

# Moral Emotions Shape the Virality of COVID-19 Misinformation on Social Media

Kirill Solovev  
kirill.solovev@wi.jlug.de  
JLU Giessen  
Germany

Nicolas Pröllochs  
nicolas.proellochs@wi.jlug.de  
JLU Giessen  
Germany

## ABSTRACT

While false rumors pose a threat to the successful overcoming of the COVID-19 pandemic, an understanding of how rumors diffuse in online social networks is – even for non-crisis situations – still in its infancy. Here we analyze a large sample consisting of COVID-19 rumor cascades from Twitter that have been fact-checked by third-party organizations. The data comprises  $N = 10,610$  rumor cascades that have been retweeted more than 24 million times. We investigate whether COVID-19 misinformation spreads more viral than the truth and whether the differences in the diffusion of true vs. false rumors can be explained by the moral emotions they carry. We observe that, on average, COVID-19 misinformation is more likely to go viral than truthful information. However, the veracity effect is moderated by moral emotions: false rumors are more viral than the truth if the source tweets embed a high number of other-condemning emotion words, whereas a higher number of self-conscious emotion words is linked to a less viral spread. The effects are pronounced both for health misinformation and false political rumors. These findings offer insights into how true vs. false rumors spread and highlight the importance of considering emotions from the moral emotion families in social media content.

## CCS CONCEPTS

• Human-centered computing → Social media; Empirical studies in collaborative and social computing; • Applied computing → Sociology.

## KEYWORDS

Social media, misinformation, COVID-19, virality, moral emotions, computational social science, explanatory modeling

### ACM Reference Format:

Kirill Solovev and Nicolas Pröllochs. 2022. Moral Emotions Shape the Virality of COVID-19 Misinformation on Social Media. In *Proceedings of the ACM Web Conference 2022 (WWW '22)*, April 25–29, 2022, Virtual Event, Lyon, France. ACM, New York, NY, USA, 12 pages. <https://doi.org/10.1145/3485447.3512266>

## 1 INTRODUCTION

Social media platforms play an ambivalent role during the COVID-19 pandemic. On the one hand, they represent an important source

of health information for large parts of society [39]. On the other hand, however, this crisis has bred a multitude of rumors [24, 26, 33, 36, 45], and verdicts of reputable fact-checking organizations (e. g., politifact.com, snopes.com) suggest that social media is rife with COVID-19 misinformation. COVID-19 misinformation on social media includes, but is not limited to, misinformation about vaccination, “miracle cures,” and supposed preventives [32]. False rumors can impact the timely and effective adoption of public health recommendations [66], the effectiveness of the countermeasures deployed by governments [50], and are sometimes even used as a political weapon [51]. Given that exposure to misinformation frequently manifests in offline consequences [45], there is an urgency to study the spread of rumors on social media in the context of COVID-19. Tedros Adhanom Ghebreyesus, director-general of WHO, and other experts speak of an “infodemic,” which must be fought [71].

While previous research – at least for non-crisis situations – suggests that false rumors on social media tend to be more viral than the truth [48, 65], the mechanism underlying its viral spread, though critical, remains unresolved. In this work, we approach this question through the lenses of morality and emotions and their role in rumor diffusion in polarized social media environments. Social media content delivers not only factual information but also carries moral ideas and sophisticated emotional signals [8]. Moral emotions provide the motivational force for humans to do good and to avoid doing bad [59] and can even serve to “moralize” actions that would otherwise be considered non-moral [68]. Since socially connected users often develop similar ideas and intuitions [8, 13, 23, 39], moral emotions are a key driver of information diffusion in polarized social media environments [8]. In the context of COVID-19, the overall discussion culture has repeatedly been characterized as highly polarized [1, 17, 21, 31, 32, 34]. For instance, people have been observed to be divided in their perceptions of government responses, confidence in scientists, and support for protective actions [31, 34]. If COVID-19 rumors are highly polarizing to social media users, then the transmission of moral emotions likely plays a key role in the rumors’ diffusion through social networks.

The principal moral emotions can be divided into two families [30]. The families are the “other-condemning” family, comprising the emotions contempt, anger, and disgust, and the “self-conscious” family comprising the emotions shame, pride, and guilt [60]. The former, other-condemning emotions, are reactions to the social behavior of others and involve a negative judgment or disapproval of others. In the context of morality, other-condemning emotions are sometimes also referred to as the “hostility triad” [52]. Other-condemning emotions are typically associated with perceived moral violations, for example, in the context of individuals’ rights and fairness [62]. While the individual emotions in the other-condemning family (i. e., anger, contempt, and disgust) are often assumed to

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

WWW '22, April 25–29, 2022, Virtual Event, Lyon, France

© 2022 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-9096-5/22/04.

<https://doi.org/10.1145/3485447.3512266>

be not particularly explosive on their own, they can become a dangerous, explosive mix when compressed together [52]. Their counterpart is the family of self-conscious emotions, which are evoked by self-reflection and self-evaluation. These emotions motivate individuals to behave in a socially acceptable fashion and are linked to prosocial behaviors such as empathy and altruism [30, 52]. As such, self-conscious emotions can enable social healing and avoid triggering the contempt, anger, and disgust of others [30]. While previous research [8] broadly distinguished moral vs. non-moral emotions in social media content, we will investigate whether these two clusters of moral emotions (self-conscious vs. other-condemning emotions) have distinct effects on the diffusion of rumors in the context of COVID-19.

**Research hypothesis:** In this work, we propose that the virality of true vs. false COVID-19 rumors can be explained by the moral emotions they carry. Although previous research suggests that false rumors are statistically more often retweeted [48, 65], not every false rumor is necessarily more viral than a truthful rumor. Rather, misinformation going viral is oftentimes spread through echo chambers with exacerbated ideological polarization [12]. In these environments, ideological identity is more salient in guiding user behavior [67] and users are moved towards more extreme positions [14]. Polarization not only reduces verification behavior [35, 41] but also makes users more receptive to hostility against others, e. g., for political attacks [61]. Here other-condemning emotions embedded in the *source tweets*, which start the rumor cascade, may function as accelerators and amplifiers [63]. In polarizing discussions about COVID-19, this would imply that radical ideas and beliefs are strengthened and are more likely to translate into action. Given increased ideological polarization for false rumors [20], the explosive mix of other-condemning emotions should thus accelerate their spread within social networks. The same reasoning suggests that false rumors embedding self-conscious emotions (that avoid triggering other-condemning emotions [30]) should be less contagious on social media. In sum, we hypothesize that rumors with a stronger combination of false content and other-condemning emotions in the source tweets reach more people, whereas the combination of false content and self-conscious emotions reaches fewer people.

**Data:** We collected a *unique* dataset of COVID-19 rumor cascades propagating on Twitter between January 2020 and the end of April 2021. Each rumor cascade was investigated and fact-checked by at least one of three independent fact-checking organizations (snopes.com, politifact.com, truthorfiction.com). Our data include 10,610 rumor cascades that have been retweeted 24.34 million times.

**Methodology:** We use textual analysis to extract fine-grained moral emotions (self-conscious and other-condemning) embedded in rumor cascades. Specifically, we employ (and validated) a dictionary-based approach to count the frequency of occurrence of self-conscious and other-condemning emotion words in the source tweets that have initiated the rumor cascades. To measure the diffusion of each rumor cascade, we employ the Twitter Historical API to obtain the number of retweets, that is, the number of users interacting with the rumor cascade. We then fit *explanatory* regression models to evaluate how variations in moral emotions are associated with differences in the number of retweets for true vs. false rumor cascades. In our regression analysis, we follow previous works

[8, 65] by controlling for variables known to affect the retweet rate independent of the main predictors, i. e., the number of followers, the account age, etc.

**Findings:** We observe that, on average, COVID-19 misinformation is more likely to go viral than truthful information. However, the veracity effect is moderated by moral emotions: false rumors are more viral than the truth if the source tweets embed a high number of other-condemning emotion words, whereas a higher number of self-conscious emotion words is linked to a less viral spread. The effects are pronounced both for health misinformation and false political rumors. These findings offer insights into how true vs. false rumors spread and highlight the importance of considering emotions from the moral emotion families in social media content.

## 2 BACKGROUND

### 2.1 Misinformation on Social Media

Social media has shifted quality control for the content from trained journalists to regular users [35]. The lack of oversight from experts makes social media vulnerable to the spread of misinformation [53]. Social media has indeed repeatedly been observed to be a medium that disseminates vast amounts of misinformation [e. g., 47, 65]. The presence of misinformation on social media also has detrimental consequences on how opinions are formed in the offline world [2, 4, 19, 42]. As a result, it not only threatens the reputation of individuals and organizations, but also society at large.

Several works have focused on the question of *why* misinformation is widespread on social media. These studies suggest that it is difficult for users to spot misinformation as it is often intentionally written to mislead others [69]. Moreover, social media users are often in a hedonic mindset and avoid cognitive reasoning such as verification behavior [41]. The vast majority of social media users do not fact-check articles they read [27, 64]. A recent study further suggests that the current platform design may discourage users from reflecting on accuracy [44]. Online social networks are also characterized by (political) polarization [38, 47, 55] and echo chambers [5]. In these information environments with low content diversity and strong social reinforcement, users tend to selectively consume information that shares similar views or ideologies while disregarding contradictory arguments [22]. These effects can even be exaggerated in the presence of repeated exposure: once misinformation has been absorbed, users are less likely to change their beliefs even when the misinformation is debunked [43].

### 2.2 Research on Rumor Spreading

Several studies have analyzed the spreading dynamics of rumors vs. non-rumors on social media. This includes analyses of summary statistics with regard to, for instance, the number of retweets [e. g., 7, 25] and the rumor lifetime [e. g., 7, 10, 19]. However, these works discern cascades from rumors vs. non-rumors, and do not focus on differences across veracity. Another stream of literature has analyzed rumors concerning specific events (e. g., the 2013 Boston Marathon bombing) with regard to the overall tweet volume or content [e. g., 18, 56, 57]. These works analyze how the user base responds to rumors but again do not analyze the diffusion dynamics of true vs. false rumors.

Only a few works have analyzed differences in the spread of true vs. false rumors. Friggeri et al. (2014) [25] classified the veracity of  $\approx 4,000$  rumors from Facebook based on fact-checking assessments from snopes.com. The authors find that a majority of resharing of false rumors occurs after fact-checking. This suggests that social media users likely do not notice the fact-checks; or intentionally ignore their verdict. Closest to our work is the study from Vosoughi et al. (2018) [65], which provides a comprehensive analysis of summary statistics of true vs. false rumors on Twitter, finding that false rumors spread significantly farther, faster, and more broadly than the truth. However, this work does not analyze the spread of true vs. false rumors in the context of COVID-19. The same dataset [65] has also been used in a recent study [48] that measures emotions embedded in the *replies* to rumor cascades. The authors find that higher frequencies of certain emotions (e. g., anger) are associated with more viral cascades for false rumors.

In the context of COVID-19, research providing large-scale quantitative analyses of the spread of true and false rumors is scant. Existing works have primarily focused on summary statistics of small sets of hand-labeled rumors or source-based approaches to identify COVID-19 misinformation [e. g., 15, 36, 54]. For example, Cinelli et al. (2020) classify news sources into reliable and non-reliable sources in order to analyze the spread of COVID-19-related content. The authors find no significant differences regarding the spreading dynamics. Notably, however, categorizations of reliable vs. non-reliable sources do not necessarily correspond to true vs. false rumors. In addition, source-based approaches ignore false rumors from influential individuals, emerging websites, and misclassify false rumors from websites that are generally considered as being reliable. Note that there are other recent papers reporting that COVID-19 misinformation is widespread on social media, characterizing COVID-19 misinformation, and expressing concerns about consequences for public health [e. g., 26, 28, 33, 36, 45]. However, these works do not focus on modeling differences in the diffusion of true vs. false COVID-19 rumor cascades.

**Our contributions:** This work makes two key contributions. (1) We collected a unique dataset of COVID-19 rumor cascades and demonstrate that misinformation is, on average, more viral than the truth. Here, our study connects to previous works [48, 65], which yielded similar conclusions, yet not in the context of COVID-19. (2) The mechanisms underlying the viral spread of false rumors, though critical, have remained largely unresolved in previous research. Our work is the first to approach the question through the lenses of morality and emotions – finding that moral emotions embedded in *source tweets* shape the diffusion of false rumors on social media.

### 3 METHODS

#### 3.1 Data Collection

**Fact-checks:** We identified three fact-checking organizations that thoroughly investigate rumors related to COVID-19. The names of the fact-checking organizations are: politifact.com, truthorfiction.com, and snopes.com. These fact-checking organizations list COVID-19 rumors in separate categories or tag them with a topic label (e. g., “COVID-19”, “Coronavirus”) which allows us to distinguish COVID-19-related rumors from other rumors (see Tbl. 1). We scraped all COVID-19-related fact-checks from these platforms.

**Table 1: Tags used to identify COVID-19-related fact-checks from fact-checking organizations.**

Fact-checking organization	Tag	#Fact-checks
politifact.org	Coronavirus	403
snopes.com	COVID-19	265
truthorfiction.com	covid-19	44

The fact-checking organizations have different ways of labeling the veracity of a rumor. For example, politifact.com articles are given a “Pants on Fire” rating for false rumors, whereas snopes.com assigns a “false” label. Consistent with Vosoughi et al. [65], we normalized the veracity labels across the different sites by mapping them to a score of 1 to 5. All rumors with a score of 1 or 2 were categorized as “false,” whereas rumors with a score of 4 or 5 were categorized as “true.” Rumors with a score of 3 were categorized as “mixed.” In some cases, the same rumors have been investigated by multiple fact-checking organizations. Previous research has shown that fact-checking websites show high pairwise agreement [65], ranging between 95 % and 98 %. Rumors classified as “true” or “false” even showed a perfect pairwise agreement of 100 % [65]. The resulting collection of fact-checks contained the following information: (i) the veracity label (“true”, “false”, “mixed”), (ii) links to the articles of the fact-checking organizations, and (iii) the headline of the article that is being verified.

**Rumor cascades on Twitter:** We followed the approach from Vosoughi et al. (2018) to identify rumor cascades on Twitter: A rumor cascade on Twitter starts with a user making an assertion about a topic such as tweeting a text message or a link to an article. Social media users then propagate the rumor by retweeting it. Oftentimes, people also reply to the original tweet. These replies sometimes contain links to fact-checking organizations that either confirm or debunk the rumor in the original tweet. We used such cascades to identify rumor cascades that are propagating on Twitter.

We employed the Twitter Historical API to map the rumors to retweet cascades on Twitter as follows. First, we collected all tweets that contain a link to any of the websites from the fact-checking organizations. Second, for each reply tweet, we extracted the original tweet and the number of retweets of the original tweet. Here, special care is needed to ensure that the replies containing a link to any of the trusted websites address the original tweet. We followed the approach from Vosoughi et al. [65] to address this important issue: (i) we considered only replies to the original tweet and exclude replies to replies. (ii) To ensure that we study how unverified and contested information diffuses on Twitter, we removed all original tweets that are directly linking to one of the fact-checking websites. Note that tweets linking to one of the fact-checking websites do not qualify as they are no longer unverified. (iii) We compared the headline of the linked article to that of the original tweet. For this purpose, we used Universal Sentence Encoder [11] to convert the headline of the fact-check and the original tweet to vector representations that capture their semantic content. We then used cosine similarity to measure the distance between the vectors. If the cosine similarity was lower than 0.4, the tweet was discarded.

The retweet cascades remaining after these filtering steps then represent rumors propagating on Twitter – for which a veracity

**Table 2: Summary statistics for tweets of rumor starters. Mean values are highlighted in bold, standard deviations are shown in parentheses. All Twitter variables were obtained from the Twitter Historical API.**

Variable	All cascades	POLITICS	HEALTH	OTHER
Dates collected	01/02/20 – 05/13/21	01/02/20 – 05/13/21	01/27/20 – 05/13/21	01/27/20 – 05/12/21
Number of cascades	10,610	8,157	4,116	1,297
Number of retweets	24,339,625	20,374,097	10,231,382	1,416,474
Retweet count range	0 – 260,637	0 – 260,637	0 – 207,155	0 – 76,092
Proportion <i>True</i>	35.3%	39.0%	34.7%	19.7%
Proportion <i>False</i>	46.9%	42.7%	48.3%	64.8%
Proportion <i>Mixed</i>	17.7%	18.3%	17.0%	15.6%
Followers	<b>2,256,095</b> (7,700,566)	<b>2,545,874</b> (8,260,175)	<b>2,526,538</b> (9,166,450)	<b>816,527.4</b> (3,859,619)
Followees	<b>9,193.9</b> (34,750.39)	<b>10,124.4</b> (37,507.33)	<b>9,320.23</b> (40,249.53)	<b>5,952.10</b> (25,541.03)
Account age	<b>3,333.35</b> (1,383.38)	<b>3,374.93</b> (1,376.82)	<b>3,321.95</b> (1,391.67)	<b>3,098.33</b> (1,386.78)
Verified users	55.1%	60.2%	56%	31.9%
Includes media	28.6%	27%	26.4%	38.2%
Other-condemning emotions	<b>0.167</b> (0.217)	<b>0.164</b> (0.201)	<b>0.153</b> (0.190)	<b>0.189</b> (0.291)
Self-conscious emotions	<b>0.300</b> (0.209)	<b>0.294</b> (0.198)	<b>0.317</b> (0.196)	<b>0.321</b> (0.256)

label is known based on the assessment from the fact-checking organization. In our data, the frequencies of fact-checking labels at cascade level are: 3,748 (=true), 4,979 (=false), and 1,883 (=mixed). These 10,610 rumor cascades have received more than 24.33 million retweets by Twitter users.

Following previous works [8, 65], we employed the Twitter API to collect a set of additional user variables for each source tweet, i. e., the number of followers, the account age, etc. These variables are known to affect the retweet rate and are later used as control variables in our regression model. Summary statistics of our dataset are reported in Tbl. 2.

### 3.2 Calculation of Emotion Scores

The “other-condemning” family of moral emotions comprises the emotions *anger*, *disgust*, and *contempt*, whereas the “self-conscious” family comprises the emotions *shame*, *pride*, and *guilt* [30, 60]. We employed text mining methods to measure the extent to which these emotions are embedded in the source tweets. For this purpose, we first applied standard preprocessing steps from text mining. Specifically, the running text was converted into lower-case and tokenized, and special characters (e. g., hashtags, emoticons) were removed. Subsequently, we applied (and validated) a dictionary-based approach analogous to earlier research [8, 48, 65].

We measured other-condemning and self-conscious emotions embedded in the source tweets based on the NRC emotion lexicon [40]. This lexicon comprises 181,820 English words that are classified according to the emotions of Plutchik’s emotion model [46]. Plutchik’s emotion model defines 8 basic emotions and 24 emotional dyads. The emotional dyads represent complex emotions, which are derived as a combination of two basic emotions [49]. We used the NRC dictionary to count the frequency of words in the tweets that belong to each of the emotions. Afterwards, we divided 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 emotions [48, 65]. In our data, 78.15% of all source tweets contained at least one emotion word from the NRC lexicon. We filtered out tweets that do not contain any emotional words since, otherwise, the denominator is not defined [48, 65]. However, our later analysis yields qualitatively identical results when including these observations (i. e., assigning zero values). Based on the scores for the 8 basic emotions and the 24 derived emotions, and the definitions of the two moral emotion families [60], we calculated other-condemning emotions by taking the sum of *anger*, *disgust*, and *contempt*. Self-conscious emotions were calculated by taking the sum of *shame*, *pride*, and *guilt*.

**User study:** In order to test the construct validity of our dictionary-based approach, we employed two trained research assistants to annotate a random subset of 200 tweets that were categorized as being more other-condemning than self-conscious based on the dictionaries; and a random subset of 200 tweets that were categorized as being more self-conscious than other-condemning. For each of the 400 tweets, the annotators were asked to what extent the tweet relates to other-condemning and self-conscious emotions on two 5-point Likert scales, ranging from 1 (“not related to [other-condemning, self-conscious] emotions at all”) to 5 (“very related to [other-condemning, self-conscious] emotions”). The annotators viewed the tweets in randomized order and were explained the difference between other-condemning and self-conscious emotions. The annotators exhibited a statistically significant inter-rater agreement according to Kendall’s  $W$  ( $p < 0.01$ ). Furthermore, the annotators rated the random subset of other-condemning tweets as more “other-condemning” than “self-conscious” [ $t = 6.53, p < 0.001$ ]; and the random subset of self-conscious tweets as more “self-conscious” than “other-condemning” [ $t = 4.50, p < 0.001$ ].

**Table 3: Exemplary tweets of rumor starters for each topic.**

Topic	Veracity	Twitter Message
POLITICS	True	Trump fired the Pandemic response team in 2018... He did not replace them... #TrumpYoureKilling
POLITICS	False	Sick: Nancy Pelosi tried to insert abortion funding measures into the Chinese Coronavirus response stimulus package I never want to hear that Donald Trump is politicizing this pandemic again while Democrats try this stunt This is a disgrace—Speaker Pelosi should be ashamed
HEALTH	True	More police officers have died from Covid-19 this year than have been killed on patrol. Gunfire is the second-highest cause of death.
HEALTH	False	80% of People Taking Moderna Vaccine Had Significant Side-Effects. While the killer Bill Gates laughs all the way to the bank. Stop this insanity now!
OTHER	True	This is the first day of school in Paulding County, Georgia.
OTHER	False	I thought this was supposed to be a conspiracy theory. But here it is, straight from Trudeau’s mouth. The pandemic is the excuse for a “Great Reset” of the world, led by the UN.

### 3.3 Rumor Topics

We employed a weakly supervised machine learning framework [70] to infer the topics in the source tweets that have initiated the rumor cascades. The benefit of this state-of-the-art approach is that (i) it is regarded as superior to conventional topic modeling (i. e., Latent Dirichlet Allocation) for short texts [70], and (ii) its weakly supervised nature allows for an ex-ante selection of topics that we perceive as being particularly relevant in the context of COVID-19. We categorized the rumor cascades into three (not mutually exclusive) topics: HEALTH (e. g., rumors about the safety of vaccines), POLITICS (e. g., allegations of political opponents), and OTHER (i. e., rumors that do not fall into one of the other categories). Example tweets for each topic are provided in Tbl. 3.

Our weakly supervised machine learning framework proceeded in three steps (see Yao et al. [70] for methodological details): (1) We started to identify topic-related tweets based on a set of manually selected keywords for each topic. For instance, for the topic HEALTH, we searched for all tweets containing words such as “vaccine,” “flu,” “mask,” etc. (see list of keywords in the Supplementary Materials). (2) We conducted clustering-assisted manual word sense disambiguation on the keyword-identified tweets [70]. Here we used the  $k$ -means clustering algorithm with Silhouette criterion to cluster the keyword-identified tweets for each topic. We then manually inspect random tweets sampled from each cluster and assessed whether the tweets in the cluster refer to the topic. We excluded each tweet cluster that does not show the pertinent meaning of the topic keyword. This allowed us to significantly clean and improve the quality of the keyword-identified tweets. (3) We used the created labeled data to train a deep neural network classifier and learn to predict whether or not individual Twitter messages belong to a certain topic. The input data for the training machine learning classifier was a vector representation of the (cleaned) keyword-identified tweets and the topic label. To create vector representations of tweets, we used neural language models in the form of the Universal Sentence Encoder [11]. In our deep neural network classifier, we treated the task of predicting topic labels for (vector representations of) tweets as a multi-label problem considering that one tweet may belong to multiple topics (i. e., HEALTH and POLITICS). In training, we used an equal number of 1000 keyword-identified Tweets for each topic as positive training instances. In addition, we used the excluded tweets from step (2) and randomly sampled unlabeled tweets equal

to the sum of labeled tweets as negative training instances, i. e., with a topic label OTHER.

**User study:** To ensure that the topic predictions are accurate, we tested for the presence of errant tweets with the help of two trained research assistants. We randomly sampled 200 tweets for each topic, and instructed the research assistants to annotate the tweets. Each annotator was asked to judge the validity of the topic label on a 5-point Likert scale, ranging from 1 (“not related to [topic] at all”) to 5 (“very related to [topic]”). When comparing the human annotations to the predicted topic labels, we found very few misclassified instances. On average, the share of tweets that were not classified as at least “somewhat related to [topic]” was lower than 8.5 % (see Supplementary Materials).

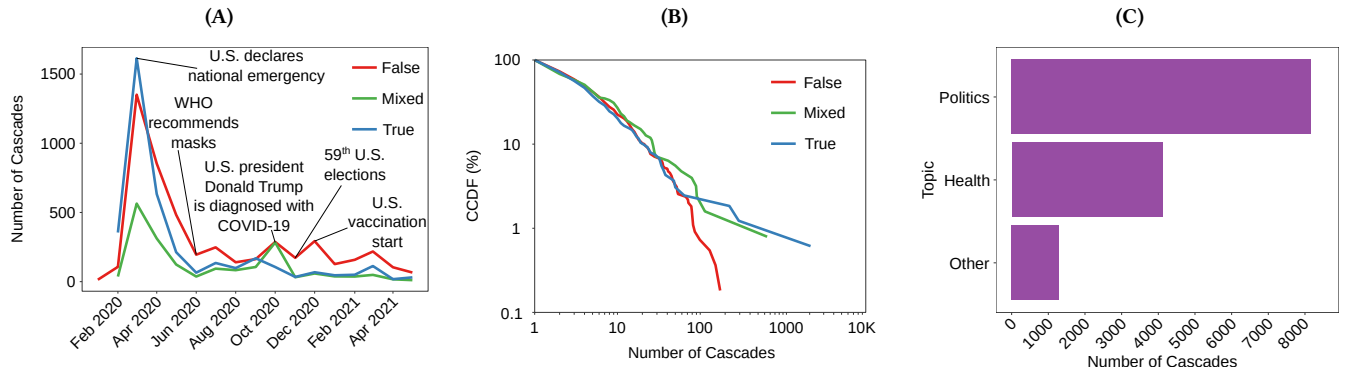
### 3.4 Model Specification

We specified regression models with interaction terms that explain the number of retweets based on rumor veracity and other-condemning emotions and self-conscious emotions. Let  $RetweetCount_i$  denote the number of retweets for rumor cascade  $i$ . Furthermore, let  $OtherCondemning_i$  denote the proportion of other-condemning emotions,  $SelfConscious_i$  the proportion of self-conscious emotions, and  $Falsehood_i$  the veracity. Here we define a true rumor as  $Falsehood_i = 0$  and a false rumor as  $Falsehood_i = 1$ . We adjusted for variables known to affect retweet rate [8, 47–49, 58, 65], which included the number of followers ( $Followers_i$ ) and followees ( $Followees_i$ ) of the author of the tweet, the account age ( $AccountAge_i$ ), whether the author was verified by Twitter ( $Verified_i$ ), and whether media was attached to the tweet ( $HasMedia_i$ ). Each of these factors was extracted from the Twitter API. We  $z$ -standardized all continuous predictors in order to facilitate interpretability.

Based on the above variables, we specified the following generalized linear model for our analysis:

$$\begin{aligned} \log(E(RetweetCount_i \mid *)) = & \beta_0 + \beta_1 Falsehood_i \\ & + \beta_2 Falsehood_i \times OtherCondemning_i \\ & + \beta_3 Falsehood_i \times SelfConscious_i \\ & + \beta_4 OtherCondemning_i + \beta_5 SelfConscious_i \\ & + \beta_6 Followers_i + \beta_7 Followees_i + \beta_8 AccountAge_i \\ & + \beta_9 HasMedia_i + \beta_{10} Verified_i \end{aligned} \quad (1)$$

with intercept  $\beta_0$ .



**Figure 1: COVID-19 rumor cascades propagating on Twitter between January 2020 and the end of April 2021. (A) Monthly counts of true, false, and mixed rumor cascades. (B) Complementary cumulative distribution functions (CCDFs) of true, false, and mixed rumor cascades. (C) Number of rumor cascades across different topics.**

*RetweetCount* is a non-negative count variable, and its variance is larger than the mean. To adjust for overdispersion, we drew upon a negative binomial regression [8, 58]. Note that because we estimate a negative binomial regression model with interaction terms, the coefficients cannot be interpreted as the change in the mean of the dependent variable for a one unit (i. e., standard deviation) increase in the respective predictor variable, with all other predictors remaining constant. The reason is that in nonlinear regression models with interaction terms, marginal effects are nonlinear functions of the coefficients and the levels of the explanatory variables [9]. Instead, the coefficients can be interpreted on a multiplicative scale by calculating the incidence rate ratio (IRR), which is equal to the exponent of the coefficient of the respective variable [9]. Here the coefficients can be interpreted as the natural logarithm of a multiplying factor by which the predicted number of retweets changes, given a one unit increase in the predictor variable, holding all other predictor variables constant [9].

## 4 RESULTS

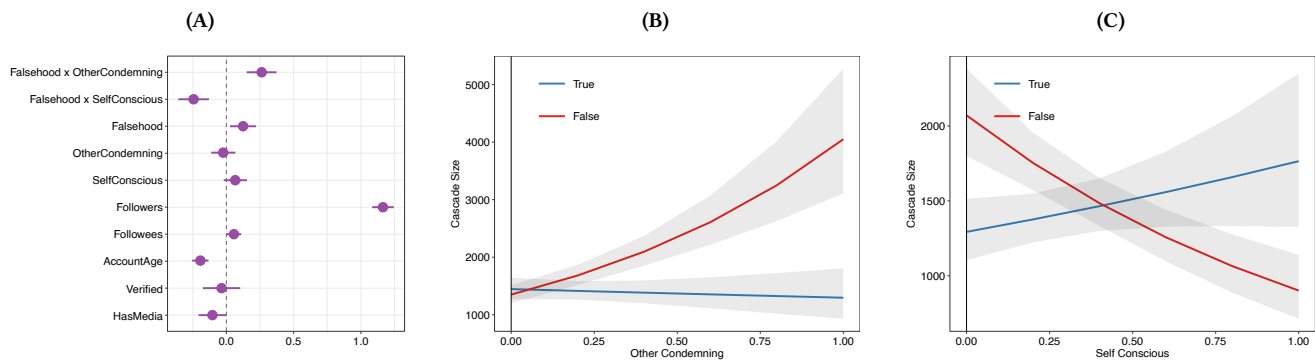
Our data include 10,610 rumor cascades that have been retweeted 24.34 million times. The total number of COVID-19 rumor cascades peaked in March 2020 when the U. S. government declared a national emergency concerning the coronavirus disease and again in October 2020, the month prior to the U. S. presidential elections (Fig. 1A). The three fact-checking organizations have categorized 46.9 % of all rumors as false, 35.3 % as true, and 17.7 % as being of mixed veracity. While the absolute number of rumor cascades has decreased over the course of the pandemic, the relative share of false vs. true rumors has increased (Fig. 1A). Compared to false rumors, a greater fraction of true rumors experienced more than 100 rumor cascades (Fig. 1B). COVID-19 rumors are not constrained exclusively to health topics (e. g., rumors about the safety of vaccines). Rather, a sizable number of COVID-19 rumors concern political topics (e. g., true or false allegations of political opponents) [17]. We thus applied topic modeling to categorize the rumor cascades in our dataset into three (not mutually exclusive) topics: POLITICS, HEALTH, and OTHER. Fig. 1C shows that a large proportion of

COVID-19 rumors were thematically related to POLITICS (76.9 %), HEALTH (38.8 %), while only 12.2 % concerned OTHER topics (e. g., conspiracy theories). A total share of 34.1 % of rumor cascades were thematically related to both POLITICS and HEALTH.

**Regression analysis:** We fitted explanatory regression models to evaluate how variations in moral emotions are associated with differences in the number of retweets for true vs. false rumor cascades. In our regression analysis, we followed previous works [8, 65] by controlling for variables known to affect the retweet rate independent of the main predictors, i. e., the number of followers, the account age, etc.

As a baseline, we started our regression analysis with a negative binomial regression explaining the number of retweets solely based on the veracity label and control variables (see Supplementary Materials). Here false rumors (Falsehood = 1) were estimated to receive 15.66 % more retweets than true rumors (IRR 1.16;  $p < 0.01$ ). Subsequently, we extended the negative binomial regression by including interaction terms between rumor veracity and other-condemning emotions, and between rumor veracity and self-conscious emotions (Fig. 2A). The coefficient estimates for these two interaction terms were statistically significant, which implies that false rumors' virality depended on the moral emotions embedded in the source tweet. Specifically, a one standard deviation increase in other-condemning emotions for false rumors was linked to a 26.99 % increase in the number of retweets (IRR 1.27;  $p < 0.01$ ). In contrast, a one standard deviation increase in self-conscious emotions for false rumors was linked to a 23.43 % decrease in the number of retweets (IRR 1.23;  $p < 0.01$ ). We found no statistically significant effect of other-condemning and self-conscious emotion words for true rumors. In sum, we observed that false rumors were more viral than the truth if the source tweet embedded a high proportion of other-condemning emotion words, whereas a high proportion of self-conscious emotion words was linked to a less viral spread (see Fig. 2B, 2C).

**Analysis across topics:** We also examined the effect of moral emotions across different topics. For each topic from Fig. 1C, we generated observation subsets and re-estimated our regression model (Fig. 3). We observed differences in the effects of moral emotions on the number of retweets. The effect of other-condemning emotions



**Figure 2: Increases in other-condemning emotions predict higher retweet counts for false rumors, whereas increases in self-conscious emotions predict less retweets. (A) Coefficient estimates for negative binomial regression with 95 % confidence intervals. The dependent variable is the number of retweets. (B–C) Predicted marginal means of the number of retweets for other-condemning emotions and self-conscious emotions. The 95 % confidence intervals are highlighted in gray.**

on the number of retweets was pronounced both for false rumors from the HEALTH category (IRR 1.34;  $p < 0.01$ ) and for false rumors from the POLITICS category (IRR 1.61;  $p < 0.01$ ). For the OTHER category (with comparatively smaller sample size), the coefficients pointed in the same directions but were not statistically significant at common significance thresholds.

**Analysis of deleted tweets:** Major social media platforms such as Twitter have intensified their efforts to combat the spread of misinformation on their platforms by deleting misinformation [6]. While the Twitter API does not provide access to the content of source tweets that have been deleted, we were still able to analyze some of their characteristics. As part of an exploratory analysis, we found that 3663 potential rumor starter tweets have been deleted (either by Twitter or by the users themselves). An overwhelming majority of those (68.35 %) are potentially false rumors. Hence, even though these numbers suggest that a relevant proportion of false rumors on Twitter has been deleted, the vast majority of false rumors continue to circulate. For those rumors, our results demonstrate that falsehood can be more viral than the truth.

**Additional checks<sup>1</sup>:** Numerous exploratory analyses and checks validated our results and confirmed their robustness: (i) Since self-conscious emotions can be regarded as the counterpart of other-condemning emotions, we tested an alternative model specification in which we included the *difference* between the emotion scores for other-condemning and self-conscious emotions instead of two individual variables. Consistent with our main analysis, we found that false rumors are more viral than the truth if the source tweet embeds a high proportion of other-condemning emotion words, whereas a high proportion of self-conscious emotion words is linked to a less viral spread. (ii) In our main analysis, we focused on rumors that are clearly true or false. However, 17.7 %, of all rumors have been categorized as being of mixed veracity by the fact-checking organizations. We tested whether counting rumors of mixed veracity as either true or false affects the validity of our results. We

find that our results are robust and that the combination of other-condemning emotion and mixed veracity is similarly viral as the combination of other-condemning emotions and false veracity. (iii) In our data, 9.48 % of rumor starters have started more than one retweet cascade. To ensure that our models are not biased due to this source of non-independence, we dropped all users with clustering and reestimated the models. The results are robust and support our findings. We also repeated our analysis with monthly fixed effects to control for differences in the virality of rumor cascades due to different start dates. Also here, the results confirmed the findings from our main analysis. (iv) We repeated our analysis for subsets of rumor cascades that have been started by users that are either verified or not verified by Twitter. We find that our main findings hold for both user groups.

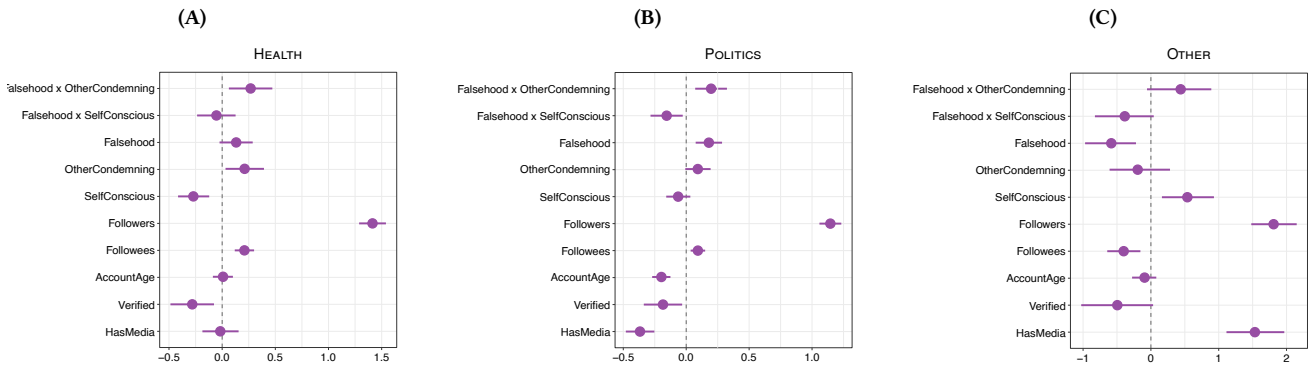
## 5 DISCUSSION

Here we provide evidence that moral emotions play a crucial role in the spread of COVID-19 misinformation on social media. Using a comprehensive dataset of COVID-19 rumors that have been fact-checked by three independent fact-checking organizations (snopes.com, politifact.com, truthorfiction.com), we establish that other-condemning emotions – also known as the hostility triad – are linked to a more viral spread of false rumors.

While false rumors pose a threat to the successful overcoming of this pandemic, an understanding of how rumors diffuse in online social networks is – even for non-crisis situations – still in its infancy. Analyzing the spreading dynamics of fact-checked rumors is to a great extent generalizable to the spread of other (non-fact-checked) rumors on social media [65]. Our finding that COVID-19 misinformation is, on average, more viral than the truth directly connects to the study from Vosoughi et al. [65], which yielded similar findings, yet outside the context of COVID-19. Previous research has also shown that misinformation on social media can have negative offline consequences. Among other instances, this has previously been confirmed to be the case during humanitarian crises [57] and elections [2–4, 29]. Our observation that COVID-19 misinformation is both widespread and viral on social media is at

<sup>1</sup>Detailed results are reported in the Supplementary Materials.





**Figure 3: Coefficient estimates for negative binomial regressions with 95 % confidence intervals for rumor cascades filtered by topic (A: HEALTH, B: POLITICS, and C: OTHER). The dependent variable is the number of retweets.**

least equally concerning. COVID-19 misinformation not only poses severe health risks to individuals but also undermines the integrity of the political discourse [50].

The results of this study highlight the role of moral emotions in rumor diffusion. Previous research [8] broadly distinguished moral vs. non-moral emotional expressions in social media content, while this work demonstrates that the two clusters of moral emotions (self-conscious vs. other-condemning emotions) have distinct effects on social transmission in the context of true and false rumors. We observe that false rumors receive more retweets than true rumors if the source tweets embed a high share of other-condemning emotions, whereas we find the opposite pattern, yet of smaller magnitude, for self-conscious emotions. Another relevant finding is that the expression of other-condemning emotion on virality is pronounced both for health misinformation and political misinformation. These findings may be partially explained by the high level of polarization of social media users in the context of COVID-19. In polarizing debates, radical ideas and beliefs are strengthened and more likely to translate into action. It thus seems plausible that the explosive mix of other-condemning emotions accelerates the spread of false rumors about those topics within social networks.

From a practical perspective, policy initiatives around the world urge social media platforms to limit the spread of false rumors [37]. While previous research has studied emotions in replies to rumor cascades [48, 49], our work highlights the importance of considering (moral) emotions in the source tweets that have initiated the rumor cascades. These findings could eventually be leveraged in machine learning models in order to detect false rumors more accurately. Emotion scores for source tweets are available immediately upon the beginning of the diffusion process – a time point at which features from propagation dynamics are scarce [16]. Our findings may also be relevant with regard to other downstream tasks such as educational applications. Altogether, considering moral emotions in social media posts might help future works to develop more effective strategies against false rumors.

This work is subject to the typical limitations of observational studies. We report associations and refrain from making causal claims. Future work should seek to corroborate our conclusions in controlled laboratory experiments and, in particular, test the causal influence of exposure to moral-emotional language on attitudes

and behavior. Our inferences are also limited by the accuracy and availability of our data, specifically those from the three different fact-checking websites. For those, however, our data comprises all COVID-19 rumor cascades on Twitter until the end of April 2021. Despite these limitations, we believe that observing and understanding how misinformation spreads is the first step toward containing it. We hope that our work inspires more research into the causes, consequences, and potential countermeasures for the spread of misinformation – both in crisis and non-crisis situations.

## 6 CONCLUSION

While false rumors pose a threat to the successful overcoming of this pandemic, an understanding of “what makes false rumors viral” is – even for non-crisis situations – still in its infancy. In this work, we approach this question through the lenses of morality and emotions and their role in rumor diffusion in polarized social media environments. For this purpose, we collected a unique dataset of COVID-19-related rumor cascades from Twitter and empirically analyze their spreading dynamics. We find that COVID-19 misinformation is, on average, more viral than the truth. However, the veracity effect is moderated by moral emotions: false rumors are more viral than the truth if the source tweets embed a high number of other-condemning emotion words, whereas a higher number of self-conscious emotion words is linked to a less viral spread. These findings offer insights into how true vs. false rumors spread and highlight the importance of considering moral emotions in social media content.

## ACKNOWLEDGMENTS

This study was supported by a grant from the German Research Foundation (DFG grant 455368471).



## REFERENCES

- [1] Hunt Allcott, Levi Boxell, Jacob Conway, Matthew Gentzkow, Michael Thaler, and David Yang. 2020. Polarization and public health: partisan differences in social distancing during the coronavirus pandemic. *Journal of Public Economics* 191 (2020), 104254.
- [2] Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of Economic Perspectives* 31, 2 (2017), 211–236. <https://doi.org/10.1257/jep.31.2.211>
- [3] Sinan Aral and Dean Eckles. 2019. Protecting elections from social media manipulation. *Science* 365, 6456 (2019), 858–861. <https://doi.org/10.1126/science.aaw8243>
- [4] Eytan Bakshy, Solomon Messing, and Lada Adamic. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science* 348, 6239 (2015), 1130–1132. <https://doi.org/10.1126/science.aaa1160>
- [5] Pablo Barberá, John T. Jost, Jonathan Nagler, Joshua A. Tucker, and Richard Bonneau. 2015. Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological Science* 26, 10 (2015), 1531–1542. <https://doi.org/10.1177/0956797615594620>
- [6] BBC. 2017. Coronavirus: World leaders' posts deleted over fake news. <https://www.bbc.com/news/technology-52106321>
- [7] Alessandro Bessi, Mauro Coletto, George Alexandru Davidescu, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. 2015. Science vs conspiracy: Collective narratives in the age of misinformation. *PLOS ONE* 10, 2 (2015), e0118093. <https://doi.org/10.1371/journal.pone.0118093>
- [8] William J. Brady, Julian A. Wills, John T. Jost, Joshua A. Tucker, and Jay J. van Bavel. 2017. Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences (PNAS)* 114, 28 (2017), 7313–7318.
- [9] Maarten L. Buis. 2010. Stata tip 87: Interpretation of interactions in nonlinear models. *The Stata Journal* 10, 2 (2010), 305–308.
- [10] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on Twitter. In *WWW*. <https://doi.org/10.1145/1963405.1963500>
- [11] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Lintiac, Rhomni St. John, Noah Constant, Guajardo-Céspedes, Steve Yuan, Cris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal Sentence Encoder. *arXiv* 1803.11175 (2018).
- [12] Daejin Choi, Selin Chun, Hyunchul Oh, Jinyoung Han, and Ted Taekyoung Kwon. 2020. Rumor propagation is amplified by echo chambers in social media. *Scientific Reports* 10, 1 (2020), 310. <https://doi.org/10.1038/s41598-019-57272-3>
- [13] Nicholas A Christakis and James H Fowler. 2009. *Connected: the surprising power of our social networks and how they shape our lives*. Little, Brown and Company.
- [14] Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. 2021. The echo chamber effect on social media. *PNAS* 118, 9 (2021).
- [15] Matteo Cinelli, Walter Quattrociocchi, Alessandro Galeazzi, Carlo Michele Valensise, Emanuele Brugnoti, Ana Lucia Schmidt, Paola Zola, Fabiana Zollo, and Antonio Scala. 2020. The COVID-19 social media infodemic. *Scientific Reports* 10, 1 (2020), 1–10.
- [16] Mauro Conti, Daniele Lain, Riccardo Lazzaretto, Giulio Lovisotto, and Walter Quattrociocchi. 2017. It's always April fools' day! On the difficulty of social network misinformation classification via propagation features. In *IEEE Workshop on Information Forensics and Security (WIFS)*. <https://doi.org/10.1109/WIFS.2017.8267653>
- [17] Alessandro Cossard, Gianmarco De Francisci Morales, Kyriaki Kalimeri, Yelena Mejova, Daniela Paolotti, and Michele Starnini. 2020. Falling into the echo chamber: the Italian vaccination debate on Twitter. In *ICWSM*.
- [18] M. de Domenico, A. Lima, P. Mougél, and M. Musolesi. 2013. The anatomy of a scientific rumor. *Scientific Reports* 3, 2980 (2013). <https://doi.org/10.1038/srep02980>
- [19] Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H. Eugene Stanley, and Walter Quattrociocchi. 2016. The spreading of misinformation online. *Proceedings of the National Academy of Sciences of the United States of America* 113, 3 (2016), 554–559. <https://doi.org/10.1073/pnas.1517441113>
- [20] Michela Del Vicario, Walter Quattrociocchi, Antonio Scala, and Fabiana Zollo. 2019. Polarization and fake news. *ACM Transactions on the Web* 13, 2 (2019), 1–22. <https://doi.org/10.1145/3316809>
- [21] James N Druckman, Samara Klar, Yanna Krupnikov, Matthew Levendusky, and John Barry Ryan. 2021. How affective polarization shapes Americans' political beliefs: a study of response to the COVID-19 pandemic. *Journal of Experimental Political Science* 8, 3 (2021), 223–234.
- [22] Ullrich K. H. Ecker, Stephan Lewandowsky, and David T. W. Tang. 2010. Explicit warnings reduce but do not eliminate the continued influence of misinformation. *Memory & Cognition* 38, 8 (2010), 1087–1100.
- [23] James H Fowler and Nicholas A Christakis. 2008. Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the Framingham Heart Study. *BMJ* 337 (2008), a2338.
- [24] Sheera Frenkel, Davey Alba, and Raymond Zhong. 2020. Surge of virus misinformation stumps Facebook and Twitter. *The New York Times* 8 (2020).
- [25] Adrien Friggeri, Lada Adamic, Dean Eckles, and Justin Cheng. 2014. Rumor cascades. In *ICWSM*.
- [26] Riccardo Gallotti, Francesco Valle, Nicola Castaldo, Pierluigi Sacco, and Manlio De Domenico. 2020. Assessing the risks of 'infodemics' in response to COVID-19 epidemics. *Nature Human Behaviour* 4, 12 (2020), 1285–1293.
- [27] Christine Geeng, Savanna Yee, and Franziska Roesner. 2020. Fake news on Facebook and Twitter: investigating how people (don't) investigate. In *CHI*.
- [28] Janessa Griffith, Husayn Marani, and Helen Monkman. 2021. COVID-19 vaccine hesitancy in Canada: Content analysis of tweets using the theoretical domains framework. *Journal of Medical Internet Research* 23, 4 (2021), e26874.
- [29] Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. 2019. Fake news on Twitter during the 2016 U.S. presidential election. *Science* 363, 6425 (2019), 374–378. <https://doi.org/10.1126/science.aau2706>
- [30] Jonathan Haidt. 2003. *The moral emotions*. Oxford University Press.
- [31] P Sol Hart, Sedona Chinn, and Stuart Soroka. 2020. Covid19? Politicization and polarization in COVID-19 news coverage. *Science Communication* 42, 5 (2020), 679–697.
- [32] Nicholas Francis Havey. 2020. Partisan public health: how does political ideology influence support for COVID-19 related misinformation? *Journal of Computational Social Science* 3, 2 (2020), 319–342.
- [33] Md Saiful Islam, Tonmoy Sarkar, Sazzad Hossain Khan, Abu-Hena Mostofa Kamal, SM Murshid Hasan, Alamgir Kabir, Dalia Yeasmin, Mohammad Ariful Islam, Kamal Ibne Amin Chowdhury, Kazi Selim Anwar, et al. 2020. COVID-19-related infodemic and its impact on public health: A global social media analysis. *The American Journal of Tropical Medicine and Hygiene* 103, 4 (2020), 1621.
- [34] Elise Jing and Yong-Yeol Ahn. 2021. Characterizing partisan political narrative frameworks about COVID-19 on Twitter. *EPJ Data Science* 10, 1 (2021), 53.
- [35] Antino Kim and Alan R. Dennis. 2019. Says who? The effects of presentation format and source rating on fake news in social media. *MIS Quarterly* 43, 3 (2019), 1025–1039. <https://doi.org/10.2530/MISQ/2019/15188>
- [36] Ramez Kouzy, Joseph Abi Jaoude, Afif Kraitem, Molly B El Alam, Basil Karam, Elio Adib, Jabra Zarka, Cindy Traboulsi, Elie W Akl, and Khalil Baddour. 2020. Coronavirus goes viral: quantifying the COVID-19 misinformation epidemic on Twitter. *Cureus* 12, 3 (2020).
- [37] David M. J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain. 2018. The science of fake news. *Science* 359, 6380 (2018), 1094–1096. <https://doi.org/10.1126/science.aao2998>
- [38] Ro'ee Levy. 2021. Social media, news consumption, and polarization: Evidence from a field experiment. *American Economic Review* 111, 3 (2021), 831–870.
- [39] Rupali Jayant Limaye, Molly Sauer, Joseph Ali, Justin Bernstein, Brian Wahl, Anne Barnhill, and Alain Labrique. 2020. Building trust while influencing online COVID-19 content in the social media world. *The Lancet Digital Health* 2, 6 (2020), 277–278.
- [40] Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence* 29, 3 (2013), 436–465. <https://doi.org/10.1111/j.1467-8640.2012.00460.x>
- [41] Patricia L. Moravec, Randall K. Minas, and Alan Dennis. 2019. Fake news on social media: People believe what they want to believe when it makes no sense at all. *MIS Quarterly* 43, 4 (2019), 1343–1360.
- [42] Onook Oh, Manish Agrawal, and H. Raghav Rao. 2013. Community intelligence and social media services: A rumor theoretic analysis of tweets during social crises. *MIS Quarterly* 37, 2 (2013), 407–426.
- [43] Gordon Pennycook, Tyrone D. Cannon, and David G. Rand. 2018. Prior exposure increases perceived accuracy of fake news. *Journal of Experimental Psychology: General* 147, 12 (2018), 1865–1880. <https://doi.org/10.1037/xge0000465>
- [44] Gordon Pennycook, Ziv Epstein, Mohsen Mosleh, Antonio A. Arechar, Dean Eckles, and David G. Rand. 2021. Shifting attention to accuracy can reduce misinformation online. *Nature* 592, 7855 (2021), 590–595.
- [45] Gordon Pennycook, Jonathon McPhetres, Yunhao Zhang, Jackson G. Lu, and David G. Rand. 2020. Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological Science* 31, 7 (2020), 770–780.
- [46] Robert Plutchik. 1984. *Emotion: Theory, research, and experience*. Academic Press, Orlando.
- [47] Nicolas Pröllochs. 2022. Community-based fact-checking on Twitter's Birdwatch platform. In *ICWSM*.
- [48] Nicolas Pröllochs, Dominik Bär, and Stefan Feuerriegel. 2021. Emotions explain differences in the diffusion of true vs. false social media rumors. *Scientific Reports* 11, 22721 (2021).
- [49] Nicolas Pröllochs, Dominik Bär, and Stefan Feuerriegel. 2021. Emotions in online rumor diffusion. *EPJ Data Science* 10, 1 (2021), 51.
- [50] David N. Rapp and Nikita A. Salovich. 2018. Can't we just disregard fake news? The consequences of exposure to inaccurate information. *Policy Insights from the Behavioral and Brain Sciences* 5, 2 (2018), 232–239.

- [51] Julie Ricard and Juliano Medeiros. 2020. Using misinformation as a political weapon: COVID-19 and Bolsonaro in Brazil. *Harvard Kennedy School Misinformation Review* 1, 3 (2020).
- [52] Paul Rozin, Laura Lowery, Sumio Imada, and Jonathan Haidt. 1999. The CAD triad hypothesis: a mapping between three moral emotions (contempt, anger, disgust) and three moral codes (community, autonomy, divinity). *Journal of Personality and Social Psychology* 76, 4 (1999), 574–586.
- [53] Chengcheng Shao, Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. 2016. Hoaxy: A platform for tracking online misinformation. In *WWW*.
- [54] Lisa Singh, Shweta Bansal, Leticia Bode, Ceren Budak, Guangqing Chi, Kornraphop Kawintiranon, Colton Padden, Rebecca Vanarsdall, Emily Vraga, and Yanchen Wang. 2020. A first look at COVID-19 information and misinformation sharing on Twitter. *arXiv* 2003.13907 (2020).
- [55] Kirill Solovev and Nicolas Pröllochs. 2022. Hate speech in the political discourse on social media: disparities across parties, gender, and ethnicity. In *WWW*.
- [56] Kate Starbird. 2017. Examining the alternative media ecosystem through the production of alternative narratives of mass shooting events on Twitter. In *ICWSM*.
- [57] Kate Starbird, Jim Maddock, Mania Orand, Peg Achterman, and Robert M. Mason. 2014. Rumors, false flags, and digital vigilantes: Misinformation on Twitter after the 2013 Boston marathon bombing. In *iConference*.
- [58] Stefan Stieglitz and Linh Dang-Xuan. 2013. Emotions and information diffusion in social media: sentiment of microblogs and sharing behavior. *Journal of Management Information Systems* 29, 4 (2013), 217–248. <https://doi.org/10.2753/MIS0742-1222290408>
- [59] June Price Tangney, Jeff Stuewig, and Debra J Mashek. 2007. Moral emotions and moral behavior. *Annu. Rev. Psychol.* 58 (2007), 345–372.
- [60] Jessica L Tracy and Richard W Robins. 2004. Putting the self into self-conscious emotions: a theoretical model. *Psychological Inquiry* 15, 2 (2004), 103–125.
- [61] Joshua A Tucker, Andrew Guess, Pablo Barberá, Cristian Vaccari, Alexandra Siegel, Sergey Sanovich, Denis Stukal, and Brendan Nyhan. 2018. Social media, political polarization, and political disinformation: A review of the scientific literature. *SSRN* 3144139 (2018).
- [62] Jacquelin Van Stekelenburg. 2017. Radicalization and violent emotions. *Political Science & Politics* 50, 4 (2017), 936–939.
- [63] Jacquelin Van Stekelenburg and Bert Klandermans. 2017. Individuals in movements: A social psychology of contention. In *Handbook of Social Movements Across Disciplines*. Springer, 103–139.
- [64] Nguyen Vo and Kyumin Lee. 2018. The rise of guardians: Fact-checking url recommendation to combat fake news. In *SIGIR*.
- [65] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359, 6380 (2018), 1146–1151. <https://doi.org/10.1126/science.aap9559>
- [66] Przemyslaw M Waszak, Wioleta Kasprzycka-Waszak, and Alicja Kubanek. 2018. The spread of medical fake news in social media—the pilot quantitative study. *Health Policy and Technology* 7, 2 (2018), 115–118.
- [67] Lilian Weng, Filippo Menczer, and Yong-Yeol Ahn. 2013. Virality prediction and community structure in social networks. *Scientific Reports* 3, 2522 (2013). <https://doi.org/10.1038/srep02522>
- [68] Thalia Wheatley and Jonathan Haidt. 2005. Hypnotic disgust makes moral judgments more severe. *Psychological Science* 16, 10 (2005), 780–784.
- [69] Liang Wu, Fred Morstatter, Kathleen M. Carley, and Huan Liu. 2019. Misinformation in social media: definition, manipulation, and detection. *ACM SIGKDD Explorations Newsletter* 21, 2 (2019), 80–90.
- [70] Wenlin Yao, Cheng Zhang, Shiva Saravanan, Ruihong Huang, and Ali Mostafavi. 2020. Weakly-supervised fine-grained event recognition on social media texts for disaster management. In *AAAI Conference on Artificial Intelligence*.
- [71] John Zarocostas. 2020. How to fight an infodemic. *The Lancet* 395, 10225 (2020), 676.

## Supplementary Materials

### A TOPIC MODELING

Table 4 shows the manually selected seed words that were used to identify topic-related tweets in weakly supervised learning.

Table 4 reports the results for our user study testing for the presence of errant tweets. On average, the share of tweets that were not classified as at least “somewhat related to [topic]” was lower than 8.5 %.

**Table 4: Seed words used to identify topic-related tweets in weakly supervised learning. Various word forms of the key-words are also considered, e. g., “masks” and “masking” are also considered for the keyword “mask”.**

Topic	Seed keywords
POLITICS	Bill, Trump, Biden, Obama, Democrats, GOP, Republicans, Tax, Administration, Red, Blue, Pelosi, Economy, Chinavirus
HEALTH	Vaccine, Flu, Mask, Fever, Ebola, SARS, Ibuprofen, Garlic, Health, Infection

**Table 5: Frequency of errors in topic labeling.**

Topic	Percent Error
POLITICS	6.0%
HEALTH	2.2%
OTHER	17.5%
Mean	8.5%

### B ANALYSIS OF CONTROL VARIABLES

We tested a model specification in which we only incorporated control variables from previous works. Table 6 shows that rumors receive a particularly high number of retweets if they are false and if they have been started by users with a larger number of followers.

**Table 6: Regression results for control variables only. The dependent variable is the number of retweets.**

Dependent Variable: <i>RetweetCount</i>	
Falsehood	0.145** (0.050,)
Followers	1.134*** (0.036,)
Followees	0.046 (0.025,)
AccountAge	-0.217*** (0.026,)
HasMedia	-0.132* (0.052,)
Verified	0.007 (0.073,)
Intercept	7.250*** (0.058,)
Observations (rumor cascades)	8727

Sign. levels: \* $p < 0.05$ , \*\* $p < 0.01$ , \*\*\* $p < 0.001$ ; standard errors in parentheses

### C VERIFIED VS. UNVERIFIED USERS

We repeated our analysis for subsets of rumor cascades that have been started by users that are verified or not-verified by Twitter. Table 7 shows that our main findings hold for both user groups.

**Table 7: Regression results for rumor cascades initiated from VERIFIED (column 1) or NON-VERIFIED (column 2) users only.**

Dependent Variable: <i>RetweetCount</i>		
	Subset: VERIFIED	Subset: NON-VERIFIED
Falsehood × OtherCondemning	0.238*** (0.058,)	0.302** (0.100,)
Falsehood × SelfConscious	-0.256*** (0.055,)	-0.262* (0.111,)
Falsehood	0.182*** (0.049,)	0.062 (0.104,)
OtherCondemning	-0.016 (0.042,)	-0.025 (0.085,)
SelfConscious	0.100** (0.037,)	0.073 (0.096,)
Followers	0.871*** (0.043,)	2.138*** (0.075,)
Followees	0.101*** (0.027,)	-0.512*** (0.055,)
AccountAge	-0.090* (0.036,)	-0.265*** (0.041,)
HasMedia	-0.344*** (0.053,)	0.142 (0.104,)
Intercept	7.412*** (0.048,)	7.969*** (0.106,)
Observations (rumor cascades)	4836	3891

Sign. levels: \* $p < 0.05$ , \*\* $p < 0.01$ , \*\*\* $p < 0.001$ ; standard errors in parentheses

### D RUMORS WITH MIXED VERACITY

**Table 8: Regression results with mixed rumors categorized as false rumors (Model 1) and mixed rumors categorized as true rumors (Model 2).**

Dependent Variable: <i>RetweetCount</i>		
	Model (1)	Model (2)
Falsehood × OtherCondemning	0.298*** (0.046,)	0.165*** (0.050,)
Falsehood × SelfConscious	-0.341*** (0.046,)	-0.100* (0.049,)
Falsehood	0.085 (0.044,)	0.156*** (0.046,)
OtherCondemning	-0.054 (0.033,)	-0.026 (0.042,)
SelfConscious	0.163*** (0.033,)	0.076 (0.040,)
Followers	1.175*** (0.033,)	1.173*** (0.033,)
Followees	0.039 (0.022,)	0.031 (0.022,)
AccountAge	-0.125*** (0.024,)	-0.132*** (0.024,)
HasMedia	-0.122* (0.048,)	-0.131** (0.048,)
Verified	-0.085 (0.066,)	-0.077 (0.066,)
Intercept	7.344*** (0.049,)	7.278*** (0.054,)
Observations (rumor cascades)	10,610	10,610

Sign. levels: \* $p < 0.05$ , \*\* $p < 0.01$ , \*\*\* $p < 0.001$ ; standard errors in parentheses

In our main analysis, we focused on rumors that are clearly true or false. However, 17.7 % of all rumors have been categorized as being of mixed veracity by the fact-checking organizations. We tested whether counting rumors of mixed veracity as either true or false affects the validity of our results. Table 8 shows that our results are robust and that the combination of other-condemning emotion and mixed veracity is similarly viral as the combination of other-condemning emotions and false veracity.

## E SENSITIVITY TO NON-INDEPENDENCE

In our data, 9.48 % of rumor starters have started more than one retweet cascade. To ensure that our models are not biased due to this source of non-independence, we dropped all users with clustering and reestimated the models. Table 9 show that the results are robust and support our findings.

We also repeated our analysis with monthly fixed effects to control for differences in the virality of rumor cascades due to different start dates (Table 9). All results confirm the findings from our main analysis.

**Table 9: Regression results without rumor cascades from users that have started more than one retweet cascade (Model 1) and with monthly fixed effects (Model 2).**

Dependent Variable: <i>RetweetCount</i>		
	Model (1)	Model (2)
Falsehood × OtherCondemning	0.428*** (0.100,)	0.225*** (0.052,)
Falsehood × SelfConscious	-0.386*** (0.104,)	-0.245*** (0.052,)
Falsehood	0.118 (0.101,)	0.226*** (0.051,)
OtherCondemning	-0.101 (0.085,)	0.005 (0.041,)
SelfConscious	0.128 (0.089,)	0.109** (0.040,)
Followers	1.937*** (0.081,)	1.289*** (0.036,)
Followees	-0.235*** (0.058,)	0.061* (0.024,)
AccountAge	-0.246*** (0.043,)	-0.273*** (0.026,)
HasMedia	0.254* (0.099,)	-0.162** (0.052,)
Verified	-0.326* (0.137,)	0.034 (0.072,)
Intercept	7.857*** (0.122,)	7.445*** (0.586,)
Monthly fixed effects	✗	✓
Observations (rumor cascades)	4139	8727

Sign. levels: \* $p < 0.05$ , \*\* $p < 0.01$ , \*\*\* $p < 0.001$ ; standard errors in parentheses

## F ALTERNATIVE EMOTION MEASURE

Since self-conscious emotions can be regarded as the counterpart of other-condemning emotions, we tested an alternative model specification in which we included the *difference* between the emotion scores for other-condemning and self-conscious emotions instead of two individual variables (Table 10). Consistent with our main analysis, we found that false rumors are more viral than the truth if the source tweet embeds a high proportion of other-condemning emotion words, whereas a high proportion of self-conscious emotion words is linked to a less viral spread.

**Table 10: Retweet count as a function of the difference of other-condemning and self-conscious emotions (OtherCondemning-SelfConscious).**

Dependent Variable: <i>RetweetCount</i>	
Falsehood × OtherCondemning-SelfConscious	0.289*** (0.048,)
Falsehood	0.127* (0.050,)
OtherCondemning-SelfConscious	-0.051 (0.038,)
Followers	1.161*** (0.036,)
Followees	0.057* (0.025,)
AccountAge	-0.196*** (0.026,)
HasMedia	-0.107* (0.052,)
Verified	-0.035 (0.073,)
Intercept	7.260*** (0.058,)
Observations (rumor cascades)	8727

Sign. levels: \* $p < 0.05$ , \*\* $p < 0.01$ , \*\*\* $p < 0.001$ ; standard errors in parentheses