Revised Manuscript

Negativity drives online news consumption

Claire Robertson[1,*], Nicolas Pröllochs[2,3,*], Kaoru Schwarzenegger[4], Phillip Parnamets[5], Jay J. Van Bavel[1,6,†] & Stefan Feuerriegel[4,7,†]

[1] Department of Psychology, New York University, New York, NY 10003

[2] University of Giessen, Giessen 35394, Germany

[3] Oxford-Man Institute, University of Oxford, Oxford OX26ED, United Kingdom

[4] ETH Zurich (Swiss Federal Institute of Technology), Zurich 8092, Switzerland

[5] Division of Psychology, Department of Clinical Neuroscience, Karolinska Institutet, Stockholm, Sweden

[6] Center for Neural Science, New York University, New York, NY 10003

[7] LMU Munich, Munich 80539, Germany

[*] These authors contributed equally

[†] These authors contributed equally

Abstract

Online media is important for society in informing and shaping opinions, hence raising the question of what drives online news consumption. Here, we analyze the effect of negative words on news consumption using a massive online dataset of viral news stories. Specifically, we conducted preliminary analyses using a large-scale, series of randomized controlled trials in the field (N = 22,743). Our final dataset will comprise ~105,000 different variations of news stories from Upworthy.com–one of the fastest growing websites of all time–that generated ~8 million clicks across more than 530 million overall impressions. As such, this dataset allows a unique opportunity to test the causal impact of negative and emotional language on consumption with millions of news readers. An analysis with preliminary data reveals that negative words in news increase consumption rates. Our results contribute to a better understanding of why users engage with online media.

1 Introduction

The newsroom phrase “if it bleeds, it leads” was coined to reflect the intuition among journalists that stories about crime, bloodshed, and tragedy sell more newspapers than stories about good news [1]. However a large portion of news readership now occurs online—the motivation to sell papers transformed into a motivation to keep readers clicking on new articles. In the United States, 89% of adults get at least some of their news online, and reliance on the Internet as a news source is increasing [2]. Even so, most users spend less than 5 minutes per month on all of the top 25 news sites put together [3]. Hence, online media is forced to compete for the extremely limited resource of reader attention [4] .

With the advent of the Internet, online media has become a widespread source of information and, subsequently, opinion formation [5, 6, 7, 8, 9]. As such, online media has a profound impact on society across domains such as marketing [10, 11], finance [12, 13, 14], health [15], and politics [16, 17, 18, 19]. Therefore, it is crucial to understand exactly what drives online news consumption. Previous work has posited that competition pushes news sources to publish “click-bait” news stories, often categorized by outrageous, upsetting, and negative headlines [20, 21, 22]. Here, we analyze the effect of negative words on news consumption using a massive online dataset of viral news stories from Upworthy.com–a website that was one of the most successful pioneers of clickbait in the history of the Internet [23].

The tendency for individuals to attend to negative news reflects something foundational about human cognition—that humans preferentially attend to negative stimuli across many domains [24, 25]. Attentional biases towards negative stimuli begin in infancy [26] and persist into adulthood as a fast and automatic response [27]. Furthermore, negative information may be

more “sticky” in our brains; people weigh negative information more heavily than positive information, when learning about themselves, learning about others, and making decisions [28, 29, 30]. This may be due to negative information automatically activating threat responses—knowing about possible negative outcomes allows for planning and avoidance of potentially harmful or painful experiences [31, 32, 33].

Prior work has explored the role of negativity for driving online behavior. In particular, negative language in online content has been linked to user engagement, that is, sharing activities [22, 34, 35, 36, 37, 38, 39]. As such, negativity embedded in online content explains the speed and virality of online diffusion dynamics (e.g., response time, branching of online cascades) [7, 34, 35, 37, 39, 40, 41]. Further, online stories from social media perceived as negative garner more reactions (e.g., likes, Facebook reactions) [42, 43]. Negativity in news increases physiological activations [44], and negative news is more likely to be remembered by users [45, 46, 47]. Some previous work has looked also into negativity effects for specific topics such as political communication and economics [34, 48, 49, 50, 51, 52]. Informed by this, we hypothesized an effect of negative words on online news consumption.

The majority of studies on online behavior are correlational (e. g., [34, 35, 36, 38, 39, 40, 41, 42]), while laboratory studies take subjects out of their natural environment. As such, there is little work examining the causal impact of negative language on real-world news consumption. Here, we analyze data from the Upworthy Research Archive [53], a repository of news consumption data that are both applied and causal. Due to the structure of this unique dataset, we are able to test the causal impact of negative (and positive) language on news engagement in an ecologically rich online context. Moreover, our dataset is large-scale, allowing for a precise estimate of the effect size of negative words on news consumption.

Data on online news consumption was obtained from Upworthy, a highly influential media website founded in 2012 that used viral techniques to promote news articles across social media [53, 54]. Upworthy has been regarded as one of the fastest growing media companies worldwide [53] and, at its peak, reached more users than established publishers such as the New York Times [55]. Content was optimized with respect to user responses through data-driven methods, specifically randomized controlled trials (RCTs) [56]. The content optimization by Upworthy profoundly impacted the media landscape (e.g., algorithmic policies were introduced by Facebook in response [23]). In particular, the strategies employed by Upworthy have also informed other content creators and news agencies.

Upworthy conducted numerous randomized control trials of news headlines on its website to evaluate the efficacy of differently worded headlines in generating article views [53]. In each experiment, Upworthy users were randomly shown different headline variations for a news story, and user responses were recorded and compared. Editors were commonly required to propose 25 different headlines from which the most promising headlines were selected for experimental testing [57].

In the current paper, we analyze the effect of negative words on news consumption. Specifically, we hypothesize that the presence of negative words in a headline will increase the click-through rate for that headline. The Design Table is in Table 2. Using a text mining framework, we extract negative words and estimate the effect on click-through rate using a multilevel regression (see Methods). We provide empirical evidence from an initial sample (N = 22,743) of large-scale, randomized controlled trials in the field. Overall, our data will contain over 105,000 different variations of news headlines from Upworthy, which have generated ~8 million clicks and more than 530 million impressions.

Negative sentiment consists of many discrete negative emotions. Prior work has proposed that certain discrete categories of negative emotions may be especially attention-grabbing [58]. For example, high-arousal negative emotions such as anger or fear have been found to efficiently attract attention and be quickly recognizable in facial expressions and body language [31, 59, 60]. This may be because of the social and informational value that high-arousal emotions like anger and fear holds – both could alert others in one's group to threats, and paying preferential attention and recognition to these emotions could help the group survive [27, 32] . This may also be why in the current age, people are more likely to share and engage with online content that is embedding anger, fear, or sadness [21, 41, 61, 62]. Therefore, in addition to examining the effect of negative words as our primary analysis, we further conduct a secondary analysis examining the effect of high and low arousal negative words. Specifically, we examine the effects of words related to anger and fear (as high-arousal negative emotions), as well as sadness (as a low-arousal negative emotion). We also examine the effects of words related to joy (positive emotion), which we predict will be associated with lower click-through rates.

2 Methods

2.1 Ethics information

The research complies with all relevant ethical regulations. Ethics approval (2020-N-151) for the main analysis was obtained from the institutional review board at ETH Zurich. For the user validation, ethics approval (IRB-FY2021-5555) was obtained from the institutional review board at New York University. Participants in the validation study were recruited from the subject pool of the Department of Psychology at New York University in exchange for 0.5 hrs of research credit for varying psychology courses. Participants provided informed consent for the user validation studies. New York University did not require IRB approval for main analysis, as it is not classified as human subjects research.

2.2 Pilot data (with large-scale field experiments)

In this research, we build upon data from the Upworthy Research Archive [53]. The data have been made available through an agreement between Cornell University and Upworthy. We have access to this unique dataset upon the condition of following the procedure of a Registered Report. In the current stage, we have access only to a subset of the dataset (i.e., the "exploratory sample"), based on which we conducted the preliminary analysis. Our final analysis will be based on data from $N = 22{,}743$ experiments (RCTs) collected on Upworthy between January 24, 2013 and April 14, 2015. Here, $N = 22{,}743$ is the size of the confirmatory sample of experiments on which we will test our pre-registered hypotheses.

Each RCT corresponds to one news story, in which different headlines for the same news story were compared. Formally, for each headline variation j in an RCT i ($i = 1, \ldots, N$), the

following statistics were recorded: (1) the number of impressions, that is, the number of users to whom the headline variation was shown ($impressions_{ij}$), and (2) the number of clicks a headline variation generated ($clicks_{ij}$). The click-through rate is then computed via $CTR_{ij} = \frac{clicks_{ij}}{impressions_{ij}}$. The experiments were conducted separately (i.e., only a single experiment was conducted at the same time for the entire website) so each test can be analyzed as independent of all other tests [53]. Examples of news headlines in the experiments are in Table 1. The Upworthy Research Archive contains data aggregated at the headline level and, thus, does not provide individual level data for users.

The data will be subject to the following filtering. First, all experiments solely consisting of a single headline variation will be discarded. Single headline variations exist because Upworthy conducted RCTs on features of their articles other than headlines, predominantly teaser images. In many RCTs where teaser images were varied, headlines were not varied at all (image data has not been made available to researchers by the Upworthy Research Archive, so we are unable to incorporate image RCTs into our analyses but we later validate our findings as part of the robustness checks). Second, some experiments contain multiple treatment arms with identical headlines, which will be merged into one representative treatment by summing their clicks and impressions. These occurred when images *and* headlines were involved in RCTs for the same story. This is relatively rare in the dataset, but for robustness checks regarding image RCTs (see Supplement F).

The analysis in the final paper will be based on the confirmatory sample of the dataset [53], which will be made available to us only after pre-registration is conditionally accepted. In the current pre-registration stage, we present the results of a preliminary analysis based on a smaller,

exploratory sample (see Supplement A). Both will be processed using identical methodology. The exploratory sample for our preliminary analysis comprises 4,873 experiments, involving 22,666 different headlines before filtering and 11,109 headlines after filtering, which corresponds to, on average, 4.27 headlines per experiment. On average, there were approximately 16,670 participants in each RCT. Additional summary statistics are given in Supplement B.

#	Headline Variation	CTR
1	If The Numbers 4 And 20 Mean Something To You, You're Gonna Want To Hear This **Shit**	0.94%
2	What He Has To Say About Pot Is Going To Make Both Sides **Angry** , But Here Goes	0.79%
3	Lots Of Things In Life Have Both A **Benefit** And A **Harm** . So Why Do We Only **Obsess** About This One?	0.60%
4	He Explains Why The Question 'What Are You Smoking' Is Actually **Kind** Of **Important** .	0.58%
1	IMAGINE: You're **Raped** At Your Job And Your Boss Intentionally Tries To Shut You Up	0.92%
2	12 Minutes. If You **Support** Our Troops, Sacrifice At Least That Much For Them.	0.21%
1	Spoofers Set Up A **Fake** Agency To Show How **Ridiculous** Some People Are When It Comes To Immigrants	0.65%
2	Something's Been **Missing** From Our **Favorite** Superhero Stories, And It Makes Reality Seem Kinda **Silly**	0.56%
3	Some Comic Book **Lovers** Might Need To Check Their Politics When They See What These Guys Have In Mind	0.53%
4	A New Agency Wants To Get Rid Of All Our **Favorite** Superheroes. I **Laughed** When I Saw Why.	0.41%
1	I Knew Which One She'd Pick, But It Still **Crushed** Me	1.10%
2	First She Points To The **Pretty** Child. Then To The **Ugly** Child. Then My Heart Breaks.	0.85%
3	1 Little Girl, 5 Cartoons And 1 **Heartbreaking** Answer	0.83%
4	What She Says About These Cartoons, Says Something Incredibly **Troubling** About The World We Live In	0.66%

Legend: **Positive** , **Negative**

Table 1: Example experiments (randomized control trials) performed by Upworthy. Shown are four experiments and each with different headline variations subject to testing. Columns report the click-through rate (CTR) and the positive and negative words as classified by the LIWC dictionary [63].

2.3 Design

We present a Design Table summarizing our methods in Table 2.

Question	Hypothesis	Sampling plan	Analysis plan	Interpretation given to different outcomes
How does the presence of negative and positive words affect the click-through rate for news headlines?	The presence of negative words in a headline will significantly increase the click-through rate for that headline. The presence of positive words in a headline will significantly decrease the click-through rate for that headline.	A power analysis suggests that the sample size of the confirmatory dataset (22, 743 RCTs) will have sufficient power to achieve 99% power to detect an effect size of 0.01, considered to be a small effect size [64]. This effect size is slightly more conservative than estimates of effect sizes from pilot studies, and is derived from theory [65].	We will conduct a multilevel binomial model examining the effects of the proportion of negative words in a headline on the click-through rate, adjusting for the proportion of negative words in a headline, the number of positive words in the headline, the complexity of the headline as measured by the Gunning Fog Index, and the age of the story relative to the age of the Upworthy platform. We include random effects grouped by RCT, and will use two models to test our hypothesis. One allows the intercept to vary, the other also allows the slope of negative words to vary.	A significant positive coefficient for negative words will be interpreted as evidence that a higher ratio of negative words in a headline is associated with a greater click-through rate. A significant negative coefficient for negative words will be interpreted as evidence that a higher ratio of negative words in a headline is associated with a lower click-through rate. A significant positive coefficient for positive words will be interpreted as evidence that a higher ratio of positive words in a headline is associated with a greater click-through rate.A significant negative coefficient for positive words will be interpreted as evidence that a higher ratio of positive words in a headline is associated with a lower click-through rate. We will consider evidence to be conclusive only in cases where both models fit to the data agree in their qualitative

				conclusions about the effect of negative words. To evaluate effects where we cannot reject the null hypothesis, we will test for equivalence [66] against an interval of [−0.001, 0.001]. If our observed confidence interval is fully contained in this interval we will consider this evidence for a null effect, otherwise we will consider the results inconclusive with respect to the null.
How does the presence of discrete emotional words affect the click-through rate for news headlines?	The presence of anger, fear, and sadness words in a headline will significantly increase the click-through rate for that headline. The presence of joy words in a headline will significantly decrease the click-through rate for that headline..	A power analysis suggests that the sample size of the confirmatory dataset (22,743 RCTs) will have sufficient power to achieve 99% power to detect effect sizes of 0.01.	We will conduct a multilevel binomial model examining the effects of the four emotions (anger, fear, joy, and sadness) on the click-through rate, adjusting for the number of words in the headline, the complexity of the headline as measured by the Gunning Fog Index, and the age of the story relative to the age of the Upworthy platform. We include the RCT as a random intercept.	A positive value for each of the emotions signifies a larger proportion of emotional words from that emotion in a headline. Therefore, a significant positive coefficient for the emotion will be interpreted as evidence that headlines containing a word from the emotion (i.e., anger, fear, joy, and sadness) is associated with a greater click-through rate. Conversely, a negative value for each of the coefficients signifies that the proportion of emotional words from the emotion is more prevalent in a headline. Therefore, a significant negative coefficient for the emotion indicates that

				headlines containing a word from emotion (i.e., anger, fear, joy, and sadness) is associated with a smaller click-through rate. To evaluate effects where we cannot reject the null hypothesis, we will test for equivalence [66] against an interval of [−0.001, 0.001]. If our observed confidence interval is fully contained in this interval we will consider this evidence for a null effect, otherwise we will consider the results inconclusive with respect to the null.

Table 2: Design table. .

2.3 Sampling plan

Given our unique opportunity to secure an extremely large sample where the *N* was predetermined, we chose to run a simulation before pre-registration to estimate what level of power we would achieve for observing an effect size represented by a regression coefficient of 0.01 (i.e., a 1% effect on the odds of clicks from a standard deviation increase in negative words). This effect size is slightly more conservative than estimates of effect sizes from pilot studies, and is derived from theory [65]. The total data size of the Upworthy data archive is N = 22,743 RCTs, with between three and twelve headlines per RCT. This thus corresponds to a total sample of between 68,229 and 227,430 headlines. Because we are not aware of the exact size, we generated datasets through a bootstrapping procedure that sampled N = 22,743 RCTs with replacement from our pilot sample of tests. We simulated 1000 such datasets and for each dataset we generated "clicks" using the estimated parameters from the pilot data. Finally, each dataset was analyzed using the model as described. This procedure was repeated for both models (varying-intercepts, and a combination of varying-intercepts and varying-slopes). We found that, under the assumptions of effect size, covariance matrix and data generating process from our pilot sample, we will have greater than 99% power to detect an effect size of 0.01 in the final sample for both models.

2.4 Analysis plan

2.4.1 Text mining framework

Text mining will be used to extract emotional words from news headlines. To prepare the data for the text mining procedure, we will apply standard preprocessing to the headlines. Specifically,

the running text will be converted into lower-case and tokenized, and special characters (i.e., punctuations and hashtags) will be removed. We will then apply a dictionary-based approach analogous to earlier research [22, 39, 40, 41].

We will perform sentiment analysis based on the Linguistic Inquiry and Word Count (LIWC) [63]. The LIWC contains word lists classifying words according to both a positive (n = 620 words, i.e. "love" and "pretty") and negative sentiment (n = 744 words, i.e. "wrong" and "bad"). A list of the most frequent positive and negative words in our dataset are in Supplement C.

Formally, sentiment analysis will be based on single words (i.e., unigrams) due to the short length of the headlines (mean length: 14.89 words). We will count the number of positive words ($n_{positive}$) and the number of negative words ($n_{negative}$) in each headline. A word is considered "positive" if it is in the dictionary of positive words (and, vice versa, for "negative" words). We will then normalize the frequency by the length of the headline, that is, the total number of words in the headline (n_{total}). This yields the two separate scores

$$Positive_{ij} = \frac{n_{positive}}{n_{total}} \text{ and } Negative_{ij} = \frac{n_{negative}}{n_{total}},$$

for headline j in experiment i. As such, the corresponding scores for each headline represent percentages. For example, if a headline has 10 words out of which one is classified as "positive" and none as "negative," the scores are $Positive_{ij} = 10\%$ and $Negative_{ij} = 0\%$. If a headline has 10 words and contains one "positive" and one "negative" word, the scores are $Positive_{ij} = 10\%$ and $Negative_{ij} = 10\%$. A headline may contain both positive and negative words, so both variables are later included in the model.

Negation words (e.g., “not,” “no”) have the ability to invert the meaning of statements and thus the corresponding sentiment. We will perform negation handling as follows. First, the text is scanned for negation terms using a predefined list, and then all positive (or negative) words in the neighborhood are counted as belonging to the opposite word list, i.e., they are counted as negative (or positive) words. In our analysis, the neighborhood (i.e., the so-called negation scope) is set to 3 words after the negation. As a result, a phrase like “not happy” is coded as negative rather than positive. Here, we will use the implementation from the *sentimentr* package (details: https://cran.r-project.org/web/packages/sentimentr/readme/README.html).

Using the above dictionary approach, our objective is to quantify the presence of positive and negative words. As such, we do not attempt to infer the internal state of a perceiver based on the language they write, consume, or share [67]. Specifically, readers’ preference for headlines containing negative words does not imply that users *felt* more negatively while reading said headlines. In contrast, we quantify how the presence of certain words is linked to concrete behavior. Following this, our pre-registered hypotheses test whether negative words increase consumption rates (see Table 2).

We validated the dictionary approach in the context of our corpus based on a pilot study [68], (Supplement D). Here, we used the positive and negative word lists from LIWC [63] and performed negation handling as described above. Perceived judgments of positivity and negativity in headlines correlate with the number of negative and/or positive words each headline contains. Specifically, we correlated the mean of the 8 human judges’ scores for a headline with NRC sentiment rating for that headline.We found a moderate but significant positive correlation ($r_s = 0.303$, $p < 0.001$). These findings validate that our dictionary approach captures significant variation in the perception of emotions in headlines from perceivers.

Two additional text statistics will be computed: First, we will determine the length of the news headline as given by the number of words. Second, we will calculate a text complexity score using the Gunning-Fog index [69]. This index estimates the years of formal education necessary for a person to understand a text upon reading it for the first time: $0.4 \times (ASL + 100 \times n_{wsy\geq3}/n_w)$, where ASL is the average sentence length (number of words), n_w is the total number of words, and $n_{wsy\geq3}$ is the number of words with three syllables or more. A higher value thus indicates greater complexity. Both headline length and the complexity score will be used as control variables in the statistical models. Results based on alternative text complexity scores are reported as part of the robustness checks.

The above text mining pipeline will be implemented with the software tool R 4.0.2 using the packages *quanteda* (version 2.0.1) and *sentimentr* (version 2.7.1) for text mining.

2.4.2 Empirical model

We will estimate the effect of emotions on online news consumption using a multilevel binomial regression. Specifically, we expect that negative language in a headline will affect the probability that users click on a news story to access its content. To test our hypothesis, we specify a series of regression models where the dependent variable is given by the click-through rate.

We will model news consumption as follows. Let $i = 1, \ldots, N$ refer to the different experiments in which different headline variations for news stories are compared through an RCT. Let $clicks_{ij}$ denote the number of clicks from headline variation j belonging to news story i. Analogously, let $impressions_{ij}$ be the corresponding number of impressions. We follow the approach in [70] and model the number of clicks to follow a binomial distribution via

$$clicks_{ij} \sim \text{Binomial}(impressions_{ij}, \theta_{ij}),$$

where $0 \leq \theta_{ij} \leq 1$ is the probability of a user clicking on a headline in a single Bernoulli trial and where θ_{ij} corresponds to the click-through rate of headline variation j from news story i.

We will estimate the effect of positive and negative words on the click-through rate θ_{ij} and capture between-experiment heterogeneity through a multilevel structure. We will further control for other characteristics across headline variations, namely length, text complexity, and the relative age of a headline (based on the age of the platform). The regression model is then given by

$$\begin{aligned}\text{logit}(\theta_{ij}) = \alpha + \alpha_i + \beta_1 Positive_{ij} + \beta_2 Negative_{ij} \\ + \gamma_1 Length_{ij} + \gamma_2 Complexity_{ij} + \gamma_3 PlatformAge_{ij},\end{aligned}$$

where α is the global intercept and α_i is an experiment-specific intercept (i.e., random effect). Both α and α_i are assumed to be independent and identically normally distributed with a mean of zero. The latter captures heterogeneity at the experiment level; that is, some news stories might be more interesting than others. In addition, we control for the length ($Length_{ij}$) and complexity ($Complexity_{ij}$) of the text in the news headline, as well as the relative age of the current experiment with regard to the platform ($PlatformAge_{ij}$). The latter denotes the number of days of the current experiment since the first experiment on Upworthy.com in 2012 and thus allows us to control for potential learning effects as well as changes in editorial practices over time. The coefficient β_2 is our main variable of interest: it quantifies the effect of negative words on the click-through rate.

In the above analysis, all variables will be z-standardized for better comparability. That is, prior to estimation, we subtract the sample mean and divide by the standard deviation. Because

of this, the regression coefficients β_1 and β_2 quantify changes in the dependent variable in standard deviations. This allows us to compare the relative effect sizes across positive and negative words (as well as emotional words later). Due to the logit link, the odds ratio is $100 \times (e^{\beta} - 1)$, which gives the percentage change in the odds of success as a result of a standard deviation change in the independent variable. In our case, a successful event is indicated by the user clicking the headline.

The above regression builds upon a global coefficient for capturing the effect of language on click-through rate, and, as such, the language reception is assumed to be equal across different RCTs. This is consistent with previous works where a similar global coefficient (without varying-slopes) was used (e.g., [22, 34, 38, 39]). However, there is reason to assume that the receptivity to language might vary across RCTs and thus among news (e.g., the receptivity of negative language might be more dominant for political news than for entertainment news, or for certain news topics over others). As such, the variance in the estimated regression coefficients is no longer assumed to be exactly zero across experiments but may vary. To do so, we augment the above random effects model by an additional varying-slopes specification. Here, a multilevel structure is used that accounts for the different levels due to the experiments $i = 1, \ldots, N$. Specifically, the coefficients β_1 and β_2 capturing the effect of positive and negative words on click-through rate, respectively, are allowed to vary across experiments. Of note, a similar varying-slopes formalization is only used for the main analysis based on positive and negative language, and an not for the subsequent extension to emotional words where it is not practical due to the fact that there would be comparatively few treatment arms in comparison to the number of varying-slopes.

Here, we pre-register a plan to conduct the analysis based on both models, that is, (i) the random effect model and (ii) the random effect model with additional varying-slopes. If the estimates from both models are in the same direction, this should underscore the overall robustness of the findings. If it is the case that estimated coefficients from the random effect model and the random effect, varying-slopes model contradict each other on the confirmatory sample (i.e., full dataset), both results will be reported but precedence in interpretation will be given to the latter due to its more flexible specification.

All models will be estimated using the *lme4* package (version 1.1.23) in R.

2.4.3 Extension to discrete emotional words

To provide further insights into how emotional language relates to news consumption, we will extend our text mining framework and perform additional secondary analyses. We are specifically interested in the effect of different emotional words (anger, fear, joy, and sadness) on the click-through rate.

Here, our analyses are based on the NRC emotion lexicon [71] due to its widespread use in academia and the scarcity of other comparable dictionaries with emotional words for content analysis [72]. The NRC lexicon comprises 181,820 English words that are classified according to the 8 basic emotions of Plutchik's emotion model [73]. Basic emotions are regarded as universally recognized across cultures, and on this basis, more complex emotions can be derived [74, 75]. The 8 basic emotions computed via the NRC are anger, anticipation, joy, trust, fear, surprise, sadness, and disgust.

We will calculate scores for basic emotions embedded in news headlines based on the NRC emotion lexicon [71]. We will count the frequency of words in the text that belong to a

specific basic emotion in the NRC lexicon (i.e., an 8-dimensional vector). A list of the most frequent emotional words in our dataset is given in Supplement C. Afterward, we will divide the word counts by the total number of dictionary words in the text, so that the vector is normalized to sum to one across the basic emotions. Following this definition, the embedded emotions in a text might be composed of, for instance, 40% *anger* while the remaining 60% are *fear*. We will omit headline variations that do not contain any emotional words from the NRC emotion lexicon (since, otherwise, the denominator will not be defined). Due to this extra filtering step, we obtained a final sample of 8,365 headlines for the pilot analysis. We again account for negations using the above approach in that the corresponding emotional words are not attributed to the emotion but skipped during the computation (as there is no defined "opposite" emotion).

As a next step, we validated the NRC emotion lexicon for the context of our study through a user study (see Supplement D). Specifically, we correlated the mean of the 8 human judges' scores for a headline with NRC emotion rating for that headline. We found that, overall, both mean user judgments on emotions and those from the NRC emotion lexicon are correlated (r_s: 0.114, $p < 0.001$). Furthermore, mean user judgements for four basic emotions were significantly correlated, namely anger (r_s: 0.22, $p = 0.005$), fear (r_s: 0.29, $p < 0.001$), joy (r_s: 0.24, $p = 0.002$), and sadness (r_s: 0.30, $p < 0.001$, respectively). The four other basic emotions from the NRC emotion lexicon showed considerably lower correlation coefficients in the validation study, namely anticipation (r_s: −0.07, $p = 0.341$), disgust (r_s: 0.01, $p = 0.926$), surprise (r_s: −0.06, $p = 0.414$), and trust (r_s: 0.12, $p = 0.122$). Because of that, we did not pre-register hypotheses for them.

The multilevel regression is specified analogous to the model above but with different explanatory variables, i.e.,

$$\text{logit}(\theta_{ij}) = \alpha + \alpha_i + \beta_1 Anger_{ij} + \beta_2 Fear_{ij} + \beta_3 Joy_{ij}$$
$$+ \beta_4 Sadness_{ij} + \gamma_1 Length_{ij} + \gamma_2 Complexity_{ij} + \gamma_3 PlatformAge_{ij},$$

where α and α_i represent the global intercept and the random effects, respectively. Specifically, α is again the global intercept and α_i captures the heterogeneity across experiments $i = 1, \ldots, N$. As above, we include the control variables, i.e., length, text complexity, and platform age. The coefficients $\beta_1, \ldots, \beta_4$ quantify the effect of the emotional words (i.e., anger, fear, joy, and sadness) on the click-through rate.

Again, all variables will be *z*-standardized for better comparability (i.e., we subtract the sample mean and divide by the standard deviation). As a result, the regression coefficients quantify changes in the dependent variable in standard deviations. This allows us to compare the relative effect sizes across different emotions.

Additional information

Acknowledgements. We are grateful to a John Templeton Foundation Grant (#61378) that funded J.V.B. . We thank Upworthy, as well as J. Nathan Matias, Kevin Munger, Marianne Aubin Le Quere, and Charles Ebersole for making the data available. We thank W. Brady for his helpful feedback.

Code availability. Upon publication of this work, code that supports the findings of our study will be made available via http://osf.io.

Data availability. Data from the randomized controlled experiments in the field is available through the Upworthy Research Archive [53]. The LIWC dictionary [63] is available for purchase (https://liwc.wpengine.com/). The NRC emotion lexicon [71] is publicly available for download (http://sentiment.nrc.ca/lexicons-for-research/). In our analysis, we used the built-in version from the "sentimentr" package.

Author information

Correspondence. Jay J. Van Bavel (jay.vanbavel@nyu.edu) and Stefan Feuerriegel (feuerriegel@lmu.de)

Author contributions. C.R., N.P., K.S., P.P., J.V.B., and S.F. conceived and designed the experiments. C.R.. N.P., and K.S. analyzed the data. C.R., N.P., K.S., P.P., J.V.B., and S.F. wrote the paper.

Competing interests. The authors declare no competing interests.

References

[1] Pooley, E.. Grins, Gore and Videotape: The Trouble with Local TV News. *New York Magazine* **22**, 36-44 (1989).

[2] Pew. Americans almost equally prefer to get local news online or on TV set. *Pew Research Center's Journalism Project* (2019)

[3] Olmstead, K., Mitchell, A., & Rosenstiel, T. Navigating news online: Where people go, how they get there and what lures them away. *Pew Research Center's Project for Excellence in Journalism* (2011).

[4] Simon, H. A. Designing Organizations for an Information-Rich World. In M. Greenberger (Ed.), *Computers, Communications, and the Public Interest*. Baltimore, MD: The Johns Hopkins Press. 38-72. (1971).

[5] Flaxman, S., Goel, S. & Rao, J. M. Filter bubbles, echo chambers, and online news consumption. *Public Opinion Quarterly* **80**, 298–320 (2016).

[6] Schmidt, A. L. et al. Anatomy of news consumption on Facebook. *Proceedings of the National Academy of Sciences (PNAS)* **114**, 3035–3039 (2017).

[7] Bakshy, E., Messing, S. & Adamic, L. A. Exposure to ideologically diverse news and opinion on Facebook. *Science* **348**, 1130–1132 (2015).

[8] Allen, J., Howland, B., Mobius, M., Rothschild, D. & Watts, D. J. Evaluating the fake news problem at the scale of the information ecosystem. *Science Advances* **6**, eaay3539 (2020).

[9] Yang, T., Majó-Vázquez, S., Nielsen, R. K. & González-Bailón, S. Exposure to news grows less fragmented with an increase in mobile access. *Proceedings of the National Academy of Sciences (PNAS)* **117**, 28678–28683 (2020).

[10] Godes, D. & Mayzlin, D. Using online conversations to study word-of-mouth communication. *Marketing Science* **23**, 545–560 (2004).

[11] Berger, J. & Schwartz, E. M. What drives immediate and ongoing word of mouth? *Journal of Marketing Research* **48**, 869–880 (2011).

[12] Antweiler, W. & Frank, M. Z. Is all that talk just noise? The information content of Internet stock message boards. *The Journal of Finance* **59**, 1259–1294 (2004).

[13] Forbes. Can 'fake news' impact the stock market? URL https://www.forbes.com/sites/kenrapoza/2017/02/26/can-fake-news-impact-the-stock-market/. (2017).

[14] Bollen, J., Mao, H. & Zeng, X. Twitter mood predicts the stock market. *Journal of Computational Science* **2**, 1–8 (2011).

[15] Garfin, D. R., Silver, R. C. & Holman, E. A. The novel coronavirus (COVID-2019) outbreak: Amplification of public health consequences by media exposure. *Health Psychology* **39**, 355–357 (2020).

[16] Aral, S. & Eckles, D. Protecting elections from social media manipulation. *Science* **365**, 858–861 (2019).

[17] Bond, R. M. et al. A 61-million-person experiment in social influence and political mobilization. *Nature* **489**, 295–298 (2012).

[18] Jones, J. J., Bond, R. M., Bakshy, E., Eckles, D. & Fowler, J. H. Social influence and political mobilization: Further evidence from a randomized experiment in the 2012 U.S. presidential election. *PLOS ONE* **12**, e0173851 (2017).

[19] Levy, R. Social media, news consumption, and polarization: Evidence from a field experiment. *American Economic Review* **111***,* 831-70. (2020).

[20] Klein, E. *Why We're Polarized.* (Avid Reader Press / Simon & Schuster, New York, NY, 2020)

[21] Crockett, M. J. Moral outrage in the digital age. *Nature Human Behaviour* **1**, 769–771 (2017).

[22] Vosoughi, S., Roy, D., & Aral, S. The spread of true and false news online. *Science* **359**, 1146-1151 (2018).

[23] Sanders, S. Upworthy was one of the hottest sites ever. You won't believe what happened next. *NPR* https://www.npr.org/sections/alltechconsidered/2017/06/20/533529538/upworthy-was-one-of-the-hottest-sites-ever-you-wont-believe-what-happened-next (2017).

[24] Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. Bad is stronger than good. *Review of General Psychology* **5**, 323-370 (2001).

[25] Rozin, P., & Royzman, E. B. Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review* **5**, 296-320 (2001).

[26] Carver, L. J., & Vaccaro, B. G. 12-Month-old infants allocate increased neural resources to stimuli associated with negative adult emotion. *Developmental Psychology* **43**, 54–69 (2007).

[27] Dijksterhuis, A., & Aarts, H. On wildebeests and humans: The preferential detection of negative stimuli. *Psychological Science* **14**, 14–18 (2003).

[28] Müller-Pinzler, L., Czekalla, N., Mayer, A. V., Stolz, D. S., Gazzola, V., Keysers, C., Paulus, F. M., & Krach, S. Negativity-bias in forming beliefs about own abilities. *Scientific Reports* **9**, 14416 (2019).

[29] Boydstun, A. E., Ledgerwood, A., & Sparks, J. A negativity bias in reframing shapes political preferences even in partisan contexts. *Social Psychological and Personality Science,* **10**, 53–61 (2019).

[30] Ito, T. A., Larsen, J. T., Smith, N. K., & Cacioppo, J. T. Negative information weighs more heavily on the brain: The negativity bias in evaluative categorizations. *Journal of Personality and Social Psychology* **75**, 887–900 (1998).

[31] Öhman, A., & Mineka, S. Fears, phobias, and preparedness: Toward an evolved module of fear and fear learning. *Psychological Review* **108**, 483–522 (2001).

[32] Öhman, A., Flykt, A., & Esteves, F. Emotion drives attention: Detecting the snake in the grass. *Journal of Experimental Psychology: General* **130**, 466–478 (2001).

[33] Shoemaker, P. J. Hardwired for news: Using biological and cultural evolution to explain the surveillance function. *Journal of Communication* **46**, 32–47 (1996).

[33] Stieglitz, S. & Dang-Xuan, L. Emotions and information diffusion in social media: Sentiment of microblogs and sharing behavior. *Journal of Management Information Systems* **29**, 217–248 (2013).

[35] Naveed, N., Gottron, T., Kunegis, J. & Alhadi, A. C. Bad news travel fast: A content-based analysis of interestingness on Twitter. In *Proceedings of the 3rd International Web Science Conference* (2011).

[36] Kim, J. & Yoo, J. Role of sentiment in message propagation: Reply vs. retweet behavior in political communication. In *International Conference on Social Informatics*, 131–136 (2012).

[37] Berger, J., & Milkman, K. L. What makes online content viral? *Journal of Marketing Research* **49**, 192-205 (2012).

[38] Chuai, Y. & Zhao, J. Anger makes fake news viral online. https://arxiv.org/pdf/2004.10399 (2020).

[39] Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences (PNAS)* **114**, 7313–7318 (2017).

[40] Pröllochs, N., Bär, D., & Feuerriegel, S. Emotions explain differences in the diffusion of true vs. false social media rumors. *Scientific Reports* **11**, 22721 (2021).

[41] Pröllochs, N., Bär, D., & Feuerriegel, S. Emotions in online rumor diffusion. *EPJ Data Science*, **10**, 51 (2021).

[42] Zollo, F., et al. Emotional dynamics in the age of misinformation. *PLOS ONE* **10**, e0138740. (2015).

[43] Rathje, S., Van Bavel, J. J., & van der Linden, S. Out-group animosity drives engagement on social media. *Proceedings of the National Academy of Sciences* **118** (2021).

[44] Soroka, S., Fournier, P., & Nir, L. Cross-national evidence of a negativity bias in psychophysiological reactions to news. *Proceedings of the National Academy of Sciences (PNAS)* **116**, 18888-18892 (2019).

[45] Trussler, M., & Soroka, S. Consumer demand for cynical and negative news frames. *The International Journal of Press/Politics* **19**, 360-379 (2014).

[46] Meffert, M. F., Chung, S., Joiner, A. J., Waks, L., & Garst, J. The effects of negativity and motivated information processing during a political campaign. *Journal of Communication* **25**, 27-51 (2006).

[47] Bradley, S. D., Angelini, J. R., & Lee, S. Psychophysiological and memory effects of negative political ADS: Aversive, arousing, and well remembered, *Journal of Advertising* **36**, 115-127 (2007).

[48] Soroka, S., & McAdams, S. News, politics, and negativity. *Political Communication* **32**, 1–22 (2012)

[49] Lengauer, G., Esser, F., & Berganza, R. Negativity in political news: A review of concepts, operationalizations and key findings. *Journalism* **13**, 179–202 (2012).

[50] Jang, S. M., & Oh, Y. W. Getting attention online in election coverage: Audience selectivity in the 2012 US presidential election. *New Media & Society* **18**, 2271–2286 (2016).

[51] Haselmayer, M., Meyer, T. M., & Wagner, M. Fighting for attention: Media coverage of negative campaign messages. *Party Politics* **25**, 412–423 (2019).

[52] Soroka, S. N. Good News and Bad News: Asymmetric Responses to Economic Information. *The Journal of Politics* **68**, 372–385 (2006).

[53] Matias, J., Munger, K., Le Quere, M.A., Ebersole, C. The Upworthy Research Archive, a time series of experiments in U.S. media. *Nature Scientific Data* **8**, 195 (2021)

[54] Karpf, D. *Analytic Activism: Digital Listening and the New Political Strategy.* Oxford Studies in Digital Politics (Oxford University Press, New York, NY, 2016).

[55] Thompson, D. I thought I knew how big Upworthy was on Facebook: Then I saw this. *The Atlantic*. https://www.theatlantic.com/business/archive/2013/12/i-thought-i-knew-how-big-upworthy-was-on-facebook-then-i-saw-this/282203/ (2012).

[56] Fitts, A. S. The king of content: How Upworthy aims to alter the web, and could end up altering the world. C*olumbia Journalism Review* **52** (2014).

[57] Upworthy. How to make that one thing go viral, *SlideShare* https://www.slideshare.net/Upworthy/how-to-make-that-one-thing-go-viral-just-kidding/25 (2012).

[58] Soroka, S., Young, L., & Balmas, M. Bad news or mad news? Sentiment scoring of negativity, fear, and anger in news content. *The ANNALS of the American Academy of Political and Social Science* **659,** 108-121 (2015).

[59] Fox, Elaine, Victoria Lester, Riccardo Russo, R. J. Bowles, Alessio Pichler, and Kevin Dutton. Facial expressions of emotion: Are angry faces detected more efficiently? *Cognition & Emotion* **14**, 61-92 (2000)

[60] De Gelder, B. Towards the neurobiology of emotional body language. *Nature Reviews Neuroscience* **7**, 242-249 (2006).

[61] Brady, W. J., Crockett, M. J., & Van Bavel, J. J. The MAD model of moral contagion: The role of motivation, attention, and design in the spread of moralized content online. *Perspectives on Psychological Science* **15**, 978-1010 (2020).

[62] Spring, V. L., Cameron, C. D., & Cikara, M.. The upside of outrage. *Trends in Cognitive Sciences* **22**, 1067-1069 (2018).

[63] Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. The development and psychometric properties of LIWC2015. http://hdl.handle.net/2152/31333 (2015).

[64] Khalilzadeh, J., & Tasci, A. D. Large sample size, significance level, and the effect size: Solutions to perils of using big data for academic research. *Tourism Management* **62**, 89-96 (2017).

[65] Cohen, J. *Statistical power analysis for the behavioral sciences.* (Academic Press, 2013).

[66] Lakens, D., Scheel, A. M., & Isager, P. M. Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science* **1**, 259-269 (2018).

[67] Kross, E., et al. Does counting emotion words on online social networks provide a window into people's subjective experience of emotion? A case study on Facebook. *Emotion* **19**, 97 (2019).

[68] Song, H. et al. In validations we trust? The impact of imperfect human annotations as a gold standard on the quality of validation of automated content analysis. *Political Communication* **37**, 550–572 (2020).

[69] Gunning, R. *The Technique of Clear Writing.* (McGraw-Hill, Toronto, CA, 1952).

[70] Richardson, M., Dominowska, E. & Ragno, R. Predicting clicks: Estimating the click-through rate for new ads. In *Proceedings of the 16th International Conference on World Wide Web*, 521–530 (2007).

[71] Mohammad, S. & Turney, P. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text,* 26–34 (2010).

[72] Mohammad, Saif M. Sentiment analysis: Detecting valence, emotions, and other affectual states from text. *Emotion Measurement*. 201-237, (2016).

[73] Plutchik, R. *Emotion: Theory, research, and experience* (Academic Press, Orlando, FL, 1984), 2 edn.

[74] Ekman, P. An argument for basic emotions. *Cognition & Emotion* **6**, 169–200 (1992).

[75] Sauter, D. A., Eisner, F., Ekman, P. & Scott, S. K. Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations. *Proceedings of the National Academy of Sciences (PNAS)* **107**, 2408–2412 (2010).

[76] Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology,* **61**, 2544-2558. (2010).

[77] Graham, J., Haidt, J., & Nosek, B. A. Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology* **96**, 1029. (2009).

[78] Curiskis, S. A., Drake, B., Osborn, T. R., & Kennedy, P. J. An evaluation of document clustering and topic modelling in two online social networks: Twitter and Reddit. *Information Processing & Management* **57**, 102034. (2020).

[79] Cer, D., et al. Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 169-174 (2018).

[80] Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., & Blei, D. M. Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems*, 288-296 (2009).

[81] Dunn, O. J. Multiple comparisons among means. *Journal of the American Statistical Association* **56**, 52–64. (1961).

Supplementary Materials

A Preliminary results from pilot analysis

The following analysis reports a pilot analysis based on the exploratory sample of the data for generating hypotheses. In stage 2 of the Registered Report, the numbers will be updated to include the results obtained for the confirmatory sample of the data (but for which access in stage 1 of the pre-registration is physically prohibited). When reporting estimates, we abbreviate standard errors via "SE" and 99% confidence intervals via "CI".

A.1 Randomized controlled trials comparing news consumption

The preliminary dataset contains a total of N = 4,873 RCTs. After applying the filtering procedure, we obtain 2,602 RCTs. Each RCT compares different variations of news headlines that all belong to the same news story. For example, the headline "*WOW: Supreme Court Have Made Millions Of Us Very, Very Happy"* and *"We'll Look Back At This In 10 Years Time And Be Embarrassed As Hell It Even Existed"* are different headlines used for the same story about the repeal of Proposition 8 in California. The headline variations are then compared with respect to the generated click-through rate, defined as the ratio of clicks per impressions (see Table 1 in the main text for examples). Overall, the 2,602 RCTs comprise 11,109 different headlines, which received ~43.38 million impressions and 586,660 clicks. The final dataset (i.e., the confirmatory sample) will be based on N = 22,743 RCTs with ~105,000 different variations that received more than 530 million impressions and ~8 million clicks [53].

In the experiments, the recorded click-through rate ranges from 0.00% to 13.60%. The average click-through rate across all experiments is 1.39%. Furthermore, the distribution among click-through rates is right-skewed, indicating that only a small proportion of news stories were associated with a high click-through rate (Figure 1A). For instance, 99% of headline variations have a click-through rate below 6%. The results lay the groundwork for identifying the drivers of high levels of news consumption.

There are considerable differences between positive and negative language in news headlines (Figure 1B). We find that positive words are more prevalent than negative words (Kolmogorov-Smirnov (KS) test: $D = 0.092$, $p < 0.001$). Overall, 3.80% of all words in news headlines are categorized as positive words, whereas 2.81% of all words are categorized as negative words. In our sample, the most common positive words are "love" ($n = 218$), "pretty" ($n = 157$), and "beautiful" ($n = 123$), and the most common negative words are "wrong" ($n = 135$), "bad" ($n = 123$), and "hate" ($n = 72$). Ninety percent (91.97%) of the news stories in our sample contain a headline with at least one positive or negative word (i.e., 2393 out of a possible 2602), and 64.52% of headlines contain at least one word from our dictionaries (i.e., 7168 out of a possible 11,109). Further statistics with word frequencies are in Supplement C.

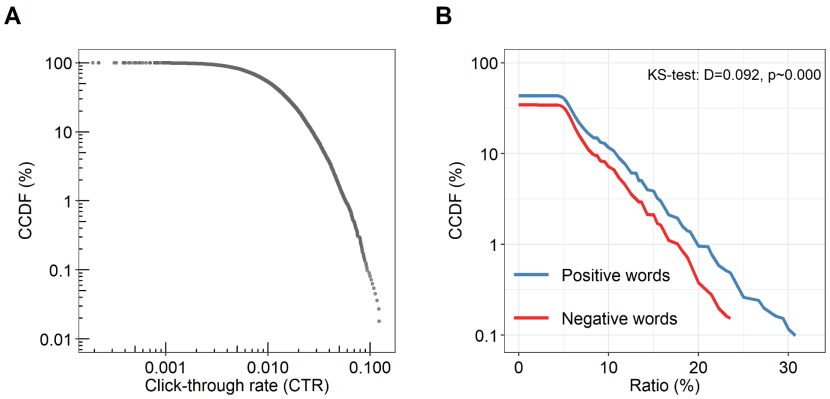

Figure 1: **(A)** Complementary cumulative distribution function (CCDF) comparing the click-through rate across all headline variations. **(B)** CCDF comparing the distribution of the ratio of positive and negative words across all headline variations. Positive words are more prevalent than negative words. A Kolmogorov-Smirnov (KS) test shows that this difference is statistically significant. The y-axes of both plots are on a logarithmic scale.

A.2 Effect of negative language on news consumption

Randomized controlled experiments are used to estimate the effect of positive and negative words on news consumption, that is, the click-through rate. We employ a multilevel binomial regression that accommodates a random effects specification to capture heterogeneity among news stories (for details, see Methods). Detailed coefficient estimates are provided in Supplement E. Further analyses are later included in the robustness checks.

Positive and negative language in news headlines are an important determinant of click-through rates (Figure 2A−C). Consistent with the negativity bias hypothesis, the effect for negative words is positive (coef: 0.010, SE = 0.002, $z = 4.901$, $p < 0.001$, CI = [0.006, 0.013]), suggesting that a larger proportion of negative words increases the propensity of users to access a news story. A one standard deviation larger proportion of negative words increases the odds of a

user clicking the headline by 1.0%. For a headline of average length (14.9 words), this implies that, for each negative word, the click-through rate increases by 1.5%. In contrast, the coefficient for positive words is negative (coef: −0.015, SE = 0.002, $z = -7.781$, $p < 0.001$, CI = [−0.018, −0.011]), implying that a larger proportion of positive words result in fewer clicks. For each standard deviation increase in the proportion of positive words per headline, the likelihood of a click decreases by 1.5%. Put differently, for each positive word in a headline of average length, the click-through rate decreases by 1.9%.

The estimated effects hold when adjusting for length and text complexity. A longer news headline increases the click-through rate (coef: 0.042, SE = 0.002, $z = 20.630$, $p < 0.001$, CI = [0.038, 0.046]). The click-through rate is also increased by a higher complexity score (coef: 0.013, SE = 0.002, $z = 6.640$, $p < 0.001$, CI = [0.009, 0.017]), yet to a smaller extent. This finding implies that lengthier and more complex formulations are appealing to users and lead to higher levels of news consumption. The control for platform age is negative (coef: −0.318, SE = 0.012, $z = -26.535$, $p < 0.001$, CI = [−0.341, −0.294]). Hence, stories published later in Upworthy's career had lower click-through rates than stories published at the beginning of Upworthy's career, implying that Upworthy headlines were most successful when its editorial practices were novel to online users.

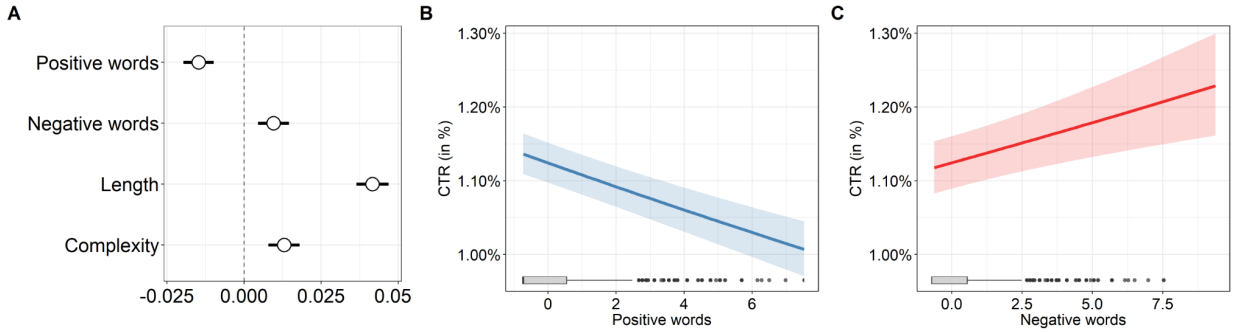

Figure 2: The effect of positive and negative words in news headlines on the click-through rate. **(A)** Shown are the estimated standardized coefficients with 99% confidence intervals for positive and negative words and for further controls. The variable *PlatformAge* is included in the model during estimation but not shown for better readability. **(B,C)** Predicted marginal effects on the click-through rate (with 99% confidence intervals). Boxplots show the distribution of the variables in our sample (center line gives the median; box limits are upper and lower quartiles; whiskers denote minimum/maximum; points are outliers defined as being beyond 1.5x of the interquartile range).

A.3 Regression analysis with varying-slopes

Following our pre-registration, we further report results from a regression analysis with random effects and additional varying-slopes in the sentiment variables (Table 3). As such, the receptivity to language is no longer assumed to be equal across all experiments but is allowed to vary. Again, the coefficients are negative for positive words and positive for negative words. This thus implies that positive language decreases the clickability of news headlines, while negative language increases it. Furthermore, this is consistent with the analysis based on a random effects model without varying-slopes.

Altogether, we find that a higher share of negative language in news headlines increases the click-through rate, whereas a higher share of positive language decreases the click-through rate. It is important to note that headlines belong to the *same* news story, and, therefore, phrasing

news—regardless of its story—in a negative language increases the rate of clicking on a headline.

	Coef	Lower CI	Upper CI	P-value
Positive	−0.022	−0.037	−0.008	< 0.001
Negative	0.015	0.001	0.029	0.006
Length	0.045	0.038	0.051	< 0.001
Complexity	0.012	0.005	0.018	< 0.001
PlatformAge	−0.319	−0.350	−0.287	< 0.001
(Intercept)	−4.502	−4.535	−4.470	< 0.001
Observations: 11,109				

Table 3: Results for the regression model with varying-slopes for the proportion of positive and negative words. Experiment-specific intercepts (i.e., random effects) are also included. Reported are standardized coefficient estimates (standard errors in parentheses).

A.4 Robustness checks

The robustness of our preliminary analysis was confirmed by a series of further checks (see Supplement F). First, we repeated the analysis with alternative sentiment dictionaries as an additional validation. We found that the coefficient estimates were in good agreement, contributing to the robustness of our results. Second, we repeated the above main analysis with an alternative approach for negation handling (i.e., a different neighborhood for inverting the polarity of words). This approach led to qualitatively identical results. Third, we repeated the analyses above using alternate text complexity measures. We found that the results remain qualitatively the same. Fourth, we controlled for quadratic effects. We still observed a dominant effect of negative language. Fifth, we repeated the analyses above but removed headlines where both positive and negative words were simultaneously present. As such, we end up with all headlines that exclusively include either positive or negative words. We found that headlines

with negative words continued to be more likely to be clicked on than headlines with positive words. Sixth, we repeated the same analyses as above, but removed all image RCTs where the teaser images were varied. This approach led to qualitatively identical results. Seventh, we computed a single sentiment score, which is given by the net difference between the proportion of positive words and the proportion of negative words. As expected, a negative sentiment increases click-through rate. Details on all robustness checks can be found in Supplement F.

We investigated moralized language as a possible moderator of emotions in driving the click-through rate (Supplement G). Previously, moralized language was identified as an important driver of the dif- fusion of social media content [39]. We extended the regression models from our main analysis with interaction terms between the proportion of moral words per headline and the variables for the proportion of positive and negative words. In addition, we included the proportion of moral words per headline as a regressor to estimate its direct effect. We found a negative and statistically significant direct effect of moralized language on click-through rate (coef: −0.023, SE = 0.003, $z = -7.650$, $p < 0.001$, CI = [−0.031, −0.015]) and negative and statistically significant effects for the interactions between the proportion of moral words and the proportion of positive (coef: −0.007, SE = 0.003, $z = -2.673$, $p = 0.007$, CI = [−0.014, −0.0003]) and negative words (coef: −0.006, SE = 0.002, $z = -2.619$, $p = 0.009$, CI = [−0.012, −0.0001]). The results thus point towards a direct and a moderating role of moralized language. Yet even when controlling for moralized language, the direct effect of negative language was still present and continues to support the negativity bias hypothesis.

A.5 Negativity effect across different news topics

We examined the effect of negative language across various news topics (see Supplement H). The rationale is that news stories in our data comprise various topics, for which the effect of

emotion on the click-through rate could potentially differ. To this end, we applied topic modeling as in earlier research (e.g., [22]). Topic modeling infers a categorization of large-scale text data through a bottom-up procedure, thereby grouping similar content into topics. The procedure led to 7 topics, which were named "Entertainment," "Government & Economy," "LGBT," "Parenting & School," "People," and "Women Rights & Feminism." Representative headlines for each topic are shown in Supplement H.

Subsequently, we validated whether the topic labels provide meaningful representations. In a user study, participants were shown headlines from each topic and were asked to select which topic the headline best fit into (topic intrusion test). Participants ($k = 10$) identified the topic from the correct headline in 51.1% of the cases. For comparison, a random guess would lead to an accuracy of 25%, implying that participants are roughly twice as good. This improvement over the random guess was further statistically significant ($\chi 2 = 249.61$, $p < 0.001$). Details are in Supplement H.

We found significant differences in the baseline click-through rate among topics. For this, we estimated a model where we additionally control for different topics via dummy variables, thus capturing the heterogeneity in how different topics generate clicks. Keeping everything else equal, we find that news generated more clicks when covering stories related to "Entertainment," "LGBT," and "People." In contrast, news related to "Government & Economy" have a lower clickability. Full results are in Supplement H.

We then controlled for how the effect of negative language might vary across different topics. Here, we found that the variables of interest (i.e., the proportion of positive and negative words) significantly interact with different topics. For example, headlines relating to "Government and Economy," "LGBT," "Parenting and School," and "People" received more

clicks when they contained a large share of negative words. We also found that headlines relating to "LGBT," "Life," "Parenting and School," and "People" received fewer clicks when they contained a large share of positive words. Overall, we found that negative language still has a statistically significant positive effect on the click-through rate (see Supplement H). In sum, these results are consistent with the main analysis.

A.6 Extension to discrete emotions

We conducted secondary analyses examining the effects of discrete emotional words on the click-through rate (see Supplement E). Prior work has suggested that certain discrete emotions such as anger [38, 41] may be particularly prevalent in online news. Furthermore, discrete emotions were found to be important determinants for various forms of user interactions (e.g., sharing [36, 37, 38, 39, 40, 41]), thus motivating that discrete emotions may also play a role for news consumption.

We report findings from four emotions (anger, fear, joy, sadness) for which we found statistically significant positive correlations between the human judgments of emotions and the dictionary scores (see Methods). We observed a statistically significant and positive coefficient for sadness (coef: 0.009, SE = 0.002, $z = 3.915$, $p < 0.001$, CI = [0.003, 0.015]). A one standard deviation increase in sadness increases the odds of a user clicking the headline by 0.9%. The coefficient estimates for anger (coef: 0.000, SE = 0.002, $z = -0.097$, $p = 0.992$, CI = [−0.006, 0.006]), fear (coef: −0.005, SE = 0.002, $z = -1.881$, $p = 0.060$, CI = [−0.011, 0.002]), and joy (coef: −0.004, SE = 0.003, $z = -1.476$, $p = 0.140$, CI = [−0.011, 0.003]) were not statistically significant at common statistical significance thresholds. Consistent with our previous findings, we observed that the click-through rate increases as the text length and complexity score increase. Again, the click-through rate was lower for headlines at the end of Upworthy's career.

The above findings are supported by additional checks. First, we controlled for different topics in our regression model and, for this, utilized the previous categorization via topic modeling. When including topic dummies, we still found statistically significant positive effects for sadness, whereas the coefficients for anger, fear, and joy were not statistically significant (see Supplement H). These results are thus consistent with the main analysis. For thoroughness, we also analyzed the effects of the four other basic emotions from the NRC emotion dictionary for which the correlation with human judgments was considerably lower in our validation study (Supplement I). Here we observed a statistically significant negative effect on the click-through rate for anticipation. Furthermore, we studied the effects of 24 emotional dyads from Plutchik's model [73], which are complex emotions composed of two basic emotions [73]. We found that several dyads such as outrage and disapproval are associated with higher click-through rates.

B Descriptive statistics

Table 4 reports descriptive statistics on the exploratory dataset used in the pilot analysis. *Clicks* denotes the raw number of clicks that a give headline received. *Impressions* denotes the number of Upworthy user that were assigned a given headline. The *CTR* gives the click-through rate, that is, the ratio of clicks per impression.

Word counts for sentiment and emotional words (i. e., before z-standardization) are as follows. $Positive$ and $Negative$ describe the percentage of words in each headline that belong to the positive and negative word lists in the LIWC dictionary, respectively. $Anger$, $Anticipation$, $Disgust$, $Fear$, Joy, $Sadness$, $Surprise$, and $Trust$ are the scores for the 8 basic emotions calculated based on the NRC emotion lexicon. These scores range between zero and one and sum up to one across the basic emotions.

Further controls are as follows. *Length* is the number of words in each headline. $Complexity$ gives the Gunning-Fog index score for each headline. $PlatformAge$ is the age of the platform, that is, the number of days since the first ever Upworthy experiment being conducted. For example, a headline with a value of 100 for $PlatformAge$ was published 100 days after the first Upworthy story.

We further elaborate the dependence structure between positive and negative words. Figure 3 visualizes the density of the proportion of positive and negative words in headlines. The figure shows that a large density in the bottom-left corner, representing headlines where both positive and negative words from the LIWC are absent (35.5 %). In addition, we find that headlines often contain exclusively positive (30.1 %) or exclusively negative words (20.9 %). This motivates later one of our robustness checks where we perform a regression analysis based on headlines with positive-only and negative-only dictionary words. Only 13.5 % of all headlines contain both positive and negative words. The overall correlation (Pearson's r) between the proportion of positive and negative words in headlines is -0.075 ($p < 0.001$).

Variable	Mean	Median	Min	Max	Std. Dev.
Outcomes					
Clicks	52.809	37.000	0.000	883.000	57.261
Impressions	3904.977	3214.000	17.000	45,907.000	2340.758
CTR (click-through rate)	0.014	0.011	0.000	0.136	0.012
Dictionary-based variables					
Positive	0.038	0.000	0.000	0.429	0.052
Negative	0.028	0.000	0.000	0.444	0.044
Anger	0.099	0.000	0.000	1.000	0.186
Anticipation	0.159	0.000	0.000	1.000	0.238
Disgust	0.069	0.000	0.000	1.000	0.159
Fear	0.135	0.000	0.000	1.000	0.213
Joy	0.142	0.000	0.000	1.000	0.211
Sadness	0.105	0.000	0.000	1.000	0.187
Surprise	0.064	0.000	0.000	1.000	0.149
Trust	0.227	0.143	0.000	1.000	0.296
Control variables					
Length	14.892	15.000	3.000	24.000	3.145
Complexity	8.501	8.277	0.600	29.067	3.700
PlatformAge	484.249	534.000	0.000	823.000	206.719

Table 4: Descriptive statistics.

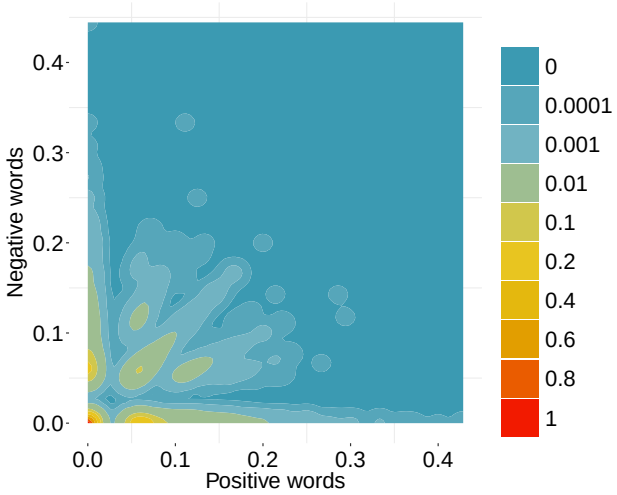

Figure 3: Dependency between positive and negative words. The density plot shows the relationship between the proportion of positive and negative words in headlines. Red (blue) corresponds to a higher (lower) density. Density is normalized to go from 0 to 1.

C Frequency of dictionary words

The most common emotional words from the NRC emotion lexicon (Figure 12) are categorized as belonging to *trust*, for which the average relative proportion of all emotional words in headlines amounts to 22.7%. This is followed by *anticipation* (15.9%) and *joy* (14.2%). In contrast, emotional words belonging to *surprise* (6.4%) and *disgust* (6.9%) are less frequent.

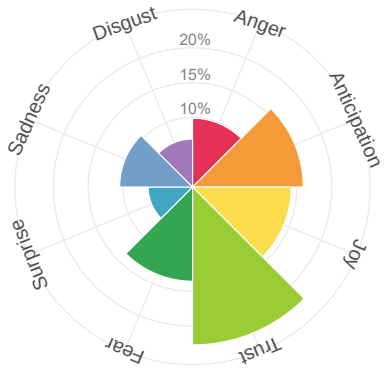

Figure 4: Average relative proportion of emotional words in news headlines. Here, the categorization involves eight basic emotions as provided by the NRC emotion lexicon [71, 72].

A list of the most frequent dictionary words in our dataset is given in Table 5 (positive and negative words) and Table 6 (basic emotions from NRC emotion lexicon). Note that words that appear unexpected at a first glance are often used in a context that is characterized by a specific emotion. For example, the term "boy" is often part of the expression "Oh boy! ..." where it is used to signal strong opposition and even *disgust*. Similarly, the term "watch" was often used in the context of "watch out" where, as a result, the headline was perceived as communicating *fear*. For details on why specific words were classified by users in a large-scale study as embedding a specific emotion, we refer to the original paper developing the NRC emotion lexicon [71, 72].

C.1 Frequency of positive and negative dictionary words

Positive		*Negative*	
Word	Frequency	Word	Frequency
love	218	wrong	135
pretty	157	bad	123
beautiful	123	hate	72
amazing	106	awful	66
funny	95	war	66
happy	93	rape	59
save	91	fight	57
awesome	90	argument	55
care	86	worst	52
super	84	tears	49

Table 5: Top 10 most frequent positive and negative words, as defined by the LIWC dictionary, in our sample.

C.2 Frequency of emotional words from NRC emotion lexicon

Anger		*Anticipation*		*Disgust*		*Fear*	
Word	Frequency	Word	Frequency	Word	Frequency	Word	Frequency
words	151	time	327	bad	123	watch	324
bad	123	watch	324	boy	76	change	159
money	115	pretty	157	hate	72	bad	123
hate	72	money	115	powerful	68	hate	72
powerful	68	white	103	awful	66	powerful	68
awful	66	sex	94	death	64	awful	66
death	64	happy	93	rape	59	war	66
rape	59	wait	80	john	54	death	64
fight	57	happen	79	congress	43	government	62
argument	55	powerful	68	sick	43	rape	59
Joy		*Sadness*		*Surprise*		*Trust*	
Word	Frequency	Word	Frequency	Word	Frequency	Word	Frequency
love	218	black	142	money	115	real	211
pretty	157	bad	123	hilarious	71	pretty	157
beautiful	123	hate	72	guess	67	school	134
money	115	music	67	death	64	money	115
food	111	awful	66	hope	52	food	111
white	103	death	64	deal	44	white	103
sex	94	rape	59	finally	42	sex	94
happy	93	die	50	break	40	word	94
save	91	crazy	46	leave	39	happy	93
found	75	sick	43	vote	35	save	91

Table 6: Top 10 most frequent words for each of the 8 basic emotions as defined by the NRC emotion lexicon, in our sample.

D User studies to validate dictionary approach

In line with best practices [68], we re-validated both the LIWC dictionary and the NRC emotion lexicon for our setting. For this, we conducted two user studies.

User study 1

In user study 1, we validated that user judgments of positivity/negativity and our computed ratings of sentiment from the LIWC dictionary were significantly correlated. Participants were recruited from the New York University subject pool, provided informed consent, and viewed a total of 213 headlines drawn randomly from the exploratory dataset. All participants were native English speakers. To avoid fatigue, participants and headlines were split into two groups, so that each had to respond to only a subset of all questions. We recruited two groups of $k = 10$ participants; after removing participants who failed to complete the study, we were left with one group of 8 raters, and one group of 10 raters, a standard number of raters for validations in prior literature [68]. The number of headlines ($N = 213$) and the number of raters were chosen based on best practices [68]. Specifically, 50 RCTs were randomly selected from the 2,602 RCTs in our filtered preliminary sample. All headlines in an RCT package were included for a total of 213 headlines to be tested. For each headline, participants were asked *"How negative or positive is this headline?"* Participants rated each headline on a -3 (very negative) to $+3$ (very positive) Likert scale. We refer to the score as "sentiment" in the following.

We first assessed the inter-rater agreement using Kendall's W. The inter-rater agreement was statistically significant ($W = 0.33$, $p < 0.001$).

Both user ratings and dictionary scores (as used in our main analysis) are not directly comparable. The reason is that user ratings refer to an *overall* sentiment (on a scale from negative to positive), whereas the independent variables are two *separate* scores for positivity and negativity. Hence, we show that both ratings and dictionary scores are related in the following ways:

- We separately compare the sentiment rating with the positivity and negativity scores (i. e.,

$Positive$ and $Negative$, respectively). Reassuringly, we use the same dictionary approach as in the main paper, including negation handling. The statistical comparison is based on Spearman's rank correlation coefficient (r_s). For positivity, the correlation is $r_s = 0.20$ and statistically significant ($p = 0.004$). For negativity, the correlation is $r_s = -0.20$ and statistically significant ($p = 0.003$). Hence, changes in the proportion of positive and negative words in a headline are also reflected in the perceived sentiment of raters.

- We map the two separate dictionary scores onto a combined sentiment score. For this, we compute the net difference between positivity and negativity in the text (i. e., $Sentiment = Positive - Negative$). We then compare the sentiment ratings against the dictionary-based sentiment scores. Specifically, we compute Spearman's rank correlation coefficient (r_s) between the mean rantings of the 8 human judges' scores with the dictionary scores. User ratings of sentiment and computed sentiment scores were weakly but significantly correlated with one another ($r_s = 0.30$, $p < 0.001$).

Altogether, this validates that, for our news headlines, user perceptions of negativity and computed negativity scores are related. Importantly, this result also confirms that dictionary words are subject to additivity, that is, that a headline that includes two negative words is perceived as being more negative than a headline that includes only one negative word.

User study 2

In user study 2, we validated that user judgments of discrete emotion and our computed emotion scores from the NRC emotion lexicon were significantly correlated. Again, participants were recruited from the NYU subject pool, were native English speakers, provided informed consent, and viewed a total of 213 headlines drawn randomly from the exploratory dataset. To avoid fatigue, four groups of participants were recruited, so that each had to respond to only a subset of questions. The number of raters ($k = 10$) and headlines ($N = 213$) were again chosen based on best practices [68]. One participant was removed for failing to complete the study, leaving three groups of 10

raters, and one group of 9 raters. For each headline, participants were asked *"How much _____ is present in this headline?"* The blank space in the question was repeatedly replaced by all of the 8 basic emotions from the NRC emotion lexicon (i. e., $Anticipation$, $Disgust$, $Fear$, Joy, $Sadness$, $Surprise$, $Trust$). This corresponds to $213 \times 8 = 1704$ questions. For each headline, participants gave ratings for all 8 emotions on a 1 (no _____) to 7 (a great deal of _____) Likert scale.

The inter-rater agreement is listed in Table 7. It was statistically significant for 7 of the discrete emotions ($Anger$, $Anticipation$, $Disgust$, $Fear$, $Sadness$, and $Disgust$).

We found that the overall correlation between NRC dictionary scores and the mean ratings of the user judgments for the 8 discrete emotions was positive and statistically significant ($r_s = 0.11$; $p < 0.001$). The correlations for the mean user user ratings of each emotion and the computed emotion score are presented in Table 8. For specific emotions, user judgments for $Anger$, $Fear$, Joy, and $Sadness$ were significantly correlated with the computed emotion scores. For these emotions, the results validate that emotion ratings from users and NRC dictionary scores are, to a large extent, meaningfully related.

In our regression analysis, we focus the four discrete emotions for which we found statistically significant positive correlation between the perceptions of emotions and the computed NRC emotion scores (i. e., *Anger*, *Fear*, *Joy*, *Sadness*). For thoroughness, we also analyze the effects of all other discrete emotions (i. e., $Anticipation$, $Disgust$, $Surprise$, $Trust$) in Supplement I.

Emotion	Kendall's W	P-value
Sentiment	0.33	< 0.001
Anger	0.24	< 0.001
Anticipation	0.17	< 0.001
Disgust	0.22	< 0.001
Fear	0.23	< 0.001
Joy	0.15	0.008
Sadness	0.23	< 0.001
Surprise	0.13	0.071
Trust	0.19	< 0.001

Table 7: Kendall's W coefficient for the inter-rater agreement between users.

Emotion	Correlation	P-value
Anger	0.22	0.005
Anticipation	-0.07	0.341
Disgust	0.01	0.926
Fear	0.29	< 0.001
Joy	0.30	0.002
Sadness	0.30	< 0.001
Surprise	-0.06	0.414
Trust	0.12	0.122

Table 8: Spearman's rank correlation coefficient (r_s) between user judgments and dictionary scores for emotional words.

E Estimation results

Detailed estimation results for all model parameters are reported for our main analysis examining the role of positive and negative words (Table 9). As secondary analyses, we study the role of discrete emotions from the NRC emotion lexicon (Table 10). Here we focus on the four discrete emotions for which we found statistically significant positive correlation between the perceptions of emotions and the computed NRC emotion scores (i. e., *Anger*, *Fear*, *Joy*, *Sadness*).

E.1 Main analysis

	Coef	Lower CI	Upper CI	P-value
Positive	-0.015	-0.020	-0.010	< 0.001
Negative	0.010	0.005	0.015	< 0.001
Length	0.042	0.036	0.047	< 0.001
Complexity	0.013	0.008	0.018	< 0.001
PlatformAge	-0.318	-0.349	-0.287	< 0.001
(Intercept)	-4.487	-4.519	-4.455	< 0.001

Observations: 11,109

Table 9: Regression model explaining click-through rate based on positive and negative words in headlines. Reported are standardized coefficient estimates and $99\,\%$ CIs. Experiment-specific intercepts (i. e., random effects) are included.

E.2 Further analysis for discrete emotions

	Coef	Lower CI	Upper CI	P-value
Anger	0.000	−0.006	0.006	0.992
Fear	−0.005	−0.011	0.002	0.060
Joy	−0.004	−0.011	0.003	0.140
Sadness	0.009	0.003	0.016	< 0.001
Length	0.043	0.037	0.050	< 0.001
Complexity	0.010	0.004	0.016	< 0.001
PlatformAge	−0.323	−0.355	−0.291	< 0.001
(Intercept)	−4.506	−4.539	−4.472	< 0.001

Observations: 8365

Table 10: Regression model explaining click-through rate based on discrete emotions in headlines. Reported are standardized coefficient estimates. Experiment-specific intercepts (i.e., random effects) are included.

F Robustness checks

F.1 Analysis with alternative sentiment dictionaries

In our main analysis, we use positive and negative words from the LIWC. We now validate our results based on alternative word lists. Specifically, we compare the estimates from the following dictionaries:

1. **LIWC** (main paper) [63] We compute scores for positive and negative words (i. e., $Positive$ and $Negative$) using the built-in dictionary from Linguistic Inquiry and Word Count (LIWC). The estimation results are those from the main paper.

2. **NRC** [71] The NRC emotion lexicon comprises 181,820 English words that are classified into positive and negative words [65]. We use the implementation from the `sentimentr` package to calculate the proportion of positive words ($Positive_{\text{NRC}}$) and negative words ($Negative_{\text{NRC}}$) in headlines.

3. **SentiStrength.** SentiStrength is a sentiment dictionary that was primarily developed for short social media texts [76] SentiStrength returns two integer scores, namely $Negative_{\text{SS}} \in [-5, -1]$ for the negative sentiment and $Positive_{\text{SS}} \in [1, 5]$ for the positive sentiment. Note that, in SentiStrength, lower values of $Negative_{\text{SS}}$ correspond to more negative sentiment. We thus multiply $Negative_{\text{SS}}$ by -1 to facilitate comparability to the other dictionaries.

Based on the above dictionaries, we then fitted separate regression models, one for each sentiment score. We again used z-standardization for better comparisons. Overall, the estimated 99 % confidence intervals (CIs) from all models are in good agreement (Figure 5). In particular, the regression models suggest that negative words increase click-through rates. This finding is consistent across all considered dictionaries.

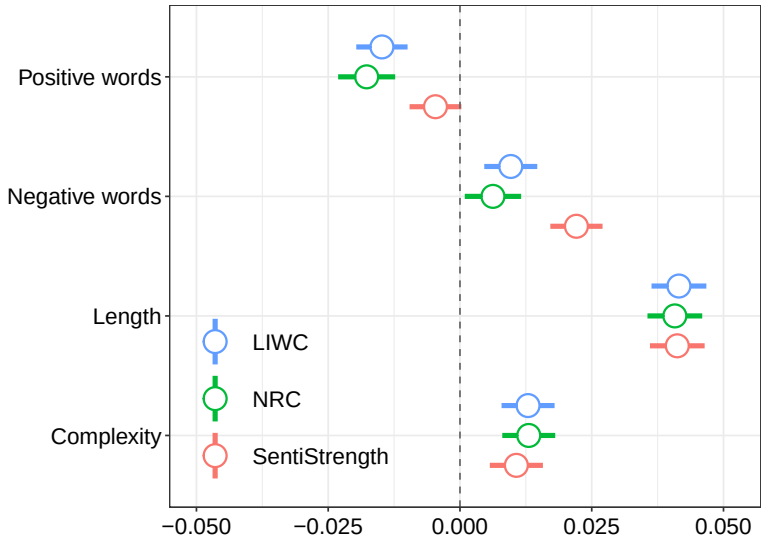

Figure 5: Comparison showing that the effect of negative words is robust across different sentiment dictionaries. The lines correspond to the 99 % confidence intervals (CIs). The variable *PlatformAge* is included in the model during estimation but not shown for better readability.

F.2 Negation handling

Our main analysis accounts for negations by counting all words in the neighborhood of a negation word (e. g., "not," "no") as belonging to the opposite word list. In our analysis, the neighborhood (i. e., the so-called negation scope) was set to 3 words after the negation. As a robustness check, we experiment with an alternative neighborhood of 5 words before and 2 words after the negation. Here, the same list of negations as in the main paper is used. We then compare the coefficient estimates for the two different approaches to negation handling. Overall, we find high agreement for positive and negative words (Figure 6). In fact, all 99 % confidence intervals (CIs) overlap and, hence, yield similar results.

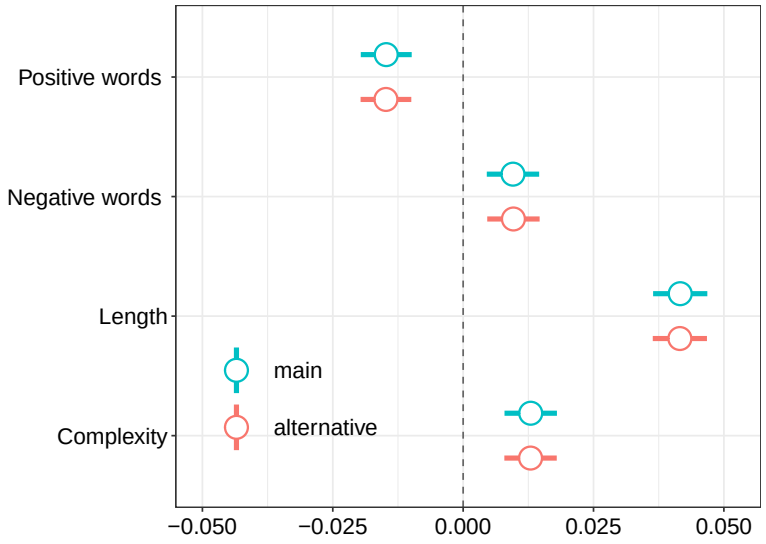

Figure 6: Comparison of the effects of emotional words across two different approaches to negation handling. The "main" approach uses a neighborhood of 3 words after the negation. The "alternative" approach uses a neighborhood of 5 words before and 2 words after the negation. The lines correspond to the 99 % confidence intervals (CIs). The variable *PlatformAge* is included in the model during estimation but not shown for better readability.

F.3 Alternative text complexity measures

The results in the main analysis use the Gunning-Fog Index as a measure of text complexity. As a robustness check, we calculate alternative text complexity measures and compare the estimates. Here, we use the implementation from the `quanteda` package (details: https://quanteda.io/reference/textstat_readability.html) to calculate the following text complexity measures:

1. **Gunning-Fog Index (Fog)** estimates the years of formal education necessary for a person to understand a text upon reading it for the first time. It is given by $0.4 \times (ASL + 100 \times n_{\text{wsy}\geq 3}/n_{\text{w}})$, where ASL is the average sentence length (number of words), n_{w} is the total number of words, and $n_{\text{wsy}\geq 3}$ is the number of words with three syllables or more. Larger values indicate greater text complexity.

2. **Automated Readability Index (ARI)** estimates an approximate representation of the US grade level needed to comprehend the text. Mathematically, it is computed via $0.5 \times ASL + 4.71AWL - 21.34$, where ASL is the average sentence length (number of words), and AWL is the average word length (number of characters). Larger values indicate greater complexity.

3. **Flesch's Reading Ease Score (Flesch)** is designed to rank how difficult a text in English is to understand. Formally, it is given by $206.835 - (1.015 \times ASL) - 84.6 \times (n_{\text{sy}}/n_{\text{w}})$, where ASL is the average sentence length (number of words), n_{w} is the total number of words, and n_{sy} is the number of syllables. Flesch's Reading Ease Score is different from the other scores here in that larger values indicate *lower* text complexity.

4. **Average Word Syllables (AWL)** measures the average word syllables in a text. It is formalized as n_{sy}/n_w, where n_{w} is the total number of words and n_{sy} is the number of syllables. Larger values indicate greater complexity.

We fitted separate regression models, one for each text complexity score. We again used z-

standardization for better comparisons. Overall, we find highly robust findings (Figure 7). We find that higher text complexity increases the click-through rate. This finding is consistent across all considered text complexity measures. Note that the coefficient for Flesch's Reading Ease Score points into the opposite direction because of its reverse interpretation (i. e., a larger value indicates *lower* complexity).

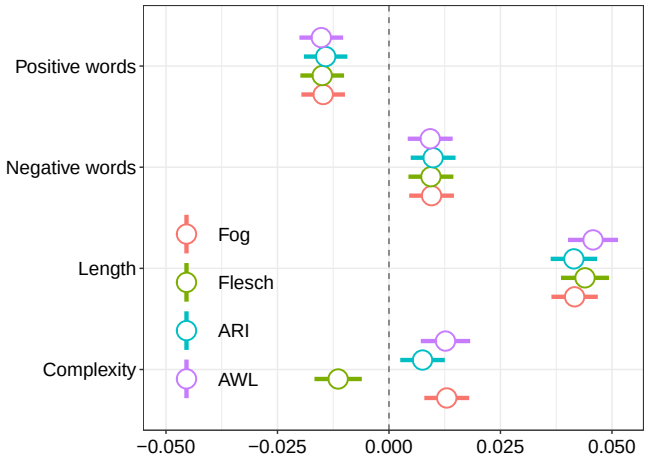

Figure 7: Regression estimates for different measures of text complexity. The lines correspond to the 99 % CIs. Larger values indicate higher text complexity for Fog, ARI, AWL and lower text complexity for Flesch. When correcting for the different interpretations and thus the opposite signs, the coefficients are in good agreement. The variable *PlatformAge* is included in the model during estimation but not shown for better readability.

F.4 Quadratic effects

We extended our models to include quadratic effects for dictionary variables, that is, for the *Positive* and *Negative* variables (Figure 8). We find a negative and statistically significant quadratic effect for negative words. The quadratic effect of positive words is not statistically significant. All direct effects are still statistically significant. This result supports the robustness of our main analysis.

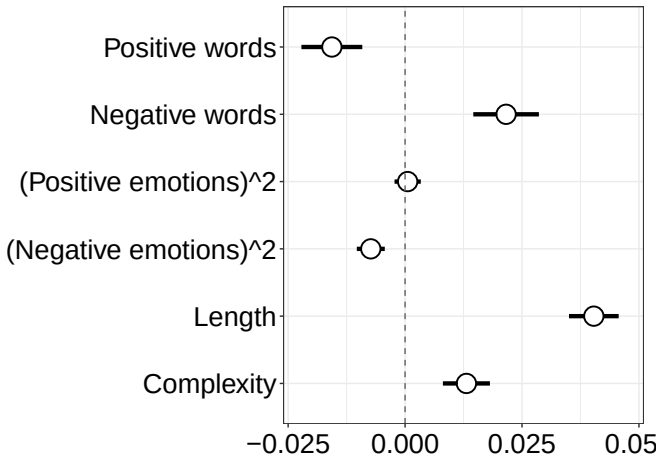

Figure 8: The effect of positive and negative words in news headlines on the click-through rate while controlling for quadratic effects in the word count variables. Shown are the estimated standardized coefficients. The lines correspond to the 99 % CIs. The variable *PlatformAge* is included in the model during estimation but not shown for better readability.

F.5 Analysis of positive-only and negative-only headlines

In some cases, headlines might contain both positive and negative language. In our main paper, these headlines were coded as containing both positive and negative words and then used for estimation. However, headlines with a combination of both positive and negative language may be perceived differently by users than a positive-only headline or a negative-only headline. Hence, we repeated the analysis from the main paper but removed headlines where both positive and negative words were simultaneously present. As such, we end up with all headlines that exclusively include either positive or negative words, that is, positive-only and negative-only headlines. This led to a sample of 9,611 headlines from 2,556 RCTs.

Overall, we find evidence that headlines with negative words are still being clicked on more than headlines with positive words. Negative words no longer predict an increase in click-through rate, but positive words predict a significantly lower click-through rate. This may be in part due to the lower number of headlines with negative words in this subsample (only 2,291 headlines in this subsample included negative words).

	Coef	Lower CI	Upper CI	P-value
Positive	-0.019	-0.025	-0.013	< 0.001
Negative	0.003	-0.004	0.009	0.301
Length	0.044	0.039	0.050	< 0.001
Complexity	0.009	0.003	0.015	< 0.001
PlatformAge	-0.317	-0.348	-0.285	< 0.001
(Intercept)	-4.500	-4.533	-4.467	< 0.001
Observations: 9611				

Table 11: Regression results excluding positive and negative mixed headlines. The dependent variable is the click-through rate. Reported are standardized coefficient estimates (standard errors in parentheses). Experiment-specific intercepts (i. e., random effects) are included.

F.6 Analysis of image RCTs

In addition to A/B testing headlines, Upworthy also A/B tested the images that were paired with each story. Most experiments tested either headlines or images, but there are occasions on which a headline RCT and an image RCT overlap. From the data in the Upworthy Research Archive, researchers can see which RCTs included an image test, but cannot see what images consisted of. We thus reran our main analyses excluding all headline RCTs that also contained image RCTs. This left 10,456 headlines in 2,389 RCTs. Overall, we find that the negativity bias is robust even when dropping headline RCTs that contained image RCTs also.

	Coef	Lower CI	Upper CI	P-value
Positive	-0.016	-0.021	-0.011	< 0.001
Negative	0.011	0.005	0.016	< 0.001
Length	0.045	0.039	0.050	< 0.001
Complexity	0.013	0.008	0.018	< 0.001
PlatformAge	-0.331	-0.364	-0.298	< 0.001
(Intercept)	-4.486	-4.519	-4.452	< 0.001
Observations: 10,456				

Table 12: Regression results for RCTs without image variations. The dependent variable is the click-through rate. Reported are standardized coefficient estimates. Experiment-specific intercepts (i. e., random effects) are included.

F.7 Regression analysis based on sentiment

We repeated the regression analysis based on a single sentiment score (as opposed two separate variables for positive and negativity). For this, we computed a single sentiment score, which is given by the net difference between the proportion of positive words and the proportion of negative words. Formally, this is given by $Sentiment = Positive - Negative$). We then estimated the model with the new sentiment variable but kept all other controls. The coefficient is negative and statistically significant ($p < 0.001$), implying that a positive sentiment decreases click-through rate while a negative sentiment increases click-through rate. This is consistent with the findings from our main analysis.

	Coef	Lower CI	Upper CI	P-value
Sentiment	-0.018	-0.023	-0.013	< 0.001
Length	0.042	0.037	0.047	< 0.001
Complexity	0.013	0.008	0.018	< 0.001
PlatformAge	-0.318	-0.349	-0.287	< 0.001
(Intercept)	-4.487	-4.519	-4.455	< 0.001

Observations: 11,109

Table 13: Regression model explaining click-through rate based on the difference between the proportion of positive and negative words in headlines ($Sentiment$). Reported are standardized coefficient estimates. Experiment-specific intercepts (i. e., random effects) are included.

G The role of moralized language as a moderator

We investigated moralized language as a possible moderator of negativity in driving the click-through rate. This is motivated by previous research, whereby moral-emotional expressions have been found to play an important role in the diffusion of moralized online content [39]. We thus investigate the role of such moral words in moderating the effect of negative words on online news consumption. Analogous to Brady et al. [39], we extract the number of moral words in each headline using a dictionary containing 411 moral words, first presented in [77].

We extend the regression models from our main analysis with interaction terms between the moral word count and the proportion of positive and negative words. In addition, we include the proportion of moral words separately as a regressor. The results (Figure 9) show that moralized language decreases the click-through rate in online news. We found a negative and statistically significant direct effect of moralized language on click-through rate (coef: -0.023, SE = 0.003, $z = -7.650$, $p < 0.001$, CI = [-0.031, -0.015]) and negative and statistically significant effects for the interactions between the proportion of moral words and the proportion of positive (coef: -0.007, SE = 0.003, $z = -2.673$, $p = 0.007$, CI = [-0.014, -0.0003]) and negative words (coef: -0.006, SE = 0.002, z = -2.619, p = 0.009, CI = [-0.012, -0.0001]). More importantly, even when controlling for a moderating role of moralized language, we find strong negativity effects consistent with those in the main analysis.

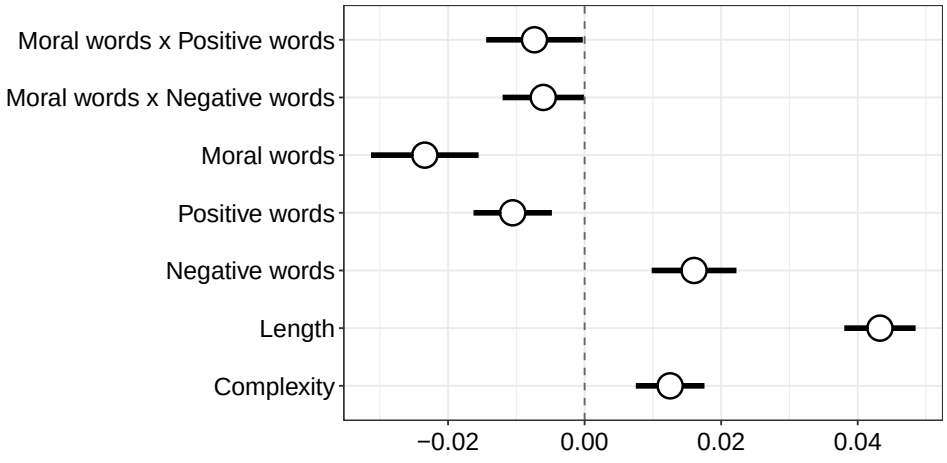

Figure 9: Regression analysis with moralized language as a potential moderator for the effect of positive and negative words on click-through rate. The lines correspond to the 99 % confidence intervals (CIs). The variable *PlatformAge* is included in the model during estimation but not shown for better readability.

H Negativity effects across topics

We use an unsupervised machine learning framework to infer the distribution of news topics in the data through a bottom-up procedure. The benefit of using machine learning is that no assumptions are made ex ante with regard to the covered topics. Our machine learning approach is further regarded as superior to conventional topic modeling (i. e., latent Dirichlet analysis) with short texts [78] Using the extracted topics, we can extend the regression from our main analysis to capture between-topic heterogeneity.

H.1 Procedure for topic modeling

Our unsupervised machine learning framework proceeds in 4 steps. (1) We treat each RCT as corresponding to a single topic. Therefore, the headline variations from each RCT are concatenated to form a single document. (2) We encode the preprocessing document through a document embedding model [79] (3) We apply k-means clustering to the document embeddings, thereby yielding k different topic clusters. (4) We assign names to the topic cluster using a systematic procedure. For this, we manually inspect characteristic words and a number of sample headlines for each cluster. To retrieve the most characteristic words, we first concatenate all headlines belonging to a cluster into a single document and then apply stemming and stop-word removal. We then take the highest-ranking words according to the term frequency–inverse document frequency (tf-idf) statistic. Informed by these, names for each topic are assigned.

H.2 Overview of generated topics

We apply the aforementioned machine learning framework to the Upworthy dataset. Upon manual inspection, the number of clusters was set to $k = 8$. This value was found to provide a suitable balance between sufficient granularity while maintaining interpretability. In particular, this value cluster headlines by overall themes (and not by individual news events). When producing topic

names, we found that two of the eight clusters were better represented using a single topic name and were therefore merged. The resulting 7 clusters and summary statistics are reported in Table 14. The summary statistics reveal that stories about people's lives are most common (29.8 %), followed by news about "Life" (16.5 %). Stories related to "Economy & Government" and other specific societal issues, such as "Woman Rights & Feminism" and "LGBT," were also frequent. Exemplary headlines for each topic are listed in Table 15.

	Topic name	Relative frequency	Characteristic words
1	Entertainment	13.54%	peopl, watch, get, stewart, black, talk, jon, white, ask, comedian
2	Government & Economy	11.99%	peopl, get, america, make, wage, us, minimum, food, work, like
3	LGBT	4.62%	gay, peopl, straight, marriag, lesbian, guy, like, ask, get, way
4	Life	16.47%	peopl, make, like, thing, world, know, video, get, see, life
5	Parenting & School	10.99%	kid, girl, school, like, get, teacher, littl, teen, make, children
6	People	29.80%	peopl, thing, make, like, get, say, guy, see, know, realli
7	Women Rights & Feminism	12.59%	women, woman, feminist, like, think, girl, look, guy, get, know

Table 14: Summary statistics for the different topics embedded in the news stories. Reported are also characteristic words of each topic defined as the top-10 words (stemmed) with regard to the tf-idf statistics.

	Topic name	Sample Headlines
1	Entertainment	“The NFL May Get A Lot Of Things Wrong, But This Former Player Is 100% Right In His Rant On Spanking” “Bill Nye Points Out The Biggest Problem With Modern Astrology” “I’m Not A Conspiracy Theorist But Learning About Movie Ratings Has Me Reaching For The Tin Foil”
2	Government & Economy	“Mr. President, I’m Not Mad. I’m Just Disappointed. No, Wait. I’m A Little Mad Too.” “Meet The Unmanned Drones Built To Fight Poverty Instead Of People” “So That’s What Hard Working Government Employees Look Like? (Pssst...Can We Send This To Congress?)
3	LGBT	“Marriage In France Just Got A Lot Gayer” “I Have A Really Secret Way To Help Protect Gay Kids From Bullying” “Lets Have The Sexuality Talk And Clear 10 Things Up”
4	Life	“Why People Risk Their Lives for For People They’ve Never Met” “A Newly Launched Camera Is Exposing Some Of Our Worst Parts” “ Finally, An Approachable Guide To Crappy Arguments We See On The Internet. Every. Day.”
5	Parenting & School	“Want To Raise A Genius? Introduce Her To Bob Dylan.” ”It’s Amazing What People Can Do When They Expect Their Children To Live Past Kindergarten” “ Band Geeks Think They’re Smarter Than The Rest Of Us. Turns Out, They’re Right.”
6	People	“She Grew Up With Privilege – And She Knows How To Use It” “It Broke Her Heart Seeing Her Daughter’s Facebook Page, Asking For Someone To Please Be Her Friend” “Have You Ever Heard ’Don’t Act Like A Typical Tourist’? Here’s Why.”
7	Women Rights & Feminism	“Sexual Objectification: What it is, Why It’s Damaging, And How We Can Change It” “A Tampon Commercial That Shows Just How Confusing Actual Tampon Commercials Are” “Calling Girls This Word May Seem Harmless — But Why Are Boys Never Called It?”

Table 15: Examples of headlines assigned to the seven topics.

H.3 Validation of topic model

Next, we validated our topic modeling approach by conducting a user study. Specifically, we followed best-practice for validating topic models by implementing a *topic intrusion* test [80]. This test allows us to validate that participants were better than chance at categorizing headlines as belonging to a certain topic in accordance with our topic model. Participants ($n = 10$) recruited from the NYU subject pool were asked to read a random subset of 70 headlines. Participants were also shown four possible topic categories – the correct topic category and 3 other topics – from which they were asked to identify which category the headline belonged to. Participants answered 51.1 % of trials correctly. This is significantly above chance which would amount to having 25 % of the trials answered correctly ($\chi^2 = 249.61$, $p < 0.01$). The user study thus confirms that the topic model generates meaningful representations. The breakdown of correct answers per topic are listed below.

Topic	Percent Correct
Entertainment	45.0%
Government & Economy	51.5%
LGTB	67.0%
Life	49.5%
Parenting & School	43.0%
People	25.3%
Women Rights & Feminism	76.0%

Table 16: Percent correct from human validators in the Topic Intrusion Task, broken down by topic.

We considered the use of a word intrusion test [80] but eventually discarded this. Word intrusion tests for a small within-topic similarity, yet this is not the focus of our topic model. On the contrary, we explicitly allow for a comparatively larger diversity among headlines within the same topic. The reason is that our topic categorization should cover thematic areas (rather than specific news events) and should thus be comparatively broad.

H.4 Regression analysis with topic controls

Using the above topics, we then repeated our main analysis while controlling for between-topic heterogeneity. Overall, the parameter estimates for the extended models are qualitatively similar for both positive and negative words (Table 17). We find that, on average, the categories "Economy & Government", "Life" and "Parenting & School" attract fewer clicks than the reference category "Entertainment."

	Coef	Lower CI	Upper CI	P-value
Positive	−0.015	−0.020	−0.010	< 0.001
Negative	0.009	0.004	0.014	< 0.001
TOPICS				
Entertainment (reference topic)	—	—	—	—
Government & Economy	−0.570	−0.681	−0.460	< 0.001
LGTB	−0.123	−0.277	0.031	0.04
Life	−0.396	−0.495	−0.297	< 0.001
Parenting & School	−0.261	−0.376	−0.146	< 0.001
People	−0.069	−0.211	0.074	0.216
Women Rights & Feminism	−0.108	−0.224	0.008	0.016
CONTROL VARIABLES				
Length	0.041	0.036	0.047	< 0.001
Complexity	0.013	0.008	0.018	< 0.001
PlatformAge	−0.328	−0.358	−0.298	< 0.001
(Intercept)	−4.210	−4.293	−4.128	< 0.001

Observations: 11,109

Table 17: Regression results estimating the effect of positive and negative words on the click-through rate. Here, dummy variables referring to the different topics are included. Reported are standardized coefficient estimates. Experiment-specific intercepts (i. e., random effects) are included.

H.5 Regression analysis with topic-specific negativity effects

We further examine interactions between topics and emotional variables. Regression estimates show that the negative effects for positive and negative words found in the main analysis are also present for the majority of topics (Table 18).

	Coef	Lower CI	Upper CI	P-value
Positive × Entertainment	0.001	−0.012	0.013	0.901
Positive × Government & Economy	−0.004	−0.018	0.011	0.484
Positive × LGTB	−0.022	−0.039	−0.006	< 0.001
Positive × Life	−0.023	−0.032	−0.015	< 0.001
Positive × Parenting & School	−0.017	−0.030	−0.004	0.001
Positive × People	−0.025	−0.046	−0.004	0.002
Positive × Women Rights & Feminism	−0.009	−0.022	0.004	0.084
Negative × Entertainment	0.003	−0.010	0.016	0.589
Negative × Government & Economy	0.024	0.012	0.037	< 0.001
Negative × LGBT	0.024	0.005	0.043	0.001
Negative × Life	−0.003	−0.013	0.006	0.387
Negative × Parenting & School	0.023	0.009	0.037	< 0.001
Negative × People	0.021	0.001	0.041	0.006
Negative × Women's Rights & Feminism	0.004	−0.009	0.016	0.431
Length	0.042	0.036	0.047	< 0.001
Complexity	0.013	0.008	0.019	< 0.001
PlatformAge	−0.318	−0.349	−0.287	< 0.001
(Intercept)	−4.487	−4.519	−4.455	< 0.001

Observations: 11,109

Table 18: Regression results estimating the effect of positive and negative words on the click-through rate. Here, we examine interactions between topic dummies and positive/negative words. Reported are standardized coefficient estimates. Experiment-specific intercepts (i. e., random effects) are included.

H.6 Further analysis with topics and emotions

Table 19 controls for topic dummies in our regression estimating the effect of discrete emotions. Even when controlling for between-topic variation in clickability, the results remain robust.

	Coef	Lower CI	Upper CI	P-value
DISCRETE EMOTIONS				
Anger	0.000	−0.006	0.006	0.997
Fear	−0.005	−0.011	0.002	0.061
Joy	−0.004	−0.011	0.003	0.127
Sadness	0.009	0.003	0.016	< 0.001
TOPICS				
Entertainment (reference topic)	—	—	—	—
Government & Economy	−0.576	−0.691	−0.461	< 0.001
LGTB	−0.146	−0.306	0.015	0.019
Life	−0.410	−0.513	−0.307	< 0.001
Parenting & School	−0.288	−0.408	−0.168	< 0.001
People	−0.081	−0.231	0.069	0.165
Women Rights & Feminism	−0.137	−0.258	−0.016	0.003
CONTROL VARIABLES				
Length	0.044	0.038	0.050	< 0.001
Complexity	0.010	0.004	0.016	< 0.001
PlatformAge	−0.334	−0.365	−0.303	< 0.001
(Intercept)	−4.215	−4.300	−4.130	< 0.001

Observations: 8365

Table 19: Regression results estimating the effect of discrete emotions on the click-through rate. Here, dummy variables referring to the different topics are included. Reported are standardized coefficient estimates. Experiment-specific intercepts (i. e., random effects) are included.

I Analysis across all basic emotions and higher-order emotions

I.1 Analysis for basic emotions

In our main regression analysis, we focused on 4 discrete emotions (i. e., *anger*, *fear*, *joy*, *sadness*) for which we found a notable correlation between the computed NRC emotion scores and human judgments, implying that humans perceive a headline to embed that emotions. For thoroughness, we performed a regression analysis based on all 8 basic emotions from the NRC emotion lexicon. This should be interpreted with caution, as humans do not necessarily read the same emotions in the headlines, and thus they should understood as "NRC dimensions".

Of note, the variables for the 8 basic emotions sum to 1 and are thus subject to linear dependency. Evidently, there are high cross-correlations among the 8 basic emotions (see Figure 10). Methodologically, they are relevant because they prohibit all 8 emotions to be examined in the same model without making the model rank deficient. To alleviate issues due to linear dependence, we performed a regression analysis based on 8 separate regression models that were estimated independently, each including one of the 8 basic emotions. The multilevel regression for the 8 basic emotions is specified analogous to our analysis from the main paper, i. e.,

$$\text{logit}(\theta_{ij}) = \alpha + \alpha_i + \beta\, BasicEmotion_{ij} + \gamma_1\, Length_{ij} + \gamma_2\, Complexity_{ij} + \gamma_3\, PlatformAge_{ij} \tag{1}$$

with a random effects specification, where α is the global intercept and α_i captures the heterogeneity among experiments $i = 1, \ldots, N$. Further, $BasicEmotion_{ij}$ denotes one of the 8 basic emotions (e. g., $Anger_{ij}$, $Anticipation_{ij}$, etc.). In addition, we again control for length, text complexity, and the age of the platform since the first overall experiment. The coefficient β then quantifies how one of the basic emotions affects the click-through rate. To account for multiple hypothesis testing, we use Bonferroni correction [81].

The estimation results confirm the findings from the main analysis (Figure 11). As in the main

paper, positive effects are found for *sadness*, while no statistically significant effects are found for *anger*, *fear*, and *joy*. In addition, a statistically significant negative effect is found for *anticipation*. The effect of *disgust* is statistically significant at the $5\,\%$ statistical significance level. Due to the estimation procedure, we refrain from comparing the effect size of the different basic emotions.

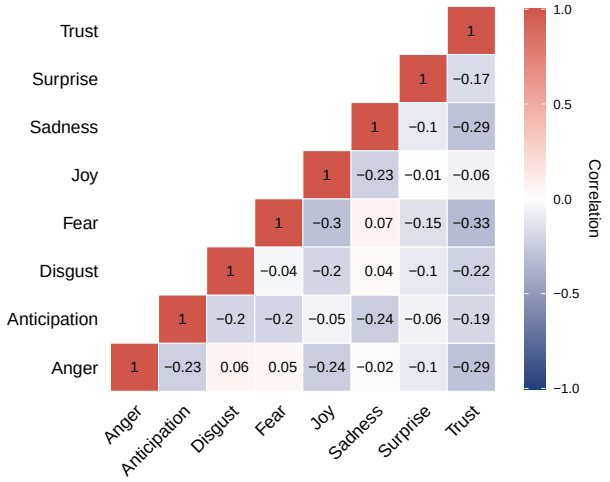

Figure 10: Cross-correlations between variables representing emotional words in news headlines. Here, the emotional variables are the proportion of emotional words as defined by NRC emotion lexicon. Pearson's correlation coefficients are reported.

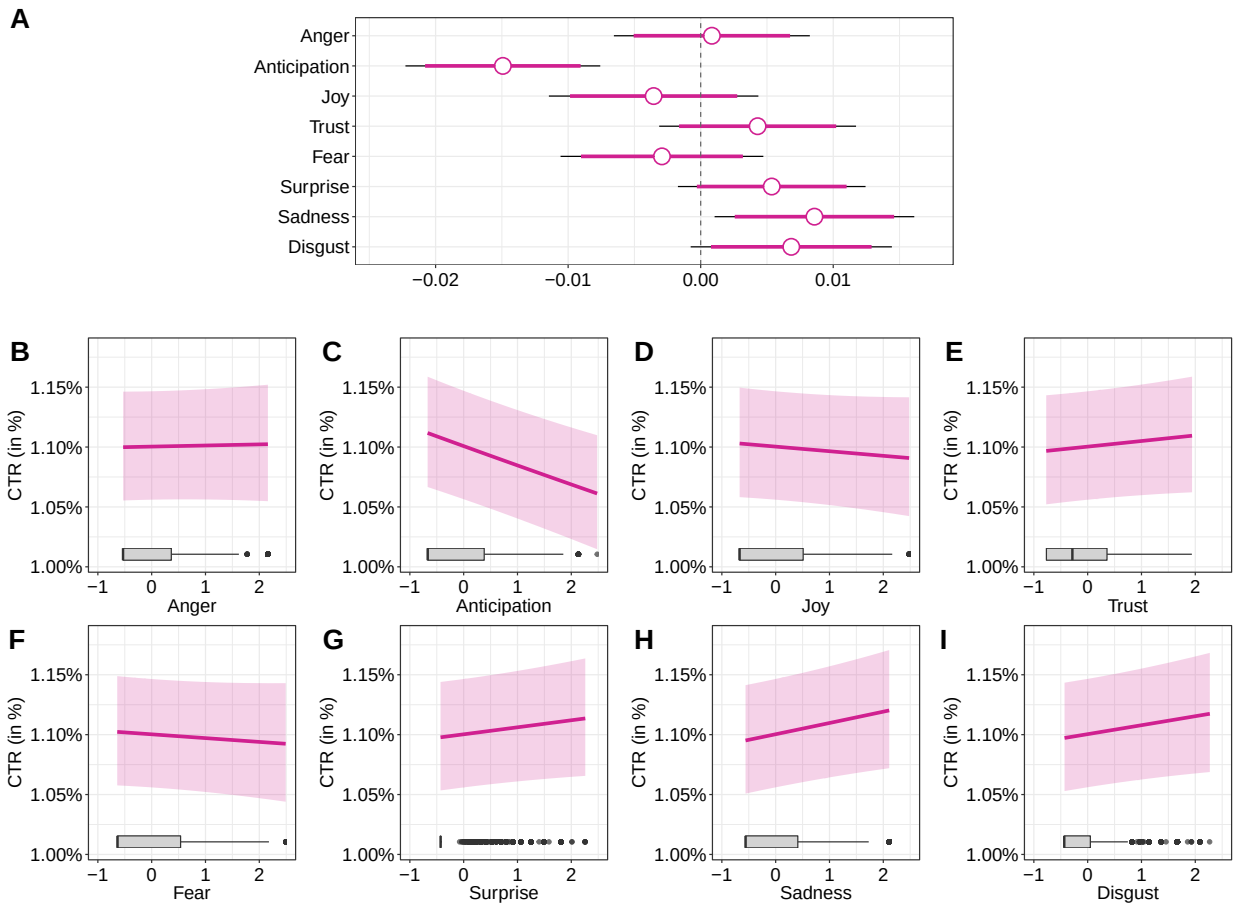

Figure 11: Effect of emotional words in news headlines on the click-through rate. **(A)** Shown are the estimates of the standardized coefficient that originate from separate regressions for the basic emotions as derived from the NRC emotion lexicon. The thick (pink) and thin (black) lines correspond to the 99 % confidence intervals (CIs) and 99 % Bonferroni-corrected [81] CIs, respectively. **(B-I)**: Predicted marginal effects of basic emotions on the click-through rate (with 99 % CIs). Boxplots show the distribution of the variables in our sample (center line gives the median; box limits are upper and lower quartiles; whiskers denote minimum/maximum; points are outliers defined as being beyond 1.5x of the interquartile range).

I.2 Analysis for bipolar emotion pairs

Following [40, 41], we analyzed the effects of bipolar emotion pairs on the click-through rate. Specifically, we arranged the basic emotions into 4 pairs of bipolar emotions (i. e., so that they represent opposite petals as in Plutchik's model [73]). The 4 bipolar emotions are *anticipation–*

surprise, *anger–fear*, *trust–disgust*, and *joy–sadness*, representing the pairs of emotions that are least similar to one another. The corresponding variables for the bipolar emotions are computed by taking the difference between the two (thereby yielding a value between -1 and 1). This yields 4 scores: $AnticipationSurprise_{ij} = Anticipation_{ij} - Surprise_{ij}$, $AngerFear_{ij} = Anger_{ij} - Fear_{ij}$, $TrustDisgust_{ij} = Trust_{ij} - Disgust_{ij}$, and $JoySadness_{ij} = Joy_{ij} - Sadness_{ij}$.

The multilevel regression is specified analogous the previous models but with additional explanatory variables, i. e.,

$$\begin{aligned}\text{logit}(\theta_{ij}) = \alpha + \alpha_i + \beta_1\, AngerFear_{ij} + \beta_2\, AnticipationSurprise_{ij} + \beta_3\, JoySadness_{ij} \\ + \beta_4\, TrustDisgust_{ij} + \gamma_1\, Length_{ij} + \gamma_2\, Complexity_{ij}\end{aligned} \tag{2}$$

where α and α_i represent the varying-intercept specification. Specifically, α is again the global intercept and α_i captures the heterogeneity across experiments $i = 1, \ldots, N$. As in the main paper, we include the control variables, i. e., length and text complexity. The coefficients $\beta_1, \ldots, \beta_4$ quantify the effect of the four bipolar emotion pairs (i. e., *anticipation–surprise*, *anger–fear*, *trust–disgust*, and *joy–sadness*) on the click-through rate.

We found a negative coefficients for words from the bipolar emotion pair *joy–sadness* (coef: -0.007, SE = 0.002, $z = -3.006$, $p < 0.01$, CI = $[-0.012, -0.002]$). The negative signs imply that a higher click-through rate is elicited by headlines containing a greater proportion of words belonging to *sadness* (Figure 12). A one standard deviation increase in the variable for the bipolar emotion pair *joy–sadness* decreases the odds of a user clicking the headline by 0.7%. The coefficient estimates for the pair *anger–fear* was not statistically significant at common significance thresholds. Consistent with our previous findings, we observed that the click-through rate increases as the text length and complexity score increase. Again, the click-through rate was lower for headlines at the end of Upworthy's career.

For thoroughness, we also analyzed emotions for which we did not found statistically significant positive correlation between the user judgments and the computed NRC emotion scores in

the validation study. Here we found a statistically significant coefficient for the bipolar emotion pair *anticipation–surprise*. The negative sign implies that a higher click-through rate is elicited by headlines containing a greater proportion of words belonging to surprise (Figure 12). The coefficient estimates for the pair *trust–disgust* was not statistically significant at common significance thresholds.

	Coef	Lower CI	Upper CI	P-value
$AngerFear$	0.002	-0.004	0.008	0.314
$AnticipationSurprise$	-0.014	-0.020	-0.008	< 0.001
$TrustDisgust$	0.002	-0.005	0.008	0.517
$JoySadness$	-0.007	-0.014	-0.001	0.003
$Length$	0.043	0.037	0.050	< 0.001
$Complexity$	0.009	0.003	0.015	< 0.001
$PlatformAge$	-0.323	-0.355	-0.291	< 0.001
(Intercept)	-4.505	-4.538	-4.472	< 0.001

Observations: 8365

Table 20: Regression model explaining click-through rate based on bipolar emotion pairs in headlines. Reported are standardized coefficient estimates. Experiment-specific intercepts (i. e., random effects) are included..

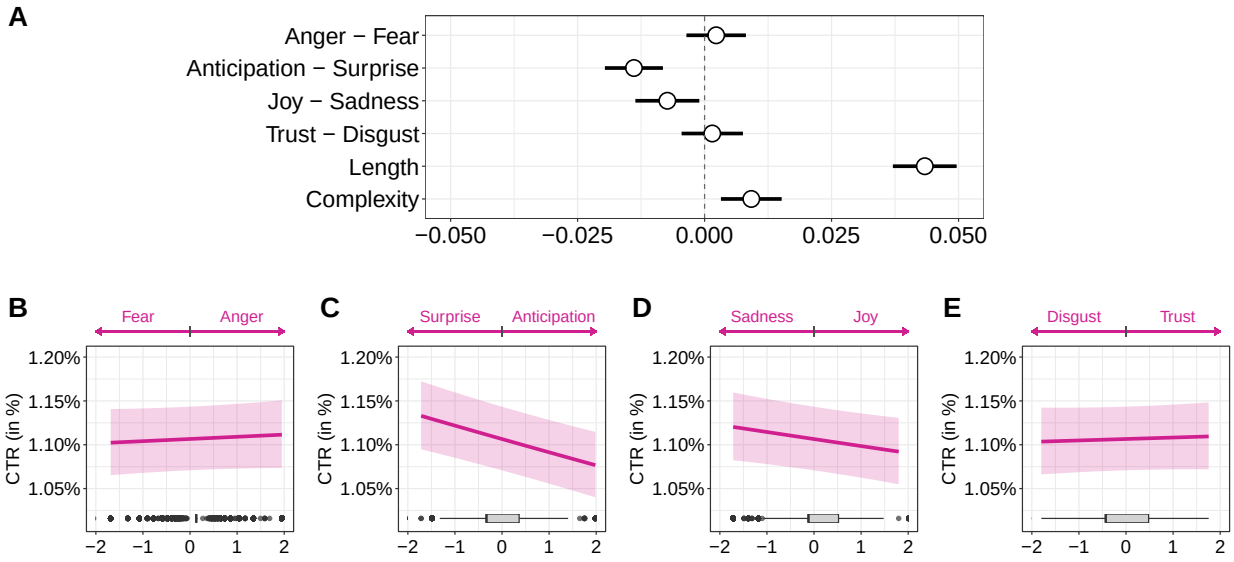

Figure 12: Effect of emotional words on the click-through rate. (A) Shown are the estimated standardized coefficients and 99% confidence intervals for each bipolar emotion derived from the NRC emotion lexicon. Overall, clicks are elicited by news headlines with words classified as *surprise* and *sadness*. The variable *PlatformAge* is included in the model during estimation but not shown for better readability. (B-E) Predicted marginal effects of bipolar emotions on the click-through rate (with 99% confidence intervals). In (B), the boxplots indicate narrow distribution for the *fear–anger* pair, suggesting that the variation in these emotions is comparatively small. Boxplots show the distribution of the variables in our sample (center line gives the median; box limits are upper and lower quartiles; whiskers denote minimum/maximum; points are outliers defined as being beyond 1.5x of the interquartile range).

I.3 Analysis for emotional dyads

Plutchik's emotions model defines 24 emotional dyads, which are more complex emotions composed of two basic emotions [67]. Following [38, 39], we compute the score for each of the 24 emotional dyads by taking the sum of two emotions (e. g., $Aggressiveness_{ij} = Anger_{ij} + Anticipation_{ij}$). Then, we will compute a score for each of the opposite pairs by taking the corresponding difference (e. g., $LoveRemorse_{it} = Love_{it} - Remorse_{it}$). Across all dyads, this will yield 12 different scores to be used in a regression analysis.

We examine the effect of emotional dyads on the click-through rate as follows. We fit twelve

separate models, that is, one for each pair among the emotional dyads, due to linear dependencies between the dyads. The underlying model is given by

$$\text{logit}(\theta_{ij}) = \alpha + \alpha_i + \beta\, EmotionalDyad_{ij} + \gamma_1\, Length_{ij} + \gamma_2\, Complexity_{ij}, \tag{3}$$

with global intercept α and varying intercept α_i and $EmotionalDyads_{ij}$ denotes one pair among the emotional dyads. We include the same control variables as in the previous models. To account for multiple hypothesis testing, we again use Bonferroni correction [81].

The regression results (Table 21 and Figure 13) show a negative coefficient for emotional words from the following dyads: *optimism–disapproval*, *anxiety–outrage*, *hope–unbelief*, and *guilt–envy*. Users thus have a propensity to respond to language expressing *disapproval*, *outrage*, *unbelief*, and *envy*, whereas the click-through rate decreases due to the presence of *optimism*, *anxiety*, *hope*, and *guilt*. There were also dyads with positive coefficients: *curiosity–cynicism* and *awe–aggressiveness*. Overall, we found that several dyads are important determinants of click-through rates.

	Coef	Lower CI	Upper CI	P-value
OptimismDisapproval	-0.016	-0.023	-0.008	< 0.001
LoveRemorse	-0.004	-0.012	0.004	0.092
SubmissionContempt	-0.001	-0.009	0.006	0.630
AweAggressiveness	0.009	0.002	0.017	< 0.001
HopeUnbelief	-0.009	-0.017	-0.001	< 0.001
GuiltEnvy	-0.007	-0.015	0.000	0.001
CuriosityCynicism	0.010	0.002	0.017	< 0.001
DespairPride	0.004	-0.004	0.012	0.129
AnxietyOutrage	-0.012	-0.019	-0.005	< 0.001
DelightPessimism	0.006	-0.002	0.013	0.016
SentimentalityMorbidness	0.004	-0.004	0.012	0.071
ShameDominance	-0.002	-0.010	0.006	0.353

Observations: 8365

Table 21: Estimation results for the model with emotional dyads. Coefficients are retrieved from separate models for each dyad pair due to linear dependence between the dyads. p-values are Bonferroni-corrected [81]. Reported are standardized coefficient estimates. Experiment-specific intercepts (i. e., random effects) are included.

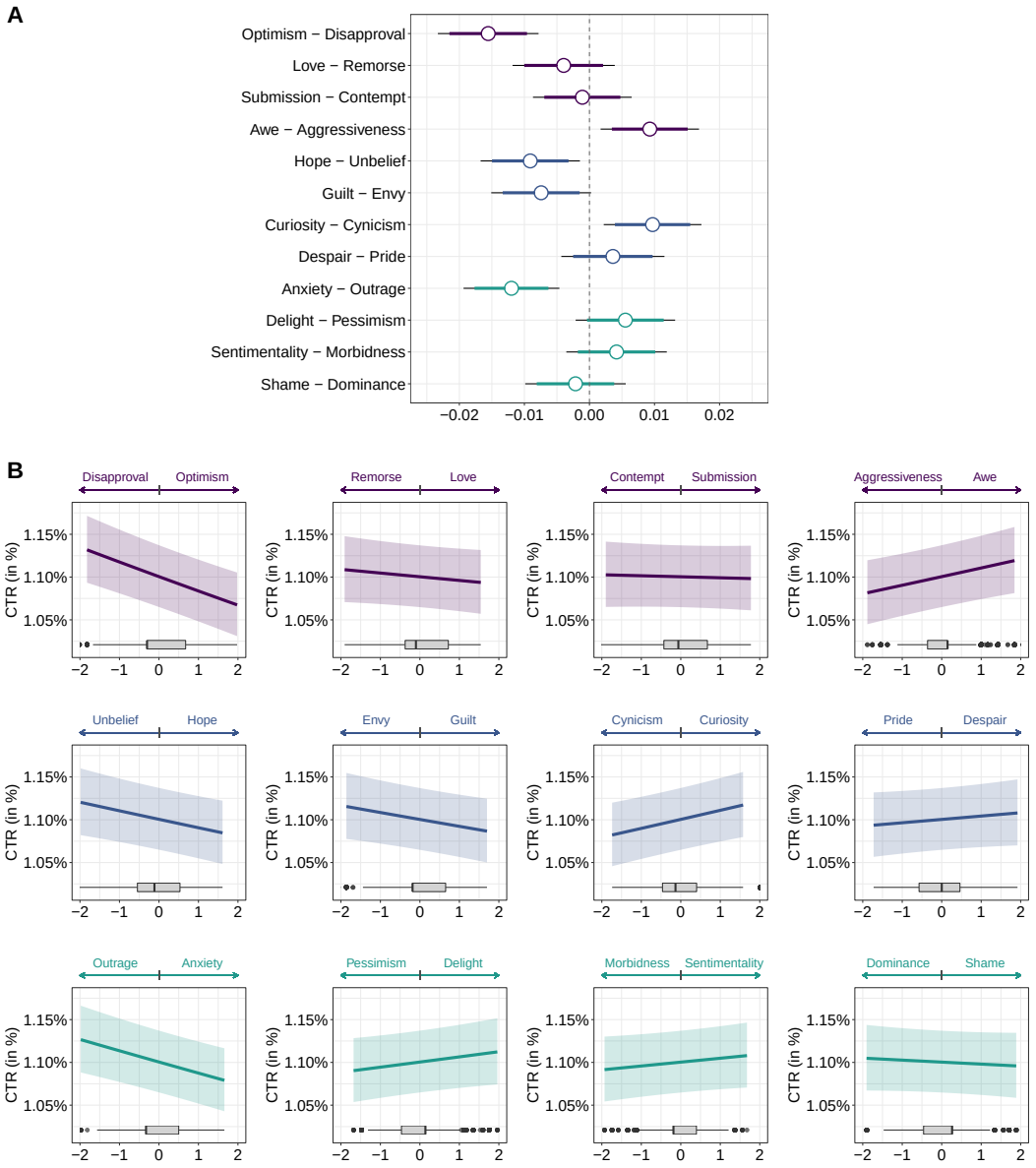

Figure 13: Effect of emotional words from emotional dyads on the click-through rate. (A) Shown are the estimated standardized coefficients for emotional dyads. The thick (colored) and thin (black) lines correspond to 99% confidence intervals and Bonferroni-corrected 99% confidence intervals, respectively. Due to linear dependencies among the dyads, the estimates originate from separate regressions. Overall, users responded most strongly to emotional words classified as *disapproval*, followed by *curiosity* and *awe*. (B) Predicted marginal effects of the emotional words from the different dyads on the click-through rate (with 99% confidence intervals). The plots are arranged by primary (top), secondary (middle), and tertiary (bottom) dyads. Boxplots show the distribution of the variables in our sample (center line gives the median; box limits are upper and lower quartiles; whiskers denote minimum/maximum; points are outliers defined as being beyond 1.5x of the interquartile range).