nature human behaviour

ARTICLES

https://doi.org/10.1038/s41562-021-01098-5

Mass-scale emotionality reveals human behaviour and marketplace success

Matthew D. Rocklage[1]✉, Derek D. Rucker[2] and Loran F. Nordgren[2]

Online reviews promise to provide people with immediate access to the wisdom of the crowds. Yet, half of all reviews on Amazon and Yelp provide the most positive rating possible, despite human behaviour being substantially more varied in nature. We term the challenge of discerning success within this sea of positive ratings the 'positivity problem'. Positivity, however, is only one facet of individuals' opinions. We propose that one solution to the positivity problem lies with the emotionality of people's opinions. Using computational linguistics, we predict the box office revenue of nearly 2,400 movies, sales of 1.6 million books, new brand followers across two years of Super Bowl commercials, and real-world reservations at over 1,000 restaurants. Whereas star ratings are an unreliable predictor of success, emotionality from the very same reviews offers a consistent diagnostic signal. More emotional language was associated with more subsequent success.

People have always looked to and relied on the opinions of those around them to make decisions[1,2]. Now, the rise and proliferation of online crowd-sourced platforms, such as Yelp and Glassdoor, have fundamentally transformed the scope and speed with which people can harness others' assessments. Given their scale, openness and availability, these platforms promise to facilitate people's ability to find the best option[3,4]. Indeed, rather than rely on trial and error or small, informal networks, people have immediate access to the experience and wisdom of crowds. In the case of movies and restaurants, for instance, this aggregated wisdom should help quickly identify success—those items that have thrived and become popular. For most platforms, the primary means to identify successful goods is through an aggregated 'star rating'—a numeric rating that measures the extent to which people's opinions are relatively positive versus negative.

Perhaps surprisingly, a striking limitation of these online rating systems has emerged: reviews are overwhelmingly positive[5]. On Amazon.com, for example, the average star rating is approximately 4.2 out of 5, with well over half of the reviews being 5-star ratings[6,7]. Nearly half of all Yelp reviews are 5-star ratings[8], and recent research indicates that nearly 90% of Uber ratings may be 5 stars[9]. A visual representation of most online ratings reveals a J-shaped distribution, with many 4- and 5-star ratings, a few 1-star ratings and few ratings in between[5]. The degree of overwhelming positivity suggests that individuals are often confronted with choosing between numerous items with similar star ratings, especially given that people will not even consider options that garner less than a 3-star rating.

A principal problem with this degree of positivity is that the ratings themselves may ultimately be an unreliable indicator of the success of that item and the human behaviour that underlies this success (for example, restaurant reservations). Specifically, two items might receive nearly identical ratings but vary vastly in their success. Indeed, past research has shown substantial variability in the link between the positivity of individuals' ratings and success[10–12]. For example, the positivity of online ratings shows little association with the underlying quality of products and fails to predict their resale value[13]. Moreover, an analysis of over 400 movies revealed that greater positivity in online ratings was associated with fewer people attending a movie, as evidenced by lower box office revenue[14]. This problem has even led companies such as Netflix to abandon standard rating systems due to their poor performance[15]. Put simply, these ratings seem not to hold the wisdom that people believe they do.

Across disciplines, behavioural scientists are beginning to recognize the problematic nature of these ratings. That is, given this large degree of positivity, a number of cases exist where items receive a similarly positive rating. Yet, when it comes to human behaviour, substantial differences exist—not all 5-star restaurants are equally popular. The high degree of positivity effectively makes the ratings ineffective signals for discriminating what are likely to be the best or most successful options. We label this challenge to discern success within the mass of positive ratings the 'positivity problem'.

Although quantitative ratings are the most salient and accessible output of online reviews, most crowd-sourced platforms include a written portion where people provide qualitative assessments. As technology has improved, researchers have embraced computational social science techniques to quantify these qualitative assessments. Perhaps the most common method to analyse text in this way is via sentiment analysis, which most often quantifies language in terms of its positivity[16]. Some words suggest greater favourability (for example, the word 'liked'), whereas others suggest greater negativity (for example, 'disliked').

Computational social science has focused primarily on the positivity (also known as valence) of people's attitudes[17]. Relatively few efforts have sought to quantify aspects of individuals' attitudes beyond positivity[18]. Nevertheless, social psychologists have long acknowledged that positivity is not always a reliable predictor of behaviour[17,19]. To address the limitations of positivity, scholars have introduced and explored additional facets of an attitude that can improve its predictive ability[17,20]. One such facet is the emotionality of an attitude—the extent to which an attitude is based on individuals' feelings or emotional reactions[21–24]. Positivity and emotionality are conceptually and empirically distinct. For example, the words 'enjoyable' and 'impeccable' imply very similar levels of positivity, but research indicates that the word 'enjoyable' is likely to be indicative of a more emotional attitude than the word 'impeccable'[25].

[1]College of Management, University of Massachusetts, Boston, MA, USA. [2]Kellogg School of Management, Northwestern University, Evanston, IL, USA. ✉e-mail: matthew.rocklage@umb.edu

Moreover, the emotionality of individuals' attitudes can now be captured via text analysis[25,26].

Attitudes based more on emotion tend to be stronger and more predictive of behaviour. In the political domain, voters' emotional reactions to a political candidate—compared with their more cognitive reactions—were better predictors of future voting behaviour[27]. Attitudes based more on emotion also tend to come to mind more quickly[28], are more extreme[25,26] and are more consistent across contexts[29] and time[30]. One reason for this relationship is that emotions provide individuals themselves with an indication that something especially impactful has occurred[31,32], and they can thereby act as a particularly clear signal to individuals regarding their own attitude[28,33,34]. This strong signal, in turn, can lead attitudes to be held more strongly in memory[28], which is an established predictor of the impact and durability of an attitude[17,35].

Outward displays of emotion also signal the importance of one's attitude to others. The social–functional approach to emotion puts forth that a primary function of emotion is to communicate the strength of one's attitudes, desires and intentions[36–38]. As social animals, understanding others' goals and intentions is vital for successful social coordination. Displays of joy and anger, for instance, provide others with strong signals regarding a person's state of mind, goals and priorities. In the context of negotiation, expressions of happiness signal that one is open to concession, whereas displays of anger signal that one is unlikely to compromise[39,40]. These findings indicate that when humans use emotion online, it is probably a signal that an experience was particularly impactful to them.

Taken together, research suggests that attitudes based on emotion are stronger and more predictive of one's own behaviour, and that people use emotion to communicate the impact of their experience to others. The consequence is that emotionality in text may be more indicative of the success of a product or service. To illustrate, consider a restaurant. From an attitudinal perspective, the ability of a restaurant to elicit a positive, emotional, feelings-based reaction is likely to lead to a more strongly held attitude in the individual. This stronger attitude could lead that restaurant to come to mind more frequently in the future and lead the individual to be more likely to visit again. From a social–functional perspective, individuals' emotional reactions may also signal to others just how impactful an experience was and thereby generate more attention for that restaurant from others. Thus, for both these reasons, more emotional language may be able to predict success where star ratings cannot.

In short, we argue that capturing the emotionality expressed in online reviews may offer one solution to the positivity problem. More specifically, we hypothesize that the emotionality of people's online reviews can predict success and the mass-scale human behaviour that underlies this success where aggregated online ratings do not. In providing evidence of both the positivity problem and the relationship between emotionality and mass-scale human behaviour across multiple domains, we aim to accomplish two objectives. First, we demonstrate the breadth of the positivity problem. Second, we offer one solution to this problem using a theory-based approach. In doing so, this work also advances our understanding of emotionality—a construct considered of great importance across the social sciences[32]—by revealing that it has the ability to predict mass-scale behaviour and marketplace success.

Results

Study 1. In Study 1, we predicted human behaviour and success in the movie industry in the form of box office revenue earned in the United States. We obtained all online reviews for all movies from Metacritic.com from 2005 to 2018—13 years of data—and used the first 30 user reviews written for each movie to measure the movie's star rating (0 to 10 stars) and text emotionality. We also measured the valence (that is, positivity) of the text to assess the unique contribution of emotionality. We selected the first 30 reviews for two reasons. First, using the first reviews written for a movie helped avoid a situation where the success of the movie is already known by reviewers, which can influence how individuals write about the movie[41]. Second, this approach helped ensure that reviewers were expressing their own opinions as opposed to echoing the consensus viewpoint of others. Prior work indicates that early reviews can systematically bias subsequent posting behaviour both in the real world and in well-controlled laboratory experiments[42,43]. By using early reviews, we sought to avoid these influences. Moreover, we used this same number of reviews consistently in all applicable studies. These results were also robust when using an alternative number of reviews (that is, the first 40 reviews) and when using all possible reviews (see the Supplementary Results for Study 1).

Across all studies, we used the Evaluative Lexicon (www.evaluativelexicon.com) to quantify the average valence and emotionality expressed[25,26]. Specifically, the Evaluative Lexicon measures the valence and emotionality implied by the words that individuals use (for example, 'amazing' or 'enjoyable'). It has been directly validated as a measure of the valence and emotionality of individuals' attitudes with both well-controlled laboratory experiments and real-world naturalistic text[25,26]. While past work using the Evaluative Lexicon has focused on the relationship between emotionality and star ratings[25,26], that work did not examine emotionality's unique relation with mass-scale behaviour, above and beyond emotionality's connection with star ratings. As overviewed earlier, while there is a relationship between emotionality and individuals' positivity, these are separable constructs. Unless noted otherwise, all results across studies utilize multiple regression with standardized coefficients (B), log-transformed dependent variables and two-tailed significance tests.

As evidence for the large number of positive ratings on this platform, 81% of movies were rated positively (that is, they received an average star rating above the midpoint of 5 stars). Given that our aim is to predict success and human behaviour within a sea of positive reviews, our analyses examined whether emotionality was predictive of box office revenue for movies that were judged as positive (those rated above 5 stars on average). There were 2,383 movies.

We first assessed whether the movie's average star rating was predictive of its box office revenue. A movie's star rating was predictive of a movie making less box office revenue ($B=-0.08$; $t(2{,}381)=3.24$; $P=0.001$; 95% confidence interval (CI), (−0.136, −0.033)). When all movies were included—even those with an initial negative rating—star ratings were not significantly predictive of box office revenue ($B=0.004$; $t(2{,}931)=0.15$; $P=0.88$; 95% CI, (−0.043, 0.050)).

We then added the average emotionality of the reviews' text to this same model and the average text valence as a control. Star ratings continued to be a significant negative predictor of the movie's box office revenue ($B=-0.13$; $t(2{,}379)=3.86$; $P<0.001$; 95% CI, (−0.193, −0.063); Fig. 1, left panel), and text valence was in the positive direction but ultimately non-significant ($B=0.06$; $t(2{,}379)=1.78$; $P=0.07$; 95% CI, (−0.006, 0.124)). Of the greatest importance, beyond these effects, emotionality was a significant positive predictor of future box office revenue ($B=0.08$; $t(2{,}379)=3.01$; $P=0.003$; 95% CI, (0.027, 0.130); Fig. 1, right panel).

These results hold when controlling for (1) movie genre, (2) the year the movie was released, (3) the length of the movie, (4) the budget of the movie and (5) the arousal implied by the text as measured by the word list in Warriner et al.[18]. Regarding the arousal of the text, although arousal and emotionality are related, arousal refers to energy level, whereas the emotionality of an attitude is the extent to which that attitude is based on emotions or feelings[25,44]. Emotionality can be high or low in arousal. For example, the adjectives 'exciting' and 'lovable' imply similar levels of emotionality but higher or lower levels of arousal, respectively. Research has shown

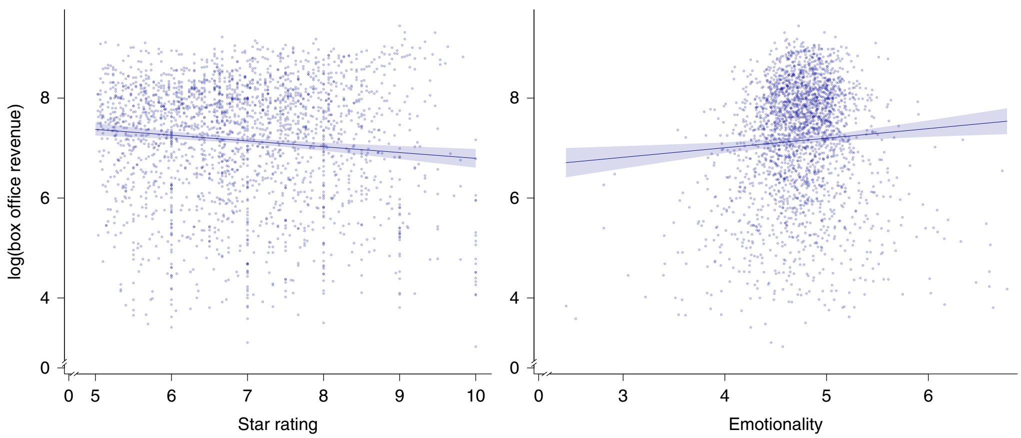

Fig. 1 | Predicting movie revenue. Scatter plots, best-fit lines and 95% CIs predicting each movie's total US box office revenue (US dollars, log transformed) from Metacritic star ratings (left) and emotionality (right; possible range: 0 to 9). The scatter points are the raw data and thus not adjusted for covariates.

that emotionality and arousal are separable in online reviews[25]. Emotionality is thus a measure of whether a movie was able to elicit a feeling or emotional reaction (for example, a movie as 'inspirational', 'enchanting' or 'adorable') rather than how 'exciting' that movie was.

To summarize, whereas the effects of star rating were inconsistent across these models, emotionality was a consistent positive predictor of box office revenue (Supplementary Table 2). Finally, emotionality was also a significant predictor when not controlling for any additional variables ($B = 0.07$; $t(2{,}381) = 2.72$; $P = 0.007$; 95% CI, (0.020, 0.122); see the Supplementary Results for Study 1 for the details of all robustness analyses).

Study 2. In Study 2, we generalized these results to a new domain. Specifically, we predicted the success of all books on Amazon.com from 1995 to 2015 (20 years of data). We again used the first 30 reviews for each book to index the book's star rating (1 to 5 stars), text valence and text emotionality. The results that follow also hold when using an alternative cut-off (that is, the first 40 reviews) and when using all possible reviews (see the Supplementary Results for Study 2). We measured the success of each book on the basis of the number of verified purchases it accrued over time.

A full 91% of the books received a positive rating by falling above the midpoint of the star rating scale (3 stars). There were 1.6 million positively rated books.

The regression results with average star ratings were mixed. Aggregated ratings were a negative predictor of the number of book purchases ($B = -0.047$; $t(1{,}576{,}840) = 164.60$; $P < 0.001$; 95% CI, (−0.047, −0.046)). When books rated as negative were also included, positive star ratings were significantly predictive of more purchases ($B = 0.015$; $t(1{,}727{,}821) = 57.54$; $P < 0.001$; 95% CI, (0.015, 0.016)). However, the overall evidence here was mixed, as star ratings were non-significant or negative predictors in 1/3 of book genres (Supplementary Table 4).

Analysing positive books, we then predicted the book's purchases on the basis of that book's average star rating and emotionality. As in Study 1, we included text valence as a control. The average star rating was a negative predictor of purchases ($B = -0.057$; $t(1{,}576{,}838) = 189.25$; $P < 0.001$; 95% CI, (−0.058, −0.057)), and the valence of the text was a significant positive predictor ($B = 0.024$; $t(1{,}576{,}838) = 78.28$; $P < 0.001$; 95% CI, (0.024, 0.025)). Beyond these effects, greater emotionality of the first 30 reviews predicted greater purchases ($B = 0.017$; $t(1{,}576{,}838) = 56.47$; $P < 0.001$; 95% CI, (0.016, 0.017)). Moreover, greater emotionality was predictive of more book purchases in 93% of genres.

We also conducted robustness analyses controlling for (1) book genre, (2) the year the book was released and (3) the arousal implied by the review text. All primary results replicated (Supplementary Table 5). Finally, emotionality was also a significant predictor when not controlling for any additional variables ($B = 0.016$; $t(1{,}576{,}840) = 54.87$; $P < 0.001$; 95% CI, (0.015, 0.016); see the Supplementary Results for Study 2 for the details of all robustness analyses).

Study 3. Study 3 examined whether the emotionality of real-time tweets in response to television commercials predicted success and human behaviour in the form of daily new followers of a brand. For both the 2016 and 2017 Super Bowls, we obtained all real-time tweets that occurred on the day of that Super Bowl that referenced a commercial shown during the Super Bowl. There were 94 commercials across 84 businesses and a total of 187,206 tweets about these commercials. We then used the Evaluative Lexicon to quantify the average valence and emotionality expressed towards each commercial across the tweets.

For the ratings of each commercial, we used the results from *USA Today*'s Ad Meter survey, which is the most popular set of Super Bowl ratings[45]. The Ad Meter survey specified to respondents that ratings between 1 and 3 indicate a 'poor' commercial, between 4 and 7 a 'good' commercial, and between 8 and 10 an 'excellent' commercial. Though the final number of survey participants is not disclosed by *USA Today*, they indicate the panel to be in the thousands[46].

We predicted the average number of daily new followers each company obtained on Facebook in the two weeks after the Super Bowl. This number of new followers reflects the number of individuals who became interested in learning more about company and its general offerings and took active steps to interact with that company. Because each company has only a single Facebook page, we aggregated the Twitter and ratings data at the level of each company by averaging across that company's commercials for each Super Bowl ($n = 84$). Given that our analysis emphasized the change in new followers that a company accrued after the Super Bowl, we controlled for the average number of daily new followers each company gained prior to the Super Bowl (see the Supplementary Methods for Study 3 for additional details).

The *USA Today* scale explicitly specifies 'good' commercials as those above 3 on the scale. Thus, unlike the rating scales in Studies 1 and 2 where we counted a movie or book as positively rated if it fell above the midpoint of the scale, using the midpoint of the *USA Today* scale would not capture all of the positive commercials. We therefore included commercials that earned a 'good' rating or higher

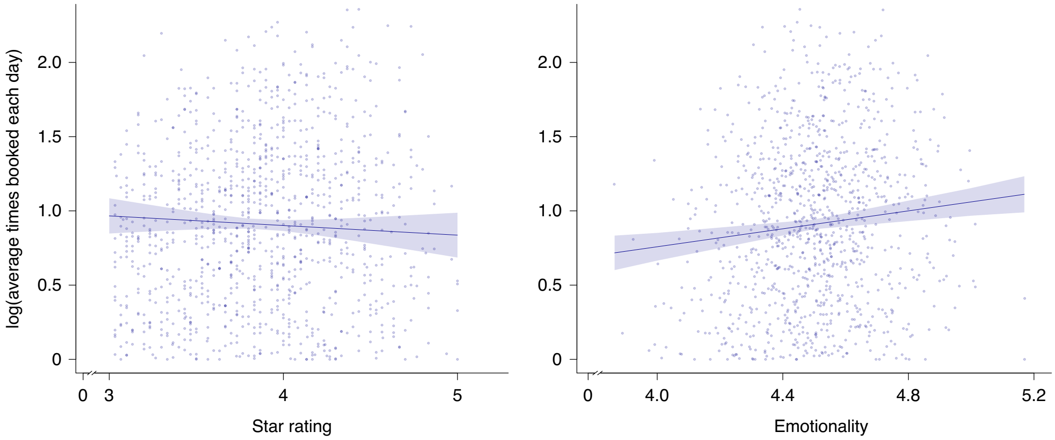

Fig. 2 | Predicting restaurant table reservations. Scatter plots, best-fit lines and 95% CIs predicting each restaurant's table reservations from Yelp star ratings (left) and emotionality (right). The scatter points are the raw data and thus not adjusted for covariates.

(that is, above 3). In fact, 100% of commercials were rated as 'good' or higher across both Super Bowls. Thus, we used all observations.

We again began with a regression model that included each commercial's average *USA Today* rating to predict the average daily new Facebook followers that a company gained in the two weeks after the Super Bowl. We additionally controlled for the average daily new Facebook followers (log transformed) that the company gained prior to the Super Bowl to assess change. The number of followers that a company accrued before the Super Bowl predicted the followers they accrued after the Super Bowl ($B=0.15$; $t(81)=14.57$; $P<0.001$; 95% CI, (0.131, 0.171)), but the *USA Today* rating was not predictive of followers ($B=0.01$; $t(81)=1.39$; $P=0.17$; 95% CI, (−0.006, 0.033)).

We then added the average emotionality of the tweets for each commercial as our primary predictor and the average valence as a control. The average *USA Today* rating ($B=0.02$; $t(79)=1.59$; $P=0.12$; 95% CI, (−0.004, 0.039)) and valence of the tweets were not predictive of the number of new followers ($B=-0.02$; $t(79)=1.49$; $P=0.14$; 95% CI, (−0.039, 0.005)). However, beyond these effects, the greater the emotionality of the tweets about a commercial, the more Facebook followers a company accrued over the next two weeks ($B=0.02$; $t(79)=2.38$; $P=0.02$; 95% CI, (0.004, 0.042)).

Past research has indicated that the relative number of positive versus negative tweets can be predictive of different outcomes[47,48]. We therefore also included this metric as a test of the robustness of the effects. Conceptually replicating previous research, the greater the number of positive (minus negative) tweets a commercial received, the more followers the company gained ($B=0.03$; $t(78)=2.62$; $P=0.01$; 95% CI, (0.007, 0.051)). As before, the *USA Today* rating was not predictive ($B=0.01$; $t(78)=0.53$; $P=0.59$; 95% CI, (−0.016, 0.028)), and the average valence of the tweets became a negative predictor of new followers ($B=-0.02$; $t(78)=2.08$; $P=0.04$; 95% CI, (−0.045, −0.001)). Beyond these effects, greater emotionality once again predicted a greater number of new followers ($B=0.02$; $t(78)=2.66$; $P=0.009$; 95% CI, (0.007, 0.043)).

In additional robustness analyses, we controlled for (1) the number of commercials a company showed, (2) the quarter in the game when the commercial was advertised and (3) the arousal implied by the tweets. All effects were similar (Supplementary Table 7). Moreover, the effects were consistent across both Super Bowls. Emotionality was also a significant predictor when controlling only for the average daily new Facebook followers each company gained prior to the Super Bowl ($B=0.02$; $t(81)=2.45$; $P=0.016$; 95% CI, (0.005, 0.042); see the Supplementary Results for Study 3 for the details of all robustness analyses).

Study 4. In Study 4, we examined success and human behaviour in the form of table reservations for restaurants on the basis of the first 30 Yelp.com reviews for all restaurants that existed in Chicago, Illinois, as of 2017. We used these reviews to index each restaurant's average star rating (1 to 5 stars), text valence and text emotionality. The results also hold when using an alternative number of reviews (that is, the first 40 reviews) and when using all possible reviews (see the Supplementary Results for Study 4). We examined the average daily table reservations across a two-month period on OpenTable.com—the most popular online table reservation service in the United States. Across this two-month period, there were 1.30 million table reservations (see the Supplementary Methods for Study 4 for additional details).

On Yelp, restaurants are rated on a 5-point star rating scale. As evidence for the large number of positive reviews, 92% of restaurants received an average star rating that was above the midpoint of 3 stars. We used the restaurants falling above this midpoint. There were 1,052 restaurants.

Unlike prior studies, the average star rating was predictive of more table reservations ($B=0.05$; $t(1{,}050)=3.06$; $P=0.002$; 95% CI, (0.019, 0.085)). This outcome was the same when including even negatively rated restaurants ($B=0.08$; $t(1137)=4.97$; $P<0.001$; 95% CI, (0.049, 0.112); see the Supplementary Results for Study 4). This positive predictive effect of star ratings allows us to examine whether emotionality continues to be a unique predictor even when ratings are initially in the positive direction.

We then added the average emotionality of the restaurant's first 30 reviews as well as the average valence to the model. The average star rating fell to non-significance ($B=-0.03$; $t(1{,}048)=0.97$; $P=0.33$; 95% CI, (−0.089, 0.030); Fig. 2, left panel), and text valence was a positive predictor ($B=0.08$; $t(1{,}048)=2.76$; $P=0.006$; 95% CI, (0.024, 0.143)). Beyond these effects, restaurants that elicited more emotion were associated with more table reservations ($B=0.06$; $t(1{,}048)=3.39$; $P<0.001$; 95% CI, (0.025, 0.092); Fig. 2, right panel).

We conducted additional analyses to assess the robustness of our findings. Specifically, we controlled for (1) how well-established the restaurant is as indexed by the relative number of years the restaurant has been open, (2) the neighbourhood where the restaurant is located, (3) the cuisine of the restaurant (for example, American, Indian or seafood), (4) the average price of a meal at the restaurant and (5) the arousal of the text. Again, an individual can use words that convey an emotional attitude (for example, describing a restaurant and its food as 'enjoyable', 'comforting' or 'alluring'), independent of whether it fosters high or low arousal in that individual. We found that, across these analyses, emotionality was a significant predictor, whereas the star rating was not (Supplementary Table 9). Finally, emotionality was again a significant predictor when not controlling for additional variables ($B = 0.07$; $t(1,050) = 4.12$; $P < 0.001$; 95% CI, (0.036, 0.102); see the Supplementary Results for Study 4 for the details of all robustness analyses).

Discussion

Across four large-scale studies, we demonstrate that anywhere from 80% to 100% of ratings were positive. The challenge of discerning success and how people will behave in this sea of positive ratings is what we term the positivity problem.

Reflecting this problem, the current research indicates that movies, books, commercials and restaurants that receive similar ratings often do not have similar levels of success. Throughout our studies, online ratings tended to provide an unreliable signal of behaviour towards, and thus success of, a large range of items. As one solution to this problem, we examined whether emotionality assessed on a massive scale using computational linguistics provided a more diagnostic signal. We found that emotionality predicted behaviour across diverse items and several distinct sources—from Metacritic, Amazon, Twitter, Yelp, Facebook and OpenTable.

This work has implications for work on online ratings and discerning the aggregated wisdom from these ratings. In line with past research, the current work further calls into question the utility of star ratings for assessing and understanding human behaviour and ultimately success. Research has indicated that the predictive ability of star ratings is at best variable[10–12] and at worst not at all or even negatively predictive of behaviour and success[14]. In the current work, we demonstrate similar outcomes: increasingly positive ratings were commonly non-diagnostic of success. Moreover, we demonstrate these outcomes across a wide range of items and online platforms. As we show, one solution to this problem is for people and organizations to pay greater attention to the emotionality of individuals' attitudes. One possibility is that organizations could consider aggregating reviewers' language and providing an 'emotional star rating' to provide more meaningful assessments to individuals. Future research could explore whether star ratings can be fruitfully replaced with other, more predictive metrics.

The aim of this research is to demonstrate the positivity problem and the predictive ability of emotionality as one solution. As such, one limitation to the current work is that we did not identify the mechanism behind emotionality's predictive ability. This research thus provides a springboard for future work where researchers can delve further into illuminating the paths through which emotionality is able to predict human behaviour. As noted earlier, attitudes based more on emotion tend to be stronger and more consistent across contexts and time[27,29,30,49]. One reason for these outcomes is that these attitudes tend to be stored more strongly in memory[28]. Stronger links in memory predict what individuals think about and what captures their attention in their environment, thereby providing a general guide for behaviour[17,35,50]. Thus, when individuals consider which restaurant to frequent, website to visit or movie to see again, attitudes based more on emotion are less likely to have changed, more likely to come to mind and consequently more likely to guide behaviour.

Additional work could explore whether attitudes based more on emotion also affect success by increasing individuals' propensity to spread information via word of mouth. This may happen either spontaneously or when individuals are directly asked for recommendations. In the former case, attitudes based on emotion may come to mind with relatively little prodding and lead individuals to spontaneously think of and talk to others about an item. In the latter case, when asked for a recommendation, individuals may think of and recommend an emotion-evoking item first, given its stronger link in memory. In line with this possibility, prior research indicates that emotion-evoking news articles are generally more likely to be shared with others[51]. Future research could explore this potential implication of attitudes based on emotion.

We show that emotionality offers one means to solve the positivity problem, but if maximizing predictive accuracy is one's final goal, a second limitation of this work is that we did not maximize predictive ability, and other solutions are possible. For example, one approach would be to use machine learning to predict success in an effort to maximize accuracy. However, the present approach benefits from offering a theory-based solution to the positivity problem. Indeed, machine learning is powerful in its predictive ability but often does not provide a clear understanding of the underlying constructs that help provide this accuracy[52]. We show that emotionality, considered of great importance across the behavioural sciences, is predictive. In doing so, we also provide a conceptual advance to the study of emotion itself. We show that mass-scale emotion can predict behaviour and marketplace success.

Whereas most past work on sentiment analysis has focused on valence, the current work builds on theorizing and empirical findings in the attitudes and affective science literatures to put forth emotionality as a unique diagnostic signal. Though the words 'enjoyable' and 'impeccable' indicate similar levels of positivity (valence), they signal higher or lower levels of emotionality, respectively. Through the current research, it is our hope to urge researchers to assess factors outside of valence in the endeavour to understand mass-scale sentiment and to use it to address issues such as the positivity problem.

Methods

Study 1. We obtained all of the online user reviews for all movies from Metacritic.com from 2005 to 2018 using Python v.2.7 (ref. [53]) in consultation with the site owners regarding the use of the data. We began with movies released in 2005 because this was the first year in which there was a meaningful number of user reviews on the platform.

We used the first 30 reviews for each movie to measure the movie's star rating (0 to 10 stars), text valence and text emotionality. We quantified text valence and emotionality using the Evaluative Lexicon[25]. Some movies garnered fewer than 30 reviews, so we used the maximum number of reviews possible for these movies. As a robustness analysis, we controlled for the number of initial reviews for each movie, and the results replicate. The results also replicate when focusing only on those movies that garnered at least 30 reviews.

We measured the success of movies using the box office revenue for each movie (total United States box office revenue). See the Supplementary Results for Study 1 for more detail.

Study 2. We obtained all book reviews from Amazon.com from its beginning in 1995 until 2015 and used those books that had an identified genre. These reviews are publicly available for download[54,55]. We used the first 30 reviews for each book to measure the book's star rating (1 to 5 stars), text valence and text emotionality. We quantified text valence and emotionality using the Evaluative Lexicon.

We measured the success of each book by the number of verified purchases that book had. See the Supplementary Results for Study 2 for more detail.

Study 3. We obtained all the tweets associated with Super Bowl commercials from both the 2016 and 2017 Super Bowls using Python v.2.7 and in line with the terms of use. We used tweets that occurred in real time on the day of each Super Bowl, that mentioned the name of the company or an affiliated keyword, and that referenced either the Super Bowl or a commercial. This helped ensure that the tweets were about the target commercials (see the Supplementary Methods for Study 3 for additional detail).

Given that Facebook did not provide easy access to long-term historical data for companies' Facebook pages, we began to collect the number of followers from each company's Facebook page in real time as soon as that company announced it would be advertising during the Super Bowl. This was done manually and in line with the terms of use. We used the Facebook page that corresponded to the most salient brand or company advertised in each commercial. As the Super Bowl is primarily viewed by those in the United States, we used the Facebook page specifically affiliated with the United States (for example, mercedesbenzusa) as opposed to its worldwide Facebook page (for example, mercedesbenz). We obtained an average of 21.85 days of daily new followers for each company before the 2016 Super Bowl (s.d. = 7.83) and 16.05 days for the 2017 Super Bowl (s.d. = 10.73). Capturing these pre–Super Bowl data was imperative to assess the change in the average number of followers for each company after the Super Bowl.

We then continued to extract the daily number of new followers for each company for the two weeks after each Super Bowl. This average number of daily new followers over these two weeks served as the dependent variable. See the Supplementary Methods and Supplementary Results for Study 3 for more detail.

Study 4. We obtained all reviews on Yelp.com for all restaurants in Chicago, Illinois, using Python v.2.7 in consultation with the site owners regarding the use of the data. To do so, we used an existing database of all zip codes in the United States and used those zip codes in the state of Illinois that directly named Chicago as the originating city ($n_{\text{zip codes}} = 91$; see the Supplementary Methods for Study 4). The reviews began in 2004 when Yelp was first founded and continued until September 2017.

To measure the success of and behaviour towards each restaurant, we obtained the number of daily table reservations made at all Chicago restaurants that used the table reservation platform from OpenTable.com—the most popular online table reservation platform in the United States[56]. We used Python v.2.7 and obtained the data in line with the terms of use. Over a two-month period (14 July to 27 September 2017), we obtained the average number of daily table reservations made at each restaurant. There was a total of 1.30 million table reservations across the Chicago restaurants at this time. See the Supplementary Methods and Supplementary Results for Study 4 for more detail.

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

Data availability

The data for Study 2 are available from Amazon (https://s3.amazonaws.com/amazon-reviews-pds/readme.html). The data from Studies 1, 3 and 4 are publicly hosted on www.metacritic.com (Study 1), www.twitter.com (Study 3), www.facebook.com (Study 3), www.opentable.com (Study 4) and www.yelp.com (Study 4). For purposes of verification and reproducibility, readers will be provided with the code and anonymized aggregated data results upon request. Although the data are publicly available, their use is governed by each site's terms of use. Those interested in the original data should contact the site administrators for permission.

Code availability

The code for these analyses is available from the authors upon request.

Received: 14 May 2019; Accepted: 10 March 2021;
Published online: 08 April 2021

References

1. Asch, S. E. Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychol. Monogr. Gen. Appl.* **70**, 1–70 (1956).
2. Sherif, M. A study of some social factors in perception. *Arch. Psychol. Columbia Univ.* **187**, 60 (1935).
3. Simonson, I. & Rosen, E. *Absolute Value: What Really Influences Customers in the Age of (Nearly) Perfect Information* (HarperBusiness, 2014).
4. Smith, A. & Anderson, M. *Online Shopping and E-Commerce* (Pew Research Center, 2016); http://assets.pewresearch.org/wp-content/uploads/sites/14/2016/12/16113209/PI_2016.12.19_Online-Shopping_FINAL.pdf
5. Hu, N., Zhang, J. & Pavlou, P. A. Overcoming the J-shaped distribution of product reviews. *Commun. ACM* **52**, 144–147 (2009).
6. Woolf, M. Playing with 80 million Amazon product review ratings using Apache Spark. *minimaxir* http://minimaxir.com/2017/01/amazon-spark/ (2017).
7. McAuley, J., Pandey, R. & Leskovec, J. Inferring networks of substitutable and complementary products. in *Proc. 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (ACM, 2015).
8. *Yelp Factsheet* (Yelp, 2017); https://www.yelp.com/factsheet
9. Athey, S., Castillo, J. C. & Knoepfle, D. Service quality in the gig economy: empirical evidence about driving quality at Uber. White Paper. https://doi.org/10.2139/ssrn.3499781 (2019).
10. Babić Rosario, A., Sotgiu, F., De Valck, K. & Bijmolt, T. H. A. The effect of electronic word of mouth on sales: a meta-analytic review of platform, product, and metric factors. *J. Mark. Res.* **53**, 297–318 (2015).
11. Floyd, K., Freling, R., Alhoqail, S., Cho, H. Y. & Freling, T. How online product reviews affect retail sales: a meta-analysis. *J. Retail.* **90**, 217–232 (2014).
12. You, Y., Vadakkepatt, G. G. & Joshi, A. M. A meta-analysis of electronic word-of-mouth elasticity. *J. Mark.* **79**, 19–39 (2015).
13. de Langhe, B., Fernbach, P. M. & Lichtenstein, D. R. Navigating by the stars: investigating the actual and perceived validity of online user ratings. *J. Consum. Res.* **42**, 817–833 (2016).
14. Holbrook, M. B. & Addis, M. Taste versus the market: an extension of research on the consumption of popular culture. *J. Consum. Res.* **34**, 415–424 (2007).
15. Fowler, G. A. When 4.3 stars is average: the Internet's grade-inflation problem; Netflix is going with simpler thumbs-up or thumbs-down reviews, while online star ratings for many products have lost their meaning. *Wall Street Journal* https://www.wsj.com/articles/when-4-3-stars-is-average-the-internets-grade-inflation-problem-1491414200 (5 April, 2017).
16. Pang, B., Lee, L. & Vaithyanathan, S. Thumbs up? Sentiment classification using machine learning techniques. in *Proc. ACL-02 Conference on Empirical Methods in Natural Language Processing* **10**, 79–86 (Association for Computational Linguistics, 2002).
17. Petty, R. E. & Krosnick, J. A. *Attitude Strength: Antecedents and Consequences* (Psychology Press, 1995).
18. 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).
19. Wicker, A. W. Attitudes versus actions: the relationship of verbal and overt behavioral responses to attitude objects. *J. Soc. Issues* **25**, 41–78 (1969).
20. Visser, P. S., Bizer, G. Y. & Krosnick, J. A. in *Advances in Experimental Social Psychology* Vol. 38 (ed. Zanna, M. P.) 1–61 (Academic Press, 2006).
21. Petty, R. E., Fabrigar, L. R. & Wegener, D. T. in *Handbook of Affective Sciences* (eds Davidson, R. J. et al.) 752–772 (Oxford Univ. Press, 2003).
22. Zanna, M. P. & Rempel, J. K. in *The Social Psychology of Knowledge* (eds Bar-Tal, D. & Kruglanski, A. W.) 315–334 (Cambridge Univ. Press, 1988).
23. Haddock, G., Zanna, M. P. & Esses, V. M. Assessing the structure of prejudicial attitudes: the case of attitudes toward homosexuals. *J. Pers. Soc. Psychol.* **65**, 1105–1118 (1993).
24. Maio, G. R. & Esses, V. M. The need for affect: individual differences in the motivation to approach or avoid emotions. *J. Pers.* **69**, 583–614 (2001).
25. Rocklage, M. D., Rucker, D. D. & Nordgren, L. F. The Evaluative Lexicon 2.0: the measurement of emotionality, extremity, and valence in language. *Behav. Res. Methods* **50**, 1327–1344 (2018).
26. Rocklage, M. D. & Fazio, R. H. The evaluative lexicon: adjective use as a means of assessing and distinguishing attitude valence, extremity, and emotionality. *J. Exp. Soc. Psychol.* **56**, 214–227 (2015).
27. Lavine, H., Thomsen, C. J., Zanna, M. P. & Borgida, E. On the primacy of affect in the determination of attitudes and behavior: the moderating role of affective-cognitive ambivalence. *J. Exp. Soc. Psychol.* **34**, 398–421 (1998).
28. Rocklage, M. D. & Fazio, R. H. Attitude accessibility as a function of emotionality. *Pers. Soc. Psychol. Bull.* **44**, 508–520 (2018).
29. Rocklage, M. D. & Fazio, R. H. On the dominance of attitude emotionality. *Pers. Soc. Psychol. Bull.* **42**, 259–270 (2016).
30. Rocklage, M. D. & Luttrell, A. Attitudes based on feelings: fixed or fleeting? *Psychol. Sci*. https://doi.org/10.1177/0956797620965532 (2021).
31. Tooby, J. & Cosmides, L. The past explains the present. *Ethol. Sociobiol.* **11**, 375–424 (1990).
32. Ekman, P. E. & Davidson, R. J. *The Nature of Emotion: Fundamental Questions* (Oxford Univ. Press, 1994).
33. Fazio, R. H. in *Attitude Strength: Antecedents and Consequences* (eds Petty, R. E. & Krosnick, J. A.) 247–282 (Lawrence Erlbaum Associates, 1995).
34. Schwarz, N. in *Handbook of Theories of Social Psychology* (eds Van Lange, P. et al.) 289–308 (Sage, 2012).
35. Fazio, R. H. Attitudes as object–evaluation associations of varying strength. *Soc. Cogn.* **25**, 603–637 (2007).
36. Frijda, N. H. & Mesquita, B. in *Emotion and Culture: Empirical Studies of Mutual Influence* (eds Kitayama, S. & Markus, H. R.) 51–87 (American Psychological Association, 1994).
37. Keltner, D. & Haidt, J. Social functions of emotions at four levels of analysis. *Cogn. Emot.* **13**, 505–521 (1999).
38. Rocklage, M. D., Rucker, D. D. & Nordgren, L. F. Persuasion, emotion, and language: the intent to persuade transforms language via emotionality. *Psychol. Sci.* **29**, 749–760 (2018).
39. Van Kleef, G. A., De Dreu, C. K. W. & Manstead, A. S. R. The interpersonal effects of anger and happiness in negotiations. *J. Pers. Soc. Psychol.* **86**, 57–76 (2004).
40. Andrade, E. B. & Ho, T.-H. Gaming emotions in social interactions. *J. Consum. Res.* **36**, 539–552 (2009).
41. Lee, Y.-J., Hosanagar, K. & Tan, Y. Do I follow my friends or the crowd? Information cascades in online movie ratings. *Manage. Sci.* **61**, 2241–2258 (2015).
42. Schlosser, A. E. Posting versus lurking: communicating in a multiple audience context. *J. Consum. Res.* **32**, 260–265 (2005).
43. Moe, W. W. & Schweidel, D. A. Online product opinions: incidence, evaluation, and evolution. *Mark. Sci.* **31**, 372–386 (2012).
44. Russell, J. A. & Barrett, L. F. Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant. *J. Pers. Soc. Psychol.* **76**, 805–819 (1999).
45. Ad Meter https://finance.yahoo.com/news/usa-today-commemorate-30th-ad-150000342.html (2018).
46. *Ad Meter 2017 FAQ* (Ad Meter, 2017); http://admeter.usatoday.com/2017/01/17/ad-meter-2017-faq/

47. Asur, S. & Huberman, B. A. Predicting the future with social media. in *Proc. 2010 IEEE/ACM International Conference on Web Intelligence-Intelligent Agent Technology (WI-IAT)* **1**, 492–499 (IEEE Computer Society, 2010).
48. O'Connor, B., Balasubramanyan, R., Routledge, B. & Smith, N. From tweets to polls: linking text sentiment to public opinion time series. in *Proc. 4th AAAI Conference on Weblogs and Social Media* **11**, 122–129 (AAAI Press, 2010).
49. Pham, M. T., Cohen, J. B., Pracejus, J. W. & Hughes, G. D. Affect monitoring and the primacy of feelings in judgment. *J. Consum. Res.* **28**, 167–188 (2001).
50. Roskos-Ewoldsen, D. R. & Fazio, R. H. On the orienting value of attitudes: attitude accessibility as a determinant of an object's attraction of visual attention. *J. Pers. Soc. Psychol.* **63**, 198–211 (1992).
51. Berger, J. & Milkman, K. L. What makes online content viral? *J. Mark. Res.* **49**, 192–205 (2012).
52. Castelvecchi, D. Can we open the black box of AI? *Nature* **538**, 20–23 (2016).
53. *Python Language Reference*, version 2.7. http://www.python.org (Python Software Foundation, 2017).
54. *Amazon Customer Reviews Dataset* (Amazon, 2020); https://s3.amazonaws.com/amazon-reviews-pds/readme.html
55. Ni, J., Li, J. & McAuley, J. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proc. 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (EMNLP-IJCNLP) 188–197 (Association for Computational Linguistics, 2019).
56. Filloon, W. In the battle for restaurant reservations, OpenTable is still way ahead. *Eater* https://www.eater.com/2018/9/24/17883688/opentable-resy-online-reservations-app-danny-meyer (2018).

Acknowledgements

We received no specific funding for this work. We thank Internet Video Archive LLC for their assistance in providing access to the movie data and metadata from Study 1.

Author contributions

M.D.R., D.D.R. and L.F.N. conceptualized the work. M.D.R. obtained and analysed the data with collaboration from D.D.R. and L.F.N. M.D.R., D.D.R. and L.F.N. wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41562-021-01098-5.

Correspondence and requests for materials should be addressed to M.D.R.

Peer review information *Nature Human Behaviour* thanks Jonah Berger, Saif Mohammad and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

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): Matthew D. Rocklage

Last updated by author(s): Feb 17, 2021

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 2.7
Data analysis	R 3.5.1, SPSS 25

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

The data for Study 2 are available from Amazon (https://s3.amazonaws.com/amazon-reviews-pds/readme.html). The data from Studies 1, 3, and 4 are publicly hosted on www.metacritic.com (Study 1), www.twitter.com (Study 3), www.facebook.com (Study 3), www.opentable.com (Study 4), and www.yelp.com (Study 4). For purposes of verification and reproducibility, readers will be provided with the code and anonymized aggregated data results upon request. Although the data are publicly available, their use is governed by each site's terms of use. Those interested in the original data should contact the site administrators for permission.

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 field data.
Research sample	Online reviews (Metacritic, Amazon, Yelp) and tweets from Twitter. Each sample is representative of online postings from the different online platforms. The online reviews are those from among the most popular online review websites for their category and we obtained all possible tweets on Twitter.
Sampling strategy	The sample size represents the available data for each domain.
Data collection	Data were collected manually or by an automated script from each corresponding website.
Timing	Data were collected from 2016 to 2020.
Data exclusions	No data were excluded.
Non-participation	No participants dropped out.
Randomization	Participants were not allocated into experimental groups.

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