear time series, *Scientific Reports* 6, Article number: 35622 (2016)
49. Paul, M. J., White, R. W. & Horvitz, E. Search and breast cancer: On episodic shifts of attention over life histories of an illness. ACM Transactions on the Web (TWEB) 10, no. 2, 13 (2016).
50. Rude, S., Gortner, E. M. & Pennebaker, J. Language use of depressed and depression-vulnerable college students. *Cogn Emotion* **18**(8), 1121–1133 (2004).
51. Detels, R. *The scope and concerns of public health*. Oxford University Press (2009).
52. Epstein, R. M. *et al*. "I didn't know what was wrong:" How people with undiagnosed depression recognize, name and explain their distress. *J Gen Intern Med* **25**(9), 954–961 (2010).
53. Fiske, S. T. & Hauser, R. M. Protecting human research participants in the age of big data. *Proc Natl Acad Sci USA* **111**(38), 13675–13676 (2014).
54. Lumb, D. Scientists release personal data for 70,000 OkCupid profiles. engt.co/2b4NnQ0 (2016).
55. Lazer, D., Kennedy, R., King, G. & Vespignani, A. The parable of Google Flu: Traps in big data analysis. *Science* **343**(6176), 1203–1205 (2014).
56. https://github.com/hmmlearn/hmmlearn.

Acknowledgements

A.G.R. was supported by the Sackler Scholar Programme in Psychobiology. P.S.D. and C.M.D. were supported by NSF grant #1447634. We thank L. Doyle for conversations and P. Mair for manuscript review.

Author Contributions

A.G.R., A.J.R., and C.M.D. contributed to design. A.G.R., A.J.R., and C.M.D. contributed to analysis. A.G.R., A.J.R., C.M.D., E.J.L., K.L.M.L., and P.S.D. contributed to writing.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-017-12961-9.

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.