ARTICLES
https://doi.org/10.1038/s41562-018-0490-5

nature human behaviour

The minute-scale dynamics of online emotions reveal the effects of affect labeling

Rui Fan[1]*, Onur Varol[2], Ali Varamesh[3], Alexander Barron[3], Ingrid A. van de Leemput[4], Marten Scheffer[4] and Johan Bollen[3,4,5]*

Putting one's feelings into words (also called affect labeling) can attenuate positive and negative emotions. Here, we track the evolution of specific emotions for 74,487 Twitter users by analysing the emotional content of their tweets before and after they explicitly report experiencing a positive or negative emotion. Our results describe the evolution of emotions and their expression at the temporal resolution of one minute. The expression of positive emotions is preceded by a short, steep increase in positive valence and followed by short decay to normal levels. Negative emotions, however, build up more slowly and are followed by a sharp reversal to previous levels, consistent with previous studies demonstrating the attenuating effects of affect labeling. We estimate that positive and negative emotions last approximately 1.25 and 1.5 h, respectively, from onset to evanescence. A separate analysis for male and female individuals suggests the potential for gender-specific differences in emotional dynamics.

People frequently share their emotions with friends and family[1], and increasingly do so on social media. This may have beneficial effects. Studies of affect labeling[2] indicate that when people are shown emotionally evocative images, their distress or negative emotions are significantly reduced by the mere act of putting their feelings into words[3–6]. Affect labeling has also been shown to lessen anxiety and fear towards phobias[7–9]. Interestingly, affect labeling exerts its effects even when people do not deliberately intend to use it to regulate their emotions nor believe in its efficacy[3]. It can therefore be considered an implicit emotion regulation mechanism.

A number of empirical methods have been employed to measure the effects of affect labeling in the laboratory. Participants can be queried about their emotional experience[7,10] after the presentation of an emotion-evoking stimulus using introspective measures (for example, survey instruments and diaries[11]). Biophysical measurements of emotions[11–18] (for example, measurements of facial musculature), in addition to brain-scanning techniques, can record individual reactions to emotionally evocative stimuli[14,15,19–21]. However, these measurements may involve extensive experimental and instrumental manipulation, which may introduce measurement and observation challenges[11].

Here, we measure the dynamics of naturally occurring (that is, not experimentally induced) individual emotions and their spontaneous expression in online language[3] at the resolution of minutes for approximately 74,487 Twitter users. As shown in Fig. 1, we focused our analyses on tweets stating "I feel…", which we consider to be instances of affect labeling. We then investigated whether those expressions of positively or negatively valenced emotions were associated with either intensification or attenuation of the original emotion at different time points before and after the expression.

Our results confirm and extend existing work in the area of affect labeling[2]. We observed rapid reductions of negative emotions and a less rapid reduction of positive emotions immediately after they have been expressed in written form. The effect generalizes across most individuals, indicating that it is probably not the result of differences in personality or social disposition. Participants were not aware of this research at the time they posted their tweets. Our study therefore extends previous laboratory studies to naturally occurring emotional expressions in an online social media platform.

Results

Figure 2 shows the time series of mean valence levels from 6 h before to 6 h after the affect labeling at time t_0 for positive (Fig. 2a) and negative emotions (Fig. 2b) separately. The positive and negative emotions both exhibit a distinct pattern of change before and after the affect labeling: a positive ramp up and negative ramp down, respectively, before the individuals' explicit positive and negative affirmations, followed by a respective ramp down and ramp up of the individuals' emotional states afterwards. Positive and negative emotions seem to follow different trajectories; negative emotions have a longer ramp-down period, possibly starting at $t_0 - 2$ h, but exhibit faster recovery immediately after the emotional expression than positive emotions.

As shown in Fig. 2a, the positive emotion shows a sharp peak of valence values before and after t_0. This indicates that the individuals' language reflects positive valence changes that match the valence of the positive affect labeling. Ramp up and ramp down of positive valence levels seems to be rapid and symmetrical around t_0, with a sharp peak and reversal located at t_0, indicating that the peak of emotional levels coincides with the emotional expression, and decays immediately after. All other fluctuations in the time series fall within a 95% confidence interval (CI) established from the distributions of valence values in the first 3 h of the time series. Similar CIs were observed when we sampled the last 3 h of the time series instead.

The negative emotion time series in Fig. 2b exhibits an equally sharp change in valence levels surrounding the emotional expression at time t_0, but with a slower and longer ramp down before a sharp negative peak at t_0, and fast reversal to the long-term mean within 10 min after the emotional expression.

[1]State Key Laboratory of Software Development Environment, Beihang University, Beijing, China. [2]Center for Complex Network Research, Northeastern University, Boston, MA, USA. [3]Center for Complex Networks and Systems Research, Indiana University, Bloomington, IN, USA. [4]Wageningen University, Wageningen, the Netherlands. [5]Cognitive Science Program, Indiana University, Bloomington, IN, USA. *e-mail: buaafanrui@gmail.com; jbollen@indiana.edu

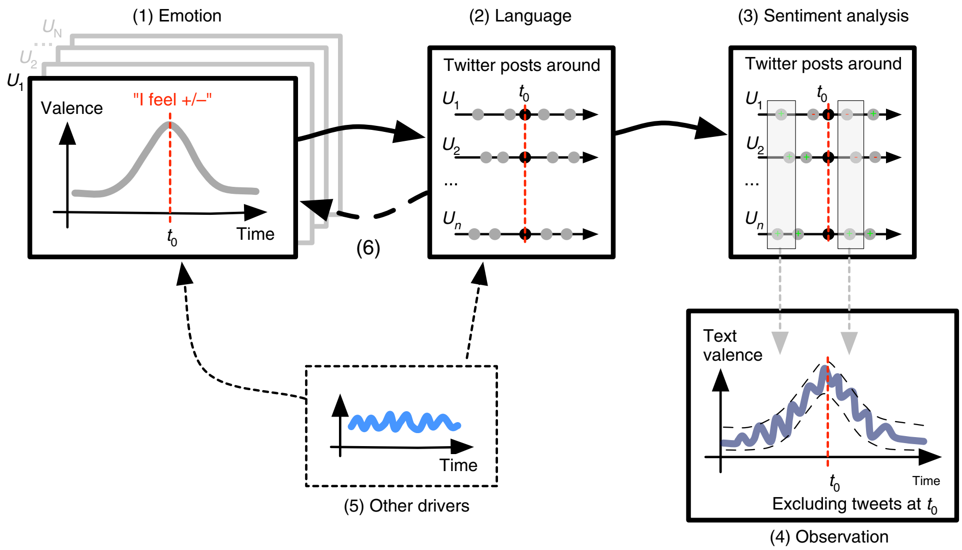

Fig. 1 | Measuring changing valence levels from language before and after affect labeling. Our method involved six steps. (1) We collected the individual Twitter timelines of N Twitter users (denoted U_1, U_2, ..., U_N) who explicitly expressed experiencing an emotion of a positive or negative valence at a specific time, t_0, by writing "I feel" followed by an adjective or adverb. In the figure, '+/−' represents a positive or negative emotion. This was deemed an act of affect labeling. (2) We aligned all individual timelines of time-ordered tweets on the time of the affect labeling (t_0). (3) We applied a sentiment analysis algorithm (VADER) to tweets posted at specific time windows before and after t_0 to detect possible changes in text valence levels. (4) We aggregated the observed valence levels within a given time window across all individuals to map changes in valence levels. (5) Emotions and language can be biased by other drivers, such as events, personal experience and dispositions. These effects were randomized across individuals (in steps 1 and 2). (6) The dashed arrow represents the assumption that language can interact with emotions. Our observations in step 4 may reflect this interaction in the timing and pattern of valence changes relative to the time of the explicit emotional expression at t_0.

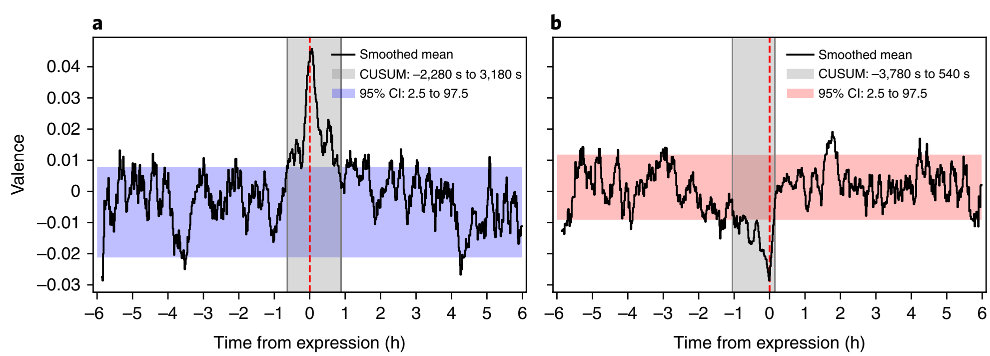

Fig. 2 | Time series of observed valence values across all individuals. a,b, Time series of observed valence values in 1 min increments, smoothed by a 10 min rolling average (positive, $n = 42{,}627$; negative, $n = 67{,}316$), for the positive (**a**) and negative (**b**) affect labeling groups. The time range extends from 6 h before to 6 h after individuals explicitly expressed experiencing an emotion at time t_0 (vertical red dashed line). The emotional expressions themselves at time t_0 were excluded from this analysis and we did not analyse timelines that contained more than 1 explicit emotional expression within a 48 h span. We used the first 3 h to estimate a baseline 95% CI for the entire time series (horizontal blue and red shaded bars). A CUSUM test was used to detect statistically significant change points in the resulting time series. Grey vertical bars mark where the CUSUM analysis indicated significant change points based on each time series' cumulative variance. Each time series was mean-centred by subtracting its mean over the 12 h span (see Supplementary Fig. 3 for raw time series showing the difference in positive and negative time series' baselines).

Detecting change points through cumulative sum testing. To confirm that the observed changes in valence were significant, we conducted a cumulative sum (CUSUM) test, which detected change points in each time series by measuring cumulative changes in its variance (see the Methods section 'CUSUM test'). The results indicate significant changes in the positive group from $t_0 - 38$ min to $t_0 + 53$ min (grey band in Fig. 2a), indicating that positive emotions typically last 92 min. A similar CUSUM test confirmed a period of significant changes in negative valence from $t_0 - 63$ min to $t_0 + 9$ min, indicating that negative emotions start earlier than positive emotions, take longer to crest and end sooner than the positive emotion after the emotional expression, with a typical duration of 73 min.

Modelling emotional dynamics. To model the dynamics of positive and negative emotions, we applied a least-squares method to determine the function that best fit the ramp up and ramp down of the positive and negative time series (see Supplementary Note 5 and Supplementary Table 1). As shown in Fig. 3a, the ramp-up and ramp-down periods of the positive emotion are best fitted with a separate exponential growth function from time $t = -38$ to -1 min ($f(t) = 0.038e^{0.157t} + 0.01$) and an exponential decay function

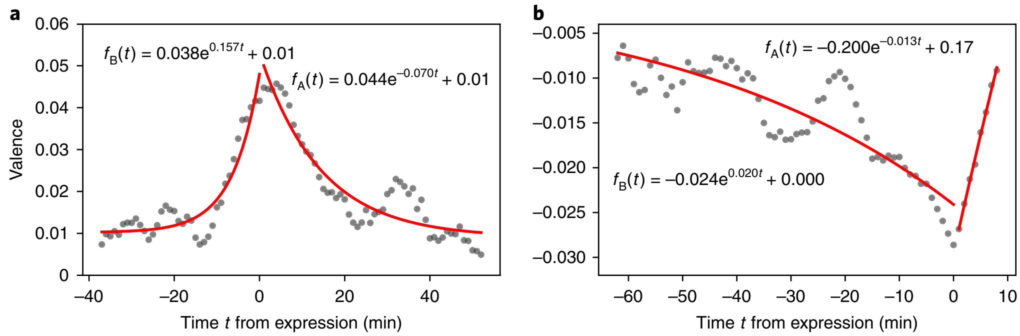

Fig. 3 | Curve-fitting results of the smoothed mean valence values. a,**b**, Mean-centered time series of smoothed mean valence values for both positive (**a**) and negative affect labeling groups (**b**) reveals a sharp peak and reversal at time t_0. The ramp up and ramp down of both positive and negative time series is best fitted by two separate exponential growth and decay functions before (f_B) and after (f_A) the affect labeling at time t_0, respectively (see Supplementary Note 5). The ramp up and ramp down of positive valence (**a**) is best fit with the functions $f_B(t) = 0.038e^{0.157t} + 0.01$ before t_0 and $f_A(t) = 0.044e^{-0.070t} + 0.01$ after t_0. The ramp down and ramp up of negative valence (**b**) is best fit with the functions $f_B(t) = -0.024e^{0.020t}$ before t_0 and $f_A(t) = -0.200e^{-0.013t} + 0.170$ after t_0.

from time $t = 1$ to 53 min ($f(t) = 0.044e^{-0.070t} + 0.01$), respectively. For the negative emotion shown in Fig. 3b, the growth before t_0 and decay period after t_0 are also best fitted with a separate exponential growth function from time $t = -63$ to -1 min ($f(t) = -0.024e^{0.020t}$) and exponential decay function from time $t = 1$ to 9 min ($f(t) = -0.200e^{-0.013t} + 0.170$), respectively.

On the basis of the fitted functions, we can calculate the half-life of positive and negative emotions. Given the peak value of an emotion p and the emotion value at the end of emotional period e (average emotion score), we define the half-life as $t_{(p+e)/2} - t_p$, indicating the time it takes for the emotion to decay from the peak to half of its baseline value. By this definition, the half-life of positive and negative emotion is 11 and 5 min, respectively, indicating a swift return from peak levels to the valence baseline.

It is suggestive that most ramp ups and declines of the individuals' emotional states—both positive as well as negative—are best fit with exponential trajectories when compared with other functions (see Supplementary Table 1). This may be indicative of the presence of feedback loops in the individuals' emotional systems that were affected or interrupted by the individuals' affect labeling at t_0.

Mean valence CIs versus a null model. Each time window, w, in our data contained a (changing) sample of tweets that were posted within a distance, k, from t_0; that is, tweets posted in the time interval $t_0 - k - w$ to $t_0 - k$. Since each tweet in that window had been assigned its individual valence rating, mean valence values for a window (such as shown in Fig. 2) were thus calculated for a distribution of n tweet valence values. The properties of this distribution could change in terms of sample size, location and variance between subsequent time windows, complicating inferences about differences in valence levels between different points of the time series. In addition, the volume of tweets for each time window could vary depending on the distance from t_0 (see Supplementary Fig. 3). Our estimates of mean valence at time t may therefore be more or less uncertain depending on the underlying sample of tweets.

To estimate the error of our estimation of mean valence for the tweets observed in a given time window w, we bootstrapped 10,000-fold (1) mean valence values and (2) null-model mean valence values (a random sample of tweets with similar diurnal, circadian and week-day features; see the Methods section 'Null model'). To increase our sample size, we grouped tweets in 10 min windows. We then compared the resulting 95% CIs between observed valence values and those produced by the null model.

The results of this comparison are shown in Fig. 4. The time series are displayed for discrete and adjacent windows of 10 min. Comparable results were obtained with 1, 5 and 15 min windows. The 95% CIs of the estimated mean valence levels are shown (red and blue bands), as well as the 95% CI for the null-model estimates (grey band) for each window.

The CIs of the positive valence time series in Fig. 4a overlap with those of the null model for most of the 12 h period under consideration, with the exception of the period from −10 to +20 min, where we observe that the CIs of the observed mean positive valence and those of the null model do not overlap. This period overlaps with the CUSUM change point detection, although it is shorter, possibly due to applying the stricter criterion of non-overlapping 95% CIs. We can draw a similar conclusion for the negative time series in Fig. 4b; that is, the 95% CIs do not overlap from −40 to 0 min, confirming a negative emotional period similar to that given by the CUSUM test. The span of the emotional period estimated by this strict criterion might be an underestimation, since areas where the CIs do overlap might still represent cases where valence levels are significantly different from null-model levels.

Estimating emotion duration. The CUSUM and 95% CIs indicate different, yet overlapping, emotional periods surrounding time t_0, as shown in Table 1. The timing of the emotional period as indicated by the CUSUM test indicates time ranges of elevated or depressed valence distributions from −38 to +53 min for the positive emotion and −63 to +9 min for the negative emotion. The CIs indicate a period of significant divergence in valence versus the null model in the ranges of −10 to +20 min and −40 to 0 min for the positive and negative emotions, respectively. As a third indication of the duration of the respective emotions, we added a less strict criterion; namely, the time period when the time series remains continuously above or below the median surrounding t_0. This criterion indicates that the positive emotional period ranges from −48 to 109 min and the negative emotion period ranges from −124 to 14 min.

Averaging the results of these three tests, we estimate that positive emotions range from −32 to +61 min (that is, a total duration of 94 min), whereas negative emotions range from −76 to +8 min (that is, a total duration of 85 min). This indicates that negative emotions do not last longer than positive emotions, but they start earlier, have a significantly longer run-up, and seem to decline and end more rapidly after their expression.

Robustness across subject samples. To assess the degree to which the time series reflected the emotional dynamics of a majority of

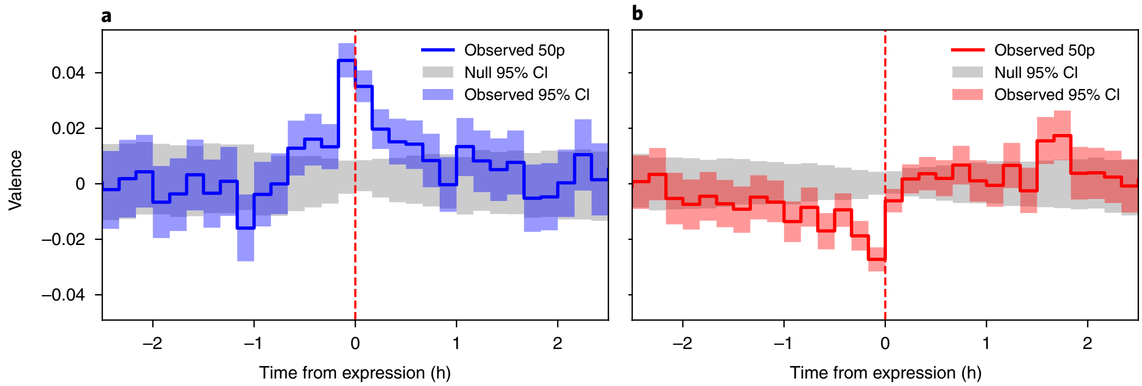

Fig. 4 | Time series of observed valence CIs versus null-model CIs at 10 min increments. **a**, Comparison of the 95% CI and median (50th percentile (50p)) of bootstrapped mean valence values (blue bands) versus the 95% CI and median (50p) of a null model (see the Methods section 'Time series construction') of randomly chosen tweets (grey bands) for the range −180 to +180 min surrounding the moment of affect labeling at time t_0 (see Table 2). CIs do not overlap surrounding time t_0 (vertical red dashed line), indicating a statistically significant increase in positive valence. **b**, Similar comparison for negative valence, indicating a statistically significant decrease of valence values surrounding time t_0 (see Table 2). Each time series was mean centred by subtracting its mean over the 12 h period shown (see Supplementary Fig. 4 for raw time series).

Table 1 | Duration and span of changes in valence levels

	Positive valence		Negative valence	
Method	**Length**	**Span**	**Length**	**Span**
CUSUM	92	−38 to +53	73	−63 to +9
95% CI	31	−10 to +20	41	−40 to 0
***v*(50p)**	158	−48 to 109	139	−124 to 14
Average	94	−32 to +61	85	−76 to +8

All values shown are in minutes. Emotion durations are estimated according to the three methods: CUSUM test; non-overlap with the 95% CI; and the length of the continuous time period within which valence levels deviate from the median (50p) before and after t_0. The bottom row shows the mean of the three methods.

Table 2 | CUSUM and CI-based estimates of emotion duration by gender

Gender (emotion)	95% CI duration (span)	CUSUM duration (span)
Female (+)	20 (−10 to +10)	94 (−41 to +53)
Male (+)	20 (−10 to +10)	97 (−48 to +49)
Female (−)	40 (−40 to 0)	48 (−41 to +7)
Male (−)	10 (−10 to 0)	40 (−32 to +8)

All values shown are in minutes. CI-based estimates were calculated at 10 min intervals and CUSUM estimates at 1 min intervals. '(+)' and '(−)' represent positive and negative emotions, respectively.

individuals, and not a minority of extreme responders, we recorded the distributions of the peak valence values (expressed in *z* scores) that surrounded t_0 for each individual valence time series (positive and negative groups separately; see Supplementary Note 6). As shown in Fig. 5a, the distribution of positive peak responses had a 95% CI of −1.240 to 2.417 and a median of +1.096, whereas the distribution of the negative peak responses (Fig. 5b) had a 95% CI of −2.434 to 1.285 and a median of −1.073, indicating that the negative and positive peak values were consistent with the positive or negative valence of the individuals' affect labeling. This result was confirmed with a two-component Gaussian mixture model (GMM), which indicated that the most significant GMM components matched the valence of the affect labeling statement (see Fig. 5a,b).

These results confirm a strong and consistent emotional response that generalizes across most individuals for both negative and positive emotions, but the distributions of peak *z* values do exhibit a positive and negative skew, respectively (see Fig. 5a,b). Although GMM component 1 of the distribution of negative peak values is centred on negative values, as expected, GMM component 2 is focused on neutral peak values, indicating attenuated or neutral peak valence *z* scores. The distribution of positive peak *z* scores exhibits a symmetrical pattern: it is centred on positive peak values with component 1, but exhibits a negative skew, due to component 2, that is centred around lower, more neutral peak sentiment values.

For negative and positive emotions, we found that 16.80 and 17.89% of peak values, respectively, ran contrary to the polarity of self-reports (that is, a positive peak value for negative emotions and a negative peak value for positive emotions). These results suggest that our observations of emotional responses around t_0 are probably an underestimation of the actual effect, since our data may include a subset of cases where the sentiment analysis yielded a neutral, attenuated or inverted sentiment signal.

Male and female responses. The above results suggest a robust and generalizable emotional response across most individuals. This is also the case when we separate our analysis for male (see Fig. 6b,d) and female individuals (see Fig. 6a,c) using a gender classifier (see the Methods section 'Gender classifier', Supplementary Note 7 and Supplementary Fig. 5).

The effect in which negative emotions dissipate rapidly after the individuals' explicit expression[22,23] appears more pronounced for women (see Fig. 6c) than men (see Fig. 6d) on the basis of a visual inspection of the relevant time series, indicating that women may experience a stronger positive effect of expressing their emotions than men do or use different regulation mechanisms[24].

When we applied the same CUSUM test and examined the timing of non-overlapping 95% CIs to detect significant changes in valence levels in the separate male and female time series, we found overlapping periods of diverging valence around t_0 in all time series (see Table 2). This indicates a significant emotional signal before and after t_0 for both male and female individuals separately.

However, this observation pertains to emotion magnitude and duration, not whether the pattern of the emotional change evolves differently for male and female individuals. To determine whether there are indications of gender differences in longitudinal dynamics, we compared the magnitudes of the male and female valence

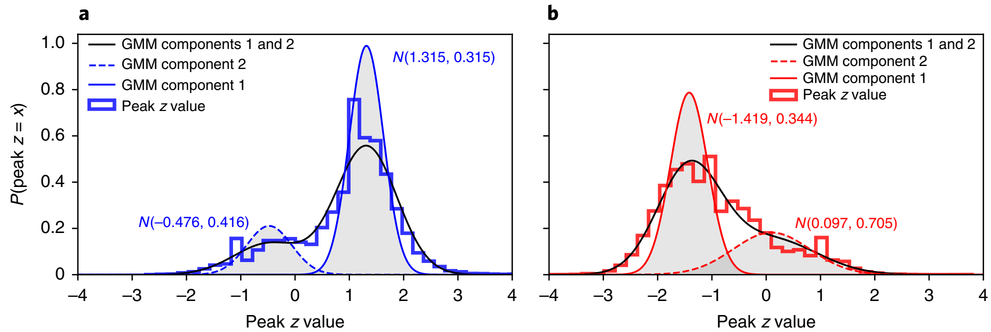

Fig. 5 | Robustness analysis. a,**b**, Distributions of peak z scores of the valence time series for positive (**a**) and negative (**b**) emotions modelled by a two-component GMM reveal that the majority of users exhibit a similar and significant emotional response matching the valence of their self-report. Both emotions (**a** and **b**) show a second GMM component that may correspond to a small subsample where the sentiment analysis tool (VADER) returns zero, insignificant or inverted sentiment scores.

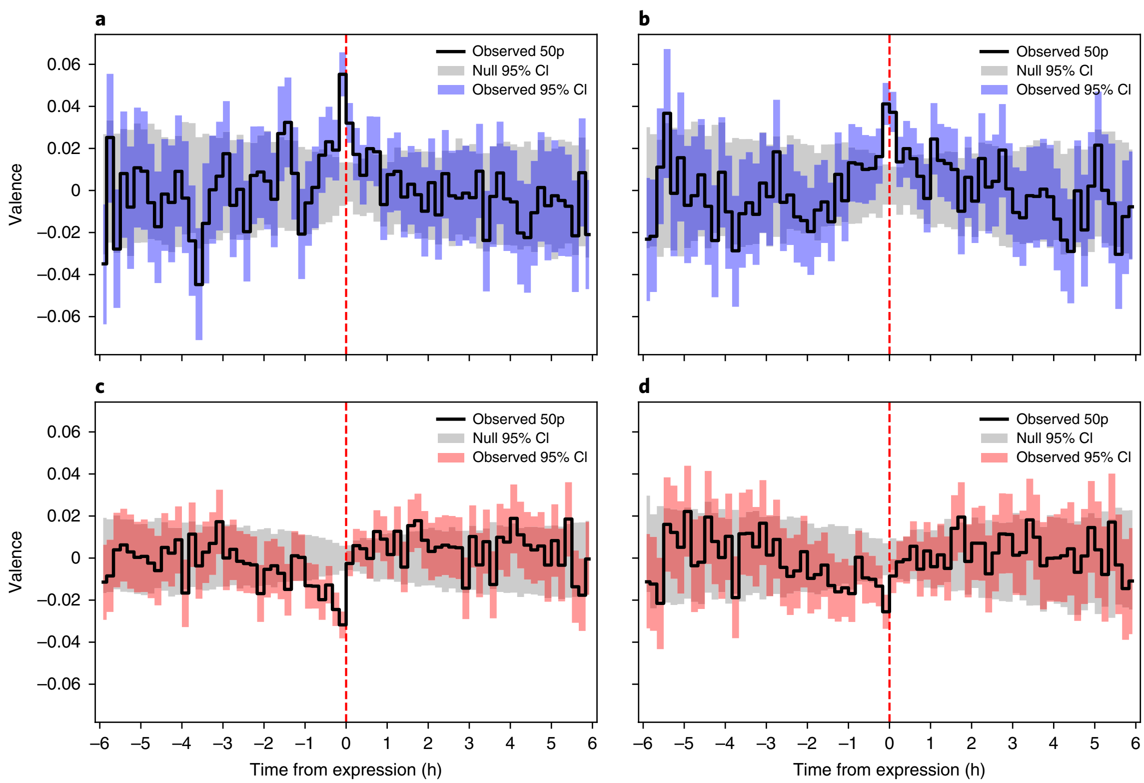

Fig. 6 | Gender differentiated time series of mean valence values. a–**d**, Time series of 95% CIs and median values (50p) determined in adjacent 10 min windows for positive (**a** and **b**) and negative (**c** and **d**) affect labeling groups, separated by gender (female, **a** and **c**; male, **b** and **d**). Blue and red bands show 95% CIs of bootstrapped mean valence levels of positive and negative time series, respectively. Grey bands show 95% CIs of valence for a null model. In **a**, there is a short period of significantly elevated sentiment surrounding t_0. We observe a sharp symmetrical peak around t_0 with a fast ramp down immediately following the affect labeling at t_0. In **b**, there is a similar short period of significantly elevated sentiment surrounding t_0. In **c**, there is a slow ramp down to a negative peak at the time of affect labeling t_0, with a rapid return to baseline valence levels. In **d**, there is also a ramp down of valence levels and a fast return to baseline levels. See the section 'Male and female responses' for a quantitative comparison of the gender-differentiated responses).

curves at different times before and after the emotional expression with a regression discontinuity analysis[25].

We plotted the difference of valence values between male and female individuals at time t (that is, $v(t) = v_{\mathrm{male}}(t) - v_{\mathrm{female}}(t)$) for positive and negative valenced emotions in Fig. 7a,b, respectively. Separate regression lines were calculated for $v(t)$ values before versus after t_0, including 95% CIs on the linear regression estimates. This allowed us to determine whether male and female valence values evolve differently before or after the time of their emotional expression, and whether the regression lines show statistically significant discontinuities.

As we observed earlier, there is a consistent and significant baseline difference between male and female valence for both positive- and negative-valence emotions. The difference between male valence and female valence is in the range −0.02 to −0.10, indicating that women generally have higher valence baselines at all points before or after the emotional expression than men, for both positive and negative emotions.

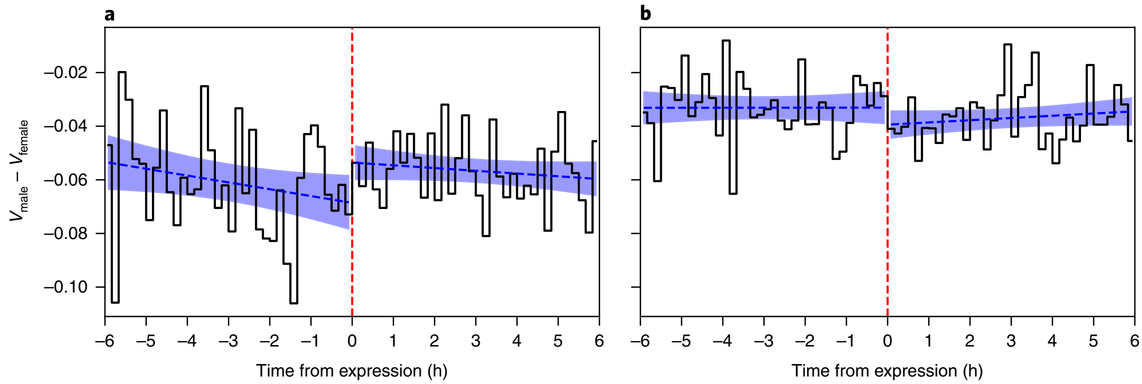

Fig. 7 | Regression discontinuity analysis of male and female time series in negative and positive affect labeling groups. a,**b**, Difference between male (v_{male}) and female (v_{female}) valence time series of mean valence values at 10 min intervals for positive (**a**) and negative valence (**b**). A regression discontinuity analysis applied to the resulting $v_{male} - v_{female}$ time series shows a divergence between male and female responses at time t_0 for both positive and negative affect labeling groups. Blue bands show the 95% CIs of the linear regression estimates.

At the time of an emotional statement (that is, t_0), we observe discontinuities for both positive and negative emotions, indicating that male and female valence diverge at the time of the expression. This pattern is most pronounced for the positive valence time series (see Fig. 7a). Before the emotional expression, the gender difference increases (the downward slope indicates that female valence increases more rapidly than male valence). This pattern is reversed immediately following the emotional expression, indicating that women experience a smaller reduction of positive emotions after the statement at time t_0 than men do.

As shown in Fig. 7b, the negative valence time series indicates a constant baseline valence difference between men and women before the expression, indicating that the difference between male and female valence does not increase or decrease before the expression of a negative emotion. However, we again observe a discontinuity at the affect labeling at time t_0, although the CIs overlap slightly and we cannot draw any firm conclusions about the statistical significance of the effect. Women seem to revert to baseline levels of negative valence emotions more slowly than men do after the affect labeling.

Discussion

Our analysis reveals significant minute-to-minute changes in valence levels before and after an individual performs an act of affect labeling on Twitter. We found that, for a majority of individuals, emotional intensity decreased rapidly after their explicit expression in an "I feel" statement. This was the case for positive and negative emotions, but the effect seemed to be strongest and most immediate for negative emotions.

We also empirically determined the parameters of how emotions changed over time because of the evolving traces they left in our individuals' language. Valence levels increased rapidly along an accelerating trajectory before the affect labeling, followed by a sharp reversal at the time of the affect labeling, and a fast return to previous baseline levels of valence afterwards. Our results indicate that emotions last approximately 1.5 h from onset to evanescence, with a decay half-life of about 11 and 5 min for positive and negative emotion, respectively.

We also found inconclusive but suggestive indications that the evolution of positive and negative emotions may differentiate by gender, and that they differentiate most strongly at the time of the affect labeling.

Our study extends the present literature on affect labeling. We did not merely observe the effects of affect labeling in attenuating induced emotions, but observed and modelled their dynamics over time at high temporal resolution, for a large sample of individuals and for naturally occurring emotions. We calculated the duration of these changes in valence values, their half-life and the trajectories along which they take place.

Our results also extend recent work[26,27] suggesting that sentiment analysis may reflect individuals' emotions as well as text sentiment. Assuming that the individuals in our sample correctly labelled their affect, our observations indicate that emotions may affect language hours before and after individuals choose to express them explicitly.

Our method has some limitations. We relied entirely on a post-hoc, data-driven analysis. Although this reduced the possibility of some forms of observer bias, it is nevertheless possible that our sample was biased by particular social media characteristics such as recruitment and interface design. In addition, since social media platforms provide a public outlet, social drivers may need to be accounted for in future research, as well as a comparative analysis of different languages, and possibly culture. However, our null model was designed to account for many such biases by sampling from the respective positive and negative subject groups separately and taking into account diurnal and circadian rhythms.

It is not clear to what degree social media provides a representative sample of the overall population, nor of the range of emotions that may naturally occur. Future investigations may require experiments that cross-validate measurements of emotions from large-scale social media data against ground truth obtained from biophysical measures, including those of facial musculature.

Although valence is one of the primary components of most models of human emotions, our sentiment analysis does not capture the rich spectrum of human emotions. Our computational indicators may therefore need to be expanded to provide a more complete picture of the dynamics of a variety of human emotions across multiple languages and cultures, including arousal, dominance, fear, amusement, calm and others that have recently been uncovered by an analysis of self-reports[28]. Here, we used an off-the-shelf sentiment analysis tool geared towards rating English text with respect to valence only, but alternative natural language-processing techniques could be designed to detect a larger variety of emotions[29].

Taken together, our results lay the foundations for the development of mechanistic models of individual as well as population-level emotional dynamics, which can provide new insights into the interaction between language and emotion and, finally, the processes through which individuals may self-regulate their emotional state.

Methods

Data and sample. We randomly chose approximately 710,000 Twitter user identification numbers and issued requests to Twitter's application programming interface to retrieve all past tweets by the given users (up to a maximum of the 3,200 most recent tweets at the time of collection). The harvesting took place in 2012; hence, the dates of our tweets range from 2006 (Twitter start date) to 2012 (time of collection). For every user, we obtained public profile information, and for every tweet, we retrieved a unique identifier, its content (140 characters) and the local coordinated universal time at which it was posted. All timelines were stored in a MongoDB database allowing fast and efficient programmatic access. Not all accounts queried were active or functional, so our final sample consisted of 665,081 individuals. In total, our database contained 1,150,447,758 tweets across all individual timelines.

As shown in Fig. 1, we then used a two-level approach. First, we identified a cohort of individuals who at some time explicitly and unambiguously reported having a positive or negative emotion through affect labeling. Second, we rated the text valence of the tweets that were submitted by these same individuals before and after their expression using the Valence Aware Dictionary and Sentiment Reasoner (VADER) sentiment analysis tool[30] to trace the evolution in language of the reported emotion over time.

Detecting affect labeling in tweets. We identified instances in which individuals explicitly and unambiguously reported their emotional state by searching for tweets that contained the statement "I feel…", including two grammatical variations: "I am feeling" or "I'm feeling"—an approach developed earlier to detect personal emotions through self-reports[29,31,32]. We assessed the statements' emotional valence (whether they were about having a positive or negative emotion) from the valence value of the adverb or adjective that followed "I feel" using the Center for Reading Research Affective Norms for English Word (ANEW)[33]—a lexicon of 13,916 valence-rated English terms. For example, "I feel so unhappy" was taken as an instance of negative affect labeling since "unhappy" has a strong negative valence rating in ANEW. Here, we focused on valence since it is a dominant component in most descriptive and nominal models of human emotions[34–40].

Note that we used the "I feel" expression (our affect labeling marker) only as a binary diagnostic, to determine whether or not the subject reported a positive or negative emotion at the specific time. The statement was neither used to analyse the emotion itself, nor to capture the fullest possible spectrum of human emotions. We considered only the clearest and most unambiguous indications that a positive or negative emotion did indeed take place.

To this end, we identified ten of the most frequent and explicit valence-related adjectives or adverbs in all "I feel" statements, ranked by their valence ratings in the Center for Reading Research ANEW lexicon[33] (see Supplementary Note 2). These were: 'good', 'happy', 'great' and 'awesome' (positive valence); and 'bad', 'unhappy', 'sad', 'terrible', 'horrible' and 'awful' (negative valence). We also included a few boosters ('so', 'very', and so on). The adjective 'unhappy' was included to balance the positive 'happy'.

This set of 10 adjectives represented 98.27% of all "I feel" statements ($n=112{,}873$) in our data (see Supplementary Fig. 1) that unambiguously pertained to high or low valence only (excluding those that did not, such as "I feel weird" or "I feel sick"). Adding more adjectives and adverbs would therefore only slightly increase our sample, but at the expense of reducing the validity of our observations (see Supplementary Note 2).

We then retrieved all tweets posted by individuals making the mentioned affect labeling statements (109,943 individuals). As mentioned, we used the valence of the adjective or adverb of the expression to separate our timelines into a positive affect labeling group ($n=42{,}627$) and a negative affect labeling group ($n=67{,}316$).

To avoid traces of other explicit or intentional emotional statements biasing our results, we excluded (1) the self-report tweet itself and (2) any timelines that contained more than 1 self-report statement within 48 h of each other. To increase the odds of each timeline in our dataset representing a single, independent positive or negative emotion, we removed: (1) 'retweets', since they did not necessarily reflect the subject's own emotional state; (2) individuals who posted more emotional expressions than 95% of individuals in our data; (3) tweets that were posted on days that were unusual in terms of their number of emotional expressions (that is, days with fewer "I feel" tweets than 5% or more "I feel" tweets than 95% of other days in the year); (4) tweets that contained "I feel", "I am feeling" or "I'm feeling", but no adjective or adverb with a known valence rating; and (5) timelines from users without time zone information, to avoid confusion with respect to circadian rhythms[41] and time of day. We also removed cases in which the same subject posted more than 1 emotional expression in the same 48 h period t_0-24 h to t_0+24 h. After these filters, 74,487 timelines remained, of which most (that is, 73,185 (98.25%)) were majority English.

Sentiment analysis of tweets posted before and after the affect labeling. For all individuals in both groups, we retrieved the tweets posted on their individual timelines 6 h before to 6 h after the time of the emotional expression (that is, a total of 12 h surrounding the expression). We centred all the resulting 12 h timelines on the time of the expression, referred to as t_0.

We used VADER—an accurate, open-source Twitter sentiment analysis algorithm introduced by Hutto and Gilbert[30]—to assign each individual tweet (again excluding the emotional statement itself) a numerical valence score between −1 (very negative) and +1 (very positive). VADER recognizes 7,516 terms of the most frequently used English words, each rated by 10 independent human raters assessing their valence value, and was designed to recognize negations, hedging, boosters, colloquial language, style, jargon and abbreviations that are commonly used on Twitter. It responds only to English; hence, our analysis automatically ruled out non-English users and content.

The F1 classification accuracy of VADER on Twitter datasets was found to be 0.96, which surprisingly exceeds that of individual human raters (0.84). A recent survey[42] also suggested that VADER produces the highest accuracy sentiment rankings for a dedicated Twitter dataset ($F1_{pos}=99.25$; $F1_{neg}=98.33$; Macro-F1 = 98.79; Coverage = 94.61) among 22 tools, surpassing commonly used tools such as Linguistic Inquiry and Word Count and SentiStrength.

It is important to stress that the performance of VADER has been vetted in the context of text sentiment analysis (that is, whether a short text evinces a positive or negative sentiment), but not its ability to measure the emotions of the subject(s) who produce the text. In fact, our results may shed light on the question of whether or not text sentiment analysis does reveal subject emotions. We also note that language has been shown to exhibit a 'positivity bias' because people prefer to use words with higher valence ratings[43] and the VADER sentiment analysis tool is not normed to operate on an absolute scale. As a result, a group of negative tweets can yield a numerically positive VADER rating. In our analysis, we did not rely on the absolute value or sign of the VADER rating, but the pattern of change of valence values over time relative to a null model.

Time series construction. Tweets were individually posted at irregular points in time within the mentioned 12 h period (that is, centred on the emotional expression at time t_0). To observe changes in emotional valence relative to t_0, we grouped tweets in discrete 1 min time windows according to the time ($t_0 \pm k$) when they were posted, from 6 h before t_0 to 6 h after t_0 (that is, −360 to +360 min), and we did so separately for both positive and negative groups.

Each 1 min window therefore contained a set of tweets that were posted at that given distance k from t_0, producing a distribution of VADER valence values. To be able to draw inferences about significant changes in valence levels over time, we obtained converging evidence through two distinct methods. First, we produced a time series of mean valence values for each 1 min window and applied a CUSUM[44] procedure that examined cumulative changes in the time series' variance to detect significant change points. Second, we bootstrapped mean valence values and compared the resulting distribution and its 95% CIs with those produced by a null model of a random sample of tweets with similar weekly, circadian and diurnal features to the tweets in the time window under consideration[41].

We thus obtained two sources of evidence with respect to significant changes in valence levels.

Null model. To determine whether we were observing significant changes in valence across our time series, we defined a null model with which we compared the observed valence values for each time window. To account for individual dispositions (individuals expressing negative emotions might be more negative overall than individuals expressing positive emotions), we randomly sampled tweets from the set of positive and negative timelines to construct separate positive and negative null models. Furthermore, we accounted for diurnal cycles and their effect on the timing of emotional expression by randomly sampling tweets from the 24 h surrounding t_0. We ensured that the random sample matched the time of day and week day (Monday, Tuesday, and so on) distribution of the observed sample, since some emotional expressions might be more common or biased at certain times of the day or on certain week days (for example, Monday morning versus Friday night). Since the volume of tweets changed in each time window (see Supplementary Fig. 2), we sampled the same number of random tweets in our null model as we observed for the specific time window. Therefore, as shown in Fig. 4, CI bands narrowed towards t_0, since the volume of tweets increased and the uncertainty of our mean valence estimates thus decreased. Finally, we computed mean valence and CIs for each time window of our null model, by repeating this sampling procedure 10,000 times (with replacement), allowing us to compute 5th, 50th and 95th percentiles for the resulting distribution of mean valence values.

Note that this procedure resulted in a null model that was more strict than an entirely random sample of all tweets and all individuals, since it required that any valence change observed in a given time window diverged significantly from one that was expected by chance for the same cohort of positive and negative individuals, the same number of tweets observed, the same time of day, the same week day, and within the same circadian and diurnal cycle as the tweets we observed in our timelines.

CUSUM test. We applied the CUSUM anomaly detection method on the $t_0 \pm 6$ h smoothed mean time series. First, we calculated the CUSUM chart, which contained the upper control limit series by $S_i^+ = \max(0, S_{i-1}^+ + x_i - (T+K))$ and the lower control limit series by $S_i^- = \min(0, S_{i-1}^- + x_i - (T-K))$, where x_i represents the series value at i. and T and K are the mean and s.d. of the corresponding null-model series, reflecting the expected mean and s.d. of the median score series. Then, points that satisfied $S_i^+ > H$ and $S_i^- < -H$ ($H=0.01$) were chosen as the upper and lower violations, respectively. Anomalies were detected by picking the increasing subsequences of the upper violations and decreasing subsequences of the lower violations. Sequences that were longer than $\lambda = 40$ min were determined to be periods of significantly elevated or attenuated valence levels, meaning that in these periods individuals' emotions were significantly higher or lower than chance levels.

Gender classifier. Men and women may differentiate with respect to how they experience and express emotions[45,46]. However, Twitter does not provide reliable gender information about its users, other than what they may sporadically self-report. Following ref. [47], we therefore used a binary classifier (random forest with 160 estimators)[48–50] that was trained on a set of Google Plus profiles with manually verified gender information (1,744 individuals, including 658 females and 1,086 males). The classifier achieved an acceptable accuracy of 86.4% and an area under the curve score of 0.916 (see Supplementary Note 7).

Separating our sample according to gender resulted in two timelines differentiated by emotional valence only (positive and negative; all individuals combined), as well as four timelines differentiated by gender (female and male) and emotional valence (negative and positive).

With respect to accuracy, note that incorrect gender predictions will have led to mixed samples and reduced the magnitude of any observed male–female differences. Hence, the gender effects we report are probably underestimations of the true effect, as our male group may contain females and vice versa.

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

Code availability. The code and related data of this study are freely available from https://github.com/onurvarol/Rui-Bollen_NHB allowing reproduction. Additional data and information are available from the authors upon reasonable request.

Data availability

The Twitter content data that support the findings of this study are publicly available from Twitter, but cannot be distributed by the authors. The authors provide the Twitter identification codes of all tweets used in this analysis to allow for retrieval of their content from the Twitter application programming interface. All other data are available from the authors upon reasonable request.

Received: 26 March 2018; Accepted: 6 November 2018;
Published online: 17 December 2018

References

1. Shariff, A. F. & Tracy, J. L. What are emotion expressions for? *Curr. Dir. Psychol. Sci.* **20**, 395–399 (2011).
2. Torre, J. B. & Lieberman., M. D. Putting feelings into words: affect labeling as implicit emotion regulation. *Emot. Rev.* **10**, 116–124 (2018).
3. Lieberman, M. D., Inagaki, T. K., Tabibnia, G. & Crockett, M. J. Subjective responses to emotional stimuli during labeling, reappraisal, and distraction. *Emotion* **11**, 468–480 (2011).
4. Constantinou, E., Van Den Houte, M., Bogaerts, K., Van Diest, I. & Van den Bergh, O. Can words heal? Using affect labeling to reduce the effects of unpleasant cues on symptom reporting. *Front. Psychol.* **5**, 807 (2014).
5. Taylor, S. F., Phan, K. L., Decker, L. R. & Liberzon, I. Subjective rating of emotionally salient stimuli modulates neural activity. *NeuroImage* **18**, 650–659 (2003).
6. Thomassin, K., Morelen, D. & Suveg, C. Motion reporting using electronic diaries reduces anxiety symptoms in girls with emotion dysregulation. *J. Contemp. Psychother.* **42**, 207–213 (2012).
7. Kircanski, K., Lieberman, M. D. & Craske, M. G. Feelings into words. *Psychol. Sci.* **23**, 1086–1091 (2012).
8. Niles, A. N., Craske, M. G., Lieberman, M. D. & Hur., C. Affect labeling enhances exposure effectiveness for public speaking anxiety. *Behav. Res. Ther.* **68**, 27–36 (2015).
9. Niles, A. N., Haltom, K. E. B., Lieberman, M. D., Hur, C. & Stanton, A. L. Writing content predicts benefit from written expressive disclosure: evidence for repeated exposure and self-affirmation. *Cogn. Emot.* **30**, 258–274 (2016).
10. Lieberman, M. D. et al. Putting feelings into words: affect labeling disrupts amygdala activity to affective stimuli. *Psychol. Sci.* **18**, 421–428 (2007).
11. Mauss, I. B. & Robinson, M. D. Measures of emotion: a review. *Cogn. Emot.* **23**, 209–237 (2009).
12. Kahneman, D. & Krueger, A. B. Developments in the measurement of subjective well-being. *J. Econ. Perspect.* **20**, 3–24 (2006).
13. Probst, T., Pryss, R., Langguth, B. & Schlee, W. Emotion dynamics and tinnitus: daily life data from the "trackyourtinnitus" application. *Sci. Rep.* **6**, 31166 (2016).
14. Phan, K. L., Wager, T., Taylor, S. F. & Liberzon, I. Functional neuroanatomy of emotion: a meta-analysis of emotion activation studies in PET and FMRI. *NeuroImage* **16**, 331–348 (2002).
15. Ochsner, K. N., Bunge, S. A., Gross, J. J. & Gabrieli., J. D. E. Rethinking feelings: an FMRI study of the cognitive regulation of emotion. *J. Cogn. Neurosci.* **14**, 1215–1229 (2002).
16. Fossati, P. et al. In search of the emotional self: an FMRI study using positive and negative emotional words. *Am. J. Psychiatry* **160**, 1938–1945 (2003).
17. Andreassi, J. L. *Psychophysiology: Human Behavior and Physiological Response* (Psychology Press, London, 2013).
18. Nummenmaa, L., Glerean, E., Hari, R. & Hietanen, J. K. Bodily maps of emotions. *Proc. Natl Acad. Sci. USA* **111**, 646–651 (2014).
19. McRae, K., Ochsner, K. N., Mauss, I. B., Gabrieli, J. J. D. & Gross, J. J. Gender differences in emotion regulation: an FMRI study of cognitive reappraisal. *Group Process. Intergroup Relat.* **11**, 143–162 (2008).
20. Koelsch, S., Fritz, T., Müller, K. & Friederici, A. D. Investigating emotion with music: an FMRI study. *Hum. Brain Mapp.* **27**, 239–250 (2006).
21. Prasad, D. K., Liu, S., Chen, S.-H. A. & Quek, C. Sentiment analysis using EEG activities for suicidology. *Expert Syst. Appl.* **103**, 206–217 (2018).
22. Pennebaker, J. W. *Emotion, Disclosure and Health* (American Psychological Association Books, Washington DC, 1995).
23. Kennedy-Moore, E. & Watson, J. C. How and when does emotional expression help? *Rev. Gen. Psychol.* **5**, 187–212 (2001).
24. Ford, B. Q., Lam, P., John, O. P. & Mauss, I. B. The psychological health benefits of accepting negative emotions and thoughts: aboratory, diary, and longitudinal evidence. *J. Pers. Soc. Psychol.* http://doi.org/10.1037/pspp0000157 (2017).
25. Thistlethwaite, D. L. & Campbell, D. T. Regression-discontinuity analysis: an alternative to the ex post facto experiment. *J. Educ. Psychol.* **51**, 309–317 (1960).
26. Beasley, A. & Mason, W. Emotional states vs. emotional words in social media. In *Proc. ACM Web Science Conference* 31 (ACM, 2015).
27. Ziemer, K. S. & Korkmaz, G. Using text to predict psychological and physical health: a comparison of human raters and computerized text analysis. *Comput. Hum. Behav.* **76**, 122–127 (2017).
28. Cowen, A. S. & Keltner, D. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proc. Natl Acad. Sci. USA* **114**, E7900–E7909 (2017).
29. Bollen, J., Mao, H. & Zeng, X. Twitter mood predicts the stock market. *J. Comput. Sci.* **2**, 1–8 (2011).
30. 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* 216–225 (AAAI, 2014).
31. Bollen, J., Mao, H. & Pepe, A. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In *Proc. Fifth International AAAI Conference on Weblogs and Social Media* 450–453 (AAAI, 2011).
32. Yang, C. & Srinivasan, P. Life satisfaction and the pursuit of happiness on twitter. *PLoS ONE* **11**, 1–30 (2016).
33. Warriner, A. B., Kuperman, V. & Brysbaert, M. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behav. Res. Methods* **45**, 1191–1207 (2013).
34. Darwin, C. *The Expression of the Emotions in Man and Animals* (John Murray, London, 1872).
35. Russell, J. A circumplex model of affect. *J. Pers. Soc. Psychol.* **39**, 1161–1178 (1980).
36. Russell, J. A. & Mehrabian, A. Evidence for a three-factor theory of emotions. *J. Res. Pers.* **11**, 273–294 (1977).
37. Mehrabian, A. *Basic Dimensions for a General Psychological Theory: Implications for Personality, Social, Environmental, and Developmental Studies* (Oelgeschlager, Gunn & Hain, Cambridge, 1980).
38. Plutchik, R. & Conte, H. R. *Circumplex Models of Personality and Emotions* (American Psychological Association, Washington DC, 1997).
39. Ekman, P. *Handbook of Cognition and Emotion* (eds Dalgleish, T. & Power, M.) Ch. 3 (John Wiley and Sons, Chichester, 1999).
40. Lang, P. J., Bradley, M. M. & Cuthbert, B. N. *International Affective Picture System (IAPS): Affective Ratings of Pictures and Instruction Manual* Technical Report A-8 (Univ. Florida, 2008).
41. Golder, S. A. & Macy, M. W. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science* **333**, 1878–1881 (2011).
42. 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).
43. Dodds, P. S. et al. Human language reveals a universal positivity bias. *Proc. Natl Acad. Sci. USA* **112**, 2389–2394 (2015).
44. Grigg, O. A., Farewell, V. T. & Spiegelhalter, D. J. Use of risk-adjusted CUSUM and RSPRT charts for monitoring in medical contexts. *Stat. Methods Med. Res.* **12**, 147–170 (2003).
45. Kring, A. M. & Gordon, A. H. Sex differences in emotion: expression, experience, and physiology. *J. Pers. Soc. Psychol.* **74**, 686–703 (1998).
46. McDuff, D., Kodra, E., Kaliouby, Rel & LaFrance, M. A large-scale analysis of sex differences in facial expressions. *PLoS ONE* **12**, 1–11 (2017).
47. Li, J., Ritter, A. & Hovy, E. Weakly supervised user profile extraction from Twitter. In *Proc. 52nd Annual Meeting of the Association for Computational Linguistics* 165–174 (ACL, 2014).

48. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
49. Niculescu-Mizil, A. & Caruana, R. Predicting good probabilities with supervised learning. In *Proc. 22nd International Conference on Machine Learning* 625–632 (ACM, 2005).
50. Pennington, J., Socher, R. & Manning, C. D. Glove: global vectors for word representation. In *Proc. 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* 1532–1543 (ACL, 2014).

Acknowledgements
R.F. acknowledges the support from the NSFC (grant number 71501005). J.B. thanks the Defense Advanced Research Projects Agency (NGS2 2016 D17AC00005), the National Science Foundation (SMA-SBE: 1636636), the Economic Development Administration (EDA/ED17HDQ3120040), Wageningen University (the Netherlands) and the ISI Foundation (Turin, Italy) for support. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. Special thanks go to F. Radicchi and C. Scheffer for insightful comments, and R. B. Correia, who kindly set up our Twitter timeline database based on B. Gonçalves' Twitter data collection.

Author contributions
R.F. and J.B. defined the research methodology, which O.V., I.A.v.d.L. and M.S. helped design. R.F. and J.B. collected the data. R.F., O.V. and J.B. conducted the analysis. A.V. designed and implemented the gender classifier. A.B. provided statistical advice. I.A.v.d.L. and M.S. conducted a literature review. R.F., O.V., I.A.v.d.L., M.S. and J.B. interpreted the results. R.F., O.V., A.B. and J.B. co-authored the manuscript text.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41562-018-0490-5.

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

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

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 AND FAN RUI

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 Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. 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
☐	☒	An indication of 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 statistics 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
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	Python 2.6
Data analysis	Python 2.7, Python modules: numpy, matplotlib, json, pandas, scipy, statsmodels, lmfit, sklearn (all open-source)

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/reviewers upon request. 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

The Twitter content data that support the findings of this study are publicly available from Twitter, but cannot be distributed by the authors. The authors will however provide the Twitter identification codes of all tweets used in this analysis to allow for retrieval of their content from the Twitter API. All other data are available from the authors upon reasonable request.

Field-specific reporting

Please select 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/authors/policies/ReportingSummary-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: a longitudinal analysis of the changing valence levels of Tweets posted by Twitter users shortly before and after they labeled their affect using existing time-coded, individual Twitter text data.
Research sample	A random selection of 710,000 Twitter users who self-selected through their participation in the Twitter social networking platform.
Sampling strategy	Sampling procedure: random, sample size of 710,000 individuals was based on availability. Sample size was chosen to be largest by 3 orders of magnitude vs existing affect labeling studies.
Data collection	Python 2.6 and Twitter API (https://developer.twitter.com/)
Timing	2012 (timeline harvest from Twitter API): resulting sample of Tweets ranges from 2006 to 2012.
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	Participants were not allocated into experimental groups.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants

Methods

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