Original Paper

Collective Response to Media Coverage of the COVID-19 Pandemic on Reddit and Wikipedia: Mixed-Methods Analysis

Nicolò Gozzi[1], MSc; Michele Tizzani[2], PhD; Michele Starnini[2], PhD; Fabio Ciulla[3], PhD; Daniela Paolotti[2], PhD; André Panisson[2], PhD; Nicola Perra[1], PhD

[1]University of Greenwich, London, United Kingdom

[2]ISI Foundation, Torino, Italy

[3]Quid Inc, San Francisco, CA, United States

Corresponding Author:
Nicolò Gozzi, MSc
University of Greenwich
Old Royal Naval College
Park Row
London, SE10 9LS
United Kingdom
Phone: 44 020 8331 8000
Email: n.gozzi@gre.ac.uk

Abstract

Background: The exposure and consumption of information during epidemic outbreaks may alter people's risk perception and trigger behavioral changes, which can ultimately affect the evolution of the disease. It is thus of utmost importance to map the dissemination of information by mainstream media outlets and the public response to this information. However, our understanding of this exposure-response dynamic during the COVID-19 pandemic is still limited.

Objective: The goal of this study is to characterize the media coverage and collective internet response to the COVID-19 pandemic in four countries: Italy, the United Kingdom, the United States, and Canada.

Methods: We collected a heterogeneous data set including 227,768 web-based news articles and 13,448 YouTube videos published by mainstream media outlets, 107,898 user posts and 3,829,309 comments on the social media platform Reddit, and 278,456,892 views of COVID-19–related Wikipedia pages. To analyze the relationship between media coverage, epidemic progression, and users' collective web-based response, we considered a linear regression model that predicts the public response for each country given the amount of news exposure. We also applied topic modelling to the data set using nonnegative matrix factorization.

Results: Our results show that public attention, quantified as user activity on Reddit and active searches on Wikipedia pages, is mainly driven by media coverage; meanwhile, this activity declines rapidly while news exposure and COVID-19 incidence remain high. Furthermore, using an unsupervised, dynamic topic modeling approach, we show that while the levels of attention dedicated to different topics by media outlets and internet users are in good accordance, interesting deviations emerge in their temporal patterns.

Conclusions: Overall, our findings offer an additional key to interpret public perception and response to the current global health emergency and raise questions about the effects of attention saturation on people's collective awareness and risk perception and thus on their tendencies toward behavioral change.

Introduction

Background

In the next influenza pandemic, be it now or in the future, be the virus mild or virulent, the single most important weapon against the disease will be a vaccine. The second most important will be communication.

This evocative sentence was written in May 2009 by John M Barry [1] in the early phases of what would soon become the 2009 H1N1 pandemic. In his essay, Barry summarized the mishandling of the deadly 1918 Spanish influenza, highlighting the importance of precise, effective, and honest information at the onset of health crises.

Eleven years later, the world is facing another pandemic. The cause is not a novel strain of influenza; however, unfortunately, Barry's words are still extremely relevant. In fact, as SARS-CoV-2 spreads worldwide and a vaccine may still be far in the future, the most important weapons to reduce the burden of the disease are nonpharmaceutical interventions [2,3]. Social distancing has become paramount, gatherings have been cancelled, and mobility within and across countries has been dramatically reduced. While these measures have been enforced to different extents across nations, they all rely on compliance. The effectiveness of these measures is linked to risk and susceptibility perception [4]; thus, the information to which citizens are exposed is of fundamental importance.

History repeats itself, and humanity appears to not be able to learn from its past mistakes. As happened in 1918, despite early evidence from China [5,6], the virus was first equated by many people with common seasonal influenza. As in 1918, many national and regional governments organized campaigns aimed at encouraging social activities (and thus local economies) while actively attempting to convince people that their cities were safe and that the spread of the disease was isolated in faraway locations. For example, the hashtag #*MilanoNonSiFerma* ("Milan does not stop") was coined to invite citizens in Milan to go out and live normally, while free aperitifs were offered in Venice. In hindsight, of course, it is easy to criticize the initial response in Italy. In fact, the country was one of the first to experience rapid growth of hospitalization [7]. However, the Mayor of London, 12 days before the national lockdown and a few days after the extension of the *cordon sanitaire* to the entire country in Italy, affirmed via his official Facebook page [8] that "we should carry on doing what we've been doing." More generally, in several western countries, news reports from other countries reporting concerning epidemic outbreaks were not considered to be relevant to the local situation. This initial phase aimed at conveying low local risk and boosting confidence in national safety was repeated, at different times, across countries. A series of surveys conducted in late February provide a glimpse of the possible effects of these approaches. These surveys report that citizens of several European countries, despite the grim news coming from Asia, were overly optimistic about the health emergency, placing their risk of infection at 1% or less [9]. As in 1918, countries that reacted earlier rather than later were better able to control the virus, with significantly fewer infections [10-14].

History repeats itself; however, the context is often radically different. In 1918, news circulated slowly via newspapers, controlled by editorial choices; of course, news also spread by word of mouth. In 2009, we witnessed the first pandemic in the social media era. Newspapers and television were still very important sources of information; however, Twitter, Facebook, YouTube, and Wikipedia started to become relevant for decentralized news consumption, boosting of peer discussions, and spreading of misinformation. Currently, these platforms and websites are far more popular and integral parts of society, and they are instrumental sources of national and international news circulation. Together with traditional news media, these platforms and websites are the principal sources of information for the public. As such, they are fundamental drivers of people's perception and opinions and thus of their behaviors. This is particularly relevant for health issues. For example, approximately 60% of adults in the United States consulted web-based sources to gather health information [15].

Furthermore, some platforms are acknowledging their growing responsibility in media consumption and have introduced specific features to increase users' awareness and levels of information.

Prior Work

With respect to past epidemics and pandemics, studies on traditional news coverage of the 2009 H1N1 pandemic highlighted the importance of framing and its effect on people's perception, behaviors (eg, vaccination intent), and stigmatization of cultures at the epicenter of the outbreak, as well as how these factors differ across countries and cultures [16-21]. During the Zika virus epidemic in 2016, public attention was synchronized across US states, driven by news coverage about the outbreak and independent of the real local risk of infection [22]. With respect to the COVID-19 pandemic itself, a recent study clearly showed how Google searches for *coronavirus* in the United States spiked significantly immediately after the announcement of the first confirmed case in each state [23]. Several studies based on Twitter data have also highlighted how misinformation and low-quality information about COVID-19, although limited overall, spread before the local outbreak and rapidly expanded once local epidemics started [24-26]. In the current landscape, this spread of misinformation has the potential to encourage irrational, unscientific, and dangerous behaviors. On the other hand, despite some important limitations [27], modern media has become a key data source to observe and monitor health. In fact, posts on Twitter [28-33], Facebook [34], and Reddit [35,36], page views in Wikipedia [37,38], and searches on Google [39,40] have been used to study, nowcast, and predict the spreading of infectious diseases as well as the prevalence of noncommunicable illnesses. Therefore, in the current full-fledged digital society, information is not only key to inform people's behavior but can also be used to develop an unprecedented understanding of these behaviors as well as of the phenomena driving them.

Goal of This Study

The context in which COVID-19 is unfolding is very heterogeneous and complex. Traditional and social media are integral parts of public perception and opinions, and they have potential to trigger behavior changes and thus influence the spread of the pandemic. This complex landscape must be characterized to understand the public attention and response to media coverage. Here, we addressed this challenge by assembling a heterogeneous data set that includes 227,768 news reports and 13,448 YouTube videos published by traditional media, 278,456,892 views of topical Wikipedia pages, and 107,898 submissions and 3,829,309 comments from 417,541 distinct users on Reddit, as well as epidemic data in four different countries: Italy, the United Kingdom, the United States, and Canada. First, we explored how media coverage and epidemic progression influence public attention and response. To achieve this, we analyzed news volume and COVID-19 incidence with respect to volumes of Wikipedia page views and Reddit comments. Our results show that public attention and response are mostly driven by media coverage rather than by disease spread. Furthermore, we observed the typical saturation and memory effects of public collective attention. Moreover, using an unsupervised topic modeling approach, we explored the different topics framed in traditional media and in Reddit discussions. We show that while the attention of news outlets and internet users toward different topics is in good accordance, interesting deviations emerge in their temporal patterns. Also, we highlight that at the end of our observation period, general interest grew toward topics about the resumption of activities after lockdown, the search for a vaccine against SARS-CoV-2, acquired immunity, and antibody tests. Overall, the research presented here offers insights to interpret public perception and response to the current global health emergency and raises questions about the effects of attention saturation on collective awareness and risk perception and thus on tendencies toward behavioral change.

Methods

Data Set

News Articles and Videos

We collected news articles using News API, a service that allows free downloads of articles published on the internet in a variety of countries and languages [41]. For each country considered, we downloaded all relevant articles published on the internet by selected sources in the period from February 7 to May 15, 2020. We selected relevant articles by considering those citing one of the following keywords: *coronavirus*, *covid19*, *covid-19*, *ncov-19*, and *sars-cov-2*. Note that for each article, we could access the title, a description, and a preview of the full text. In total, our data set consisted of 227,768 news articles; 71,461 were published by Italian media, 63,799 by UK media, 82,630 by US media, and 9878 by Canadian media.

Additionally, we collected all videos published on YouTube by major news organizations in the four countries under investigation via their official YouTube channels using the official application programming interface (API) [42]. In this process, we downloaded the titles and descriptions of all the videos and selected as relevant those that mentioned one of the following keywords: *coronavirus*, *virus*, *covid*, *covid19*, *sars*, *sars-cov-2*, and *sarscov2*. The reach of each channel (measured by the number of subscribers) varied drastically, from more than 9 million for CNN (United States) to approximately 12,000 for Ansa (Italy). In total, the YouTube data set consisted of 13,448 videos; 3325 were published by Italian channels, 3525 by UK channels, 6288 by US channels, and 310 by Canadian channels.

It is important to underline that while there is good overlap between the sources of news articles and videos, some do not match. This is due to the fact that not all news organizations have a YouTube channel, while others do not produce traditional articles. In Multimedia Appendix 1, we provide a complete list of the news outlets and YouTube channels we considered.

Reddit Posts

Reddit is a social content aggregation website on which users can post, comment, and vote on content. It is structured in subcommunities (ie, subreddits) that are centered around a variety of topics. Reddit has already been proven to be suitable for a variety of research purposes, ranging from the study of user engagement and interactions between highly related communities [43,44] to postelection political analyses [45]. Moreover, it has been used to study the impact of linguistic differences in news titles [46] and to explore recent web-related issues such as hate speech [47] and cyberbullying [48] as well as health-related issues such as mental illness [49]; it also provides insights into the opioid epidemic [50].

We used the Reddit API to collect all submissions and comments published in Reddit under the subreddit r/Coronavirus from February 15 to May 15, 2020. After cleaning the data by removing entries deleted by authors and moderators, we retained only submissions with scores >1 to avoid spam. We removed comments with <10 characters and with >3 duplicates to avoid including automatic messages from moderators. The final data set contained 107,898 submissions and 3,829,309 comments from 417,541 distinct users.

To characterize the topics discussed on Reddit, we then selected entries with links to English-language news outlets. The contents of the URLs were extracted using the available implementation of the method described in [51], resulting in 66,575 valid documents.

Reddit does not provide explicit information about users' locations; therefore, we used self-reporting via regular expression to assign locations to users. Reddit users often declare geographical information about themselves in submissions or comment texts. We used the same approach described in [50], in which the use of regular expressions was found to be reliable, resulting in high correlation with census data in the United States; however, we acknowledge a potential higher bias at the country level due to heterogeneities in Reddit population coverage and user demographics. We selected all English-language texts containing expressions such as "I am from" or "I live in" and extracted candidate expressions from the text that followed the expressions to identify texts that

represented country locations. By removing inconsistent self-reporting, we were able to assign a country to 789,909 distinct users, among which 41,465 had posted at least one comment in the subreddit r/Coronavirus (13,811 from the United States, 6870 from Canada, 3932 from the United Kingdom, and 445 from Italy).

Wikipedia Page Views

Wikipedia has become a popular digital data source to study health information–seeking behavior [52] and to monitor and forecast the spreading of infectious diseases [53,54]. Here, we used the Wikimedia API [55] to collect the number of visits per day to Wikipedia articles and the total monthly visits to a specific project from each country. We considered language to be indicative of a specific country, suggesting that the relevant projects for our analysis would be written in English and Italian (ie, *en.wikipedia* and *it.wikipedia*, respectively). We chose articles directly related to COVID-19 and those in the "See also" section of each page at the time of the analysis (February 7 to May 15, 2020), including country-specific articles (see Multimedia Appendix 1 for the full list of webpages considered).

Except for Italian, where the language is highly indicative of the location, the number of visits to English pages is almost evenly distributed among English-speaking countries. To normalize the signal related to each country, we weighted the number of daily visits to a single article from a specific project p, $S_p(d)$, with the total number of monthly visits from a country c, to the related Wikipedia project $T_p^c(d)$, such that the number of daily page views for a given Wikipedia project and country is:

$$y_{a,p}^c(d) = \frac{S_p(d)T_p^c(d)}{\sum_c T_p^c(d)}$$

(1)

where the denominator is the total number of views of the specific Wikipedia project. The total volume of views on day d from country c is then given by the sum over all the articles a and projects p, namely:

$$y^c(d) = \sum_{a,p} y_{a,p}^c(d)$$

(2)

Media Coverage and Collective Web-Based Response

With our data set, we aimed to provide an overview of media coverage and a proxy of public attention and response. On the one hand, the study of news articles and videos enabled us to estimate the exposure of the public to information about the COVID-19 pandemic in traditional news media. On the other hand, the study of users' discussions and responses on social media (through Reddit) and information-seeking (through Wikipedia page views) allowed us to quantify the reaction of individuals both to the COVID-19 pandemic and to news exposure. As mentioned in the Introduction, previous studies showed the usefulness of social media, internet use, and search trends to analyze health-related information streams and monitor public reaction to infectious diseases [56-60]. Hence, we considered the volume of comments of geolocalized users on

the subreddit r/Coronavirus to explore the public discussion in reaction to media coverage of the epidemic in the various countries; meanwhile, we considered the number of views of relevant Wikipedia pages about the COVID-19 pandemic to quantify users' interest. It is important to stress that Reddit and Wikipedia provide different aspects of internet users' behavior and collective response. In fact, while Reddit posts can be regarded as a general indicator of the web-based discussion surrounding the global health emergency, the number of visits to COVID-19–related Wikipedia pages is a proxy of health information–seeking behavior. Health information–seeking behavior is the act by which individuals retrieve and acquire new knowledge about a specific topic related to health [61,62]; it is likely to be triggered on a population scale by a disrupting event, such as the threat of a previously unknown disease [63,64].

Linear Regression Approach to Model Collective Attention

To analyze the relationship between media coverage, epidemic progression, and users' collective web-based response, we considered a linear regression model that predicts the public response for each country given the amount of news exposure. To include "memory effects" in the public response to media coverage, we also considered a modified version of this simple model, in which we weight a cumulative news articles volume time series with an exponential decay term [22]. Formally, we define the new variable as:

$$newsMEM = \sum_{\Delta t=1}^{\tau} e^{-\frac{\Delta t}{\tau}} news(t - \Delta t)$$

(3)

where τ is a free parameter that sets the memory time scale and is tuned by comparing different variants of the linear regression with $\tau \in [1,45]$ in terms of the adjusted coefficient of determination R^2 [65] (results for the best τ are displayed). These two models were compared to a linear regression that considers only COVID-19 incidence to predict public collective attention. Then, the models considered are:

Model I: $y_t = \alpha_1 incidence_t + u_t$

Model II: $y_t = \alpha_1 news_t + u_t$

Model III: $y_t = \alpha_1 news_t + \alpha_2 newsMEM_t + u_t$ (4)

where y_t can be the volume of Reddit comments of geolocalized users or of country-specific Wikipedia visits, and u_t is the error term. In Multimedia Appendix 1, we provide more details on the model diagnostics and fitting procedure.

Topic Modeling

Topic modeling has emerged as one of the most effective methods for classifying, clustering, and retrieving textual data, and it has been the object of extensive investigation in the literature. Many topic analysis frameworks are extensions of well-known algorithms that are considered to be state-of-the-art for topic modeling. Latent Dirichlet allocation (LDA) [66] is the reference for probabilistic topic modeling. Nonnegative matrix factorization (NMF) [67] is the counterpart of LDA for

matrix factorization. Although there are many approaches to temporal and hierarchical topic modeling [68-70], we chose to apply NMF to the data set and then build time-varying intensities for each topic using the publication dates of the articles. Starting from a data set D containing the news articles shared in Reddit, we extracted words and phrases with the methodology described in [71], discarding terms with frequencies >10, to form a vocabulary V with approximately 60,000 terms. Each document was then represented as a vector of term counts in a bag-of-words approach. We applied term frequency–inverse document frequency (TF-IDF) normalization [72] and extracted a total of $K=64$ topics through NMF:

$$\min_{W,\,H} \|X - WH\|_F^2$$

(5)

where $\|\,\|_F^2$ is the Frobenius norm and $X \in R^{|D| \times |V|}$ is the matrix resulting from TF-IDF normalization, subject to the constraint that the values in $W \in R^{|D| \times K}$ and $H \in R^{K \times |V|}$ must be nonnegative. The nonnegative factorization was achieved using the projected gradient method with sparseness constraints, as described in [73,74]. The matrix H was then used as a transformation basis for other data sets (eg, with a new matrix $\tilde{X}$, we fixed H and calculated a new $\tilde{W}$ according to Equation 5). For each topic k, we built a time series s_k for each data set D, where $s_k^{(t)}$ is the strength of topic k at time t. For the news outlets data set, $s_k^{(t)} = \Sigma_{i \in D^{(t)}} w_{ik}$, where $D^{(t)}$ is the set of all documents shared at time t in news outlets. For Reddit, we weighted each shared document by its number of comments, and $s_k^{(t)} = \Sigma_{i \in D^{(t)}} w_{ik} \cdot c_i$, where $D^{(t)}$ is the set of all documents shared at time t in Reddit and c_i is the number of comments associated with document i. Finally, we defined the relevance R of a topic as the integral in time of the strength. Therefore, given t_0 and t_f as the start and end of our analysis interval, $R = \int_{t_0}^{t_f} dt\, s_k^{(t)}$. In Multimedia Appendix 1, we show that choosing $K=64$ as the number of extracted topics provides a good balance between sufficient captured topic strength and good topic coherence.

Results

### Impact of Media Coverage and Epidemic Progression on Collective Attention

To answer the important question of how collective attention is shaped by news media coverage and epidemic progression,

we started by comparing the weekly volumes of news stories and videos published on YouTube, Wikipedia page views, and Reddit comments of geolocalized users with the weekly COVID-19 incidence in the four countries considered (Figure 1). It can be seen that as COVID-19 spread, both media coverage and public interest grew with time. However, public attention, quantified by the number of Reddit comments and Wikipedia page views, sharply decreased after reaching a peak, even though the volume of news stories and the incidence of COVID-19 remained high. Furthermore, the peak in public attention consistently anticipated the maximum media exposure and maximum COVID-19 incidence.

The correlation between media coverage, public attention, and progression of the epidemic is quantified in more detail in Table 1. The table shows that news coverage of each country is strongly correlated with COVID-19 incidence (both global and domestic) and slightly less correlated with the volumes of Reddit comments and Wikipedia views, which in turn are much less correlated with COVID-19 incidence (both global and domestic). This result was observed for all countries under consideration; it highlights how the spread of COVID-19 triggered media coverage as well as how public response was more likely to be driven by news exposure in each country than by the progression of COVID-19.

Beyond these observations, Figure 2 shows the share of citations of Chinese versus home country locations by Italian, UK, US, and Canadian news outlets before and after the first COVID-19 death occurred in those countries; the geographic locations were extracted from the text using the methods described in [75,76]. Interestingly, Italy is the only country where the news volume shows a higher correlation with domestic incidence than with global incidence (ie, news references to China). This suggests that Italian media coverage follows internal evolution more closely than global evolution, in contrast to other countries. This is probably due to the fact that Italy is the location of the first COVID-19 outbreak outside Asia. This observation is supported by Figure 2, which shows the citation share of Italian locations by Italian news media before and after the first COVID-19 death was confirmed in Italy on February 23, 2020. After this date, Italian locations represent about 74% of all places cited by Italian media (in our data set), with an increase of 45% with respect to the same statistics calculated before. Similar effects, although generally less intense, were observed in the other countries. Therefore, while media coverage is generally well synchronized with global COVID-19 incidence, the media attention gradually shifts toward the internal evolution of the pandemic as soon as domestic outbreaks erupt.

Figure 1. Normalized weekly volumes of news articles and YouTube videos, Reddit comments, and Wikipedia page views related to the COVID-19 pandemic and the incidence of COVID-19 in different countries.
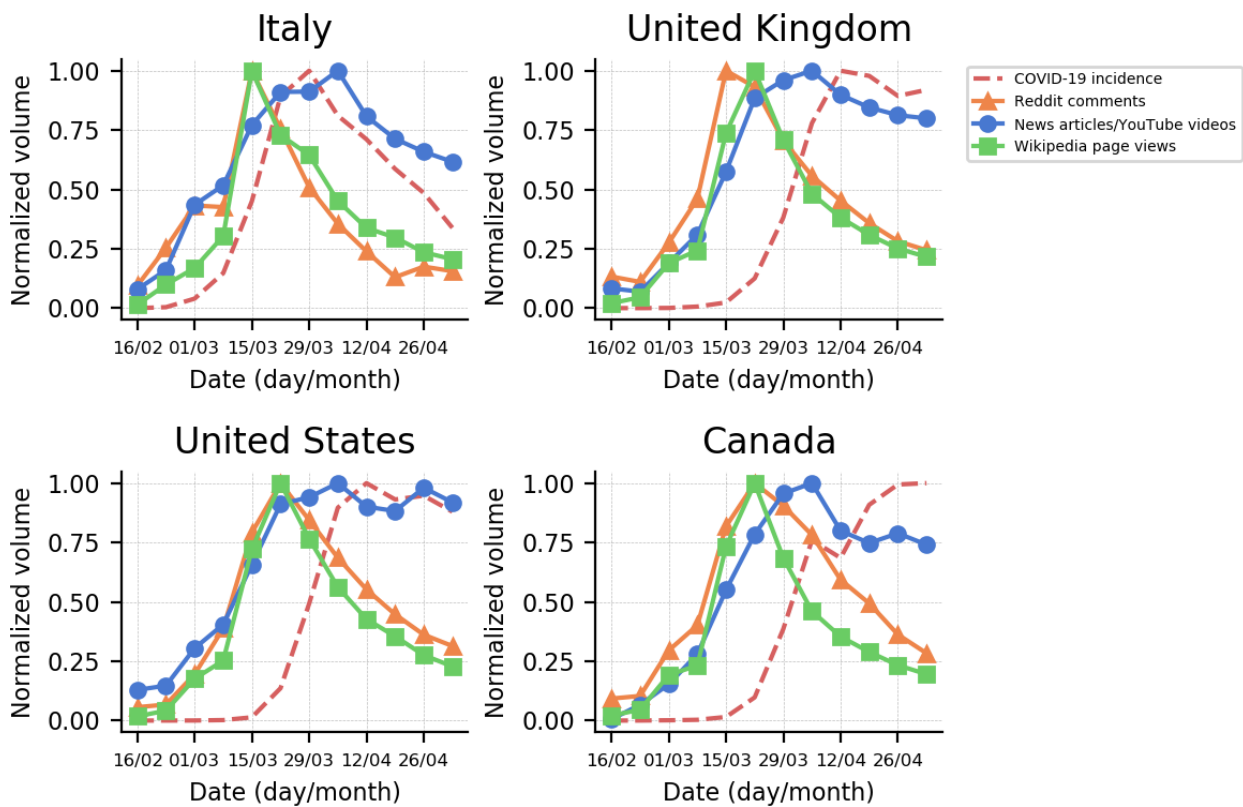
Table 1. Country-specific Pearson correlation coefficients for news coverage and global and domestic COVID-19 incidence, volumes of Reddit comments, and volumes of Wikipedia page views; domestic COVID-19 incidence and volumes of Reddit comments and Wikipedia views; and global COVID-19 incidence and volumes of Reddit comments and Wikipedia views.

Country	Global incidence of COVID-19	P value	Country incidence of COVID-19	P value	Reddit comments	P value	Wikipedia page views	P value
Italy								
News	0.59	.04	0.92	<.001	0.43	.17	0.71	.009
Global incidence of COVID-19	1	N/A[a]	—[b]	N/A	–0.42	.18	–0.01	.97
Country incidence of COVID-19	—	N/A	1	N/A	0.30	.34	0.64	.02
United Kingdom								
News	0.83	<.001	0.74	.006	0.50	.10	0.62	.03
Global incidence	1	N/A	—		–0.04	.90	0.09	.77
Country incidence	—	N/A	1	N/A	–0.15	.64	–0.04	.91
United States								
News	0.84	<.001	0.79	.002	0.70	.01	0.64	.03
Global incidence	1	N/A	—	N/A	0.25	.44	0.17	.60
Country incidence	—	N/A	1	N/A	0.16	.62	0.08	.81
Canada								
News	0.82	.001	0.71	.01	0.73	.007	0.59	.04
Global incidence	1	N/A	—	N/A	0.23	.46	0.06	.85
Country incidence	—	N/A	1	N/A	0.05	.87	–0.10	.76

[a]N/A: not applicable.

[b]—: not determined.

Figure 2. Shares of citations of China versus home country locations by Italian, UK, US, and Canadian news outlets before and after the first COVID-19 death occurred in each country.
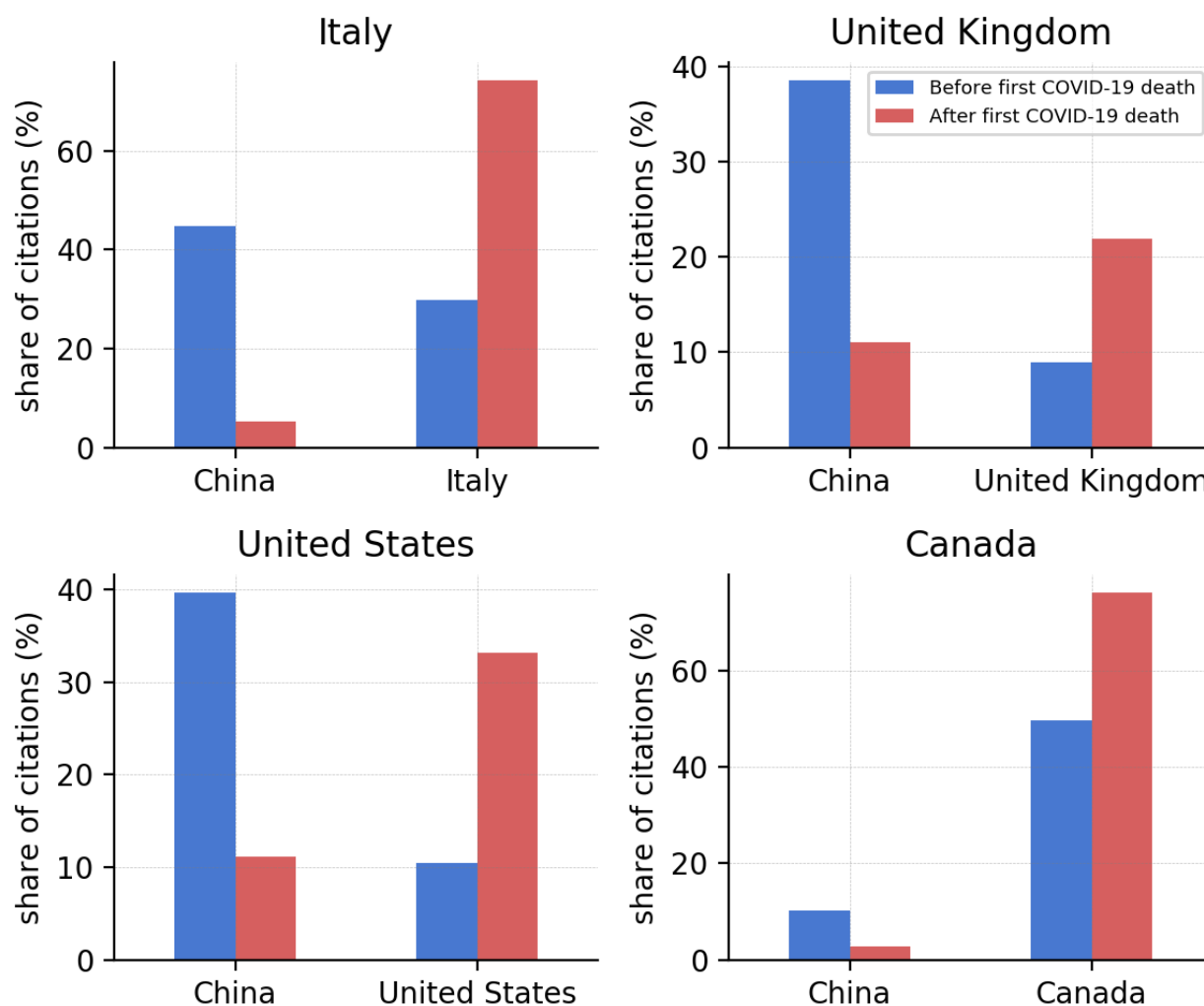
To more systematically explore the relationship between media coverage, public attention, and epidemic progression, we considered a linear regression model to nowcast the collective public attention for each country (quantified by the number of comments by geolocalized Reddit users or visits to relevant Wikipedia pages) using the volume of media coverage or the COVID-19 incidence as independent variables. We also included "memory effects" on the public attention by considering an exponential decaying term in the news time series [22]. We compared the three models, where the independent variables are the domestic incidence, the news volume, and the news volume plus a memory term, using the adjusted coefficient of determination (R^2) [65]. We found that the model that considered only COVID-19 incidence performed worse than the models that considered media coverage (Table 2). This enforces the idea that collective attention is mainly driven by media coverage rather than by COVID-19 incidence. In addition, we found that including memory effects improved the model performance.

Table 2. Adjusted R^2 values for the three linear regression models applied to predict Reddit comments and Wikipedia page views ($P<.001$).

Country	Model I		Model II		Model III	
	Reddit comments	Wikipedia page views	Reddit comments	Wikipedia page views	Reddit comments	Wikipedia page views
Italy	0.52	0.65	0.68	0.73	*0.82* [a]	*0.79*
United Kingdom	0.27	0.27	0.72	0.74	*0.82*	*0.85*
United States	0.42	0.35	0.82	0.74	*0.89*	*0.82*
Canada	0.35	0.23	0.83	0.71	*0.90*	*0.82*

[a]Italics indicate the superior performance of Model III.

More formally, we compared Model I to Model III using the Cox test [77] for nonnested models, and we compared Model II to Model III using the *F* test [78] for nested models. In all cases we obtained *P* values <.001, providing strong statistical evidence that Model III actually outperforms the other models. Not surprisingly, the coefficients of the "memory effects" term reported in Table 3 are negative for all countries. This implies

that public attention actually saturates in response to news exposure and enables us to quantify the rate at which this phenomenon occurs.

In the next section, we will characterize the media coverage and internet users' response more specifically in terms of content produced and consumed.

Table 3. Coefficient estimates (95% CI) for Model III (news plus memory effects). All coefficients are significant with *P*<.001.

Country	News		News plus memory effects	
	Reddit comments	Wikipedia page views	Reddit comments	Wikipedia page views
Italy	0.87 (0.60 to 1.14)	0.43 (0.29 to 0.58)	–0.41 (–0.59 to –0.23)	–0.15 (–0.26 to –0.04)
United Kingdom	0.95 (0.62 to 1.27)	0.99 (0.68 to 1.30)	–0.44 (–0.71 to –0.18)	–0.47 (–0.70 to –0.23)
United States	1.03 (0.79 to 1.27)	0.83 (0.58 to 1.09)	–0.51 (–0.77 to –0.24)	–0.46 (–0.73 to –0.19)
Canada	1.12 (0.89 to 1.36)	1.06 (0.67 to 1.44)	–0.40 (–0.59 to –0.22)	–0.45 (–0.72 to –0.18)

Dynamics of Content Production and Consumption

While collective attention and media coverage are well correlated in terms of volume, the content and topics discussed by media and consumed by internet users may not be as synchronized [79,80]. To shed light on this issue, we adopted an unsupervised topic modeling approach to extract prevalent topics in the news articles mentioned and discussed on Reddit. Indeed, users on Reddit often post submissions containing news articles, and discussion unfolds in the comments under the submissions. Differently from the previous section and to provide a comprehensive overview of the topics discussed, in this section, we do not take any geographical context into account. However, in Multimedia Appendix 1, we provide some additional insights into the specific topics discussed by users in different countries.

We characterized the main topics discussed on Reddit by considering all submissions that included a news article in English. We then applied a topic modelling approach to the content of this news article set. Specifically, we extracted topics using NMF [67], a popular method for this type of task. In this way, we extracted the 64 most relevant topics in the news articles shared on Reddit. As a second step, we applied the model trained on the Reddit news to the set of articles published

by mainstream media. That is, we characterized the news published by media outlets in terms of the topics discussed on Reddit. This choice enabled us to directly compare the topics covered by the media to the public discussion around this news exposure. A complete list of the 64 topics extracted with the most frequent words is provided in Multimedia Appendix 1. We considered the number of articles published on a certain topic as a proxy of general interest of traditional media in that topic; meanwhile, we measured the collective interest of Reddit users by the number of comments under the news articles on a specific topic. Figure 3 shows an overview of the topics extracted and a comparison of the interest of media and Reddit users. We obtained a diverse and heterogeneous set of topics, including the global spread of the virus (Outbreaks, WHO [World Health Organization], CDC [US Centers for Disease Control and Prevention]); COVID-19 symptoms, treatment, hospitals and care facilities (Symptoms, Medical Treatment, Medical Staff, Care Facilities); the economic impact of the pandemic and responses from the governments to the upcoming crisis (Economy, Money); different societal aspects (Sports, Religious Services, Education); and possible interventions to mitigate the spread of the virus (Face Masks, Social Distancing, Tests, Vaccine).

Figure 3. Differences in interest percentage shares of different topics by traditional media outlets and Reddit users. For example, +2% on the x-axis indicates that traditional media dedicates proportionally 2% more attention to that specific topic than Reddit users. CDC: US Centers for Disease Control and Prevention; UK: United Kingdom; WHO: World Health Organization.
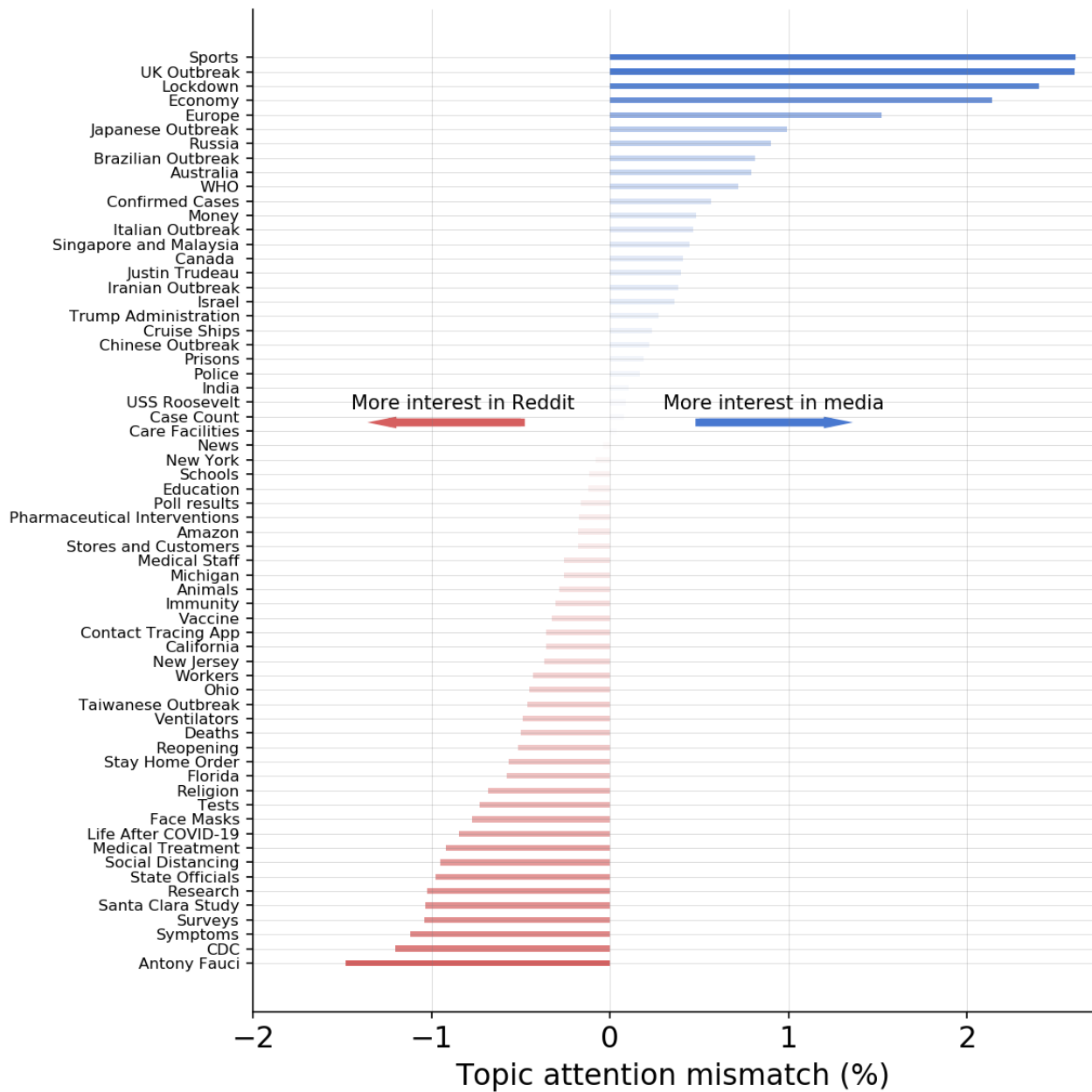
Overall, the levels of attention of traditional media outlets and Reddit users toward the different topics are in good accordance. Indeed, in Figure 3, we represent the difference between interest shares toward different topics in media and Reddit submissions. That is, we computed the percentage share of attention dedicated by news outlets and Reddit users to each topic, and we subtracted these two quantities. We observed a maximum absolute mismatch in interest share of 2.61%. However, we observed that Reddit users are slightly more interested in topics regarding health (Symptoms, Medical Treatment), nonpharmaceutical interventions and personal protective equipment (Social Distancing, Face Masks), studies and information on the epidemic (Research, Surveys, Santa Clara Study, CDC), and specific public figures such as Anthony Fauci. Interestingly, the Santa Clara Study topic refers to the discussion

about a controversial scientific paper suggesting that a much higher fraction of the population in Santa Clara County was infected with respect to what was originally thought [81]. Because the study suggested a lower mortality rate, the preprint was quickly leveraged to support protest against lockdowns [82]; meanwhile, substantial flaws have been detected in the scientific methodology of the paper [83].

The overview of topics presented here does not take temporal dynamics of interest into account. However, topics showing similar overall statistics may present a mismatch in temporal patterns. Hence, in the following, we take into account the temporal evolution of interest toward different topics. In Figure 4, we represent each topic as a single point: the x-coordinate and y-coordinate indicate the $t_{1/2}$ when the topic reached 50% of its total relevance R in news outlets and on Reddit,

respectively, during the analysis interval. Therefore, topics in the bottom left region became relevant very early in the public discussion. Among these, we recognize themes centered on early COVID-19 outbreaks (ie, Chinese, Japanese, Iranian, and Italian outbreaks), events related to cruise ships, specific countries (ie, Israel, Singapore, and Malaysia), and topics regarding (early) health issues (ie, Symptoms, Confirmed Cases, and CDC). In contrast, topics in the top right region became relevant toward the end of the analysis interval (early May).

Reasonably, this region contains topics about the resumption of activities after lockdown (ie, Reopening), the feasibility and timing of a possible vaccine against SARS-CoV-2 (ie, Vaccine), and discussions regarding acquired immunity and antibody tests (ie, Immunity). All other topics are clustered around the end of March and mid-April 2020, which is the period in which the general discussion surrounding the COVID-19 pandemic increased sharply, as also shown in Figure 1.

Figure 4. Scatter plot with the 64 topics extracted via nonnegative matrix factorization. The x-axis and y-axis coordinates indicate when a topic achieved 50% of its relevance in news outlets and on Reddit, respectively, during our analysis interval. CDC: US Centers for Disease Control and Prevention.
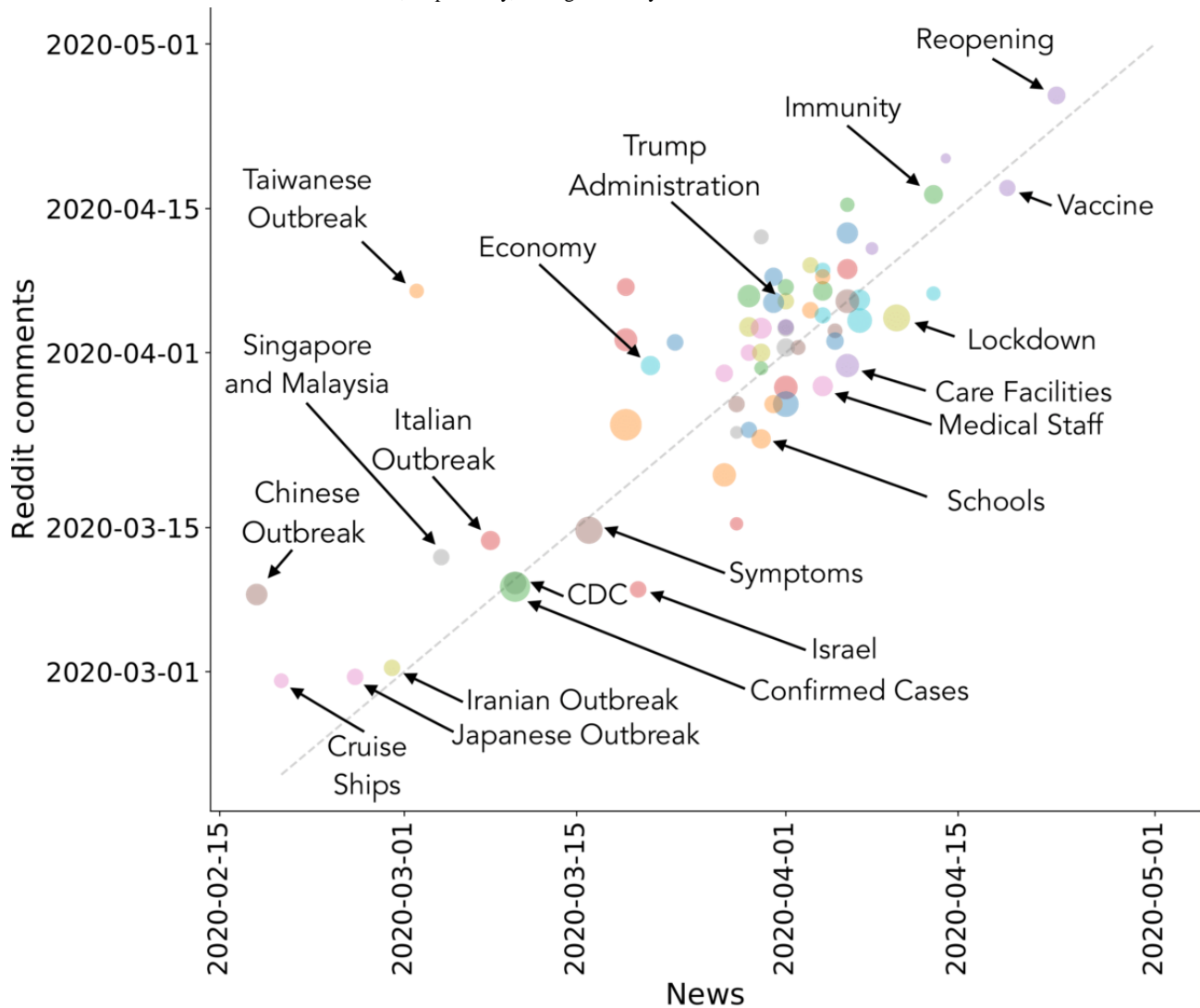
Note that the diagonal in Figure 4 (plotted as a dashed line) separates topics according to their temporal evolution. Above and below the diagonal, we find topics in which interest on Reddit grows slowly and quickly, respectively, with respect to the media coverage. Therefore, above the diagonal, the interest of Reddit users is mainly triggered by media exposure, while below it, the interest grows faster and declines rapidly despite sustained media exposure. The top left and bottom right regions are empty, indicating that as a first approximation, temporal patterns of attention by traditional media and Reddit users are well synchronized; however, interesting deviations from the diagonal are observable. For example, above the diagonal, one can mainly find topics related to various outbreaks, economics,

and politics, for which the interest on Reddit follows the media coverage. Below the diagonal, we observe topics more related to everyday life, such as Schools, Medical Staff, Care Facilities, and Lockdown, for which the attention on Reddit accelerates with respect to media coverage and then declines rapidly. Note that our view of the topics discussed on Reddit is limited, as we only considered topics from news articles shared in submissions and did not explicitly take content expressed in comments into account. This ensures a proper comparison with the topics extracted from published news reports and explains the absence of points in the bottom right corner of Figure 4.

Discussion

Principal Results

In this work, we characterized the response of internet users to both media coverage and COVID-19 pandemic progression. As a first step, we focused on the impact of media coverage on collective attention in different countries, characterized as volumes of country-specific Wikipedia page views and comments of geolocalized Reddit users. We showed that collective attention was mainly driven by media coverage rather than epidemic progression, rapidly became saturated, and decreased despite media coverage and COVID-19 incidence remaining high. These results are in very good accordance with findings obtained in previous contexts related to epidemics and pandemics. Indeed, a similar media-driven spiky unfolding of public attention, measured through the information-seeking and public discussions of internet users, was observed during the 2009 H1N1 influenza pandemic [84,85], the 2016 Zika virus outbreak [86], influenza season [87], and more localized public health emergencies such as the 2013 measles outbreak in the Netherlands [88]. Our findings confirm the central role of the media, showing how media exposure is capable of shaping and driving collective attention during a national and global health emergency. Media exposure is another important factor that can influence individual risk perception as well [79,89-91]. The timing and framing of the information disseminated by media can actually modulate the attention and, ultimately, the behavior of individuals [2]. This becomes an even greater concern in a context where the most effective strategy to fight the spread of disease involves containment measures based on individuals' behavior.

Also, we showed how media coverage sharply shifted to the domestic situation as soon as the first death was confirmed in the home country. Arguably, this may have played an important role in individual risk perception. We can speculate that reframing the emergency within a national dimension can amplify the perceived susceptibility of individuals [92,93] and thus increase the adoption of behavioral changes [4,94]. Indeed, previous studies showed that at the beginning of February 2020, people were overly optimistic regarding the risks associated with the new virus circulating in Asia, and their perception sharply changed after the first cases were confirmed in their countries [9,95].

As a second step, we focused on the dynamics of content production and consumption. We modeled topics published in mainstream media and discussed on Reddit, showing that Reddit users were generally more interested in health, data regarding the new disease, and interventions needed to halt the spreading with respect to media exposure. By taking into account the dynamics of the extracted topics, we showed that while their temporal patterns are generally synchronized, the public attention to topics related to politics and economics is mainly triggered by media exposure, while the interest in topics more related to daily life increases on Reddit with respect to media coverage.

Limitations

Of course, our research comes with limitations. First, we characterized the exposure of individuals to the COVID-19 pandemic by considering only news articles and YouTube videos published on the internet by major news outlets. However, individuals are also exposed to relevant information through other channels, with television being the most important [96]. Second, a 2013 Pew Internet Study found that Reddit users are more likely to be young men [97]; it was shown that around 15% of male internet users aged 18 to 29 years report using Reddit, compared to 5% of women in the same age range and 8% of men aged 30 to 49 years. Similarly, informal surveys proposed to users [98] showed that most respondents were males in their "late teens to mid-20s" and that female users were "very much in the minority." Furthermore, Reddit is much more popular among urban and suburban residents than among individuals living in rural areas [97]. In addition to sociodemographic biases, other studies have suggested that Reddit has become an increasingly self-referential community, reinforcing the tendency to focus on its own contents rather than external sources [99]. Thus, the perceptions, interests, and behaviors of Reddit users may differ from those of the general population. A similar argument can be raised for Wikipedia searches. Indeed, the use of the internet, especially for information-seeking purposes, can vary across people with different sociodemographic backgrounds [100-102]. Additionally, we extracted Reddit users' geographic location using a method based on regular expressions that has been successfully used in previous work [50]. However, because we have no ground truth data for comparison, we must consider the quality of location detection to be a possible limitation. Finally, our view on internet users' reactions is partial. Indeed, we did not consider other popular digital data sources, such as Twitter. The reasons for this choice are twofold. First, many studies have already characterized public response during current and past health emergencies through the lens of Twitter [25,58,60,85,86,103,104]. Second, several studies have reported a high prevalence of bots as drivers of low-quality information and discussions on COVID-19 on this platform [24,25,105-107]. Thus, careful and challenging additional steps would be necessary to isolate, identify, and distinguish organic Twitter discussions and reactions that originated from traditional media from those sparked by social bots. We leave this for future work.

Conclusions

Our work offers further insights to interpret public response to the current global health emergency and raises questions about possible undesired effects of communication. On one hand, our results confirm the pivotal role of media during health emergencies, showing how collective attention is mainly driven by media coverage. Therefore, because people are highly reactive to the news they are exposed to at the beginning of an outbreak, the quality and type of information provided may have critical effects on risk perception and behaviors, which will ultimately affect the unfolding of the outbreak. However, we also found that collective internet attention saturates and declines rapidly, even when media exposure and disease circulation remain high. Attention saturation has the potential to affect collective awareness and perceived risk, which

ultimately affects the propensity toward virtuous individual behavioral changes aimed at mitigating the spread of disease. Furthermore, especially in the case of unknown viruses, attention saturation may exacerbate the spreading of low-quality information, which is likely to spread in the early phases of the outbreak when the characteristics of the disease are uncertain. Future work is needed to characterize the actual effects of attention saturation on human perceptions during a global health emergency. Our findings suggest that public health authorities should consider reinforcing specific communication channels, such as social media platforms, to compensate for the natural phenomenon of attention saturation. Indeed, these channels have the potential to create more durable engagement with people through a continuous loop of direct interactions. Currently, public health authorities are regularly issuing declarations on social media. However, the CDC did not even have a Twitter account in 2009 during the H1N1 pandemic (the account was created in May 2010). While this is just one example, it underlines that the communication of these global health emergencies through social media platforms is relatively new. Therefore, there is great need to further reinforce these channels and engage people through them. Simultaneously, public health authorities should consider strengthening additional communication channels. One example of this is the participatory surveillance platforms that are appearing worldwide, such as Influenzanet, Flu Near You, and FluTracking [108-110], which can deliver in-depth targeted information to individuals during public health emergencies and promote the exchange of information between people and public health authorities; this has potential to enhance the level of engagement in the community [111].

Acknowledgments

Authors' Contributions

NG, MS, DP, AP, and NP conceptualized the study. NG, NP, AP, and MT collected the data. NG, AP, and FC performed analyses. NG, MS, and NP wrote the initial draft of the manuscript. NG and AP provided visualization. All authors (NG, NP, DP, MS, AP, MT, and FC) discussed the research design, reviewed, edited, and approved the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Supplementary information.
[DOCX File , 12651 KB-Multimedia Appendix 1]

References

1. Barry JM. Pandemics: avoiding the mistakes of 1918. Nature 2009 May 21;459(7245):324-325 [FREE Full text] [doi: 10.1038/459324a] [Medline: 19458695]
2. Funk S, Salathé M, Jansen VAA. Modelling the influence of human behaviour on the spread of infectious diseases: a review. J R Soc Interface 2010 Sep 06;7(50):1247-1256 [FREE Full text] [doi: 10.1098/rsif.2010.0142] [Medline: 20504800]
3. Verelst F, Willem L, Beutels P. Behavioural change models for infectious disease transmission: a systematic review (2010-2015). J R Soc Interface 2016 Dec;13(125):20160820 [FREE Full text] [doi: 10.1098/rsif.2016.0820] [Medline: 28003528]
4. Rosenstock IM, Strecher VJ, Becker MH. Social learning theory and the Health Belief Model. Health Educ Q 1988 Sep 04;15(2):175-183. [doi: 10.1177/109019818801500203] [Medline: 3378902]
5. Wu JT, Leung K, Bushman M, Kishore N, Niehus R, de Salazar PM, et al. Estimating clinical severity of COVID-19 from the transmission dynamics in Wuhan, China. Nat Med 2020 Apr 19;26(4):506-510 [FREE Full text] [doi: 10.1038/s41591-020-0822-7] [Medline: 32284616]
6. WHO Timeline - COVID-19 (Archived). World Health Organization. URL: https://www.who.int/news-room/detail/27-04-2020-who-timeline---covid-19 [accessed 2020-05-11]

XSL•FO

RenderX

7. Coronavirus disease (COVID-19) Weekly Epidemiological Update and Weekly Operational Update. World Health Organization. URL: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports/ [accessed 2020-05-11]

8. #Coronavirus: Today I met with England's Chief Medical Officer Professor Chris Whitty @CMO_England for the latest update. Sadiq Khan Facebook page. 2020 Mar 11. URL: https://www.facebook.com/sadiqforlondon/posts/3025766374142796 [accessed 2020-05-11]

9. Raude J, Debin M, Souty C, Guerrisi C, Turbelin C, Falchi A, et al. Are people excessively pessimistic about the risk of coronavirus infection? PsyArXiv. Preprint posted online March 8, 2020 . [doi: 10.31234/osf.io/364qj]

10. Kraemer MUG, Yang C, Gutierrez B, Wu C, Klein B, Pigott DM, Open COVID-19 Data Working Group, et al. The effect of human mobility and control measures on the COVID-19 epidemic in China. Science 2020 May 01;368(6490):493-497 [FREE Full text] [doi: 10.1126/science.abb4218] [Medline: 32213647]

11. Maier BF, Brockmann D. Effective containment explains subexponential growth in recent confirmed COVID-19 cases in China. Science 2020 May 15;368(6492):742-746 [FREE Full text] [doi: 10.1126/science.abb4557] [Medline: 32269067]

12. Anderson RM, Heesterbeek H, Klinkenberg D, Hollingsworth TD. How will country-based mitigation measures influence the course of the COVID-19 epidemic? Lancet 2020 Mar;395(10228):931-934. [doi: 10.1016/s0140-6736(20)30567-5]

13. Bedford J, Enria D, Giesecke J, Heymann DL, Ihekweazu C, Kobinger G, et al. COVID-19: towards controlling of a pandemic. Lancet 2020 Mar;395(10229):1015-1018. [doi: 10.1016/s0140-6736(20)30673-5]

14. Colbourn T. COVID-19: extending or relaxing distancing control measures. Lancet Public Health 2020 May;5(5):e236-e237. [doi: 10.1016/s2468-2667(20)30072-4]

15. Fox S, Duggan M. Health Online 2013. Pew Research Center. 2013 Jan 15. URL: https://www.pewresearch.org/internet/2013/01/15/health-online-2013/ [accessed 2020-09-21]

16. Lee ST. Predictors of H1N1 Influenza Pandemic News Coverage: Explicating the Relationships between Framing and News Release Selection. Int J Strateg Commun 2014 Sep 09;8(4):294-310. [doi: 10.1080/1553118x.2014.913596]

17. McCauley M, Minsky S, Viswanath K. The H1N1 pandemic: media frames, stigmatization and coping. BMC Public Health 2013 Dec 03;13:1116 [FREE Full text] [doi: 10.1186/1471-2458-13-1116] [Medline: 24299568]

18. Lin CA, Lagoe C. Effects of News Media and Interpersonal Interactions on H1N1 Risk Perception and Vaccination Intent. Commun Res Rep 2013 Apr;30(2):127-136. [doi: 10.1080/08824096.2012.762907]

19. Lee ST, Basnyat I. From press release to news: mapping the framing of the 2009 H1N1 A influenza pandemic. Health Commun 2013;28(2):119-132. [doi: 10.1080/10410236.2012.658550] [Medline: 22439616]

20. Jung Oh H, Hove T, Paek H, Lee B, Lee H, Kyu Song S. Attention cycles and the H1N1 pandemic: a cross-national study of US and Korean newspaper coverage. Asian J Commun 2012 Apr;22(2):214-232. [doi: 10.1080/01292986.2011.642395]

21. Keramarou M, Cottrell S, Evans MR, Moore C, Stiff RE, Elliott C, et al. Two waves of pandemic influenza A(H1N1) 2009 in Wales--the possible impact of media coverage on consultation rates, April-December 2009. Euro Surveill 2011 Jan 20;16(3):19772 [FREE Full text] [Medline: 21262184]

22. Tizzoni M, Panisson A, Paolotti D, Cattuto C. The impact of news exposure on collective attention in the United States during the 2016 Zika epidemic. PLoS Comput Biol 2020 Mar;16(3):e1007633. [doi: 10.1371/journal.pcbi.1007633] [Medline: 32163409]

23. Bento AI, Nguyen T, Wing C, Lozano-Rojas F, Ahn Y, Simon K. Evidence from internet search data shows information-seeking responses to news of local COVID-19 cases. Proc Natl Acad Sci USA 2020 May 26;117(21):11220-11222 [FREE Full text] [doi: 10.1073/pnas.2005335117] [Medline: 32366658]

24. Gallotti R, Valle F, Castaldo N, Sacco P, De Domenico M. Assessing the risks of. medRxiv. Preprint posted online on April 16, 2020 . [doi: 10.1101/2020.04.08.20057968]

25. Singh L, Bansal S, Bode L, Budak C, Chi G, Kawintiranon K, et al. A first look at COVID-19 information and misinformation sharing on Twitter. ArXiv. Preprint posted online on March 31, 2020 . [Medline: 32550244]

26. Cinelli M, Quattrociocchi W, Galeazzi A, Valensise CM, Brugnoli E, Schmidt AL, et al. The COVID-19 Social Media Infodemic. arXiv. Preprint posted online on March 10, 2020 [FREE Full text]

27. Lazer D, Kennedy R, King G, Vespignani A. The parable of Google Flu: traps in big data analysis. Science 2014 Mar 14;343(6176):1203-1205. [doi: 10.1126/science.1248506] [Medline: 24626916]

28. Culotta A. Towards detecting influenza epidemics by analyzing Twitter messages. In: SOMA '10: Proceedings of the First Workshop on Social Media Analytics. 2020 Jul Presented at: SOMA 2010: Workshop on Social Media Analytics; May 7-21, 2010; Washington, DC p. 115-122. [doi: 10.1145/1964858.1964874]

29. Lampos V, Cristianini N. Tracking the flu pandemic by monitoring the social web. 2010 Presented at: 2nd International Workshop on Cognitive Information Processing; June 14-16, 2010; Elba, Italy p. 411-416. [doi: 10.1109/cip.2010.5604088]

30. Zhang Q. Forecasting Seasonal Influenza Fusing Digital Indicators and a Mechanistic Disease Model. In: WWW '17: Proceedings of the 26th International Conference on World Wide Web. 2017 Apr Presented at: 26th International Conference on World Wide Web; April 3-7, 2017; Perth, Australia. [doi: 10.1145/3038912.3052678]

31. De Choudhury M, Gamon M, Counts S, Horvitz E. Predicting Depression via Social Media. In: Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media. 2013 Presented at: Seventh International AAAI Conference on Weblogs and Social Media; July 8-11, 2013; Cambridge, MA.

XSL•FO
RenderX

32. De Choudhury M, Counts S, Horvitz E. Social media as a measurement tool of depression in populations. In: WebSci '13: Proceedings of the 5th Annual ACM Web Science Conference. 2013 May Presented at: 5th Annual ACM Web Science Conference; May 2-4, 2013; Paris, France p. 47-56. [doi: 10.1145/2464464.2464480]

33. Broniatowski DA, Paul MJ, Dredze M. National and local influenza surveillance through Twitter: an analysis of the 2012-2013 influenza epidemic. PLoS One 2013 Dec 9;8(12):e83672 [FREE Full text] [doi: 10.1371/journal.pone.0083672] [Medline: 24349542]

34. Araujo M, Mejova Y, Weber I, Benevenuto F. Using Facebook Ads Audiences for Global Lifestyle Disease Surveillance: Promises and Limitations. In: WebSci '17: Proceedings of the 2017 ACM on Web Science Conference. 2017 Jun Presented at: 2017 ACM Web Science Conference; June 25-28, 2017; Troy, NY p. 253-257. [doi: 10.1145/3091478.3091513]

35. Park A, Conway M. Tracking Health Related Discussions on Reddit for Public Health Applications. AMIA Annu Symp Proc 2017:1362-1371 [FREE Full text] [Medline: 29854205]

36. Kumar M, Dredze M, Coppersmith G, De Choudhury M. Detecting Changes in Suicide Content Manifested in Social Media Following Celebrity Suicides. HT ACM Conf Hypertext Soc Media 2015 Sep;2015:85-94 [FREE Full text] [doi: 10.1145/2700171.2791026] [Medline: 28713876]

37. Generous N, Fairchild G, Deshpande A, Del Valle SY, Priedhorsky R. Global disease monitoring and forecasting with Wikipedia. PLoS Comput Biol 2014 Nov 13;10(11):e1003892 [FREE Full text] [doi: 10.1371/journal.pcbi.1003892] [Medline: 25392913]

38. Hickmann K, Fairchild G, Priedhorsky R, Generous N, Hyman JM, Deshpande A, et al. Forecasting the 2013-2014 influenza season using Wikipedia. PLoS Comput Biol 2015 May;11(5):e1004239 [FREE Full text] [doi: 10.1371/journal.pcbi.1004239] [Medline: 25974758]

39. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature 2009 Feb 19;457(7232):1012-1014. [doi: 10.1038/nature07634] [Medline: 19020500]

40. Dugas AF, Jalalpour M, Gel Y, Levin S, Torcaso F, Igusa T, et al. Influenza forecasting with Google Flu Trends. PLoS One 2013;8(2):e56176 [FREE Full text] [doi: 10.1371/journal.pone.0056176] [Medline: 23457520]

41. News API. URL: https://newsapi.org [accessed 2020-05-11]

42. YouTube Data API. URL: https://developers.google.com/youtube/v3 [accessed 2020-05-11]

43. Tan C, Lee L. All Who Wander: On the Prevalence and Characteristics of Multi-community Engagement. In: WWW '15: Proceedings of the 24th International Conference on World Wide Web. 2015 Presented at: 24th International Conference on World Wide Web; May 18-22, 2015; Florence, Italy. [doi: 10.1145/2736277.2741661]

44. Hessel J, Tan C, Lee L. Science, askscience, and badscience: On the coexistence of highly related communities. In: Proceedings of the 10th International AAAI Conference on Web and Social Media. 2016 Presented at: 10th International AAAI Conference on Web and Social Media; May 17–20, 2016; Cologne, Germany p. 171-180.

45. Barthel M. How the 2016 presidential campaign is being discussed on Reddit. Pew Research Center. 2020 May 26. URL: https://www.pewresearch.org/fact-tank/2016/05/26/how-the-2016-presidential-campaign-is-being-discussed-on-reddit/ [accessed 2020-09-24]

46. Horne B, Adali S. The impact of crowds on news engagement: A Reddit case study. ArXiv. Preprint posted online on March 03, 2017 [FREE Full text]

47. Saleem H, Dillon K, Benesch S, Ruths D. A web of hate: Tackling hateful speech in online social spaces. ArXiv. Preprint posted online on September 28, 2017 [FREE Full text]

48. Bin Abdur Rakib T, Soon LK. Using the Reddit Corpus for Cyberbully Detection. In: Nguyen N, Hoang D, Hong TP, Pham H, Trawiński B, editors. Intelligent Information and Database Systems. ACIIDS 2018. Lecture Notes in Computer Science, vol 10751. Cham, Switzerland: Springer; 2018.

49. Choudhury MD, De S. Mental Health Discourse on reddit: Self-disclosure, Social Support, and Anonymity. In: Proceedings of the Eighth International Conference on Weblogs and Social Media. 2014 Presented at: Eighth International Conference on Weblogs and Social Media; June 1-4, 2014; Ann Arbor, MI.

50. Balsamo D, Bajardi P, Panisson A. Firsthand Opiates Abuse on Social Media: Monitoring Geospatial Patterns of Interest Through a Digital Cohort. In: WWW '19: The World Wide Web Conference. 2019 May Presented at: WWW '19: The World Wide Web Conference; May 13-17, 2019; San Francisco, CA p. 2572-2579. [doi: 10.1145/3308558.3313634]

51. Radford A, Wu J, Child R, Luan D, Amodei S, Sutskever I. Language Models are Unsupervised Multitask Learners. SemanticsScholar. URL: https://www.semanticscholar.org/paper/Language-Models-are-Unsupervised-Multitask-Learners-Radford-Wu/9405cc0d6169988371b2755e573cc28650d14dfe [accessed 2020-09-24]

52. Laurent MR, Vickers TJ. Seeking health information online: does Wikipedia matter? J Am Med Inform Assoc 2009;16(4):471-479 [FREE Full text] [doi: 10.1197/jamia.M3059] [Medline: 19390105]

53. McIver DJ, Brownstein JS. Wikipedia usage estimates prevalence of influenza-like illness in the United States in near real-time. PLoS Comput Biol 2014 Apr;10(4):e1003581 [FREE Full text] [doi: 10.1371/journal.pcbi.1003581] [Medline: 24743682]

54. Generous N, Fairchild G, Deshpande A, Del Valle SY, Priedhorsky R. Global disease monitoring and forecasting with Wikipedia. PLoS Comput Biol 2014 Nov;10(11):e1003892 [FREE Full text] [doi: 10.1371/journal.pcbi.1003892] [Medline: 25392913]

55. Wikimedia API. URL: https://wikimedia.org/api/rest [accessed 2020-05-11]

56. Eysenbach G. Infodemiology and infoveillance tracking online health information and cyberbehavior for public health. Am J Prev Med 2011 May;40(5 Suppl 2):S154-S158. [doi: 10.1016/j.amepre.2011.02.006] [Medline: 21521589]

57. Milinovich GJ, Williams GM, Clements ACA, Hu W. Internet-based surveillance systems for monitoring emerging infectious diseases. Lancet Infect Dis 2014 Feb;14(2):160-168 [FREE Full text] [doi: 10.1016/s1473-3099(13)70244-5]

58. Park HW, Park S, Chong M. Conversations and Medical News Frames on Twitter: Infodemiological Study on COVID-19 in South Korea. J Med Internet Res 2020 May 05;22(5):e18897 [FREE Full text] [doi: 10.2196/18897] [Medline: 32325426]

59. Park A, Conway M. Tracking Health Related Discussions on Reddit for Public Health Applications. AMIA Annu Symp Proc 2017;2017:1362-1371 [FREE Full text] [Medline: 29854205]

60. Lamb A, Paul M, Dredze M. Investigating Twitter as a Source for Studying Behavioral Responses to Epidemics. AAAI Fall Symposium - Technical Report. 2012 Jan 01. URL: https://www.researchgate.net/publication/266506521_Investigating_Twitter_as_a_Source_for_Studying_Behavioral_Responses_to_Epidemics [accessed 2020-09-24]

61. Lewis N. Information Seeking and Scanning. In: The International Encyclopedia of Media Effects. Hoboken, NJ: Wiley; Mar 08, 2017.

62. Lambert SD, Loiselle CG. Health information seeking behavior. Qual Health Res 2007 Oct;17(8):1006-1019. [doi: 10.1177/1049732307305199] [Medline: 17928475]

63. Walter D, Bohmer MM, Reiter S, Krause G, Wichmann O. Risk perception and information-seeking behaviour during the 2009/10 influenza A(H1N1)pdm09 pandemic in Germany. Euro Surveill 2012 Mar 29;17(13):pii=20131 [FREE Full text] [Medline: 22490383]

64. Pang NLS. Crisis-based information seeking: Monitoring versus blunting in the information seeking behaviour of working students during the Southeast Asian Haze Crisis. Inform Res 2014 Dec;19(4):online [FREE Full text]

65. Miles J. R‐Squared, Adjusted R‐Squared. In: Everitt BS, Howell DC, editors. Encyclopedia of Statistics in Behavioral Science. Hoboken, NJ: Wiley; 2005.

66. Blei DM, Ng AY, Jordan MI. Latent Dirichlet Allocation. J Mach Learn Res 2003 Mar;3:993-1022 [FREE Full text]

67. Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. Nature 1999 Oct 21;401(6755):788-791. [doi: 10.1038/44565] [Medline: 10548103]

68. Blei DM, Lafferty JD. Dynamic topic models. In: ICML '06: Proceedings of the 23rd International Conference on Machine Learning. 2006 Jun Presented at: 23rd International Conference on Machine Learning; June 25-29, 2006; Pittsburgh, PA p. 113-120. [doi: 10.1145/1143844.1143859]

69. Gobbo B, Balsamo D, Mauri M, Bajardi P, Panisson A, Ciuccarelli P. Topic Tomographies (TopTom): a visual approach to distill information from media streams. Comput Graph Forum 2019 Jul 10;38(3):609-621. [doi: 10.1111/cgf.13714]

70. Dou W, Yu L, Wang X, Ma Z, Ribarsky W. HierarchicalTopics: Visually Exploring Large Text Collections Using Topic Hierarchies. IEEE Trans Visual Comput Graphics 2013 Dec;19(12):2002-2011. [doi: 10.1109/tvcg.2013.162]

71. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. In: NIPS'13: Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2. 2013 Dec Presented at: 26th International Conference on Neural Information Processing Systems; December 5-8, 2013; Lake Tahoe, NV p. 3111-3119.

72. Sparck Jones K. A statistical interpretation of term specificity and its application in retrieval. J Doc 1972 Jan;28(1):11-21. [doi: 10.1108/eb026526]

73. Lin C. Projected gradient methods for nonnegative matrix factorization. Neural Comput 2007 Oct;19(10):2756-2779. [doi: 10.1162/neco.2007.19.10.2756] [Medline: 17716011]

74. Hoyer PO. Non-negative matrix factorization with sparseness constraints. J Mach Learn Res 2004 Dec;5:1457-1469. [doi: 10.1109/icmlc.2011.6016966]

75. Chen Y, Skiena S. False-Friend Detection and Entity Matching via Unsupervised Transliteration. ArXiv. Preprint posted online on November 21, 2016 [FREE Full text]

76. Python Geocoder. URL: https://geocoder.readthedocs.io [accessed 2020-05-29]

77. Greene WH. Econometric Analysis. Upper Saddle River, NJ: Prentice Hall; 2002.

78. Allen MP. Testing hypotheses in nested regression models. In: Understanding Regression Analysis. Boston, MA: Springer; 1997:113-117.

79. Zhao WX, Jiang J, Weng J, He J, Lim EP, Yan H, et al. Comparing Twitter and Traditional Media Using Topic Models. In: Clough P, editor. Advances in Information Retrieval. ECIR 2011. Lecture Notes in Computer Science, vol 6611. Berlin, Germany: Springer; 2011.

80. Diao Q, Jiang J, Zhu F, Lim EP. Finding bursty topics from microblogs. In: ACL '12: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1. 2012 Jul Presented at: 50th Annual Meeting of the Association for Computational Linguistics; July 8-14, 2012; Jeju Island, Korea p. 536-544.

81. Bendavid E, Bendavid E, Mulaney B, Sood N, Shah S, Ling E, et al. COVID-19 Antibody Seroprevalence in Santa Clara County, California. medArxiv. Preprint posted online on April 04, 2020 . [doi: 10.1101/2020.04.24.20078824]

82. Bajak A, Howe J. A Study Said Covid Wasn't That Deadly. The Right Seized It. New York Times. 2020 May 14. URL: https://www.nytimes.com/2020/05/14/opinion/coronavirus-research-misinformation.html [accessed 2020-09-24]

83. Yong E. Why the Coronavirus Is So Confusing. The Atlantic. 2020 Apr 29. URL: https://www.theatlantic.com/health/archive/2020/04/pandemic-confusing-uncertainty/610819 [accessed 2020-09-24]

84. Tausczik Y, Faasse K, Pennebaker JW, Petrie KJ. Public anxiety and information seeking following the H1N1 outbreak: blogs, newspaper articles, and Wikipedia visits. Health Commun 2012;27(2):179-185. [doi: 10.1080/10410236.2011.571759] [Medline: 21827326]

85. Chew C, Eysenbach G. Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak. PLoS One 2010 Nov 29;5(11):e14118 [FREE Full text] [doi: 10.1371/journal.pone.0014118] [Medline: 21124761]

86. Pruss D, Fujinuma Y, Daughton AR, Paul MJ, Arnot B, Albers Szafir D, et al. Zika discourse in the Americas: A multilingual topic analysis of Twitter. PLoS One 2019;14(5):e0216922 [FREE Full text] [doi: 10.1371/journal.pone.0216922] [Medline: 31120935]

87. Smith M, Broniatowski DA, Paul MJ, Dredze M. Towards Real-Time Measurement of Public Epidemic Awareness: Monitoring Influenza Awareness through Twitter. 2015 Presented at: AAAI Spring Symposium on Observational Studies through Social Media and Other Human-Generated Content; March 21-23, 2015; Stanford, CA.

88. Mollema L, Harmsen IA, Broekhuizen E, Clijnk R, De Melker H, Paulussen T, et al. Disease detection or public opinion reflection? Content analysis of tweets, other social media, and online newspapers during the measles outbreak in The Netherlands in 2013. J Med Internet Res 2015 May 26;17(5):e128 [FREE Full text] [doi: 10.2196/jmir.3863] [Medline: 26013683]

89. Wahlberg AAF, Sjoberg L. Risk perception and the media. J Risk Res 2000 Jan;3(1):31-50 [FREE Full text] [doi: 10.1080/136698700376699]

90. Klemm C, Das E, Hartmann T. Swine flu and hype: a systematic review of media dramatization of the H1N1 influenza pandemic. J Risk Res 2014 Jun 20;19(1):1-20. [doi: 10.1080/13669877.2014.923029]

91. Tchuenche JM, Dube N, Bhunu CP, Smith RJ, Bauch CT. The impact of media coverage on the transmission dynamics of human influenza. BMC Public Health 2011 Feb 25;11(S5):online. [doi: 10.1186/1471-2458-11-s1-s5]

92. Johnson BB. Explaining Americans' responses to dread epidemics: an illustration with Ebola in late 2014. J Risk Res 2016 Mar 03;20(10):1338-1357 [FREE Full text] [doi: 10.1080/13669877.2016.1153507]

93. Sell TK, Boddie C, McGinty EE, Pollack K, Smith KC, Burke TA, et al. Media Messages and Perception of Risk for Ebola Virus Infection, United States. Emerg Infect Dis 2017 Jan;23(1):108-111. [doi: 10.3201/eid2301.160589]

94. Gozzi N, Perrotta D, Paolotti D, Perra N. Towards a data-driven characterization of behavioral changes induced by the seasonal flu. PLoS Comput Biol 2020 May;16(5):e1007879 [FREE Full text] [doi: 10.1371/journal.pcbi.1007879] [Medline: 32401809]

95. Wise T, Zbozinek T, Michelini G, Hagan C. Changes in risk perception and protective behavior during the first week of the COVID-19 pandemic in the United States. PsyArXiv. Preprint posted online on March 19, 2020 . [doi: 10.31234/osf.io/dz428]

96. Wang W, Ahern L. Acting on surprise: emotional response, multiple-channel information seeking and vaccination in the H1N1 flu epidemic. Soc Influ 2015 Feb 23;10(3):137-148 [FREE Full text] [doi: 10.1080/15534510.2015.1011227]

97. Duggan M, Smith A. 6% of Online Adults are reddit Users. Pew Research Center. 2103 Jul 03. URL: https://www.pewresearch.org/internet/2013/07/03/6-of-online-adults-are-reddit-users/ [accessed 2020-09-24]

98. Finlay SC. Age and Gender in Reddit Commenting and Success. J Inf Sci Theory Pract 2014 Sep 30;2(3):18-28. [doi: 10.1633/jistap.2014.2.3.2]

99. Singer P, Flock F, Meinhart C, Zeitfogel E, Strohmaier M. Evolution of reddit: from the front page of the internet to a self-referential community? In: WWW '14 Companion: Proceedings of the 23rd International Conference on World Wide Web. 2014 Apr Presented at: 23rd International Conference on World Wide Web; April 7-11, 2014; Seoul, Korea p. 517-522. [doi: 10.1145/2567948.2576943]

100. van Deursen AJ, van Dijk JA. The digital divide shifts to differences in usage. New Media Soc 2013 Jun 07;16(3):507-526. [doi: 10.1177/1461444813487959]

101. van Deursen A, van Dijk J. Internet skills and the digital divide. New Media Soc 2010 Dec 06;13(6):893-911. [doi: 10.1177/1461444810386774]

102. Robinson L, Cotten SR, Ono H, Quan-Haase A, Mesch G, Chen W, et al. Digital inequalities and why they matter. Inf Commun Soc 2015 Mar 16;18(5):569-582 [FREE Full text] [doi: 10.1080/1369118x.2015.1012532]

103. Guidry JP, Jin Y, Orr CA, Messner M, Meganck S. Ebola on Instagram and Twitter: How health organizations address the health crisis in their social media engagement. Public Relat Rev 2017 Sep;43(3):477-486. [doi: 10.1016/j.pubrev.2017.04.009]

104. Martínez-Rojas M, Pardo-Ferreira MDC, Rubio-Romero JC. Twitter as a tool for the management and analysis of emergency situations: A systematic literature review. Int J Inf Manag 2018 Dec;43:196-208. [doi: 10.1016/j.ijinfomgt.2018.07.008]

105. Ferrara E. #COVID-19 on Twitter: Bots, Conspiracies, and Social Media Activism. ArXiv. Preprint posted online on April 20, 2017 [FREE Full text]

XSL•FO
RenderX

106. Yang K, Torres-Lugo C, Menczer F. Prevalence of Low-Credibility Information on Twitter During the COVID-19 Outbreak. ArXiv. Preprint posted online on April 29, 2020 [FREE Full text]

107. Ahmed W, Vidal-Aiaball J, Downing J, López Seguí F. COVID-19 and the 5G Conspiracy Theory: Social Network Analysis of Twitter Data. J Med Internet Res 2020 May 06;22(5):e19458 [FREE Full text] [doi: 10.2196/19458] [Medline: 32352383]

108. Guerrisi C, Turbelin C, Blanchon T, Hanslik T, Bonmarin I, Levy-Bruhl D, et al. Participatory Syndromic Surveillance of Influenza in Europe. J Infect Dis 2016 Dec 01;214(suppl_4):S386-S392. [doi: 10.1093/infdis/jiw280] [Medline: 28830105]

109. Smolinski MS, Crawley AW, Baltrusaitis K, Chunara R, Olsen JM, Wójcik O, et al. Flu Near You: Crowdsourced Symptom Reporting Spanning 2 Influenza Seasons. Am J Public Health 2015 Oct;105(10):2124-2130. [doi: 10.2105/ajph.2015.302696]

110. Dalton C, Durrheim D, Fejsa J, Francis L, Carlson S, Tursan d'Espaignet E, et al. Flutracking: A weekly Australian community online survey of influenza-like illness in 2006, 2007 and 2008. Commun Dis Intell 2009 Dec 07;33(3):316-322 [FREE Full text]

111. Wójcik OP, Brownstein JS, Chunara R, Johansson MA. Public health for the people: participatory infectious disease surveillance in the digital age. Emerg Themes Epidemiol 2014;11:7 [FREE Full text] [doi: 10.1186/1742-7622-11-7] [Medline: 24991229]

112. COVID-19 News Tracker. Scops. URL: https://covid19.scops.ai/scops/home/ [accessed 2020-09-24]

Abbreviations

API: application programming interface
CDC: US Centers for Disease Control and Prevention
LDA: latent Dirichlet allocation
NMF: nonnegative matrix factorization
TF-IDF: term frequency–inverse document frequency
WHO: World Health Organization