JMIR Publications
The Leading eHealth Publisher

Collective response to the media coverage of COVID-19 Pandemic on Reddit and Wikipedia

Nicolò Gozzi, Michele Tizzani, Michele Starnini, Fabio Ciulla, Daniela Paolotti, André Panisson, Nicola Perra

Submitted to: Journal of Medical Internet Research
on: June 19, 2020

Preprint
JMIR Publications

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Preprint JMIR Publications

Collective response to the media coverage of COVID-19 Pandemic on Reddit and Wikipedia

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 GB
[2]ISI Foundation Torino IT
[3]Quid Inc San Francisco US

Corresponding Author:
Nicolò Gozzi MSc
University of Greenwich
Old Royal Naval College
Park Row
London
GB

Abstract

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

Objective: The goal of this work is to provide a characterization of media coverage and online collective response to COVID-19 pandemic in four countries: Italy, United Kingdom, United States, and Canada.

Methods: We collect a heterogeneous dataset including 227’768 online news articles and 13’448 YouTube videos published by mainstream media, 107’898 users posts and 3’829’309 comments on the social media platform Reddit, and 278’456’892 views to COVID-19 related Wikipedia pages.

Results: Our results show that public attention, quantified as users activity on Reddit and active searches on Wikipedia pages, is mainly driven by media coverage and declines rapidly, while news exposure and COVID-19 incidence remain high. Furthermore, by using an unsupervised, dynamical topic modeling approach, we show that while the attention dedicated to different topics by media and online 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 collective awareness, risk perception and thus on tendencies towards behavioural changes.

(JMIR Preprints 19/06/2020:21597)
DOI: https://doi.org/10.2196/preprints.21597

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Preprint
JMIR Publications

Original Paper

Collective response to the media coverage of COVID-19 Pandemic on Reddit and Wikipedia

Abstract

Background: The exposure and consumption of information during epidemic outbreaks may alter risk perception, trigger behavioral changes, and ultimately affect the evolution of the disease. It is thus of the uttermost importance to map information dissemination by mainstream media outlets and public response. However, our understanding of this exposure-response dynamic during COVID-19 pandemic is still limited.
Objective: The goal of this work is to provide a characterization of media coverage and online collective response to COVID-19 pandemic in four countries: Italy, United Kingdom, United States, and Canada.
Methods: We collect a heterogeneous dataset including 227'768 online news articles and 13'448 YouTube videos published by mainstream media, 107'898 users posts and 3'829'309 comments on the social media platform Reddit, and 278'456'892 views to COVID-19 related Wikipedia pages.
Results: Our results show that public attention, quantified as users activity on Reddit and active searches on Wikipedia pages, is mainly driven by media coverage and declines rapidly, while news exposure and COVID-19 incidence remain high. Furthermore, by using an unsupervised, dynamical topic modeling approach, we show that while the attention dedicated to different topics by media and online 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 collective awareness, risk perception and thus on tendencies towards behavioural changes.

Keywords: social media; news coverage; digital epidemiology; data science; topic modeling; pandemic; covid19

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" [1]. This evocative sentence was written in May 2009 by John M. Barry, in the early phases of what soon after became the H1N1 2009 pandemic. In his essay, Barry summarized the mishandling of the deadly 1918 Spanish flu highlighting the importance of precise, effective and honest information in the onset of health crises.
Eleven years later we find ourselves dealing with another pandemic. The cause is not a novel strain of influenza, but these words are, unfortunately, still extremely relevant. In fact, as the SARS-CoV-2 sweeps the world and the vaccine is just a far vision of hope, the most important weapons to reduce the burden of the disease are non-pharmaceutical interventions [2, 3]. Social distancing became

paramount, gatherings have been cancelled, mobility within and across countries have been dramatically reduced. While such measures have been enforced to different extents across nations, they all rely on compliance. Their effectiveness is linked to risk and susceptibility perception [4], thus the information that citizens are exposed to is fundamental.

History repeats itself and we seem not be able to learn from our past mistakes. As happened in 1918, despite early evidences from China [5, 6], the virus was first equated, by many, to the normal seasonal flu. As happened in 1918, many national and regional governments organized campaigns aimed at boosting social activities (and thus local economies) actively trying to convince people that their cities were safe and that the spreading 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. Free aperitifs were offered in Venice. In hindsight, of course, is easy to criticize the initial response in Italy. In fact, the country has been one of the first to experience rapid growth of hospitalizations [7]. However, the Mayor of London, twelve days before the national lockdown, and few days after the extension of the cordon sanitaire to the entire country in Italy, affirmed via his official Facebook page “we should carry on doing what we’ve been doing” [8]. More in general, in several western countries, the news coming from others reporting worrying epidemic outbreaks were not considered as relevant for the internal situation. This initial phase aimed at conveying low local risk and boosting confidence about national safety has been repeated, at different times, across countries. A series of surveys conducted in late February provide a glimpse of the possible effects of these approaches. They 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 infections to be 1% or less [9]. As happened in 1918, the countries that reacted earlier rather than later were able to control the virus with significant less victims [10–14].

History repeats itself, but the context often is radically different. In 1918, news circulated slowly via newspapers, controlled by editorial choices, and of course words of mouth. In 2009, we witnessed the first pandemic in the social media era. Newspapers and TV were still very important source of information, but Twitter, Facebook, YouTube, Wikipedia started to become relevant for decentralized news consumption, boosting peer discussions, and misinformation spread. Today these platforms and websites are far more popular, integral part of society and instrumental pieces of the national and international news circulations. Together with traditional news media, they are the principal sources of information for the public. As such, they are fundamental drivers of people perception, opinions, and thus behaviors. This is particularly relevant for health issues. For example, about 60% of adults in the USA consulted online sources to gather health information [15].

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 (such as vaccination intent), stigmatization of cultures at the epicenter of the outbreak, and how these factors differ across countries/cultures [16–21]. During Zika epidemic in 2016, public attention was synchronized across US states, driven by news coverage about the outbreak and independently of the real local risk of infection [22]. With respect to COVID-19 pandemic itself, a recent study clearly shows how Google searches for “coronavirus” in the USA spiked significantly right after the announcement of the first confirmed case in each state [23]. Several studies based on Twitter data also highlight how misinformation and low-quality information about COVID-19, although overall limited, spread before the local outbreak and rapidly took off once the local epidemic started. In the current landscape, this has the potential to boost irrational, unscientific, and

dangerous behaviors [24–26]. 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 be used to develop an unprecedented understanding of such behaviors, as well as of the phenomena driving them.

Goal of This Study

The context where COVID-19 is unfolding is thus very heterogeneous and complex. Traditional and social media are integral parts of our perception and opinions, have the potential to trigger behavior change and thus influence the pandemic spreading. Such complex landscape must be characterized in order to understand the public attention and response to media coverage. Here, we tackle this challenge by assembling an heterogeneous dataset which includes 227'768 news and 13'448 YouTube videos published by traditional media, 278'456'892 views of topical Wikipedia pages, 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, United Kingdom, United States, and Canada. First, we explore how media coverage and epidemic progression influence public attention and response. To achieve this, we analyze news volume and COVID-19 incidence with respect to Wikipedia page views volume and Reddit comments. Our results show that public attention and response are mostly driven by media coverage rather than disease spreading. Furthermore, we observe typical saturation and memory effects of public collective attention. Moreover, using an unsupervised topic modeling approach, we explore the different topics framed in traditional media and in Reddit discussions. We show that, while attentions of news outlets and online users towards different topics are in good accordance, interesting deviations emerge in their temporal patterns. Also, we highlight that, at the end of our observation period, general interest grows towards topics about the resumption of activities after lockdown, the search for a vaccine against Sars-Cov-2, acquired immunity and antibodies tests. Overall, the research presented here offers insights to interpret public perception and response to the current global health emergency and raises interrogatives about the effects of attention saturation on collective awareness, risk perception and thus on tendencies towards behavioral changes.

Methods

Dataset

News Articles and Videos

We collect news articles using News API, a service that allows to download articles published online in a variety of countries and languages [41]. For each of the country considered, we download all relevant articles published online by selected sources in the period 2020/02/07 - 2020/05/15. We select "relevant" articles considering those citing one of the following keywords: 'coronavirus', 'covid19', 'covid-19', 'ncov-19', 'sars-cov-2'. Note that for each article we have access to title, description and a preview of the whole text. In total, our dataset consists in 227'768 news: 71'461 published by Italian, 63'799 by UK, 82'630 by US, and 9'878 by Canadian media.

Additionally, we collect all videos published on YouTube by major news organizations, in the four countries under investigation, via their official YouTube channels using the official API [42]. In doing so, we download title and description of all videos and select as relevant those that mention

one of the following keywords: ‘coronavirus’, ‘virus’, ‘covid’, ‘covid19’, ‘sars’, ‘sars-cov-2’, ‘sarscov2’. The reach of each channel (measured by number of subscribers) varies quite drastically from more than 9 million for CNN (USA) to about 12 thousand for Ansa (Italy). In total, the YouTube dataset consist of 13, 448 videos: 3’325 by Italian, 3’525 by British, 6’288 by American, and 310 by Canadian channels.

It is important to underline that, while there is a 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 run a YouTube channel and others do not produce traditional articles. In the Supplementary Information, we provide a complete list of news outlets and YouTube channels considered.

Reddit Posts

Reddit is a social content aggregation website where users can post, comment and vote content. It is structured in sub-communities (i.e. *subreddits*), centred around a variety of topics. Reddit has already 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 post-election political analyses [45]. Also, 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] or cyberbullying [48] as well as health related issues like mental illness [49], also providing insights about the opioid epidemics [50, 51].

We use the Reddit API to collect all submissions and comments published in Reddit under the subreddit /r/Coronavirus from 15/02/2020 to 15/05/2020. After data clean-up by removing entries deleted by authors and moderators, we keep only submissions with score > 1 to avoid spam. We remove comments with less than 10 characters and with more than 3 duplicates, to avoid using automatic messages from moderation. Final data contains 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 news outlets. The content of the URLs was extracted using the available implementation of the method described in [52], resulting in 66’575 valid documents.

Reddit does not provide any explicit information about users’ location; therefore, we use self-reporting via regular expression to assign a location to users. Reddit users often declare geographical information about themselves in submissions or comment texts. We use the same approach as described in [51], that found the use of regular expressions as reliable, resulting in high correlation with census data in the US, although we acknowledge a potential higher bias at country level due to heterogeneities in Reddit population coverage and users’ demographics. We select all texts containing expressions such as ‘I am from’ or ‘I live in’ and extract candidate expressions from the text that follows the expression, to identify which ones represent country locations. By removing inconsistent self-reporting, we are able to assign a country to 789’909 distinct users, from which 41’465 have written at least one comment in the subreddit r/Coronavirus (13’811 from USA, 6’870 from Canada, 3’932 from UK and 445 from Italy).

Wikipedia Pages Views

Wikipedia has become a popular digital data source to study health information seeking behaviour [53, 54], and to monitor and forecast the spreading of infectious diseases [55, 56]. Here, we use the Wikimedia API [57] to collect the number of visits per day of Wikipedia articles and the total monthly accesses to a specific project from each country. We consider the language as indicative of a specific country, suggesting the relevant projects for our analysis to be in English and Italian, i.e. *en.wikipedia* and *it.wikipedia* respectively. We choose the articles directly related to COVID-19 and

the ones in the *'see also'* section of each page at the time of the analysis, 2020/02/07 - 2020/05/15, including country-specific articles (see Supplementary Information for full list of web pages considered).

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

$$y_{a,p}^c(d) = \frac{S_p(d) T_p^c(d)}{\sum_c T_p^c(d)} \quad (1)$$

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

$$y^c(d) = \sum_{a,p} y_{a,p}^c(d) \quad (2)$$

Media Coverage and Online Collective Response

The dataset just described aims 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 allows us to estimate the exposure of the public to COVID-19 pandemic in traditional news media. On the other hand, the study of users' discussions and response on social media (through Reddit) and information seeking (through Wikipedia page views) allows us to quantify the reaction of individuals to both the COVID-19 pandemic and 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 [68–72]. Hence, we consider volume of comments of geolocalized users on the subreddit /r/Coronavirus to explore the public discussion in reaction to media covering the epidemic in the various countries, while we consider the number of views of relevant Wikipedia pages about COVID-19 pandemic to quantify users' interest. It is important to stress how Reddit and Wikipedia provide different aspects of online users' behavior and collective response. In fact, while Reddit posts can be regarded as a general indicator of the online discussion surrounding the global health emergency, the number of access to COVID-19 related Wikipedia pages is a proxy of health information seeking behavior (HISB). HISB is the act through which individuals retrieve and acquire new knowledge about a specific topic related to health [73, 74], and it is likely to be triggered on a population scale by a disrupting event, such as the threaten of a previously unknown disease [75, 76].

Linear Regression Approach to Model Collective Attention

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

new variable:

$$newsMEM = \sum_{\Delta t=1}^{\tau} e^{\frac{-\Delta t}{\tau}} news(t-\Delta t) (3)$$

Where τ is a free parameter that sets the memory time scale and is tuned with cross-validation (more details in the Supplementary Information). These two models are 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 either the volume of Reddit comments of geolocalized users or country specific Wikipedia visits, and u_t is the error term.

Topic Modeling

Topic modeling has emerged as one of the most effective methods for classifying, clustering, and retrieving textual data, and has been the object of extensive investigation in the literature. Many topic analysis frameworks are extensions of well-known algorithms, considered as state-of-the-art for topic modeling. Latent Dirichlet Allocation (LDA) [59] is the reference for probabilistic topic modeling. Nonnegative matrix factorization (NMF) [60] is the counterpart of LDA for the matrix factorization community.

Although there are many approaches to temporal and hierarchical topic modeling [61–63], we choose to apply NMF to the dataset, and then build time-varying intensities for each topic using the articles publication date. Starting from a dataset D containing the news articles shared in Reddit, we extract words and phrases with the methodology described in [64], discarding terms with frequency below 10, to form a vocabulary V with around 60k terms. Each document is then represented as a vector of term counts, in a bag-of-words approach. We apply TF-IDF normalization [65] and extract 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 is achieved using the projected gradient method with sparseness constraints, as described in [66, 67]. The matrix H is then used as a transformation basis for other datasets, e.g. with a new matrix $\tilde{X}$ we fix H and calculate a new $\tilde{W}$ according to Eq. 5.

For each topic k we build a time series s_k for each dataset D, where $s_k^{(t)}$ is the strength of topic k at time t. For the news outlets dataset $s_k^{(t)} = \sum_{i \in D^{(t)}} w_{ik}$, where $D^{(t)}$ is the set of all documents shared at time t in news outlets. For Reddit, we weight each shared document by its number of

comments, and $s_k^{(t)} = \sum_{i \in D^{(t)}} w_{ik} \cdot c_i$, where $D^{(t)}$ is the set of all documents shared at time t in Reddit, and c_i is the number of comments associated to document *i*. Finally, we define the relevance R of a topic as the integral in time of the strength. Therefore, given t_0 and t_f as the start/end of our analysis interval $R = \int_{t_0}^{t_f} dt\, s_k^{(t)}$.

Results

Impact of Media Coverage and Epidemic Progression on Collective Attention

How is collective attention shaped by news media coverage and epidemic progression? To tackle this important question, we start by comparing, in Figure 1, the weekly volume of news and videos published on YouTube, Wikipedia views, and Reddit comments of geolocalized users in comparison with the weekly COVID-19 incidence in the four countries considered. It can be seen how, as COVID-19 spreads, both media coverage and public interest grow in time. However, public attention, quantified by the number of Reddit comments and Wikipedia views, sharply decreases after reaching a peak, despite the volume of news and COVID-19 incidence remaining high. Furthermore, the peak in public attention consistently anticipates the maximum media exposure and maximum COVID-19 incidence.

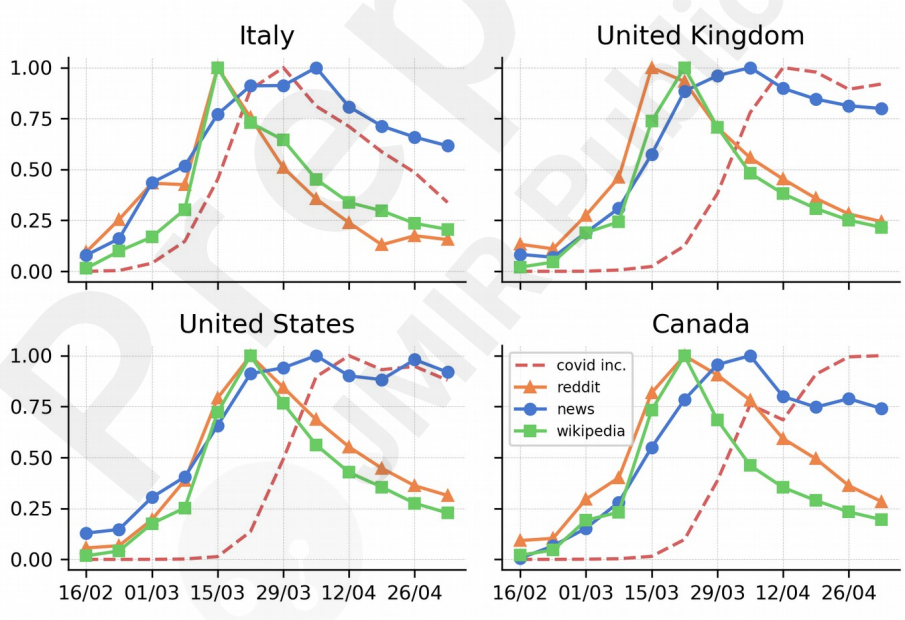


Figure 1. Normalized weekly volume of news articles and YouTube videos (news), Reddit comments (reddit), Wikipedia views (wikipedia) related to COVID-19 pandemic and COVID-19 incidence (covid inc.) in different countries.

The correlation between media coverage, public attention, and the epidemic progression is quantified more in details in Figure 2. The plot shows that news coverage of each country is strongly correlated with COVID-19 incidence (both global and domestic), and slightly less with the volume of Reddit comments and Wikipedia views, which, in turn, are much less correlated

with COVID-19 incidence (both global and domestic). This holds for all countries

under consideration and highlights how the disease spreading triggers media coverage, and how the public response is more likely driven by such news exposure in each country rather than COVID-19 progression.

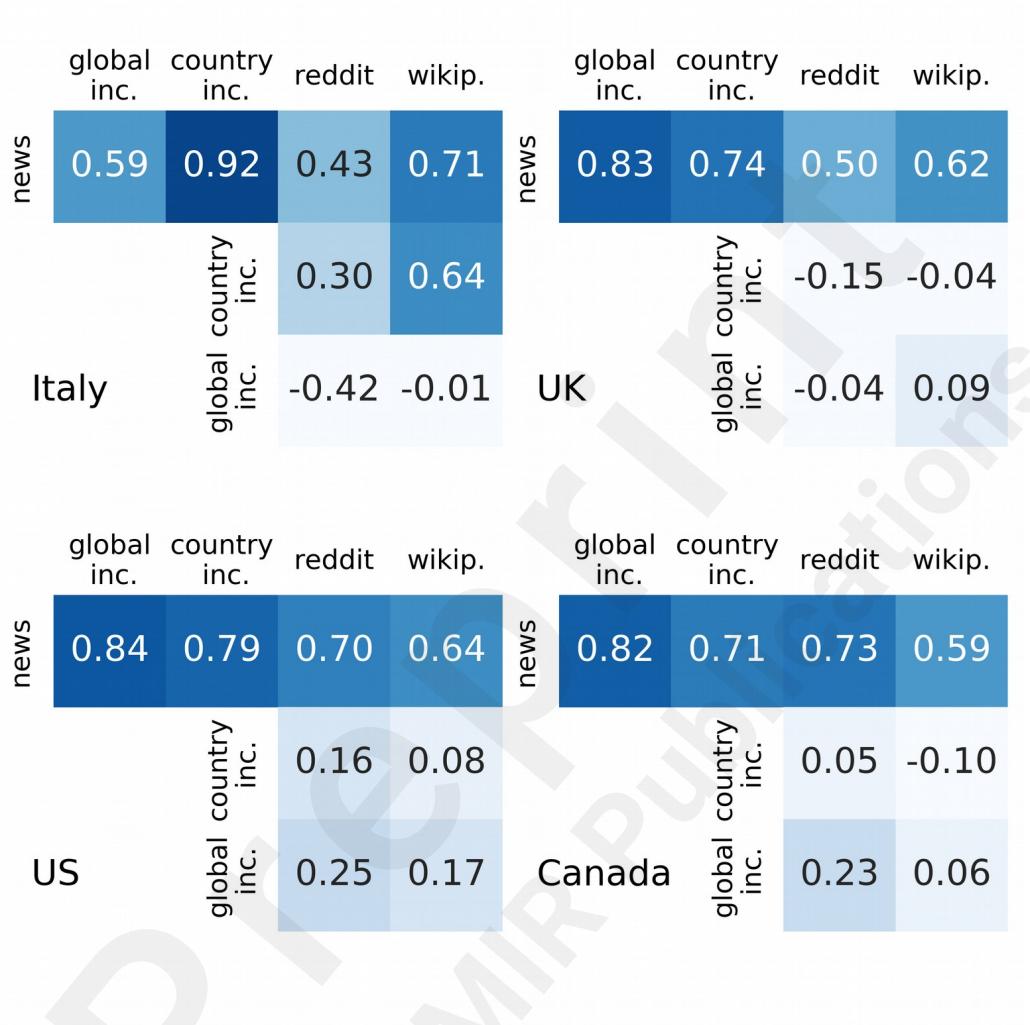


Figure 2. Country specific Pearson correlation coefficients between 1) news coverage and global/domestic COVID-19 incidence, volumes of Reddit comments and Wikipedia views; 2) domestic COVID-19 incidence and volumes of Reddit comments and Wikipedia views; 3) global COVID-19 incidence and volumes of Reddit comments and Wikipedia views.

Beyond these observations, it is interesting to notice from Figure 2 that Italy is the only country where news volume shows higher correlation with domestic rather than global incidence. This suggests that Italian media coverage follows more closely the internal evolution rather than the global one, at odds with respect to other countries. This is probably due to Italy being the location of the first COVID-19 outbreak outside Asia. This observation is supported by Figure 3, showing the citation share of Italian locations by Italian news media, before and after the first COVID-19 death was confirmed in Italy on 2020/02/20. After this date, Italian locations represent about 74% of all places cited by Italian media (in our dataset), with an increase of 45% with respect to the same statistics calculated before. Similar effects, though generally less

intense, can be observed also in the other countries. Therefore, while media coverage is generally well synchronized with the global COVID-19 incidence, the media attention gradually shifts towards the internal evolution of the pandemic as soon as domestic outbreaks erupt. Arguably, this may have played an important role in individual risk perception. We can speculate that re-framing the emergency within a national dimension had the potential to amplify the perceived susceptibility of individuals [77, 78] and thus increase the adoption of behavioural changes [4, 79]. Indeed, previous studies showed how at the beginning of February 2020 people were overly optimistic regarding the risks associated with the new virus circulating in Asia, and how their perception sharply changed after first cases were confirmed in their countries [9, 80].

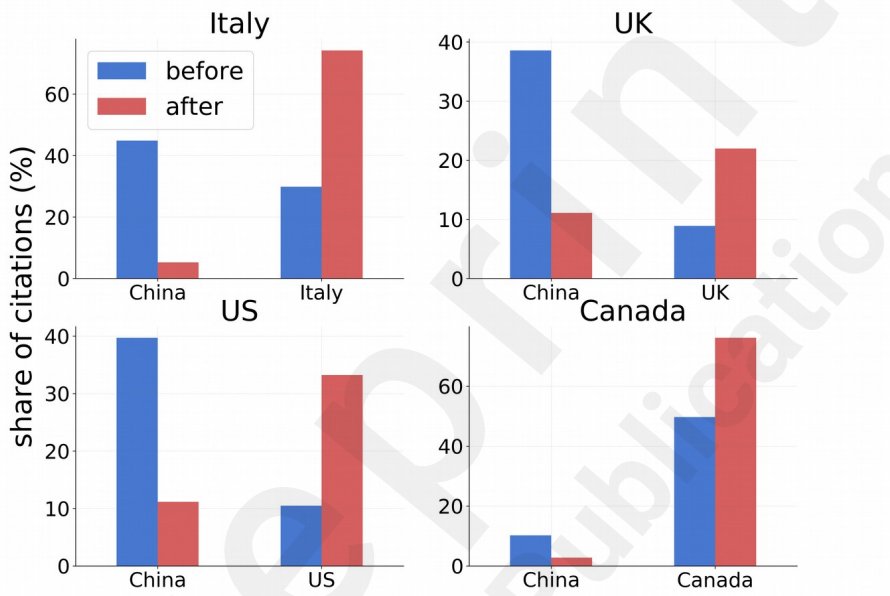


Figure 3. Share of citations of China versus home country locations by Italian/UK/US/Canadian news outlets before and after first COVID-19 death occurred in different countries considered. Geographic locations are extracted from text using [55, 56]

To explore more systematically the relationship between media coverage, public attention and epidemic progression, we consider a linear regression model to nowcast, separately for each country, collective public attention (quantified with the number of comments by geolocalized Reddit users or visits to relevant Wikipedia pages) given the volume of media coverage or the COVID-19 incidence as independent variables. We include also “memory effects” in the public attention by considering an exponential decaying term in the news time series [22]. We compare the three models, where the independent variable(s) are the domestic incidence, the news volume, the news volume plus a memory term, by using the Akaike Information Criterion (AIC) [58] and coefficient of determination (R^2). We find that the model considering only COVID-19 incidence has much less predictive power than the ones considering media coverage (Table 1). This enforces the idea that collective attention is mainly driven by media coverage rather than COVID-19 incidence. In addition, we found that including memory effects improves significantly the model performance. Not surprisingly, the coefficients of the “memory effects” term reported in Table 2 are negative for all countries. This implies that public attention actually saturates in response to news exposure and gives us the chance to quantify the rate at

which this phenomenon happens.

Table 1. Akaike Information Criterion for the three linear regression models applied to predict Reddit comments and Wikipedia visits. As a practical rule, a model *i* is preferred to model *j* if $AIC(j) - AIC(i) \leq 3$.

	incidence		*news*		*news + newsMEM*	
	reddit	wikipedia	reddit	wikipedia	reddit	wikipedia
Italy	3.71	-98.73	-29.68	-121.3	-76.44	-138.49
UK	64.24	65.85	-16.03	-20.46	-51.97	-65.24
US	86.02	61.85	-11.76	-14.67	-48.39	-44.93
Canada	84.53	68.65	-28.53	-13.99	-69.26	-53.08

Table 2. Coefficient estimates for news plus memory effects linear regression model and related R^2. It is worth underlying that, despite the simplicity of this approach, we obtain relatively high values of R^2.

	news		*newsMEM*		R^2	
	reddit	wikipedia	reddit	wikipedia	reddit	wikipedia
Italy	0.87	0.43	-0.41	-0.15	0.82	0.73
UK	0.95	0.99	-0.44	-0.47	0.82	0.85
US	1.07	0.87	-0.48	-0.44	0.88	0.82
Canada	1.12	1.06	-0.40	-0.45	0.90	0.82

The results presented so far 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 online users, has been observed during the 2009 H1N1 influenza pandemic [53, 83], the 2016 Zika outbreak [84], the seasonal flu [85] and during more localized public health emergency such as the 2013 measles outbreak in Netherlands [86]. Our findings confirm the central role of media, showing how media exposure is capable of shaping and driving collective attention during a national and global health emergency. Media exposure is an important factor that can influence individual risk perception as well [87–89]. 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 spreading are containment measures based on individuals' behavior. For this reason, in the next section we characterize media coverage and online users' response more specifically in terms of content produced and consumed.

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 online users may not be as synchronized [90, 91]. To shed light on this issue, we adopt an unsupervised topic modeling approach to extract prevalent topics in the news articles mentioned and discussed on Reddit. Often, indeed, users on Reddit post a submission containing a news article, and discussion unfolds in comments under such submission.

Differently from the previous section and to provide a comprehensive overview of the topics discussed, here we do not take into account any geographical context. Nonetheless, in the Supplementary Information we provide some insights also on the specific topics discussed by users in different countries.

We characterize the main topics discussed on Reddit by considering all submissions that include a news article in English. We then apply a topic modelling approach on the content of this news article set. Specifically, we extract topics by means of non-Negative Matrix Factorization (NMF) [60], a popular method for this kind of tasks. In this way, we extract the $n = 64$ most relevant topics in the news shared on Reddit. As a second step, we apply the model trained on the Reddit news to the set of articles published by mainstream media. That is, we characterize the news published by media in terms of the topics discussed on Reddit. This choice allows us to directly compare the topics covered by media with the public discussion around such news exposure. A complete list of the 64 topics extracted with the most frequent words is provided in the Supplementary Information. We consider the number of articles published on a certain topic as a proxy of general interest of traditional media towards it, while we measure the collective interest of Reddit users by the number of comments under the news articles on a specific topic.

Figure 4 shows an overview of the topics extracted and a comparison of the interest of media and Reddit users. We find a diverse and heterogeneous set of topics. Among others, we recognize topics about the global spreading of the virus (Outbreaks, WHO, CDC), 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 also the possible interventions to mitigate the spreading of the virus (Face Masks, Social Distancing, Tests, Vaccine).
Overall, the attention of traditional media and Reddit users towards different topics are in good accordance. Indeed, in Figure 4 we represent the difference between interest share towards different topics in media and Reddit submissions. That is, we compute the percentage share of attention dedicated by news outlets and Reddit users to each topic, and we subtract these two quantities. We observe a maximum absolute mismatch in interest share of 2.61%. Nonetheless, we observe that Reddit users are slightly more interested to topics regarding health (Symptoms, Medical Treatment), non-pharmaceutical interventions and personal protective equipment (Social Distancing, Face Masks), studies and information on the epidemic (Research, Surveys, Santa Clara Study, CDC), and also to 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 the Santa Clara County was infected respect to what originally thought [92]. Since the study suggests a lower mortality rate, the preprint has been quickly leveraged to support protest against lockdowns [111], while substantial flaws have been detected in the scientific methodology of the paper [112].

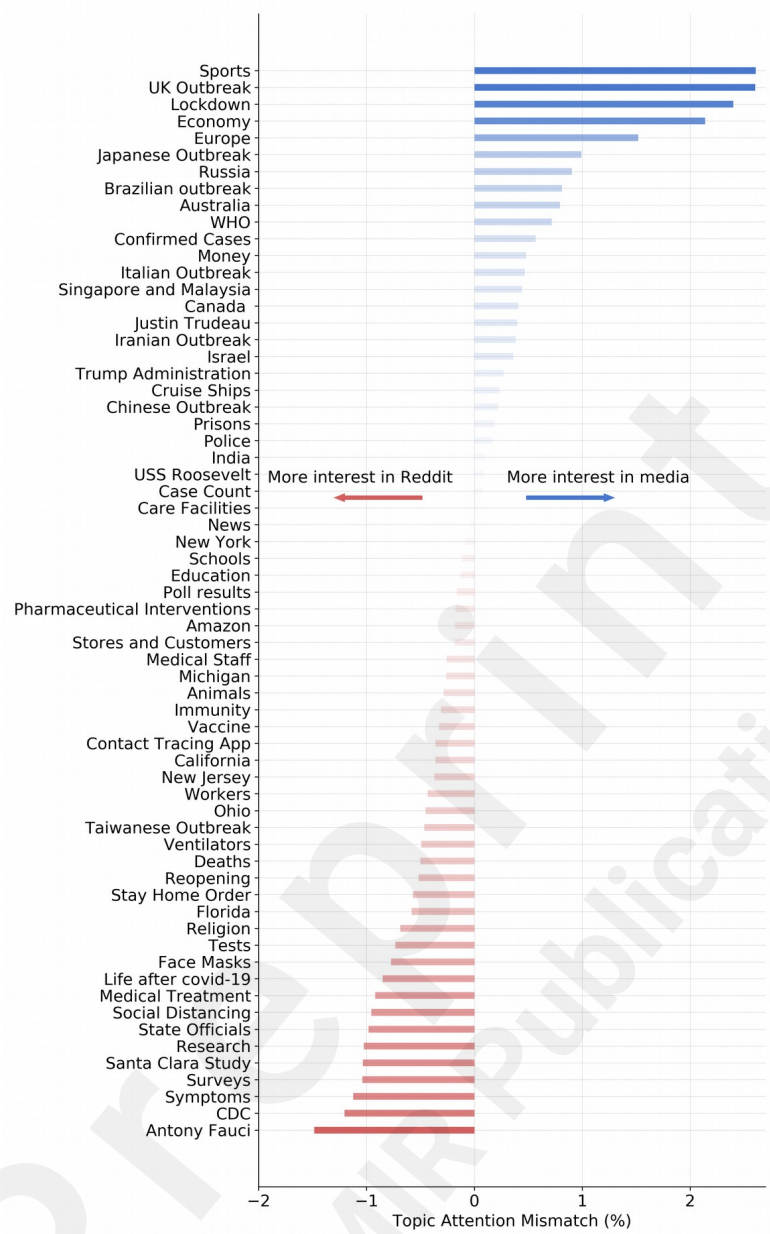


Figure 4. Difference in interest percentage share of different topics by traditional media and Reddit users. For example, +2% on the x-axis indicates that traditional media dedicates proportionally 2% more attention to that specific topic with respect to Reddit users.

The topics overview presented so far does not take into account any temporal dynamics of interest. However, topics showing a similar overall statistic may present a mismatch in temporal patterns. Hence, in the following, we take into account the temporal evolution of interest towards different topics. In Figure 5 we represent each topic as a single point: its x-coordinate (y-coordinate) indicates the $t_{1/2}$ when such topic reached 50% of its total relevance R in news outlets (on Reddit) during the analysis interval. Therefore, topics at the bottom left became relevant very early in the public discussion. Among these, we recognize themes centred on early COVID-19 outbreaks (i.e., Chinese, Japanese, Iranian and Italian outbreaks), the events related to cruise ships, specific countries (i.e., Israel, Singapore and Malaysia), and also topics regarding (early) health issues such as Symptoms, Confirmed Cases and the CDC. On the contrary, topics in the top right became relevant toward the end of the analysis interval (early May). Reasonably, we find here topics about the resumption of

activities after lockdown (i.e. Reopening), the feasibility and timing of a possible vaccine against Sars-Cov-2 (i.e. Vaccine), and discussions regarding acquired immunity and antibodies tests (i.e. Immunity). In-between, we find all other topics clustered around end of March and mid-April 2020, the period when the general discussion surrounding COVID-19 pandemic aroused sharply, as also shown in Figure 1.

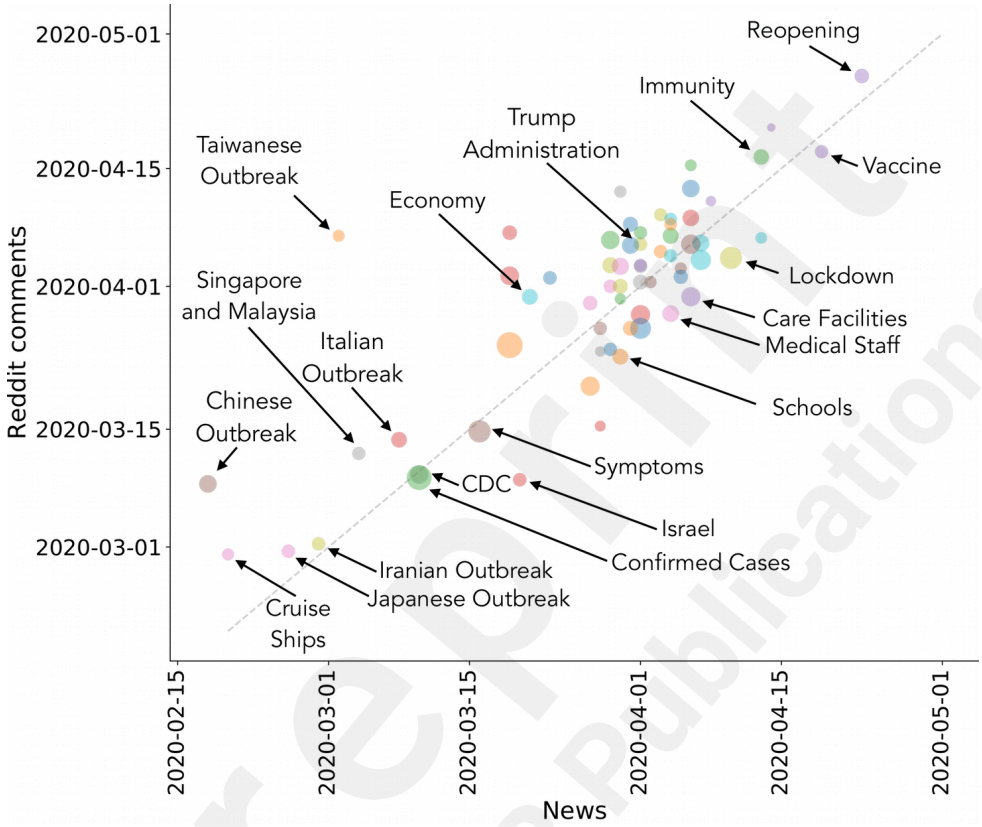


Figure 5. Scatter plot with the 64 topics extracted via NMF. X-axis (Y-axis) coordinate indicates when the topic achieved 50% of its relevance in news outlets (Reddit) during our analysis interval.

Note that the diagonal (plotted as a dashed line) in Figure 5 separates topics according to their temporal evolution. Above (below) the diagonal, we find topics whose interest on Reddit grows slowly (quickly) 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. While 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, interesting deviations from the diagonal are observable. For example, above the diagonal one can find mainly topics related to various outbreaks, economics and politics, for which the interests 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 topics discussed on Reddit is limited, since we only consider topics from news articles shared in submissions and do not explicitly take into account content expressed in comments. This ensures a proper comparison with topics extracted from news published and explains the absence of points in the bottom right corner of Figure 5.

Discussion

Principal Results

In this work, we characterized the response of online 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 pages views and comments of geolocalized Reddit users. We showed that collective attention was mainly driven by media coverage rather than epidemic progression, rapidly saturated, and decreased despite media coverage and COVID-19 incidence remaining high. This trend is very similar to that observed during other outbreaks [53, 83–86]. Also, we showed how media coverage sharply shifted to the domestic situation as soon as the first death was confirmed in the home country, discussing the implications for re-shaping individuals perception of risk [9, 80]. 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 topics extracted, we show that, while their temporal patterns are generally synchronized, the public attention for topics related to politics and economics is mainly triggered by media exposure, while the interests for topics more related to daily life accelerates on Reddit with respect to media coverage.

Limitations

Of course, our research comes with limitations. First, we characterized the exposure of individuals to COVID-19 pandemic by considering only news articles and YouTube videos published online by major news outlets. However, individuals are also exposed to relevant information through other channels, with television on top of these [93]. Second, a 2013 Pew Internet Study found that Reddit users are more likely young males [94], showing that around 15% of male internet users aged between 18 and 29 declare to use Reddit, compared to the 5% of women in the same age range and to the 8% of men aged between 30 and 49. Similarly, informal surveys proposed to users showed that most of respondents were males in their “late teens to mid-20s”, and that female users were “very much in the minority” [95]. Furthermore, Reddit is much more popular among urban and suburban residents rather than individuals living in rural areas [94]. Besides socio-demographic biases, other works suggested also that Reddit has become more and more a self-referential community, reinforcing the tendency to focus on its own contents rather than external sources [96]. Thus, perceptions, interests, and behaviors of Reddit users may differ from those of the general population. A similar argument may be raised for Wikipedia searches. Indeed, the usage of Internet, especially for information seeking purposes, can vary across people with different socio-demographic backgrounds [97–100].

Finally, our view on online users’ reaction is partial. Indeed, we do not consider other popular digital data sources such as, for example, Twitter. The reason behind this choice is twofold. First, many studies already characterized public response during the current and past health emergencies through the lens of Twitter [25, 50, 70, 72, 83, 84, 101–103]. Second, several studies have reported high prevalence of bots as drivers of low-quality information and discussions on COVID-19 on this platform [24, 25, 104–106]. Thus, careful and challenging extra steps would be necessary to isolate, identify, and distinguish organic discussions/reactions possibly originated from traditional media from those sparked by social

bots. We leave this for future work.

Conclusions

In conclusion, 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, since people are highly reactive to the news they are exposed to, in the beginning of an outbreak, the quality and type of information provided might have critical effects on risk perception, behaviors, and ultimately on the unfolding of the disease. On the other hand, however, we found that collective online attention saturates and declines rapidly despite media exposure and disease circulation remaining high. Attention saturation has the potential to affect collective awareness, perceived risk and ultimately propensity towards virtuous individual behavioral changes aimed at mitigating the spreading. Furthermore, especially in case of unknown viruses, attention saturation might 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 works are 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 to reinforce specific communication channels, such as social media platforms, in order to compensate the (natural) phenomenon of attention saturation. Indeed, these channels have the potential to create a more durable engagement with people, through a continuous loop of direct interactions. Currently, we see public health authorities issuing regularly declarations on social media. However, the CDC didn't even have a Twitter account in 2009 during H1N1 pandemic (the account was created in May 2010). While this is just an example, it underlines how we are relatively new to communicating such global health emergencies through social medias. Therefore, there is great need to further reinforce and engage people through these channels. Alongside, public health authorities should consider to strengthen additional communication channels. An example can be represented by participatory surveillance platforms all over the world such as Influenzanet, Flu Near You and FluTracking [107–109], which have the potential of delivering in-depth targeted information to individuals during public health emergencies, to promote the exchange of information between people and public health authorities, with the potential to enhance the level of engagement in the community [110].

Acknowledgements

Authors would like to thank the startup Quick Algorithm for providing the platform https://covid19.scops.ai/scops/home/, where the data collected during COVID-19 pandemic were visualized in real-time. D.P. and M.T. acknowledge support from the Lagrange Project of the Institute for Scientific Interchange Foundation (ISI Foundation) funded by Fondazione Cassa di Risparmio di Torino (Fondazione CRT). M.T. acknowledges support from EPIPOSE - "Epidemic intelligence to minimize COVID-19's public health, societal and economical impact" H2020-SC1-PHE-CORONAVIRUS-2020 call. M.S/ and A.P. acknowledge support from the Research Project "Casa Nel Parco" (POR FESR 14/20 - CANP - Cod. 320 - 16 - Piattaforma Tecnologica "Salute e Benessere") funded by Regione Piemonte in the context of the Regional Platform on Health and Wellbeing.
A.P. acknowledges partial support from Intesa Sanpaolo Innovation Center. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

N.G. acknowledges support from the Doctoral Training Alliance.

Authors Contribution Statement

N.G, M.S., D.P., A.P. and N.P. conceptualized the study. N.G., N.P., A.P. and M.T. collected the data. N.G., A.P. and F.C. performed analyses. N.G., M.S. and N.P. wrote the initial draft of the manuscript. N.G. and A.P. provided visualization. All authors (N.G., N.P., D.P., M.S., A.P., M.T., F.C.) discussed the research design, reviewed, edited, and approved the manuscript.

Conflicts of Interest

Authors declare no conflicts of interests.

References

1. Barry, J. M. Pandemics: avoiding the mistakes of 1918. Nature 459, 324–325 (2009).
2. Funk, S., Salathe´, M. & Jansen, V. A. A. Modelling the influence of human behaviour on the spread of infectious diseases: a review. Journal of The Royal Society Interface 7, 1247–1256, DOI: 10.1098/rsif.2010.0142 (2010). https://royalsocietypublishing.org/doi/pdf/10.1098/rsif.2010.0142.
3. Verelst, F., Willem, L. & Beutels, P. Behavioural change models for infectious disease transmission: a systematic review (2010-2015). Journal of The Royal Society Interface 13, DOI: 10.1098/rsif.2016.0820 (2016). https://royalsocietypublishing.org/doi/pdf/10.1098/rsif.2016.0820.
4. Rosenstock, I. M., Strecher, V. J. & Becker, M. H. Social learning theory and the health belief model. Health Education Quarterly 15, 175–183, DOI: 10.1177/109019818801500203 (1988). PMID: 3378902, https://doi.org/10.1177/109019818801500203.
5. Wu, J. T. et al. Estimating clinical severity of covid-19 from the transmission dynamics in Wuhan, China. Nature Medicine 26, 506–510 (2020).
6. Covid19 timeline. https://www.who.int/news-room/detail/27-04-2020-who-timeline---covid-19. Accessed: 2020-05-11.
7. Who covid19 official situation reports. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports/. Accessed: 2020-05-11.
8. Mayor of London announcement March 11th. https://www.facebook.com/sadiqforlondon/posts/3025766374142796. Accessed: 2020-05- 11.
9. Raude, J. et al. Are people excessively pessimistic about the risk of coronavirus infection? (2020).
10. Kraemer, M. U. et al. The effect of human mobility and control measures on the covid-19 epidemic in china. Science 368, 493–497 (2020).
11. Maier, B. F. & Brockmann, D. Effective containment explains sub-exponential growth in recent confirmed covid-19 cases in china. Science 368, 742–746, DOI: 10.1126/science.abb4557 (2020). https://science.sciencemag.org/content/368/6492/742.full.pdf.
12. Anderson, R. M., Heesterbeek, H., Klinkenberg, D. & Hollingsworth, T. D. How will country-based mitigation measures influence the course of the covid-19 epidemic? The Lancet 395, 931–934 (2020).
13. Bedford, J. et al. Covid-19: towards controlling of a pandemic. The Lancet 395, 1015–1018 (2020).
14. Colbourn, T. Covid-19: extending or relaxing distancing control measures. The Lancet Public Health (2020).

15. Fox, S. & Duggan, M. Health online 2013. Health 2013, 1–55 (2013).
16. Lee, S. T. Predictors of h1n1 influenza pandemic news coverage: Explicating the relationships between framing and news release selection. International Journal of Strategic Communication 8, 294–310 (2014).
17. McCauley, M., Minsky, S. & Viswanath, K. The H1N1 pandemic: media frames, stigmatization and coping. BMC Public Health 13, 1116 (2013).
18. Lin, C. A. & Lagoe, C. Effects of news media and interpersonal interactions on h1n1 risk perception and vaccination intent. Communication Research Reports 30, 127–136 (2013).
19. Lee, S. T. & Basnyat, I. From press release to news: mapping the framing of the 2009 h1n1 a influenza pandemic. Health Communication 28, 119–132 (2013).
20. Jung Oh, H. et al. Attention cycles and the h1n1 pandemic: A cross-national study of us and Korean newspaper coverage. Asian Journal of Communication 22, 214–232 (2012).
21. Keramarou, M. et al. Two waves of pandemic influenza a (H1N1) 2009 in Wales–the possible impact of media coverage on consultation rates, April–December 2009. Eurosurveillance 16, 19772 (2011).
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 computational biology 16, e1007633 (2020).
23. Bento, A. I. et al. Evidence from internet search data shows information-seeking responses to news of local covid-19 cases. Proceedings of the National Academy of Sciences (2020).
24. Gallotti, R., Valle, F., Castaldo, N., Sacco, P. & De Domenico, M. Assessing the risks of “infodemics” in response to covid-19 epidemics. arXiv preprint arXiv:2004.03997 (2020).
25. Singh, L. et al. A first look at covid-19 information and misinformation sharing on twitter. arXiv preprint arXiv:2003.13907 (2020).
26. Cinelli, M. et al. The covid-19 social media infodemic. arXiv preprint arXiv:2003.05004 (2020).
27. Lazer, D., Kennedy, R., King, G. & Vespignani, A. The parable of google flu: traps in big data analysis. Science 343, 1203–1205 (2014).
28. Culotta, A. Towards detecting influenza epidemics by analyzing twitter messages. In Proceedings of the first workshop on social media analytics, 115–122 (2010).
29. Lampos, V. & Cristianini, N. Tracking the flu pandemic by monitoring the social web. In 2010 2nd international workshop on cognitive information processing, 411–416 (IEEE, 2010).
30. Zhang, Q. et al. Forecasting seasonal influenza fusing digital indicators and a mechanistic disease model. In Proceedings of the 26th international conference on world wide web, 311–319 (2017).
31. De Choudhury, M., Gamon, M., Counts, S. & Horvitz, E. Predicting depression via social media. In Seventh international AAAI conference on weblogs and social media (2013).
32. De Choudhury, M., Counts, S. & Horvitz, E. Social media as a measurement tool of depression in populations. In Proceedings of the 5th Annual ACM Web Science Conference, 47–56 (2013).
33. Broniatowski, D. A., Paul, M. J. & Dredze, M. National and local influenza surveillance through twitter: an analysis of the 2012-2013 influenza epidemic. PloS one 8 (2013).
34. Araujo, M., Mejova, Y., Weber, I. & Benevenuto, F. Using Facebook ads audiences for global lifestyle disease surveillance: Promises and limitations. In Proceedings of the 2017 ACM on Web Science Conference, 253–257 (2017).
35. Park, A. & Conway, M. Tracking health related discussions on reddit for public health applications. In AMIA Annual Symposium Proceedings, vol. 2017, 1362 (American Medical Informatics Association, 2017).
36. Kumar, M., Dredze, M., Coppersmith, G. & De Choudhury, M. Detecting changes in suicide content manifested in social media following celebrity suicides. In Proceedings of the 26th

ACM conference on Hypertext & Social Media, 85–94 (2015).
37. Generous, N., Fairchild, G., Deshpande, A., Del Valle, S. Y. & Priedhorsky, R. Global disease monitoring and forecasting with Wikipedia. PLoS computational biology 10 (2014).
38. Hickmann, K. S. et al. Forecasting the 2013–2014 influenza season using Wikipedia. PLoS computational biology 11 (2015).
39. Ginsberg, J. et al. Detecting influenza epidemics using search engine query data. Nature 457, 1012–1014 (2009).
40. Dugas, A. F. et al. Influenza forecasting with google flu trends. PloS one 8 (2013).
41. Eysenbach, G. Infodemiology and infoveillance tracking online health information and cyber-behavior for public health. American journal of preventive medicine 40, S154–8, DOI: 10.1016/j.amepre.2011.02.006 (2011).
42. Milinovich, G. J., Williams, G. M., Clements, A. C. A. & Hu, W. Internet-based surveillance systems for monitoring emerging infectious diseases. The Lancet Infectious Diseases 14, 160 – 168, DOI: https://doi.org/10.1016/S1473-3099(13)70244-5 (2014).
43. Park, H. W., Park, S. & Chong, M. Conversations and medical news frames on twitter: Infodemiological study on covid-19 in South Korea. J Med Internet Res 22, e18897, DOI: 10.2196/18897 (2020).
44. Park, A. & Conway, M. Tracking health related discussions on reddit for public health applications. AMIA ... Annual Symposium proceedings. AMIA Symposium 2017, 1362–1371 (2018).
45. Lamb, A., Paul, M. & Dredze, M. Investigating twitter as a source for studying behavioral responses to epidemics. AAAI Fall Symposium - Technical Report (2012).
46. Lewis, N. Information Seeking and Scanning, 1–10 (American Cancer Society, 2017).
47. Lambert, S. D. & Loiselle, C. G. Health information—seeking behavior. Qualitative Health Research 17, 1006–1019, DOI: 10.1177/ 1049732307305199 (2007). PMID: 17928475, https://doi.org/10.1177/1049732307305199.
48. Walter, D., Bohmer, M. M., Reiter, S., Krause, G. & Wichmann, O. Risk perception and information-seeking behaviour during the 2009/10 influenza a(h1n1) pandemic in Germany. Eurosurveillance 17, DOI: https://doi.org/10.2807/ese.17.13.20131-en (2012).
49. Pang, N. L.-S. Crisis-based information seeking: monitoring versus blunting in the information seeking behaviour of working students during the Southeast Asian haze crisis. Inf. Res. 19 (2014).
50. Johnson, B. B. Explaining Americans' responses to dread epidemics: an illustration with Ebola in late 2014. Journal of Risk Research 20, 1338–1357, DOI: 10.1080/13669877.2016.1153507 (2017). https://doi.org/10.1080/13669877.2016.1153507.
51. Sell, T. et al. Media messages and perception of risk for Ebola virus infection, united states. Emerging Infectious Diseases 23, DOI: 10.3201/eid2301.160589 (2017).
52. Gozzi, N., Perrotta, D., Paolotti, D. & Perra, N. Towards a data-driven characterization of behavioral changes induced by the seasonal flu. PLOS Computational Biology 16, e1007879 (2020).
53. 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, DOI: 10.31234/ osf.io/dz428 (2020).
54. Akaike, H. Information theory and an extension of the maximum likelihood principle. In Selected papers of Hirotugu Akaike, 199–213 (Springer, 1998).
55. Chen, Y. & Skiena, S. False-friend detection and entity matching via unsupervised transliteration. arXiv preprint arXiv:1611.06722 (2016).
56. Python geocoder. https://geocoder.readthedocs.io. Accessed: 2020-05-29.
57. Tausczik, Y., Faasse, K., Pennebaker, J. W. & Petrie, K. J. Public anxiety and information seeking following the h1n1 outbreak: Blogs, newspaper articles, and Wikipedia visits. Health

Communication 27, 179–185, DOI: 10.1080/10410236.2011.571759 (2012). PMID: 21827326, https://doi.org/10.1080/10410236.2011.571759.
58. Chew, C. & Eysenbach, G. Pandemics in the age of twitter: Content analysis of tweets during the 2009 h1n1 outbreak. PLOS ONE 5, 1–13, DOI: 10.1371/journal.pone.0014118 (2010).
59. Pruss, D. et al. Zika discourse in the Americas: A multilingual topic analysis of twitter. PLOS ONE 14, 1–23, DOI: 10.1371/journal.pone. 0216922 (2019).
60. Smith, M. C. & Broniatowski, D. A. Towards real-time measurement of public epidemic awareness monitoring influenza awareness through twitter (2015).
61. Mollema, L. 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 17, e128, DOI: 10.2196/jmir.3863 (2015).
62. Wahlberg, A. A. F. & Sjoberg, L. Risk perception and the media. Journal of Risk Research 3, 31–50, DOI: 10.1080/136698700376699 (2000). https://doi.org/10.1080/136698700376699.
63. Klemm, C., Das, E. & Hartmann, T. Swine flu and hype: A systematic review of media dramatization of the h1n1 influenza pandemic. Journal of Risk Research 19, 1–20, DOI: 10.1080/13669877.2014.923029 (2014).
64. Tchuenche, J., Dube, N., Bhunu, C., Smith, R. & Bauch, C. The impact of media coverage on the transmission dynamics of human influenza. BMC public health 11 Suppl 1, S5, DOI: 10.1186/1471-2458-11-S1-S5 (2011).
65. Zhao, W. X. et al. Comparing twitter and traditional media using topic models. In Clough, P. et al. (eds.) Advances in Information Retrieval, 338–349 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2011).
66. Diao, Q., Jiang, J., Zhu, F. & Lim, E.-P. Finding bursty topics from microblogs. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1, ACL '12, 536–544 (Association for Computational Linguistics, USA, 2012).
67. Lee, D. D. & Seung, H. S. Learning the parts of objects by non-negative matrix factorization. Nature 401, 788 (1999).
68. Bendavid, E. et al. Covid-19 antibody seroprevalence in Santa Clara county, California. medRxiv DOI: 10.1101/2020.04.14.20062463 (2020).
69. Wang, W. & Ahern, L. Acting on surprise: emotional response, multiple-channel information seeking and vaccination in the h1n1 flu epidemic. Social Influence 10, 137–148, DOI: 10.1080/15534510.2015.1011227 (2015). https://doi.org/10.1080/15534510.2015.1011227.
70. Duggan, M. & Smith, A. C. 6% of online adults are reddit users (2013).
71. Finlay, C. Age and gender in reddit commenting and success. Journal of Information Science Theory and Practice 2, 18–28, DOI: 10.1633/JISTaP.2014.2.3.2 (2014).
72. 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 Proceedings of the 23rd International Conference on World Wide Web, WWW '14 Companion, 517–522, DOI: 10.1145/2567948.2576943 (Association for Computing Machinery, New York, NY, USA, 2014).
73. van Deursen, A. J. & van Dijk, J. A. The digital divide shifts to differences in usage. New Media & Society 16, 507–526, DOI: 10.1177/1461444813487959 (2014). https://doi.org/10.1177/1461444813487959.
74. van Deursen, A. & van Dijk, J. Internet skills and the digital divide. New Media & Society 13, 893–911, DOI: 10.1177/1461444810386774 (2011). https://doi.org/10.1177/1461444810386774.
75. Robinson, L. et al. Digital inequalities and why they matter. Information, Communication & Society 18, 569–582, DOI: 10.1080/ 1369118X.2015.1012532 (2015). https://doi.org/10.1080/1369118X.2015.1012532.

76. van Deursen, A. J., van Dijk, J. A. & Peters, O. Rethinking internet skills: The contribution of gender, age, education, internet experience, and hours online to medium- and content-related internet skills. Poetics 39, 125 – 144, DOI: https://doi.org/10.1016/j.poetic. 2011.02.001 (2011).
77. Park, A. & Conway, M. Towards tracking opium related discussions in social media. Online journal of public health informatics 9 (2017).
78. Guidry, J. P., Jin, Y., Orr, C. A., Messner, M. & Meganck, S. Ebola on Instagram and twitter: How health organizations address the health crisis in their social media engagement. Public relations review 43, 477–486 (2017).
79. Martinez-Rojas, M., del Carmen Pardo-Ferreira, M. & Rubio-Romero, J. C. Twitter as a tool for the management and analysis of emergency situations: A systematic literature review. International Journal of Information Management 43, 196–208 (2018).
80. Park, H. W., Park, S. & Chong, M. Conversations and medical news frames on Twitter: Infodemiological study on covid-19 in South Korea. Journal of Medical Internet Research 22, e18897 (2020).
81. Ferrara, E. # covid-19 on twitter: Bots, conspiracies, and social media activism. arXiv preprint arXiv:2004.09531 (2020).
82. Yang, K.-C., Torres-Lugo, C. & Menczer, F. Prevalence of low-credibility information on twitter during the covid-19 outbreak. arXiv preprint arXiv:2004.14484 (2020).
83. Ahmed, W., Vidal-Alaball, J., Downing, J. & Seguı, F. L. Covid-19 and the 5g conspiracy theory: Social network analysis of twitter data. Journal of Medical Internet Research 22, e19458 (2020).
84. Guerrisi, C. et al. Participatory Syndromic Surveillance of Influenza in Europe. The Journal of Infectious Diseases 214, S386–S392, DOI: 10.1093/infdis/jiw280 (2016). https://academic.oup.com/jid/article-pdf/214/suppl 4/S386/7717566/jiw280.pdf.
85. Smolinski, M. et al. Flu near you: Crowdsourced symptom reporting spanning 2 influenza seasons. American Journal of Public Health 105, 2124–2130, DOI: 10.2105/AJPH.2015.302696 (2015).
86. Dalton, C. et al. Flutracking: A weekly Australian community online survey of influenza-like illness in 2006, 2007 and 2008. Communicable diseases intelligence 33, 316–22 (2009).
87. Wojcik, O. P., Brownstein, J. S., Chunara, R. & Johansson, M. A. Public health for the people: participatory infectious disease surveillance in the digital age. Emerging Themes in Epidemiology 11 (2014).
88. News API. https://newsapi.org. Accessed: 2020-05-11.
89. YouTube API. https://developers.google.com/youtube/v3. Accessed: 2020-05-11.
90. Tan, C. & Lee, L. All who wander: On the prevalence and characteristics of multi-community engagement. In Proceedings of the 24th International Conference on World Wide Web, 1056–1066 (International World Wide Web Conferences Steering Committee, 2015).
91. Hessel, J., Tan, C. & Lee, L. Science, askscience, and badscience: On the coexistence of highly related communities. In Proc. of the 10th Intl AAAI Conf. on Web and Social Media, ICWSM 2016, 171–180 (The AAAI Press, 2016).
92. Barthel, M. How the 2016 presidential campaign is being discussed on reddit (2016).
93. Horne, B. D. & Adali, S. The impact of crowds on news engagement: A reddit case study. ArXiv abs/1703.10570 (2017).
94. Saleem, H. M., Dillon, K. P., Benesch, S. & Ruths, D. A web of hate: Tackling hateful speech in online social spaces. arXiv preprint arXiv:1709.10159 (2017).
95. Rakib, T. B. A. & Soon, L.-K. Using the reddit corpus for cyberbully detection. In Asian Conference on Intelligent Information and Database Systems, 180–189 (Springer, 2018).
96. Choudhury, M. D. & De, S. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In Proceedings of the Eighth Intl´ Conf. on Weblogs and Social Media,

ICWSM 2014 (The AAAI Press, 2014).
97. Balsamo, D., Bajardi, P. & Panisson, A. Firsthand opiates abuse on social media: monitoring geospatial patterns of interest through a digital cohort. In The World Wide Web Conference, 2572–2579 (2019).
98. Radford, A. et al. Language models are unsupervised multitask learners. OpenAI Blog 1, 9 (2019).
99. Laurent, M. R. & Vickers, T. J. Seeking Health Information Online: Does Wikipedia Matter? Journal of the American Medical Informatics Association 16, 471–479, DOI: 10.1197/jamia.M3059 (2009). https://academic.oup.com/jamia/article-pdf/16/4/471/2292889/16-4- 471.pdf.
100. McIver, D. J. & Brownstein, J. S. Wikipedia usage estimates prevalence of influenza-like illness in the united states in near real-time. PLOS Computational Biology 10, 1–8, DOI: 10.1371/journal.pcbi.1003581 (2014).
101. Generous, N., Fairchild, G., Deshpande, A., Del Valle, S. Y. & Priedhorsky, R. Global disease monitoring and forecasting with Wikipedia. PLOS Computational Biology 10, 1–16, DOI: 10.1371/journal.pcbi.1003892 (2014).
102. Wikimedia api. https://wikimedia.org/api/rest v1/. Accessed: 2020-05-11.
103. Blei, D. M., Ng, A. Y. & Jordan, M. I. Latent Dirichlet allocation. Journal of machine Learning research 3, 993–1022 (2003).
104. Blei, D. M. & Lafferty, J. D. Dynamic topic models. In Proceedings of the 23rd international conference on Machine learning, 113–120 (ACM, 2006).
105. Dou, W., Yu, L., Wang, X., Ma, Z. & Ribarsky, W. Hierarchical Topics: Visually exploring large text collections using topic hierarchies. IEEE Transactions on Visualization and Computer Graphics 19, 2002–2011, DOI: 10.1109/TVCG.2013.162 (2013).
106. Gobbo, B. et al. Topic tomographies (toptom): a visual approach to distill information from media streams. In Computer Graphics Forum, vol. 38, 609–621 (Wiley Online Library, 2019).
107. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. & Dean, J. Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems, 3111–3119 (2013).
108. Jones, K. S. A statistical interpretation of term specificity and its application in retrieval. Journal of documentation (1972).
109. Lin C. Projected gradient methods for nonnegative matrix factorization. Neural computation 19, 2756–2779 (2007).
110. Hoyer, P. Non-negative matrix factorization with sparseness constraints. The Journal of Machine Learning Research 5, 1457–1469 (2004).
111. https://www.nytimes.com/2020/05/14/opinion/coronavirus-research-misinformation.html
112. https://www.theatlantic.com/health/archive/2020/04/pandemic-confusing-uncertainty/610819

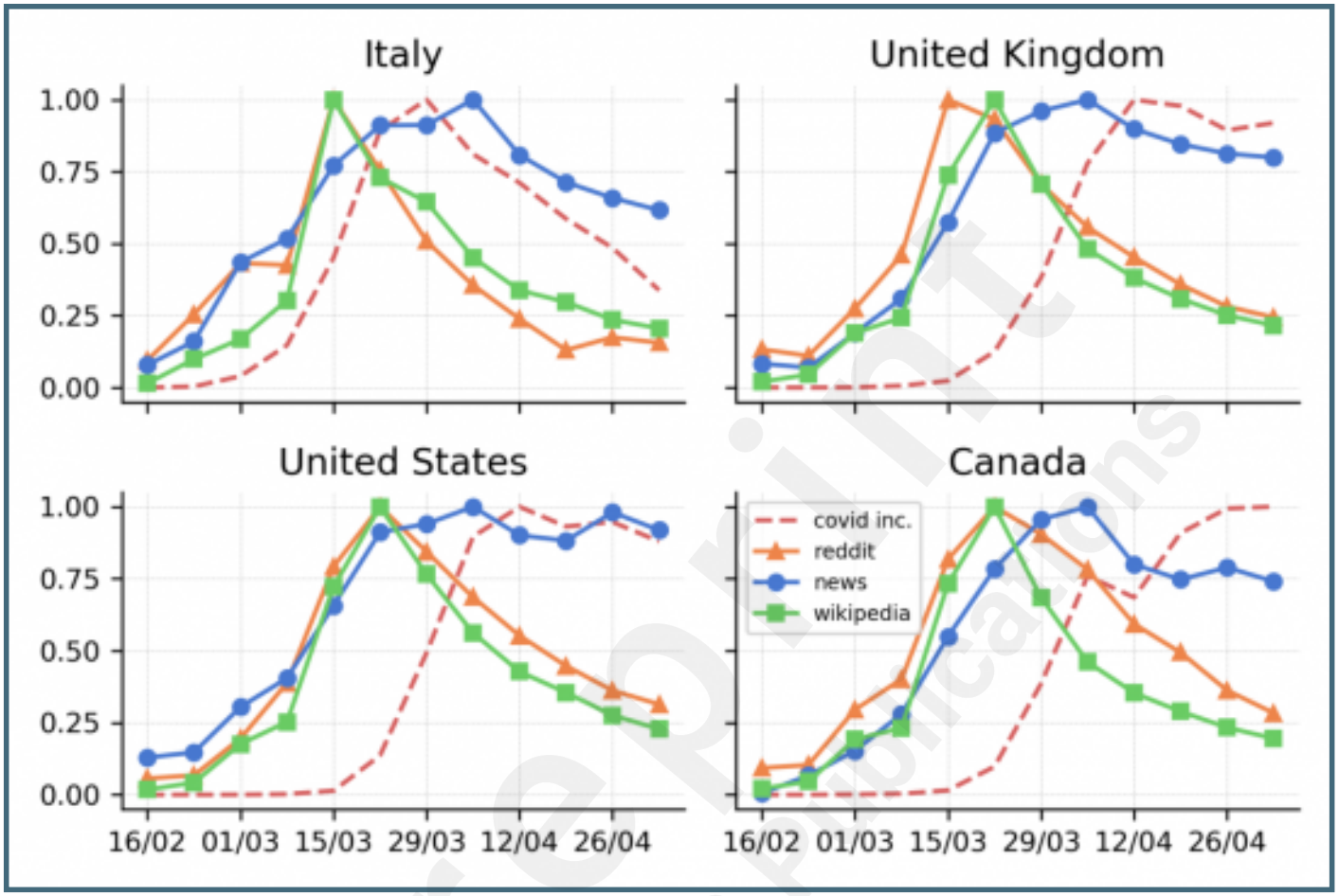
Supplementary Files

Preprint JMIR Publications

Figures

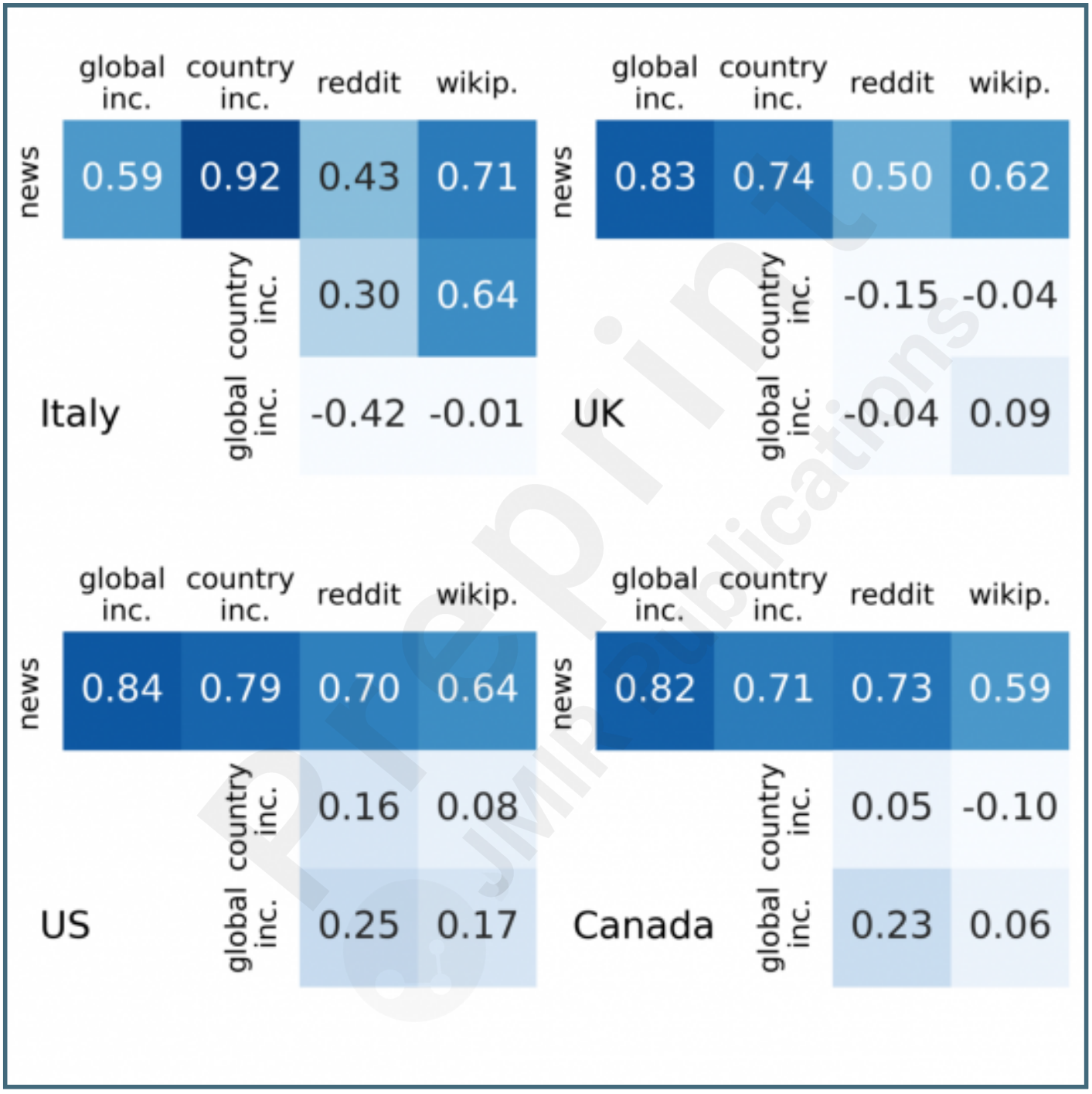
Preprint JMIR Publications

Normalized weekly volume of news articles and YouTube videos (news), Reddit comments (reddit), Wikipedia views (wikipedia) related to COVID-19 pandemic and COVID-19 incidence (covid inc.) in different countries.

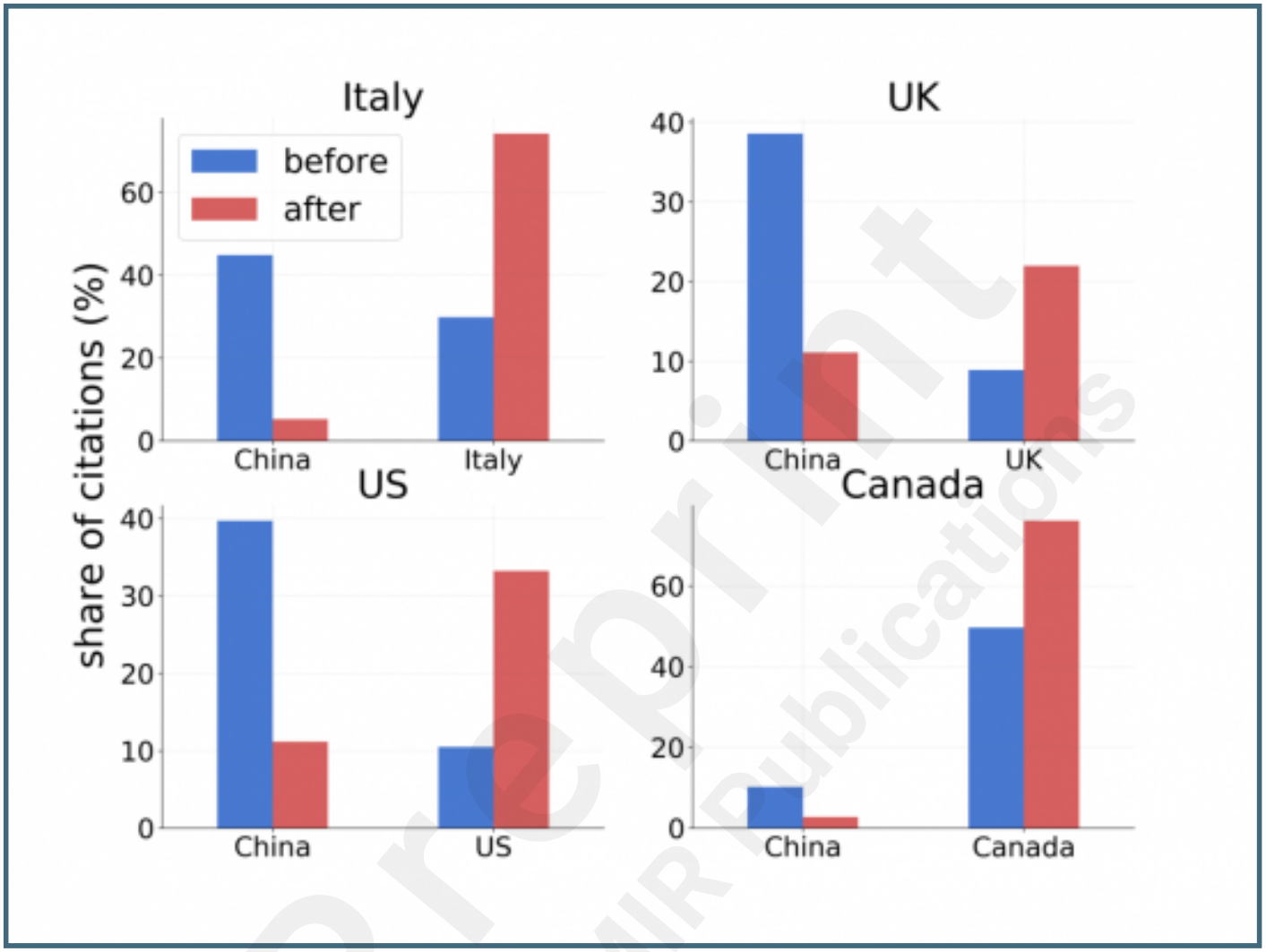


PrePrint JMIR Publications

Country specific Pearson correlation coefficients between 1) news coverage and global/domestic COVID-19 incidence, volumes of Reddit comments and Wikipedia views; 2) domestic COVID-19 incidence and volumes of Reddit comments and Wikipedia views; 3) global COVID-19 incidence and volumes of Reddit comments and Wikipedia views.

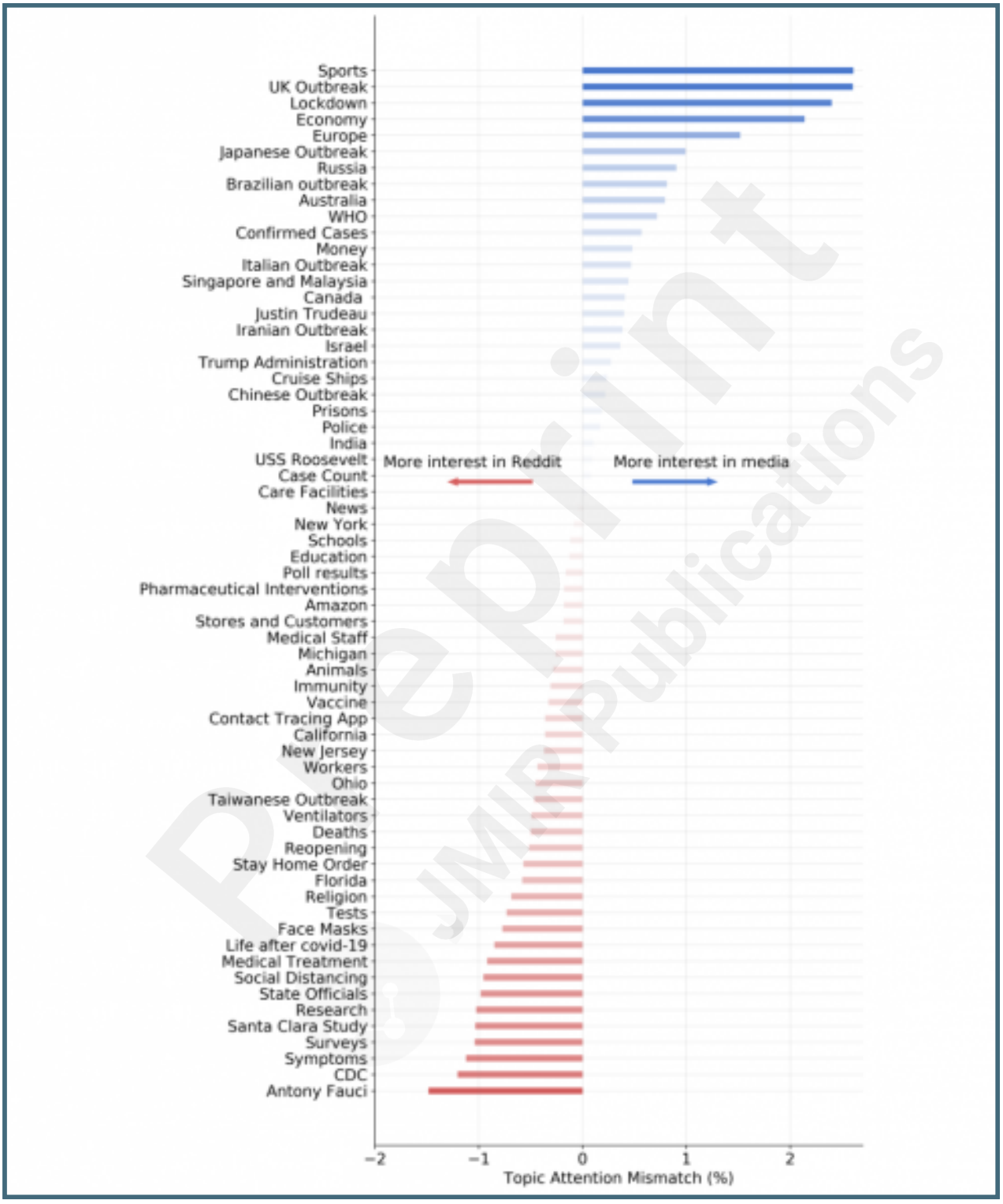


Share of citations of China versus home country locations by Italian/UK/US/Canadian news outlets before and after first COVID-19 death occurred in different countries considered. Geographic locations are extracted from text using [55, 56].



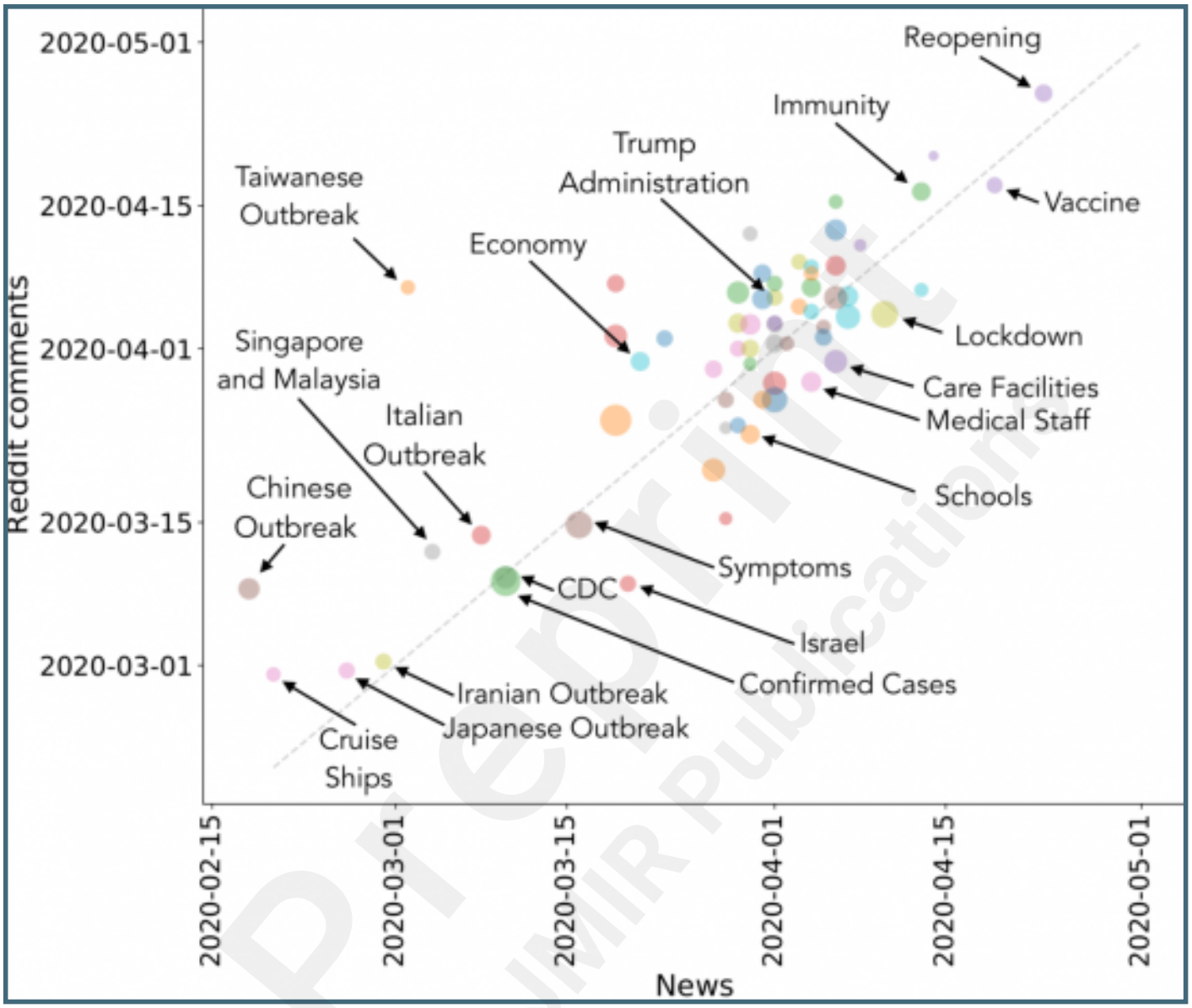
Preprint JMIR Publications

Difference in interest percentage share of different topics by traditional media and Reddit users. For example, +2% on the x-axis indicates that traditional media dedicates proportionally 2% more attention to that specific topic with respect to Reddit users.



Preprint JMIR Publications

Scatter plot with the 64 topics extracted via NMF. X-axis (Y-axis) coordinate indicates when the topic achieved 50% of its relevance in news outlets (Reddit) during our analysis interval.



Preprint JMIR Publications

Multimedia Appendixes

Preprint JMIR Publications

Supplementary Information.
URL: https://asset.jmir.pub/assets/e9888f33ff54f711dcb6f9c44273300e.docx

Preprint JMIR Publications