ARTICLES
https://doi.org/10.1038/s41562-021-01050-7

nature human behaviour

Individuals with depression express more distorted thinking on social media

Krishna C. Bathina[1], Marijn ten Thij[1], Lorenzo Lorenzo-Luaces[2], Lauren A. Rutter[2] and Johan Bollen[1]✉

Depression is a leading cause of disability worldwide, but is often underdiagnosed and undertreated. Cognitive behavioural therapy holds that individuals with depression exhibit distorted modes of thinking, that is, cognitive distortions, that can negatively affect their emotions and motivation. Here, we show that the language of individuals with a self-reported diagnosis of depression on social media is characterized by higher levels of distorted thinking compared with a random sample. This effect is specific to the distorted nature of the expression and cannot be explained by the presence of specific topics, sentiment or first-person pronouns. This study identifies online language patterns that are indicative of depression-related distorted thinking. We caution that any future applications of this research should carefully consider ethical and data privacy issues.

Depression is a leading contributor to the burden of disability worldwide[1,2], and there is some evidence that disability attributed to depression is rising, particularly among young people[3,4]. A key challenge in reducing the prevalence of depression has been that it is often under-recognized[5] as well as undertreated[6].

Cognitive behavioural therapy (CBT) is the most widely researched psychotherapy for depression. It is equivalent to antidepressant medications in its short-term efficacy and evidences superior outcomes in the long-term[7,8]. In CBT, therapists work with their clients to identify depressogenic thinking patterns by identifying lexical or verbal markers of rigid, distorted or overly negative interpretations[9,10]. For example, statements that include 'should' or 'must' are often challenged as reflecting overly rigid rules about the world ('I shouldn't be lazy', 'I must never fail'). This process often entails a series of conversations with the client to uncover and address statements that reflect these cognitive distortions.

The cognitive theory that underlies CBT argues that the ways in which individuals process and interpret information about themselves and their world is directly related to the onset, maintenance and recurrence of their depression[11,12]. This model is consistent with information-processing accounts of mood regulation[13] and its dynamics[14], as well as basic research that supports the role of cognitive reappraisal and language in emotion regulation[15–18].

However, the critical assumption at the foundation of CBT, namely that depression is associated with changes in language that are indicative of distorted thinking, has not been directly confirmed from studies of the language of individuals with depression in real-world settings.

The idea that depression is associated with changes in language is supported by previous research. Specifically, it has been shown that individuals with depression more frequently use a variety of terms that describe negative emotions[19–21], first-person pronouns (FPPs)[21–25], common symptoms[26] and linguistic inquiry and word count (LIWC) categories deemed to correspond to 'absolutist' language[27]. Machine learning approaches have shown good performance with respect to predicting whether social media users have depression[28–30], identifying the most useful term features to render a prediction.

In this Article, we refine and expand on these data-driven approaches along several fronts. First, we empirically verified a crucial tenet of CBT theory, namely that individuals with depression, in their thinking, exhibit higher levels of cognitive distortions as conceived by CBT. This is distinct from attempting to estimate the morbidity of depression itself in the general population or to algorithmically discover any set of features that is useful to predict depression. Second, rather than sampling from data obtained in a clinical setting, possibly confounding the context of a specific therapeutic approach, we relied on naturalistic language recorded in an ex post hoc manner from large samples of social media users. Third, we conducted our analysis on the basis of a set of context-free semantic schemata (*n*-grams) that encode the semantics of patterns of thought, that is, cognitive distortions as hypothesized by CBT, not individual terms or features. In other words, we captured the structure of thought behind CBT's notion of cognitive distortions. This is distinct from previous research that used term features that are either derived from general lexicons or discovered by supervised machine learning algorithms.

We compared the prevalence of a set of 241 cognitive distortion schemata (CDS)—patterns of thought represented by sequences of words (*n*-grams)—in the language of a large cohort of individuals with depression versus a random sample on social media (Twitter), excluding institutions and organizational accounts (see the 'Data and sample construction' section in the Methods). We show a set of examples of these CDS in Table 1. We designed our method to be platform-independent, but we chose Twitter because it (1) is a fast-paced real-time medium with hundreds of millions of active users (posting daily or regularly) who use colloquial language in a short text format that is especially suitable for our approach and (2) has been active since 2006, providing comprehensive longitudinal data spanning more than a decade.

For our analysis, we built two cohorts of individuals: individuals with depression (D cohort) and a cohort of randomly selected individuals (R cohort). For our D cohort, following Coppersmith et al.[31], we identified a cohort of social media users who (1) received a clinical diagnosis of depression and (2) posted an explicit report of this diagnosis on Twitter, that is, by stating a variant of 'I was

[1]Luddy School of Informatics, Computing, and Engineering, Indiana University Bloomington, Bloomington, IN, USA. [2]Department of Psychological and Brain Sciences, Indiana University Bloomington, Bloomington, IN, USA. ✉e-mail: jbollen@indiana.edu

Table 1 | Common types of cognitive distortions associated with depression[65] and their definitions

Category	Definition	Examples
Catastrophizing	Exaggerating the importance of negative events	'The evening will be a disaster'
Dichotomous reasoning	Thinking that an inherently continuous situation can fall into only two categories	'No one will ever like me'
Disqualifying the positive	Unreasonably discounting positive experiences	'OK but[1] my grade was not that good[2]'
Emotional reasoning	Thinking that something is true on the basis of how one feels, ignoring the evidence to the contrary	'My grades are good but it still feels[1] like I will fail[2]'
Fortune-telling	Making predictions, usually negative ones, about the future	'Whatever I try I will not be successful'
Labelling and mislabelling	Labelling yourself or others while discounting evidence that could lead to less disastrous conclusions	'I am a[1] total[2] loser[3]'
Magnification and minimization	Magnifying negative aspects or minimizing positive aspects	'My good grades are really not important'
Mental filtering	Paying too much attention to negative details instead of the whole picture	'If I only worked harder, I would be more successful'
Mindreading	Believing you know what others are thinking	'Everyone believes[1] I am a[2] failure[3]'
Overgeneralizing	Making sweeping negative conclusions on the basis of a few examples	'Nobody ever cares for me'
Personalizing	Believing others are behaving negatively because of oneself, without considering more plausible or external explanations for behaviour	'Everyone thinks[1] I am a loser[2] for calling her'
Should statements	Having a fixed idea on how you and/or others should behave	'I have to[1] to do this or I will not[2] make it to the weekend'

Some of the examples contain more than one type of CDS, indicated by superscript numbers.

diagnosed with depression by my doctor' (Methods). An overview of this approach is shown in Fig. 1.

Supporting an important assumption underlying CBT, our results indicate that there is a significantly higher prevalence of most types of distorted thinking, marked by a set of CDS *n*-grams, in the individuals with depression, both at the within-individual and between-cohort level. Notably, CDS in the 'personalizing' and 'emotional reasoning' types occur approximately two times more frequently in the online language of individuals with depression. We verified whether our results could be explained by gender or age differences, random variations in our user sample, our particular choice of CDS *n*-grams, the sentiment loadings of our CDS set and the known propensity of individuals with depression to make self-referential statements (see the 'Robustness' section). In all cases, we continued to find much higher levels of certain types of distorted thinking in the language of individuals with depression compared with in the random sample of online individuals.

We emphasize that, in contrast to some previous research, our goal was not to detect or classify users with depression on Twitter, but to compare the prevalence of expressions of cognitive distortions in the language of users who personally report having a diagnosis with those who do not.

Results

Sample demographics. The age and gender distributions of our D and R cohorts align with previous studies[32–34] as indicated by the M3 classifier[35] that we used to predict individual's gender (M3 Macro-F1: 0.915) and age (M3 Macro-F1: 0.425) categories. As shown in Table 2, our D cohort has a similar 2:1 female-to-male ratio as observed in clinical depression studies[32,33], indicating that the demographics of our Twitter cohort closely match previous clinical findings that women are twice as likely to be diagnosed with depression compared with men. Note that this gender disparity was not found to be associated with differences in language used to express depression or depressive symptomologies in women versus men[36,37]. The indicated age distribution of our D cohort (although less reliable, Macro-F1: 0.425) is also consistent with clinical studies[32,34]; specifically, we found a decreasing number of individuals in each age-group as the age of the group increases in the D cohort. Our subsequent analysis accounts for the observed distributions of gender and age between the D and R cohorts by performing comparisons across identical demographics (men versus men, women versus women and so on), amounting to a stratified sampling approach.

Within-individual CDS prevalence. First, we compared the within-individual CDS prevalence between the D and R cohorts. For each individual, we counted the number of their tweets containing any of the 241 CDS and divided it by their total number of tweets, resulting in a single within-individual CDS prevalence (Methods). We next compared the density distribution of individual prevalence values between all of the individuals from the D and R cohorts as shown in Fig. 2a,b.

In Fig. 2b, we observed that the distribution of within-individual CDS prevalence is shifted to the right for the D cohort relative to that of the R cohort, indicating that individuals in the D cohort express significantly more CDS (mean prevalence, $\bar{P}_D = 0.232$) than individuals in the R cohort (mean prevalence, $\bar{P}_R = 0.173$). On the basis of a two-sided Welch's unequal variances *t*-test, we rejected the null hypothesis that the two samples have equal means ($t_{1,619} = 21.20$, $P < 0.001$, Cohen's $d = 0.56$). Data distribution was assumed to be normal, but this was not formally tested. Note that 9.756% of the individuals in the R cohort have no tweets with CDS, whereas only 0.386% of the individuals in the D cohort express no CDS.

After comparison of the distribution of within-individual prevalence between the subgroups on the basis of demographic information, as shown in Fig. 2a, we found that all distributions differ based on Welch's unequal variances *t*-test (male: $t_{335} = 9.82$, $P < 0.001$, Cohen's $d = 0.53$; female: $t_{1,127} = 16.81$, $P < 0.001$, Cohen's $d = 0.62$; aged 18 and under: $t_{208} = 9.35$, $P < 0.001$, Cohen's $d = 0.71$; aged 19–29: $t_{580} = 13.49$, $P < 0.001$, Cohen's $d = 0.67$; aged 30–39: $t_{217} = 7.73$, $P < 0.001$, Cohen's $d = 0.59$; aged 40 and over: $t_{103} = 3.49$, $P < 0.001$, Cohen's $d = 0.30$). Excluding individuals that have no tweets with CDS from our analysis led to similar results across all demographic subgroups.

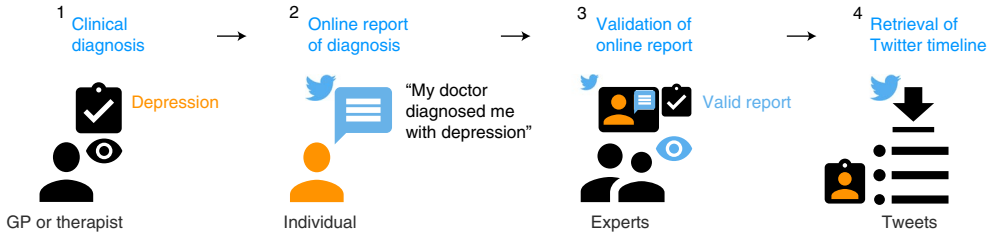

Fig. 1 | Cohort of individuals with depression. We identified a cohort of individuals with depression who (1) received a clinical diagnosis of depression and (2) explicitly reported this diagnosis on social media using a variant of the statement 'I was diagnosed with depression by my doctor'. (3) A team of experts rated each statement to ensure that the statement actually reports a personal, clinical diagnosis of depression, after which (4) the individual's timeline (all tweets up to the limit allowed by the Twitter data service: the 3,200 most recent tweets) was downloaded and added to our analysis cohort. Twitter, tweet, retweet, and the Twitter logo are trademarks of Twitter, Inc. or its affiliates.

Table 2 | Demographic information predicted using the M3 Twitter-trained classifier for the D and R cohorts

		D cohort		R cohort	
		Number of accounts	Number of tweets	Number of accounts	Number of tweets
Total no. of individuals		1,035 (100.00%)	1,510,359	7,349 (100.00%)	6,783,353
Gender	All	887 (85.70%)		6,231 (84.79%)	
	Male	268 (30.21%)	400,444	3,313 (53.17%)	3,403,224
	Female	619 (69.79%)	908,850	2,918 (46.83%)	2,504,347
Age (years)	All	687 (66.38%)		4,934 (67.14%)	
	≤18	152 (22.13%)	158,595	1,200 (24.32%)	694,398
	19–29	318 (46.29%)	463,811	1,648 (33.40%)	1,483,615
	30–39	135 (19.65%)	245,245	845 (17.13%)	998,023
	≥40	82 (11.94%)	134,323	1,241 (25.15%)	1,401,708

Note that the totals for the gender and age dimension do not add up to the total cohort size due to a strict classification threshold to achieve high precision (Methods).

Between-cohort CDS prevalence. We conducted a between-cohort analysis to compare the prevalence of CDS between the D and the R cohorts. We did this by calculating the prevalence of CDS for all tweets from each cohort and calculating the prevalence ratio (PR) between the two cohorts (see the 'PR values' section in the Methods). A PR value of higher than 1 indicates that the presence of CDS in the tweets written by the D cohort is greater than the R cohort. To assess the sensitivity of our results to changes in our cohort samples, for example a few 'high-power' users biasing our analysis, we repeatedly calculated the estimated PR over 10,000 random resamples (with replacement) of both groups, resulting in a distribution of PR values shown in Fig. 2 (see the 'Bootstrapping estimates' section in the Methods)

We found narrow distributions of the number of tweets in each resample (D cohort: 95% confidence interval (CI) = 1,454,068.75–1,566,230.325; R cohort: 95% CI = 6,630,441.375–6,941,408.2) indicating that our results are not biased by the presence of exceptionally active or inactive users in either cohort sample. Note that PR values express the relative difference in CDS prevalence between the two cohorts, not the absolute difference.

We observed in Fig. 2c that the median of this distribution of PR values for all of the individuals in the D and R cohorts is much greater than 1, and that its 95% CI does not include 1, indicating that we found a statistically significant higher prevalence of CDS in the D cohort (1.129×) compared with the R cohort. This result is robust to random changes in our cohort samples, indicating that outliers or exceptionally active or inactive users are not biasing our results. Furthermore, when we performed a between-cohort comparison within each of the gender and age categories, as shown in Fig. 2c, in all cases, we found a statistically significant higher prevalence of CDS in the D cohort, with median values ranging from 1.102× for individuals aged 40 and over to 1.164× for individuals aged 19–29.

To investigate the possible influence of the difference in the time intervals that are spanned by both cohorts, we performed stratified sampling by month whereby tweets were placed as a time-matched control and found similar results for all individual months (Supplementary Information Section 2). We found no indications of a time-dependent effect on CDS prevalence.

CDS prevalence by cognitive distortion type. The between-cohort PR values shown in Fig. 2c do not reflect specific distortion types; all CDS are equally and independently matched to all tweets. However, CDS types may differ in their prevalence between our cohorts. We therefore repeated the above analysis with CDS separated by cognitive distortion type.

As shown in Table 3 and Fig. 2d, the prevalence of CDS is significantly higher for nearly all cognitive distortion types in the tweets of the D cohort compared with those of the *R* cohort; PR values ranged from 2.084× to 1.056×, with the exception of 'fortune-telling', 'mind-reading' and 'catastrophizing', which produce a PR that is not significantly different from parity. However, PR values vary by cognitive distortion type. The cognitive distortion types 'personalizing' and 'emotional reasoning' have the greatest PR values of 2.084× and 1.983×, respectively, followed by 'overgeneralizing' (1.441×), 'mental filtering' (1.296×), 'disqualifying the positive' (1.229×), 'labelling and mislabelling' (1.207×) and 'dichotomous reasoning' (1.131×). The cognitive distortion types 'should statements' and 'magnification and minimization' have significant PR values of lower than 1.1×. Table 4 shows the number and ratios of schemata for each

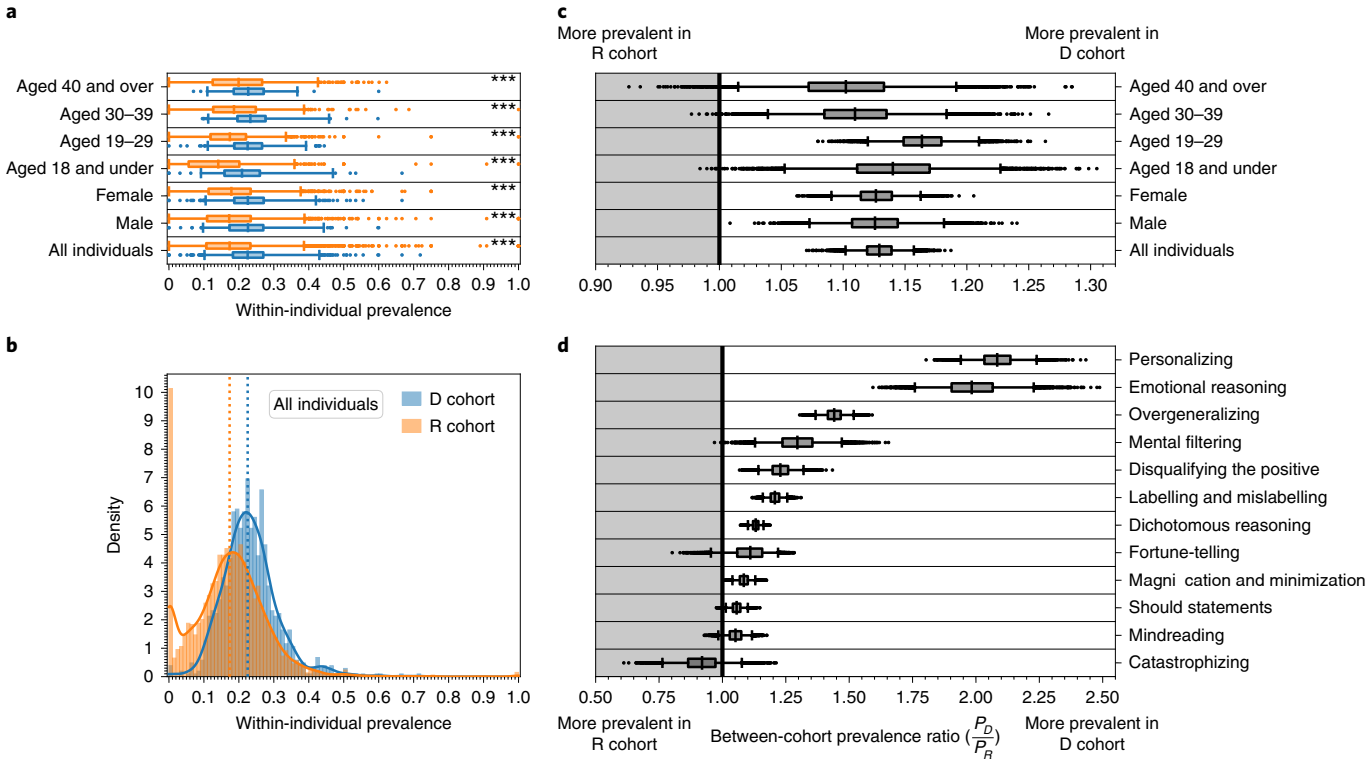

Fig. 2 | Within-individual CDS prevalence and between-cohort PR values. a, Box and whisker (box, 50% CI; whiskers, 95% CI; vertical line, median value) plots for within-individual CDS prevalence distributions compared between all individuals in the D and R cohorts and within the same demographic group (age and gender). All points that fall outside the 95% CI are indicated by dots. For all of the demographic subgroups (gender and age categories), we can reject the null hypothesis that the two distributions have the same mean on the basis of Welch's unequal variances *t*-test. ***$P < 0.001$. **b**, The density of within-individual prevalence of tweets containing CDS for the D cohort (blue, $\bar{P}_D = 0.232$) versus the R cohort (orange, $\bar{P}_R = 0.173$). The dashed vertical lines indicate the median value for each cohort. A large fraction of individuals in the R cohort (9.756%) have no tweets that contain any CDS. **c**, Box and whisker (box, 50% CI; whiskers, 95% CI; vertical line, median value) plots of bootstrapped between-cohort PR values between the D and R cohort (exact median and 95% CI values are provided in Table 3). All points that fall outside the 95% CI are indicated by dots. The 95% CI of the distribution does not include 1.00 (vertical line), indicating a significantly higher prevalence of all CDS for the D cohort. **d**, Box and whisker (box, 50% CI; whiskers, 95% CI; vertical line, median value) plots of CDS PR values between the D and R cohort for each cognitive distortion type. All points that fall outside the 95% CI are indicated by dots. The D cohort showed a significantly higher use of CDS than the R cohort for most CDS types separately (PR≫1) with the exception of 'catastrophizing', 'mindreading' and 'fortune-telling'. Further details about the PR values are provided in Table 3.

Table 3 | PR and 95% CIs of CDS between the D and R cohort

	PR_A		PR_I		PR_C	
	Median	**95% CI**	**Median**	**95% CI**	**Median**	**95% CI**
All CDS	1.129*	1.102–1.157	1.110*	1.082–1.137	1.231*	1.168–1.320
Personalizing	2.084*	1.940–2.239	–	–	2.403*	1.676–3.043
Emotional reasoning	1.983*	1.759–2.228	1.815*	1.467–2.217	2.316*	2.013–3.158
Overgeneralizing	1.441*	1.367–1.518	1.344*	1.271–1.420	1.605*	1.414–1.776
Mental filtering	1.296*	1.129–1.471	1.191	0.931–1.491	1.466*	1.171–1.924
Disqualifying the positive	1.229*	1.142–1.320	1.229*	1.142–1.320	1.401*	1.203–1.536
Labelling and mislabelling	1.207*	1.159–1.256	1.090*	1.041–1.139	1.336*	1.176–1.554
Dichotomous reasoning	1.131*	1.101–1.162	1.131*	1.101–1.162	1.217*	1.159–1.305
Fortune-telling	1.110	0.955–1.219	0.908	0.735–1.037	1.177	0.855–1.506
Magnification and minimization	1.084*	1.039–1.130	1.084*	1.039–1.130	1.085*	1.020–1.412
Should statements	1.056*	1.013–1.100	1.056*	1.013–1.100	1.116	0.837–1.409
Mindreading	1.052	0.984–1.117	1.052	0.984–1.117	1.127	0.894–1.259
Catastrophizing	0.920	0.763–1.077	0.920	0.763–1.077	0.979	0.859–1.046

PR≫1 indicates a significantly higher prevalence in the D cohort (indicated by asterisks). Values were calculated under three distinct conditions, labelled PR_A (values for the entire set of CDS), PR_I (values for CDS without FPPs ('I', 'me', 'my', 'mine' and 'myself')) and PR_C (values with a 95% CI resulting from resampling the set of CDS instead of our sample of individuals) (Methods).

Table 4 | Statistics with respect to significance for our set of CDS, grouped in 12 cognitive distortion categories

Cognitive distortion category	N_{CD}	N^*	N_r^* (%)
Personalizing	14	8	57.1
Emotional reasoning	7	4	57.1
Overgeneralizing	21	12	57.1
Mental filtering	14	3	21.4
Disqualifying the positive	14	3	21.4
Labelling and mislabelling	44	15	34.1
Dichotomous reasoning	23	14	60.9
Fortune-telling	8	2	25.0
Magnification and minimization	8	3	37.5
Should statements	5	1	20.0
Mindreading	72	7	9.7
Catastrophizing	11	1	9.1
Total	241	73	30.3

The N^* and N_r^* columns show the number and ratio of *n*-grams, respectively, for which we found a statistically significantly greater prevalence in the D cohort compared with the R cohort.

cognitive distortion type that have PR values for which we can conclude that the D cohort uses these schemata more.

We observed the individually highest PR scores for the CDS 'if it only', 'because my' and 'because I feel' and the individually lowest PR scores for 'she will not believe', 'we will not think' and 'nobody will believe' (which belong to the non-reflexive 'mindreading' type).

Robustness. In the following text, we discuss our efforts to verify whether our results may be explained by random variations in our sample of individuals, our particular choice of CDS *n*-grams, the sentiment loadings of our CDS set and the known propensity of individuals with depression to make self-referential statements. When accounting for these factors, in all cases, we continued to find much higher levels of distorted thinking in the language of the individuals in the D cohort compared with individuals in the R cohort. However, we caution that possible biases resulting from our data collection (for example, the veracity of the diagnosis statements or the degree to which individuals are willing to disclose a diagnosis) are difficult to assess, and are part of an ongoing discussion in the literature[38,39].

Absence of sentiment effect. Previous research has shown that the language of individuals with depression is less positive (lower text valence) and contains higher levels of self-referential language[19,40–44]. To determine the degree to which our results can be explained by text sentiment or self-referential statements instead of distorted thinking, we examined the valence loadings of our collection of tweets and CDS, and reproduced our results with and without CDS containing self-referential statements.

First, we determined the valence values of each CDS *n*-gram in our set using the VADER sentiment analysis tool[45], which was shown in a recent survey to outperform other available sentiment analysis tools for social media language[46]. VADER is particularly appropriate for this use, as its sentiment ratings take into account grammatical context, such as negation, hedging and boosting. We found that 75.9% of our CDS have either no sentiment-loaded content or are rated to have zero valence (neutral sentiment scores). The average valence rating of all CDS is −0.05 ($n=241$) on a scale from −1.0 to +1.0. Figure 3a shows the VADER sentiment distribution of only CDS *n*-grams with non-zero ratings. Here we observed only a small negative skew of CDS sentiment for this small minority of CDS *n*-grams (24.1%).

Furthermore, as shown in Fig. 3b, the sentiment distributions of all tweets for the D and R cohorts are both skewed towards positive sentiment (right side of distribution). This matches earlier findings that human language exhibits a Pollyanna effect[47], which is a near-universal phenomenon that skews human language towards positive valence. VADER sentiment ratings in the range 0.70–1.00 seem to be slightly more prevalent among the tweets of the D cohort (Fig. 3b), possibly indicating an increased emotionality (higher levels of both negative and positive affect). We found nearly identical distributions of sentiments for the tweets of the two cohorts, whether we performed the comparison for all tweets (Fig. 3b) or for only tweets containing at least one CDS (Fig. 3c). One particular deviation in the sentiment range of 0.40–0.45 was found to be uniquely associated with the use of the 'face with tears of joy' emoji (VADER sentiment = 0.4404) more often by individuals in the R cohort compared with individuals in the D cohort.

Taken together, these findings strongly suggest that the higher prevalence of CDS in the language of the D cohort can neither be attributed to a negative valence skew in the CDS set nor the sentiment distribution of the tweets produced by either the D or R cohorts.

Absence of personal pronoun effect. Research has shown that FPPs are more prevalent in the language of individuals with depression[19,23]. As many CDS contain FPPs (Supplementary Table 1, FPP(%)), our results may to a degree reflect this phenomenon instead of the 'distorted' nature of our CDS. To test the sensitivity of our results to the presence of FPPs in our set of CDS, we repeated our analysis entirely without CDS containing the FPPs 'I' (upper case), 'me', 'my', 'mine' and 'myself'. As shown in Table 3 (PR_1), we found that their removal does not alter the observed effect, except for the cognitive distortion type 'fortune-telling', which is not significantly different from parity in this case. The respective CIs resulting from our removal of FPP schemata changed slightly, but most overlap with those obtained from the analysis that included the full set of CDS (Table 3, PR_A versus PR_1), demonstrating that the presence of FPPs does not alter our results. Note that we could not determine any values for 'personalizing' because, by definition, its CDS all contain FPPs.

Robustness to CDS changes. To determine the sensitivity of our results to the particular choice of CDS, we recalculated PR values between the D and R cohorts but, instead of resampling our D and R cohort, we randomly resampled (with replacement) the set of 241 CDS *n*-grams. The 95% CI of the resulting distribution of PR values indicates how sensitive our results are to random changes in our CDS set. The results of this analysis are shown in Table 3 (PR_C). We observed small changes in the dispersion of the resulting distribution of PR values, but the median values and 95% CIs remain largely unchanged. As before, the 95% CIs continue to exclude 1.000 for all of the cognitive distortion types except for 'mindreading', 'should statements', 'fortune-telling' and 'catastrophizing', and we can continue to reject the null hypothesis that PR values are similar between the D and R cohort for nearly all cognitive distortion types. Furthermore, as shown in Table 3, the 95% CIs of PR_C and PR_A largely overlap across all cognitive distortion types, indicating that our results are robust to random changes in our CDS set as well as our D and R cohort samples. Furthermore, we examined whether URLs in tweets may bias CDS prevalence rates as they could be indicative of externally generated content that does not reflect the individual's own state. However, we found similar PR values regardless of whether we included or excluded tweets with URLs (Supplementary Information).

Discussion

In a sample of online individuals, we used a theory-driven approach to measure the prevalence of linguistic markers that may indicate

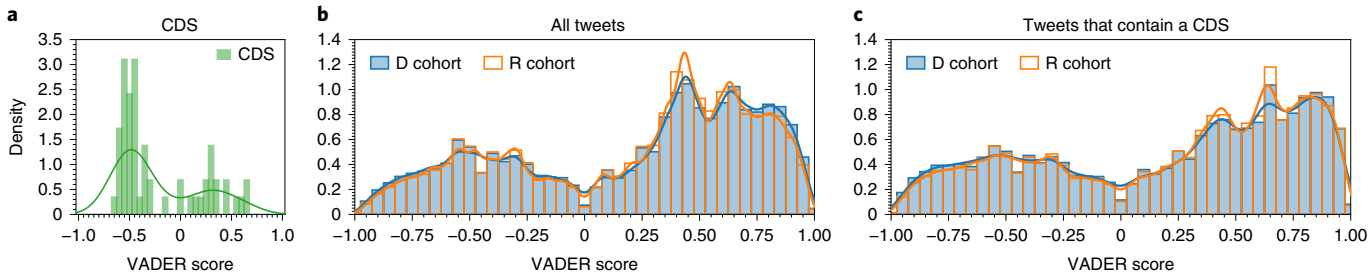

Fig. 3 | CDS and tweet sentiment scores (VADER). a, Density of VADER scores for CDS with non-zero sentiment values (58 out of 241 schemata). Most CDS carried no valence loading (75.9%). The average rating for the complete set CDS is −0.05 ($n = 241$). **b**, Distributions and kernel density estimates of the VADER valence ratings for all individual tweets. Both cohorts indicate a clear right-hand skew towards positive sentiment. The D cohort has slightly more-extreme positive and negative sentiment values compared with the R cohort, but distributions are largely comparable, indicating that there is only a small difference in sentiment values between the two cohorts. **c**, Distributions and kernel density estimates of the VADER valence ratings for all individual tweets that contain at least one CDS.

cognitive vulnerability to depression, according to CBT theory. We defined a set of CDS that we grouped along 12 widely accepted types of distorted thinking and compared their prevalence between two cohorts of Twitter users—the first included individuals who reported that they received a clinical diagnosis of depression and the second was a similar random sample.

As hypothesized, the individuals in the D cohort use significantly more CDS in their online language compared with individuals in the R cohort, particularly schemata associated with 'personalizing' and 'emotional reasoning'. We observed significantly increased levels of CDS across nearly all cognitive distortion types, sometimes more than twice as much, but did not find a statistically significant increase in prevalence among the D cohort for two specific types, namely 'fortune-telling' and 'catastrophizing'. This may be due to the difficulty of capturing these specific cognitive distortions in the form of a set of 1–5-grams—their expression in language can involve an interactive process of conversation and interpretation. Notably, our findings are not explained by the use of FPPs or more negatively loaded language. These results shed a light on the degree to which depression-related language of cognitive distortions are manifested in the colloquial language of social media platforms. This is of social relevance given that these platforms are specifically designed to propagate information through the social ties that connect individuals on a global scale.

An advantage of studying theory-driven differences between the language of individuals with and without depression, in contrast to a purely data-driven or machine learning approach, is that we can explicitly use the principles underpinning CBT to understand the cognitive and lexical components that may shape depression. Cognitive behavioural therapists have developed a set of strategies to challenge the distorted thinking patterns that are characteristic of depression. Preliminary findings suggest that specific language can be related to specific therapeutic practices and seems to be related to outcomes[48]. However, these practices have been largely shaped by a clinical understanding and not necessarily informed by objective measures of how patterns of language reflect cognitive distortions, which could be harnessed to facilitate the path of recovery.

Our results suggest a path for mitigation and intervention, including applications that engage individuals with mood disorders, such as major depressive disorder, through social media platforms and that challenge particular expressions and types of depression-related language. Future characterization of the relationship between depression-related language and mood may help in the development of automated interventions (such as 'chatbots') or suggest promising targets for psychotherapy. Another approach that has shown promise in leveraging social media for the treatment of mental health problems involves crowdsourcing the responses to cognitively distorted content[49]. These types of applications have the potential to be more-scalable mental health interventions compared with existing approaches such as face-to-face psychotherapy[50]. The extent to which user CDS prevalence can be used as a passive index of vulnerability to depression that may be expected to change with treatment could also be explored. Insofar as online language can be considered to be an index of cognitive vulnerability to depression, a better understanding of online language may help to tailor treatments, especially internet-based treatments, to the more-specific needs of individuals. For example, interventions that target depression-related thinking and language may be well-suited for individuals with depression who express relatively higher levels of these distortions, whereas interventions that target other mechanisms (such as physical activity, circadian rhythm) may be better suited for individuals who do not show relatively higher levels of CDS. More research towards understanding differences in language patterns in depression and related disorders, such as anxiety disorders, is recommended. However, when implementing these types of approaches, ethical considerations and privacy issues have to be adequately addressed[38,39].

Several limitations of our theory-driven approach should be considered. First, we relied on individuals reporting their personal clinical depression diagnoses on social media. Although we verified that the statement indeed pertains to a clinical diagnosis, we do not have verification of the diagnosis itself nor of its accuracy. This may introduce individuals into the D cohort who might not have been diagnosed with depression or accurately diagnosed. Vice versa, we have no verification that individuals in our random sample do not suffer from depression. However, the potential inaccuracy of this inclusion criterion will probably reduce the difference in depression rates between the two cohorts and, therefore, reduce the observed effect sizes (PR values between cohorts) due to the larger heterogeneity of our sample. As a consequence, our results are probably not an artefact of the accuracy of our inclusion criterion. Second, our approach is limited to discovering only individuals who are willing to disclose their diagnosis on social media. As this might skew our D cohort to a subgroup of individuals suffering from depression, we recommend caution when generalizing our findings to the level of all individuals who have depression. Third, our lexicon of CDS was composed and approved by a panel of ten experts who may have been only partially successful in capturing all of the *n*-grams used to express distorted ways of thinking. On a related note, the use of CDS *n*-grams implies that we measure distorted thinking by proxy, namely through language, and our observations may be therefore be affected by linguistic and cultural factors. Common idiosyncratic

or idiomatic expressions may syntactically represent a distorted form of thinking, but no longer do so in practice. For example, an expression such as 'literally the worst' may be commonly used to express dismay, without necessarily involving the speaker experiencing a distorted mode of thinking. Thus, the presence of a CDS does not point to a cognitive distortion per se. Fourth, both cohorts were sampled from Twitter, one of the leading social media platforms, the use of which may be associated with higher levels of psychopathology and reduced well-being[51–53]. We may therefore be observing increased or biased rates of distorted thinking in both cohorts as a result of platform effects. However, we report relative prevalence numbers with respect to a carefully construed random sample also taken from Twitter, which probably compensates for this effect and the effect that individuals with depression might be more active than their random counterparts. Furthermore, recent analysis indicates that representative samples with respect to psychological phenomena can be obtained from social media content[54]. This is an important discussion in computational social science that will continue to be investigated. Data-driven approaches that analyse natural language in real-time will continue to complement theory-driven work such as ours.

As we analysed individuals on the basis of inferred health-related information, we want to stress some additional considerations regarding ethical research practices and data privacy[30,38,39]. We limited our investigation strictly to comparing, in the aggregate, the publicly shared language of two deidentified cohorts of individuals (individuals who report that they have been diagnosed with depression and a random sample). We carefully deidentified all obtained data to protect user privacy and performed our analysis under the constraints of two IRB protocols (IU IRB Protocols 2010371843 and 1707249405). Whereas the outcomes of our analysis could contribute to a better understanding of depression as a mental health disorder, they could also inform approaches that detect traces of mental health issues in the online language of individuals, and as such contribute to future detection, diagnostics and intervention efforts. This may raise important ethical and user privacy concerns as well as risk of harm, including but not limited to the right to privacy, data ownership and transparency. For example, even though social media data are technically public, individuals do not necessarily realize nor consent to particular retrospective analysis when they share information on their public accounts[55] nor can they consent to how these data may be leveraged in future approaches that may involve individualized interactions and inventions. Considering existing evidence that individuals are more willing to share biomedical data than social media data[56], in future research, we hope to reach a larger sample of individuals who understand public data availability and increase transparency through a carefully managed consent process. We acknowledge that these considerations are part of an active and ongoing discussion in our community that we encourage and that we hope our research may contribute to.

We emphasize that not all use of CDS *n*-grams reflects depressive thinking, as these phrases are part of normal English usage, and it would therefore be wrong to try to diagnose depression merely on the basis of use of one or more such phrases. Such an approach would, as well as being inaccurate, potentially lead to harm in terms of stigmatizing individuals.

Methods

Data privacy and handling. Throughout our analysis we adhered to two Indiana University (IU) Institutional Review Board (IRB) protocols, namely IU IRB Protocol 2010371843 'Depressed individuals express more distorted thinking on social media', which was reviewed specifically for this entire study and its research team, and IU IRB Protocol 1707249405, which previously covered the data collection and analysis. As this study analyses individuals on the basis of inferred health-related information, additional steps were taken to ensure the privacy of all of the individuals in our cohorts. We deidentified all data by assigning each tweet and each user a unique label, for example D2345960 or R17156599, in both cohorts, to remove all identifying information from our analysis. All raw data are stored on a protected IU server that is accessible only to members of the study team.

Demographic information. Twitter accounts are not generally associated with detailed demographic information about the individuals in question, other than what individuals may choose to self-report in their profiles and the content that they post. However, demographic information can be reliably inferred from a variety of account characteristics, such as the individual's name and 'screen name', profile photograph and biographies. To infer the demographic information of all Twitter accounts, we used the M3 system[35], which is a highly accurate deep learning classifier that was trained on a massive Twitter dataset using profile images, screen names, names and biographies. The classifier is built to classify an account along three categories; (1) gender (male/female, Macro-F1: 0.915), (2) age ('18 and below', '19–29', '30–39' and '40 and up', Macro-F1: 0.425) and (3) organization (individual versus organizational account, Macro-F1: 0.898). To assure precision, we used a high threshold to assign a label to each account on the basis of the output of the M3 system. For the gender and organization categories, we set the threshold at 0.8. For age, we set the threshold at 0.6.

Data and sample construction. Using the Twitter application program interface (API) and the IUNI OSoMe[57] (a service that provides searchable access to the Twitter Gardenhose, a 10% sample of all daily tweets), we searched for tweets that matched both 'diagnos*' and 'depress*'. The resulting set of tweets were then filtered for matching the expressions 'i', 'diagnos*', 'depres*' in that order in a case-insensitive manner, allowing insertions to match the greatest variety of diagnosis statements; for example, a tweet that states 'I was in fact just diagnosed with clinical depression' would match. Finally, to ensure that we are including only true self-referential statements of a depression diagnosis, a team of three experts manually excluded quotes, jokes and external references. The members of this team assessed the collection of tweets to verify that we included only explicit statements that the individual had received a clinical diagnosis. All quotes, retweets and external references to depression (that is, 'My friend and I were practically diagnosed with depression over the Game of Thrones finale') were removed. A similar approach was deemed to be most accurate in a comparative analysis of social media sampling methods[58]. As recommended previously[59], we avoided the use of data-driven supervised machine learning approaches to draw conclusions with respect to the language features and population morbidity of depression[58].

We do not have certainty that the reported clinical depression diagnoses are in fact accurate. However, although the clinical recognition of depression is poor in some settings[5], patients who are recognized as being depressed tend to, on average, have higher levels of depression compared with those who are not recognized[60]. This observation, along with research suggesting that depression is best understood as existing on a continuum (reviewed previously[61]), supports our use of an explicit report of a clinical depression diagnosis as the inclusion criteria for the D cohort.

For each qualifying diagnosis tweet, we retrieved the timeline of the corresponding Twitter user using the Twitter 'user_timeline' API endpoint (https://developer.twitter.com/en/docs/tweets/timelines/api-reference/get-statuses-user_timeline). Subsequently, we excluded all non-English tweets (Twitter API machine-detected 'lang' field), all retweets and all tweets containing 'diagnos*' or 'depress*'. As we wanted to analyse only personal accounts belonging to individuals, we excluded all accounts that M3 predicted to belong to an organization or institution, leading to a final D cohort of 1,035 individuals and 1,510,359 tweets.

To compare CDS prevalence rates of the D cohort to a baseline, we constructed a random sample of individuals (R cohort). To do so, we collected a large sample of random tweets in 3 weeks (that is, 1–8 September 2017, 1–8 March 2018 and 1–8 September 2018) from the IUNI OSoMe[57]. We extracted all Twitter user identifiers from these tweets ($n = 588{,}356$), and included only those that specified their geographical location and were not already included in our D cohort. To equalize platform, interface and behavioural changes over time, we selected a subsample of these individuals such that the distribution of their account creation dates matches those of the D cohort, resulting in an initial set of 9,525 random individuals. Finally, we collected the Twitter timelines of these users and filtered the obtained data in the same manner as described for the D cohort, again excluding accounts that the M3 classifier predicted to be an institution or organization, resulting in a final R cohort consisting of 7,349 individuals and 6,783,353 tweets.

No statistical methods were used to predetermine sample sizes, but our sample sizes are similar to those reported in previous publications[21,29].

Construction of set of CDS *n*-grams. A. T. Beck introduced the concept of cognitive distortions to characterize the thinking of individuals with depression[62,63]. Subsequently, other clinicians expanded on his typology of distortions[64]—notably, clinical psychologist and CBT expert, J. Beck[65]. We drew on these latest lists, which consist of 12 types of cognitive distortions, that may characterize the thinking of individuals with depression.

A panel of CBT experts (three co-authors and seven experts consulted) engaged in a process of collaborative design, followed by a consensus voting procedure (unanimous decision) to map a set of 241 CDS *n*-grams, each geared to express at least one type of cognitive distortion. The schemata in each

category were formulated to capture the minimal semantic building blocks of distorted thinking for a particular type, avoiding expressions that are specific to depression-related topics, such as poor sleep or health issues. For example, the common 3-gram 'I am a' was included as a building block of expressing a variety of 'labelling and mislabelling' cognitive distortions, because it would be a highly likely (and nearly unavoidable) n-gram to express many self-referential ('I') expressions of labelling ('am a'). We show a set of examples in Table 1. Where possible, higher-order n-grams were chosen to capture as much of the semantic structure of one or more distorted schemata as possible, for example, the 3-gram 'everyone will believe' captures both 'overgeneralizing' and 'mindreading'. We did include 1-grams, such as 'nobody' and 'everybody', as they strongly correspond to the expression of 'dichotomous reasoning'. The number of schemata per category in our CDS set along with the average n-gram size, as well as a number of relevant grammatical features, are provided in Supplementary Table 1. The complete set of CDS is provided in Supplementary Table 2.

PR values. For each Twitter user u in our sample, we retrieved a timeline T_u of their time-ordered k most recent tweets, $T_u = \{t_1, t_2, ..., t_k\}$. We also defined a set $C = \{c_1, c_2, ..., c_n\}$ of n-grams where $n = 241$ (Table 4) with varying $n \in [1,5]$ number of terms. The elements of set C are intended to represent the lexical building blocks of expressing cognitive distortions (Table 4 and Supplementary Table 2. We introduced a CDS matching function $\mathcal{F}_C(t) \rightarrow \{0, 1\}$, which maps each individual tweet t to either 0 or 1 according to whether a tweet t contains one or more of the schemata in set C. Note that the range of $\mathcal{F}_C(t)$ is binary; therefore, a tweet that contains more than one CDS still counts as 1.

The within-individual prevalence of tweets for individual u is defined as the ratio of tweets that contain a CDS in C over all tweets in their timeline T_u:

$$P_C(u) = \frac{\sum_{t \in T_u} \mathcal{F}_C(t)}{|T_u|}$$

Our sample is separated into two cohorts—one of 1,035 individuals with depression and another of 7,349 randomly sampled individuals. We denoted the set of all individuals with depression $D = \{u_1, u_2, ..., u_{1,035}\}$ and random sample cohort $R = \{u_1, u_2, ..., u_{7,349}\}$. Thus, the sets of all tweets written by users in the D and R cohorts are defined as:

$$T_D = \bigcup_{u \in D} T_u \text{ and } T_R = \bigcup_{u \in R} T_u \quad (1)$$

We can then define the prevalence (P) of tweets with CDS C for each the D and R cohorts as follows:

$$P_C(D) = \frac{\sum_{t \in T_D} \mathcal{F}_C(t)}{|T_D|} \text{ and } P_C(R) = \frac{\sum_{t \in T_R} \mathcal{F}_C(t)}{|T_R|} \quad (2)$$

or, informally, the ratio of tweets that contain any CDS over all tweets written by the individuals of that cohort.

As a consequence, the PR of CDS in set C between the two cohorts D and R, denoted $\mathrm{PR}_C(D,R)$, is defined simply as the ratio of their respective CDS prevalence $P_C(T_D)$ and $P_C(T_R)$ in the tweet sets T_D and T_R, respectively:

$$\mathrm{PR}_C(D, R) = \frac{P_C(D)}{P_C(R)} \quad (3)$$

If $\mathrm{PR}_C(D,R) \simeq 1$, the prevalence of CDS in the tweets of the D cohort are comparable to their prevalence in the tweets of the R cohort. However, any value $\mathrm{PR}_C(D,R) \ll 1$ or $\mathrm{PR}_C(D,R) \gg 1$ may indicate a significantly higher prevalence in each respective cohort. Here we used $\gg 1$ and $\ll 1$ to signify that a PR value is significantly higher or lower than 1 respectively, which we asses on the basis of whether its 95% CI includes 1 or not (see the 'Bootstrapping estimates' section below).

Bootstrapping estimates. The estimated P and PR values can vary with the particular composition of (1) set C (our CDS n-grams) or (2) the set of individuals in our D and R cohorts. We verified the reliability of our results by randomly resampling either C or both D and R, with replacement. This was repeated $B = 10,000$ times, leading to a set of resampled cognitive distortion sets or cohort samples. Each of these B number of resamples of either (1) the set of CDS C or (2) or the sets D and C of all individuals in our D and R cohorts results in B number of corresponding P or PR values:

$$P^* = \{P_1^*, P_2^*, ..., P_B^*\} \text{ and } PR^* = \{PR_1^*, PR_2^*, ..., PR_B^*\} \quad (4)$$

The distributions of P^* and PR^* were then characterized by their median (μ_{50}) and their 95% CI ($\mu_{2.5}$–$\mu_{97.5}$). A 95% CI of a PR that does not contain 1 is held to indicate a significant difference in prevalence between the two cohorts.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All data used in this study are available in deidentified form in a dedicated and open GitHub repository (https://github.com/mctenthij/CDS_paper). Any additional information with respect to the data used in this study will be made available from the corresponding author upon reasonable request, provided this information can be made available in deidentified form. Any additional data and information are available from the corresponding author on reasonable request.

Code availability

The code and related data of this study are freely available at GitHub (https://github.com/mctenthij/CDS_paper) enabling reproduction.

Received: 20 August 2020; Accepted: 7 January 2021;
Published online: 11 February 2021

References

1. Greenberg, P. E., Fournier, A.-A., Sisitsky, T., Pike, C. T. & Kessler, R. C. The economic burden of adults with major depressive disorder in the United States (2005 and 2010). *J. Clin. Psychiatry* **76**, 155–162 (2015).
2. World Health Organization. *Depression and Other Common Mental Disorders: Global Health Estimates* (World Health Organization, 2017).
3. Case, A. & Deaton, A. Rising morbidity and mortality in midlife among white non-Hispanic Americans in the 21st century. *Proc. Natl Acad. Sci. USA* **112**, 15078–15083 (2015).
4. Mojtabai, R., Olfson, M. & Han, B. National trends in the prevalence and treatment of depression in adolescents and young adults. *Pediatrics* **138**, e20161878 (2016).
5. Mitchell, A. J., Vaze, A. & Rao, S. Clinical diagnosis of depression in primary care: a meta-analysis. *Lancet* **374**, 609–619 (2009).
6. Wang, P. S. et al. Twelve-month use of mental health services in the United States: results from the National Comorbidity Survey replication. *Arch. Gen. Psychiatry* **62**, 629–640 (2005).
7. Hofmann, S. G., Asnaani, A., Vonk, I. J., Sawyer, A. T. & Fang, A. The efficacy of cognitive behavioral therapy: a review of meta-analyses. *Cognit. Ther. Res.* **36**, 427–440 (2012).
8. Cuijpers, P. et al. Does cognitive behaviour therapy have an enduring effect that is superior to keeping patients on continuation pharmacotherapy? A meta-analysis. *BMJ Open* **3**, e002542 (2013).
9. Lorenzo-Luaces, L., German, R. E. & DeRubeis, R. J. It's complicated: the relation between cognitive change procedures, cognitive change, and symptom change in cognitive therapy for depression. *Clin. Psychol. Rev.* **41**, 3–15 (2015).
10. Ozdel, K. et al. Measuring cognitive errors using the cognitive distortions scale (CDS): psychometric properties in clinical and non-clinical samples. *PloS ONE* **9**, e105956 (2014).
11. Beck, A. T. & Haigh, E. A. Advances in cognitive theory and therapy: the generic cognitive model. *Annu. Rev. Clin. Psychol.* **10**, 1–24 (2014).
12. Clark, D. A. & Beck, A. T. Cognitive theory and therapy of anxiety and depression: convergence with neurobiological findings. *Trends Cognit. Sci.* **14**, 418–424 (2010).
13. Foland-Ross, L. C. & Gotlib, I. H. Cognitive and neural aspects of information processing in major depressive disorder: an integrative perspective. *Front. Psychol.* **3**, 489 (2012).
14. van de Leemput, I. A. et al. Critical slowing down as early warning for the onset and termination of depression. *Proc. Natl Acad. Sci. USA* **111**, 87–92 (2014).
15. Webb, T. L., Miles, E. & Sheeran, P. Dealing with feeling: a meta-analysis of the effectiveness of strategies derived from the process model of emotion regulation. *Psychol. Bull.* **138**, 775–808 (2012).
16. Fan, R. et al. The minute-scale dynamics of online emotions reveal the effects of affect labeling. *Nat. Hum. Behav.* **3**, 92–100 (2019).
17. DeRubeis, R. J., Siegle, G. J. & Hollon, S. D. Cognitive therapy versus medication for depression: treatment outcomes and neural mechanisms. *Nat. Rev. Neurosci.* **9**, 788–796 (2008).
18. Troy, A. S., Wilhelm, F. H., Shallcross, A. J. & Mauss, I. B. Seeing the silver lining: cognitive reappraisal ability moderates the relationship between stress and depressive symptoms. *Emotion* **10**, 783–795 (2010).
19. Rude, S., Gortner, E.-M. & Pennebaker, J. Language use of depressed and depression-vulnerable college students. *Cogn. Emot.* **18**, 1121–1133 (2004).
20. Tackman, A. M. et al. Depression, negative emotionality, and self-referential language: a multi-lab, multi-measure, and multi-language-task research synthesis. *J. Pers. Soc. Psychol.* **116**, 817–834 (2019).
21. Coppersmith, G., Dredze, M., Harman, C. & Hollingshead, K. From ADHD to SAD: analyzing the language of mental health on Twitter through self-reported diagnoses. In *Proc. 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, CLPsych* (eds Mitchell, M., Coppersmith, G. & Hollingshead, K.) 1–10 (Association for Computational Linguistics, 2015).

22. Bernard, J. D., Baddeley, J. L., Rodriguez, B. F. & Burke, P. A. Depression, language, and affect: an examination of the influence of baseline depression and affect induction on language. *J. Lang. Soc. Psychol.* **35**, 317–326 (2016).
23. Smirnova, D. et al. Language patterns discriminate mild depression from normal sadness and euthymic state. *Front. Psychiatry* **9**, 105 (2018).
24. Zimmermann, J., Brockmeyer, T., Hunn, M., Schauenburg, H. & Wolf, M. First-person pronoun use in spoken language as a predictor of future depressive symptoms: preliminary evidence from a clinical sample of depressed patients. *Clin. Psychol. Psychother.* **24**, 384–391 (2017).
25. Cacheda, F., Fernandez, D., Novoa, F. J. & Carneiro, V. Early detection of depression: social network analysis and random forest techniques. *J. Med. Internet Res.* **21**, e12554 (2019).
26. Cavazos-Rehg, P. A. et al. A content analysis of depression-related tweets. *Comput. Hum. Behav.* **54**, 351–357 (2016).
27. Al-Mosaiwi, M. & Johnstone, T. In an absolute state: elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation. *Clin. Psychol. Sci.* **6**, 529–542 (2018).
28. Eichstaedt, J. C. et al. Facebook language predicts depression in medical records. *Proc. Natl Acad. Sci. USA* **115**, 11203–11208 (2018).
29. Choudhury, M. D., Gamon, M., Counts, S. & Horvitz, E. Predicting depression via social media. In *Proc. Seventh International AAAI Conference on Weblogs and Social Media, ICWSM* (eds Emre Kiciman, E., Ellison, N. B., Hogan, B., Resnick, P. & Soboroff, I.) 128–137 (2013).
30. Reece, A. G. et al. Forecasting the onset and course of mental illness with Twitter data. *Sci. Rep.* **7**, 13006 (2017).
31. Coppersmith, G., Dredze, M. & Harman, C. Quantifying mental health signals in Twitter. In *Proc. Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, CLPsych* (eds Resnik, P., Resnik, R. & Mitchell, M.) 51–60 (Association for Computational Linguistics, 2014).
32. Hasin, D. S. et al. Epidemiology of adult DSM-5 major depressive disorder and its specifiers in the United States. *JAMA Psychiatry* **75**, 336–346 (2018).
33. Rnic, K., Dozois, D. J. & Martin, R. A. Cognitive distortions, humor styles, and depression. *Eur. J. Psycholol.* **12**, 348–362 (2016).
34. Fiske, A., Wetherell, J. L. & Gatz, M. Depression in older adults. *Ann. Rev. Clin. Psychol.* **5**, 363–389 (2009).
35. Wang, Z. et al. Demographic inference and representative population estimates from multilingual social media data. In *Proc. 2019 World Wide Web Conference, WWW* (eds Liu, L. & White, R.) 2056–2067 (Association for Computing Machinery, 2019).
36. Albert, P. R. Why is depression more prevalent in women? *J. Psychiatry Neurosci.* **40**, 219–221 (2015).
37. Kuehner, C. Why is depression more common among women than among men? *Lancet Psychiatry* **4**, 146–158 (2017).
38. Chancellor, S., Birnbaum, M. L., Caine, E. D., Silenzio, V. M. B. & De Choudhury, M. A taxonomy of ethical tensions in inferring mental health states from social media. In *Proc. Conference on Fairness, Accountability, and Transparency, FAT* (eds Boyd, D. & Morgenstern, J.) 79–88 (Association for Computing Machinery, 2019).
39. Benton, A., Coppersmith, G. & Dredze, M. Ethical research protocols for social media health research. In *Proc. First ACL Workshop on Ethics in Natural Language Processing, EthNLP* (eds Hovy, D. et al) 94–102 (Association for Computational Linguistics, 2017).
40. Molendijk, M. L. et al. Word use of outpatients with a personality disorder and concurrent or previous major depressive disorder. *Behav. Res. Ther.* **48**, 44–51 (2010).
41. Fast, L. A. & Funder, D. C. Gender differences in the correlates of self-referent word use: authority, entitlement, and depressive symptoms. *J. Pers.* **78**, 313–338 (2010).
42. Al-Mosaiwi, M. & Johnstone, T. In an absolute state: elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation. *Clin. Psychol. Sci.* **6**, 529–542 (2018).
43. Brockmeyer, T. et al. Me, myself, and I: self-referent word use as an indicator of self-focused attention in relation to depression and anxiety. *Front. Psychol.* **6**, 1564 (2015).
44. Ingram, R. E. Self-focused attention in clinical disorders: review and a conceptual model. *Psychol. Bull.* **107**, 156–176 (1990).
45. Hutto, C. J. & Gilbert, E. VADER: a parsimonious rule-based model for sentiment analysis of social media text. In *Proc. Eighth International AAAI Conference on Weblogs and Social Media, ICWSM* (eds Adar, E. & Resnick, P.) (Association for Computing Machinery, 2014).
46. Ribeiro, F. N., Araújo, M., Gonçalves, P., Gonçalves, M. A. & Benevenuto, F. Sentibench—a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Sci.* **5**, 23 (2016).
47. Dodds, P. S. et al. Human language reveals a universal positivity bias. *Proc. Natl Acad. Sci. USA* **112**, 2389–2394 (2015).
48. Ewbank, M. P. et al. Quantifying the association between psychotherapy content and clinical outcomes using deep learning. *JAMA Psychiatry* **77**, 35–43 (2019).
49. Morris, R. R., Schueller, S. M. & Picard, R. W. Efficacy of a web-based, crowdsourced peer-to-peer cognitive reappraisal platform for depression: randomized controlled trial. *J. Med. Internet Res.* **17**, e72 (2015).
50. Kazdin, A. E. & Blase, S. L. Rebooting psychotherapy research and practice to reduce the burden of mental illness. *Perspect. Psychol. Sci.* **6**, 21–37 (2011).
51. Lin, L. Y. et al. Association between social media use and depression among U.S. young adults. *Depress. Anxiety* **33**, 323–331 (2016).
52. Keles, B., McCrae, N. & Grealish, A. A systematic review: the influence of social media on depression, anxiety and psychological distress in adolescents. *Int. J. Adolesc. Youth* **25**, 79–93 (2020).
53. Kelly, Y., Zilanawala, A., Booker, C. & Sacker, A. Social media use and adolescent mental health: findings from the UK Millennium Cohort Study. *EClinicalScience* **6**, 59–68 (2018).
54. Kalimeri, K., Beiro, M. G., Bonanomi, A., Rosina, A. & Cattuto, C. Traditional versus Facebook-based surveys: evaluation of biases in self-reported demographic and psychometric information. *Demogr. Res.* **42**, 133–148 (2020).
55. McKee, R. Ethical issues in using social media for health and health care research. *Health Pol.* **110**, 298–301 (2013).
56. Pratap, A. et al. Contemporary views of research participant willingness to participate and share digital data in biomedical research. *JAMA Netw. Open* **2**, e1915717 (2019).
57. Davis, C. A. et al. OSoMe: the IUNI observatory on social media. *PeerJ Comput. Sci.* **2**, e87 (2016).
58. Ernala, S. K. et al. Methodological gaps in predicting mental health states from social media: triangulating diagnostic signals. In *Proc. 2019 CHI Conference on Human Factors in Computing Systems* (eds Brewster, S. & Fitzpatrick, G.) 134 (Association for Computing Machinery, 2019).
59. Chancellor, S. & Choudhury, M. D. Methods in predictive techniques for mental health status on social media: a critical review. *npj Digit. Med.* **3**, 43 (2020).
60. Kamphuis, M. H. et al. Does recognition of depression in primary care affect outcome? The PREDICT-NL study. *Fam. Pract.* **29**, 16–23 (2011).
61. Ruscio, A. M. Normal versus pathological mood: implications for diagnosis. *Ann. Rev. Clin. Psychol.* **15**, 179–205 (2019).
62. Beck, A. T. Thinking and depression: I. Idiosyncratic content and cognitive distortions. *Arch. Gen. Psychiatry* **9**, 324–333 (1963).
63. Beck, A. T. Thinking and depression: II. Theory and therapy. *Arch. Gen. Psychiatry* **10**, 561–571 (1964).
64. Burns, D. *The Feeling Good Handbook* (Harper-Collins Publishers, 1989).
65. Beck, J. S. & Beck, A. T. *Cognitive Therapy: Basics and Beyond* (Guilford Press, 1995).

Acknowledgements

We thank L. M. Rocha for his feedback on the general methodology and terminology, as well as K. Dobson, R. DeRubeis, C. Webb, S. Hoffman, N. Kazantzis, J. Garber and R. Jarrett for their feedback on the content of our list of CDS. J.B. thanks NSF grant SMA/SME1636636, COVID-19 funding from IU's Office of the Vice President for Research, The Urban Mental Health institute at the University of Amsterdam, Wageningen University and Research, and the ISI Foundation for their support. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

M.t.T. and J.B. conceptualized the analysis; M.t.T. and J.B. designed the methodology; L.A.R., L.L.-L. and J.B. constructed the CDS lexicon; K.C.B. and M.t.T. constructed the datasets; K.C.B., M.t.T. and J.B. performed data analysis; and K.C.B., M.t.T., L.A.R., L.L.-L. and J.B. wrote the manuscript.

Competing interests

L.L.-L. received an honorarium for consulting from Happify, Inc. in September 2020. Happify, Inc. creates behavioural change technologies (for example, apps) for mental health. Happify had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41562-021-01050-7.

Correspondence and requests for materials should be addressed to J.B.

Peer review information *Nature Human Behaviour* thanks Glen Coppersmith, Christopher Danforth and David Dozois for their contribution to the peer review of this work. Primary Handling Editor: Jamie Horder.

Reprints and permissions information is available at www.nature.com/reprints.

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

nature research

Corresponding author(s): JOHAN BOLLEN

Last updated by author(s): 12/08/2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Python 3.6 and Twitter API (https://developer.twitter.com/)
Data analysis	Data analysis was performed with open source software only, namely Python 3.6 with Python modules: numpy, matplotlib, json, pandas, sand cipy, VADER (https://github.com/cjhutto/vaderSentiment), and M3inference (https://github.com/euagendas/m3inference). Customized source code used in our analysis is made freely accessible and available by the authors from a dedicated Github repository (https://github.com/mctenthij/CDS_paper). All source code and software specifies exactly which modules and libraries were used to conduct the analysis.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All data and source code used in this study are available in deidentified form in a dedicated and open GitHub repository: https://github.com/mctenthij/CDS_paper. Any additional information with respect to the data used in this study will be made available from the corresponding author upon reasonable request, provided this information can be made available in deidentified form.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Quantitative, using existing, public text data: we searched Twitter for individuals who explicitly stated that they received a clinical diagnosis of depression (N=1,035) and compared the prevalence of a set of 241 cognitive distortion schemata (represented by 1, 2, 3, 4, and 5-grams) in their language vs the language of a set of random individuals on Twitter (N=7,349). Comparisons were conducted between both groups, and sub-samples separated by gender and age category.
Research sample	A random selection of Twitter users who self-selected through their participation in the Twitter social networking platform and their public text content.
Sampling strategy	Sampling procedure: 1,035 individuals in Depressed group and 7,349 individuals in random sample group based on online availability. Sample size either conform to and exceed existing studies in this domain (N = [100,1000])
Data collection	Python python3.6 and Twitter API (https://developer.twitter.com/)
Timing	All Twitter text data was collected in September 2018 and February 2019.
Data exclusions	No data were excluded from our analysis.
Non-participation	Participants were not invited to participate since the study used pre-existing, publicly available Twitter posts.
Randomization	Individuals were not allocated into experimental groups, but separated into two cohorts on the basis of the content of their existing, public Twitter posts.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging