Journal of Computational Social Science
https://doi.org/10.1007/s42001-024-00257-8

RESEARCH ARTICLE

Long-term assessment of social amplification of risk during COVID-19: challenges to public health agencies amid misinformation and vaccine stance

Ali Unlu[1,2] · Sophie Truong[2] · Nitin Sawhney[2] · Jonas Sivelä[1] · Tuukka Tammi[1]

Received: 28 September 2023 / Accepted: 12 February 2024

Abstract

This study employs the Social Amplification of Risk Framework to investigate the stance on COVID-19 vaccines and the spread of misinformation on Twitter in Finland. Analyzing over 1.6 million tweets and manually annotating 4150 samples, the research highlights the challenges faced by the Finnish Institute for Health and Welfare (THL) in steering online vaccination communication. Using BERT models, Botometer, and additional computational methods, the study classifies text, identifies bot-like accounts, and detects malicious bots. Social network analysis further uncovers the underlying social structures and key actors in Twitter discussions during the pandemic. The THL remained a primary source of COVID-19 information throughout the pandemic, maintaining its influence despite challenges posed by malicious bots spreading misinformation and adopting negative vaccine stances. However, THL ceased its Twitter activity at the end of 2022 because its posts were being exploited to gain visibility and traction for misinformation and negative vaccine stance. The study also identifies key influencers in online vaccine discussions, suggesting avenues for improving public health communication. Overall, the research underscores the need to understand social media dynamics to counter misinformation and foster accurate public communication on COVID-19 and vaccination.

✉ Ali Unlu
ali.unlu@thl.fi
https://scholar.google.com/citations?user=2lCMVtYAAAAJ&hl=en

Nitin Sawhney
https://scholar.google.com/citations?user=ohJ44iMAAAAJ&hl=en&oi=ao

Jonas Sivelä
https://scholar.google.com/citations?hl=sv&user=IicWW6AAAAAJ

Tuukka Tammi
https://scholar.google.com/citations?user=1ellVdwAAAAJ&hl=en

[1] The Cultural, Behavioral and Media Insights Centre (CUBE), Finnish Institute for Health and Welfare (THL), Mannerheimintie 166, 00271 Helsinki, Finland

[2] Department of Computer Science, Aalto University, Espoo, Finland

Published online: 22 March 2024

Keywords Social amplification of risk framework (SARF) · Misinformation · Vaccine stance · COVID-19 · Twitter · Finland

Introduction

The COVID-19 pandemic has been a multifaceted global crisis, significantly impacting health and information dissemination across populations and countries. The health outcomes vary widely among individuals who have had COVID-19, with the majority recovering without long-term consequences. While some long-term effects present challenges to healthcare systems, they do not constitute a widespread chronic crisis, affecting only a portion of those who have contracted the virus [1–3]. The resultant spread of rumors and contradictions exacerbates this challenge [4]. Governments and public health authorities face significant hurdles, including promoting public health literacy, navigating uncertain institutional communication, and managing media coverage [5]. Furthermore, the pandemic has exerted extensive social repercussions, affecting areas like social vulnerability, education, social capital, relationships, mobility, and welfare [6].

Social media has played a crucial role during the pandemic, both positively and negatively. It has provided sustained connection and access to information [7]. However, it has also been characterized by an overburden of information, accurate and inaccurate, which has made it difficult for people to find trustworthy sources and reliable guidance [8, 9]. False reports about COVID-19 vaccines may undermine public confidence in vaccination [7].

Furthermore, social media bots have become a significant concern due to their potential to distort communication [10, 11]. These automated accounts are capable of posting or retweeting content without human intervention and can be used for various purposes such as spreading propaganda, influencing public opinion, or disrupting online discussions [12–14]. Bots can magnify the impact of trolls, who intentionally post offensive content to provoke emotional reactions or derail discussions. Trolls, along with malicious bots, can direct harassment, abuse, or ridicule towards public health agencies (PHA) and their supporters, harming their reputation and credibility [12].

As people contended with fear and panic due to the sudden outbreak of COVID-19, accompanied by narratives of widespread deaths and the shock of severe measures like border closures and lockdowns, these psychological factors impacted societal behaviors in terms of information access and consumption [15]. Furthermore, the dissemination of information related to COVID-19 and its vaccinations was shaped by various elements, such as political ideology, the anti-vax movement, the participation of malicious actors in online discourse, and the proliferation of misinformation [16–19].

In this context, *misinformation* refers to false or inaccurate information that is intentionally or unintentionally disseminated, while a *negative vaccine stance* is defined as the intentional avoidance of vaccination, regardless of the reason for the avoidance. This can include refusal, delay, or discontinuation of vaccination [20–22]. Both phenomena can undermine public trust and confidence in PHA, as well as influence individual and collective health behaviors and outcomes [23–25].

PHA and governments have faced challenges in implementing effective monitoring systems and communication strategies [26]. The need for massive public health literacy and collaboration between governments, health institutions, and the media has been emphasized [26]. Various countries have implemented different public health responses, including monitoring, public education, and the establishment of healthcare facilities [26]. During this period, PHA have been leveraging social media to mitigate public panic and enhance knowledge about pandemic prevention. The effectiveness of these efforts, however, hinges not only on the quality of the disseminated information but also on various factors, such as public trust in the government [27].

To communicate with the public throughout the COVID-19 pandemic, PHA have adopted various social media strategies. These include utilizing digital communication tools like Twitter (new X) to disseminate information and engage with the public [28]. Their focus encompasses public information, and addressing specific pandemic-related issues, such as specific vaccines and virus variants [29, 30].

The Finnish Institute for Health and Welfare (THL) is one of the main government agencies in Finland to provide information and guidance regarding COVID-19 [31]. The THL had used Twitter among other channels to reach a bigger audience through social media platforms [32]. The THL account became the target of a series of coordinated attacks, involving a mix of genuine users and automated bot accounts. These participants disseminated a high volume of misinformation and anti-vaccine sentiment. The primary objective of these attacks appears to have been the exploitation of THL's communication channel. This increased the visibility of these accounts, thereby extending their influence on a wider audience. Among other repercussions, this systematic exploitation led THL to suspend its Twitter account by the end of 2022 [33].

The withdrawal of a PHA from a major social media platform like Twitter raises critical questions about the efficacy of these platforms. The reasons behind this decision go beyond the user profile of Twitter and include the challenge of addressing these attacks with limited organizational resources.

The significance of this study lies in its detailed examination of the communication patterns between a PHA and its followers, particularly focusing on the evolving nature of this engagement during a crisis. It also scrutinizes the challenges faced by PHA in such scenarios, including how they navigate the risks and opportunities of maintaining an online presence. By analyzing these aspects, the study aims to provide valuable insights into optimizing public health communication strategies in the digital era, especially during times of crisis.

The paper evaluates the Twitter communication patterns during COVID-19, examining the role of THL, bots, malicious bots, and networks of digital communities that shared MINVS. This paper makes a noteworthy contribution by utilizing three consecutive years of data, offering a more comprehensive representation of the evolving patterns in COVID-19 discourse. This approach distinguishes itself from previous studies, which have relied on shorter-term assessments. To achieve this, the study uses a mixed-methods approach, integrating NLP text analysis of Twitter data with network analysis of Twitter activities of THL, bots, and other groups.

This study utilizes the Social Amplification of Risk Framework (SARF) to analyze the communication dynamics between the Twitter communities and THL. The growing divide between these two entities contributes to vaccine hesitancy [34, 35]

and mistrust to authorities [24, 36, 37]. The study further merges SARF with Social Network Analysis (SNA) to scrutinize communication and knowledge exchange patterns in Twitter networks involving PHA and other stakeholders, focusing on misinformation and vaccine stance. This method can guide public health communication strategies and foster effective responses to these emerging phenomena.

The present study addresses several research questions, which are central to understanding the dynamics of misinformation and negative vaccine stance. The inquiries are as follows: How did the THL lead online discussions about COVID-19 and how did the Twitter audience react? How has the sentiment of the Twitter audience toward THL's COVID-19 posts changed over time? How did bot accounts interact with THL account, and what was their post coverage and influence on risk perceptions? What are the key differences between networks that spread misinformation and those that express negative vaccine stance? What is the scope of these networks within the Finnish-speaking community on Twitter discussing COVID-19? This study aims to elucidate the intricate interplay between PHA, misinformation networks, and social media within the SARF.

Theoretical framework

The SARF serves as a conceptual framework in this research, focusing on the intricate dynamics surrounding the development of risk perceptions, particularly as they relate to risk events such as the COVID-19 pandemic that are disseminated through various social actors as social stations, including government agencies, media outlets, interest groups, and individuals, through multiple communication channels [38, 39]. With social media, these actors actively contribute to the transmission of risk signals rather than simply being passive recipients of them, changing and reshaping the narrative and story [40–42]. Within the SARF framework, risk amplification is delineated into two distinct phases: information transmission and response mechanism.

Information transmission within SARF concerns the dissemination and contouring of risk signals [38]. In this stage, two types of arguments were frequently observed: health promotions and counter-arguments against preventive measures (Fig. 1). The COVID-19 pandemic presents a significant health risk, necessitating responses such as vaccination, which PHA advocate as the most effective solution. However, MINVS introduces an additional risk, as it can amplify vaccine hesitancy. This hesitancy, fueled by MINVS, may lead individuals to make risky decisions, such as refusing vaccines or opting for ineffective treatments, with potential consequences for public health such as outbreaks of preventable diseases. Consequently, the Twitter community, guided by their communication preferences and trust in various sources, attempts to navigate these perceived risks.

In alignment with the SARF's conceptualization of information transmission, which encompasses elements such as volume, ambiguity, dramatization, and symbolic connotations [38]. We operationalized this phase by conducting a rigorous stance detection and misinformation identification analysis of user posts produced

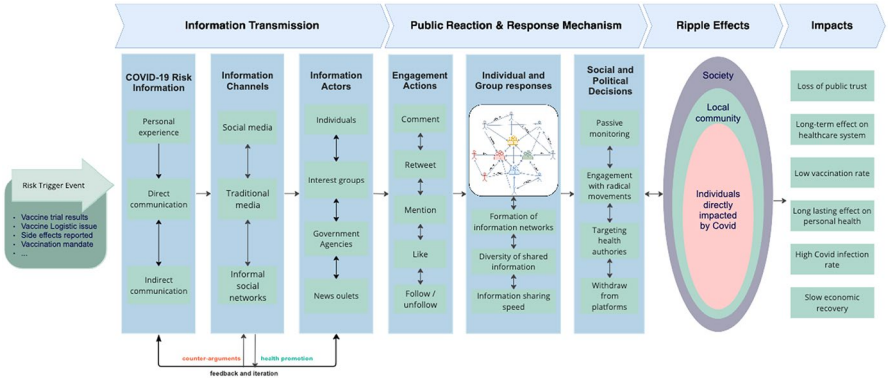

Fig. 1 SARF analysis of COVID-19 information flow

by a variety of actors on social media, as well as the volume of these posts. Specifically, these methods were employed to identify counterarguments against measures promoted by health authorities, including vaccination. These posts frequently exhibited ambiguities about the impact of such measures, dramatizations of the situation, and the repackaging of symbolic connotations of known conspiracy theories within the COVID-19 framework, all of which are indicative of the complex dynamics surrounding the development of risk perceptions as delineated by the SARF [40–42].

In addition, the rapid spread of risk information on social media is often driven by negative emotions, especially anger. This anger can lead to the phenomenon of replacing blame, in which people shift or redirect responsibility for a risk event, particularly towards authorities [42], such as unpreparedness, availability of facemasks, insufficient intensive care units, and strict lockdowns. This process can be amplified by social media, as it allows people to share their anger and frustration with a large audience.

As Zhang and Cozma [42] noted, blame and anger combined can create ripple effects, heightening public attention to the risk event. Within this dynamic, only a few sources emerge as information brokers, guided by public trust, which is often situational or contextual. When authorities are seen as failing to manage risk events effectively, the public's inclination to replace blame increases, escalating risk amplification and leading to secondary impacts beyond the primary risk event [38, 41, 43]. By analyzing the tone and sentiment of tweets in terms of vaccine stance and, as well as the prevalence of public responses to specific themes, namely misinformation here, text classification analysis of tweets provides insights into how risks associated with COVID-19 are framed and amplified on social media [42, 44].

The response mechanism phase is characterized by the audience's processing and reactions to risk information, taking into account factors such as individuals' heuristics and values, social group connections, signal value, stigmatization, and feedback to the risk [38, 45]. In alignment with these components, we have defined the response mechanism by evaluating behaviors such as user engagements (i.e., retweet, mention), decisions regarding the frequency of posts, network connections, targeting health authorities, or decisions to withdraw from the platform altogether.

Risk involves individuals' perceptions and information obtained from social sources, which can result in a lay understanding of the concept. This interactive process of risk interpretation can lead individuals to selectively transmit information that aligns with their pre-existing beliefs while downplaying opposing views or factual evidence [43, 46]. In this stage, social media emerges as a pivotal factor due to its unique design and functionality, altering traditional paradigms of risk communication during the COVID-19 pandemic. Users, guided by their risk perceptions, selectively follow certain narratives or accounts to receive updates, engaging with the content through various behaviors such as reading, reposting, commenting, and selectively following or unfollowing specific accounts. The cumulative volume of these engagements, along with the emergence of different information networks, plays a crucial role in the amplification or attenuation of risk perceptions.

More specifically, this study examines not only the sheer volume of retweets and mentions but also the structural properties of information networks, including centrality measures, density, and reciprocity. This multifaceted analysis provides insight into the interconnected and participatory nature of social media, highlighting its potential to both amplify and mitigate risk signals. The inclusion of social amplification proxies, such as retweets and mentions, further allows for the measurement of public attention to risk, predicated on the observation that social media users tend to share information more readily when faced with uncertainty or in the absence of definitive information. Thus, this complex interplay of factors within the social media landscape contributes to a nuanced understanding of risk communication and amplification during the pandemic, as delineated within the SARF [41, 42, 44, 47].

The SARF systematically explores the mechanisms through which risk information is disseminated, delving into the rationale, methods, and underlying processes of risk amplification and attenuation [45]. In alignment with this framework, our study conducts an incisive examination of various networks to understand how different types of information are propagated. We assess public attention through two key dimensions of user engagement: (a) the volume of posts containing MINVS, and (b) the number of Twitter users actively circulating MINVS in their interactions, a category encompassing retweets and mentions. Such a comprehensive exploration facilitates an understanding of the degree to which messages permeate within specific networks, an aspect vital to potential retransmission and the subsequent amplification of both message visibility and public attention [24].

The rigorous alignment of operationalization with the SARF framework equips this study to navigate the complex landscape of risk amplification within social media. By synergizing SARF with Social Network Analysis (SNA), it empowers the investigation of intricate phenomena such as echo chambers, as well as the influential role of algorithms in determining content exposure [48]. This methodological integration serves as a robust foundation for probing the nuanced interactions of misinformation, moral outrage, and consensus building within the sociotechnical fabric of social media platforms [39, 49]. Consequently, this research contributes a refined and comprehensive perspective to the understanding of how risk amplification and communication are molded and articulated within the dynamic landscape of contemporary digital discourse.

Methods

Dataset

We extracted vaccination-related conversations from Twitter (https://twitter.com) between December 1st, 2019, and October 24th, 2022. From a set of 14 Finnish terms,[1] 147 query keywords were constructed using word compounding and variation. Our keyword strategy focuses primarily on the misinformation and vaccine-related views surrounding the COVID-19 debate in Finland. Utilizing the *AcademictwitteR* R programming package [50] for collecting Twitter data, we compiled a dataset of 1,683,700 tweets. The dataset contains 724,214 retweets, 57,865 quotes, and 901,621 original tweets.

Text classification

During the initial phase, 4150 tweets were randomly sampled from the dataset. These tweets were manually annotated by four Finnish-speaking research assistants. To discern instances of vaccine stance, the study assessed posts based on their expression of attitudes or opinions towards vaccination. This analysis employed a methodology pre-established for vaccine stance detection [51, 52]. This methodology involves a pre-defined code book that outlines how a tweet may exhibit negative, positive, or neutral/unclear stances on the COVID-19 vaccines. We combined these categories in the analysis stage into binary classifications, with a score of 1 representing a negative stance and a score of 0 representing a positive or neutral/unclear stance.

Similarly, our annotators underwent training using the misinformation codebook developed by Memon and Carley [53], which has been previously used in other studies [54]. Misinformation and disinformation are both false or misleading information, but they differ in intent. Misinformation is shared without malicious intent, while disinformation is deliberately created and spread to deceive [53, 55]. While the distinction between the two is not always clear, it is not crucial for this research, as both misinformation and disinformation pose challenges to PHA and require similar strategies to address. Thus, our study categorized tweets based on their reference to misinformation, without differentiating between misinformation types (i.e., conspiracy theories, counterfeit treatments/cures, panic buying) or between disinformation and misinformation.

It is also important to keep in mind that coding for misinformation using a pre-established codebook from 2020 becomes a challenge, as the nature and context of misinformation can evolve over time. Considering the dynamic and ever-changing landscape of misinformation, we have adapted certain criteria or definitions in the codebook to align with the realities of 2022. We utilize several fact-checking

[1] Korona, rokote (vaccine), mrna, Pfizer, Biontech, Moderna, piikki (spike), Astra Zeneca, kuole (death), injektio, rokotushaitta (vaccination disadvantage), myrkkypiikki (poison spike), covid, toinen rokotus (second vaccination).

websites, including the THL pages, to support our analysis. However, we remain aware of the potential risks and difficulties in maintaining consistent and relevant annotations across the time.

The tweet annotation process involved six rounds of training, and Krippendorff's Alpha was employed to evaluate the inter-rater agreement. The results indicated good consistency among the raters,[2] with average scores of 0.693 and 0.668 for stance detection and misinformation, respectively. These results indicate that annotating misinformation from posts poses a significant challenge due to the complex and evolving nature of the information landscape surrounding the pandemic. For the text classification task, we utilized BERT, a state-of-the-art large pre-trained language model developed by Google AI Language [56]. BERT models are capable of being fine-tuned for diverse language recognition tasks. Because the study language is Finnish, Turku University FinBERT pre-trained embeddings model was fine-tuned on the annotated samples [57]. We applied text classification to a pretrained dataset of 4150 tweets, categorizing them into two classes: misinformation and negative vaccine stance. The model was then fine-tuned to create an algorithm capable of classifying the remaining posts.

Malicious bot classification

Initially, the study utilized the Botometer software to differentiate between bot-like accounts and those that are human-like among Twitter users in our dataset. The fourth version of Botometer is built on top of a supervised machine-learning approach that trains separated classifiers for each type of bot and combines the results from all classifiers to generate a bot score [13, 58]. The algorithm extracts over 1,200 features from each Twitter account, such as metadata, content information, and sentiment from the 200 most recent tweets [59]. Botometer returns a bot score between 0 and 1 for an account and a Complete Automation Probability (CAP) score for a probabilistic interpretation of the bot score [60]. Following Botometer's recommendation, we used a CAP score of 0.804 or above, equivalent to a bot score of 0.43 or higher, to identify bot-like accounts. Furthermore, we employed *default* CAP scores for English-speaking users and *universal* scores for Finnish-speaking users.

We defined a malicious account as a bot account that spreads MINVS intending to influence Twitter conversations. In addition to the bot labels provided by Botometer, the study [61] used additional features to differentiate between malicious and non-malicious bots. These features included the account's MINVS ratio, COVID-related tweet ratio, account age concerning the first COVID occurrence in Finland, and account status (active, Twitter suspended or deleted). Each feature was assigned a penalty score, which was then summed up to determine the total score. Accounts with a total score above 0.5 were classified as malicious bots.

[2] According to Krippendorff (1980), the minimum alpha value for acceptable reliability is 0.667. Tentative conclusion can be drawn from data with alpha values between 0.667 to 0.8. Reliable conclusion can be drawn from data with alpha value above 0.8.

Social network analysis

SNA is a method of exploring social structures using networks and graph theory. To create a network graph, nodes were used to represent Twitter users, while edges represented interactions between them, such as mentions, replies, or retweets. We employed four network measures: density, diameter, reciprocity, and assortativity. These metrics respectively evaluate the connectedness, dispersion, balance, and degree of similarity among users within the networks, providing valuable insights into the networks' cohesiveness, fragmentation, mutual ties, and homophily. Additionally, we used centrality metrics, including degree, betweenness, closeness, and Page Rank, to determine the centrality of THL within the network on a full data set [62, 63].

We examined COVID-19 communication on Twitter at two levels: all discourse and THL-linked discourse. We assessed user interactions through retweets and mentions and analyzed communication patterns across all tweets, with a separate focus on tweets with MINVS. This resulted in 12 distinct networks, analyzed using R and the *igraph package* [64]. Each network, from general to THL-specific, retweet to mention activities, and overall, to misinformation tweets, provided insights into different facets of user interactions and content.

Results

Descriptive statistics

The dataset comprises 1,683,700 tweets from 60,560 unique Twitter accounts, providing a large and diverse sample of tweets for our analysis. Botometer identified 13,522 accounts as bots, underlining the presence and potential impact of automated entities in the conversation. Our refined FinBERT model, employed for stance detection, classified 446,400 tweets (26.5%) as negative, indicating a considerable subset of users holding or promoting negative views on vaccination. This finding highlights the need to understand the characteristics and influence of such networks in the Finnish-speaking Twitter community. The misinformation classification model further identified 460,087 tweets (27.3%) as containing misinformation. (See Table 1 and Table 2 in the Online Appendix for more details.)

Notably, our malicious bot classification function categorized 4894 (8%) bots as malicious. The remaining accounts were classified as 8628 (14%) normal bots and 47,038 (78%) non-bots, which we presumed to represent the actual human accounts. These results offer a glimpse into the distribution of various user types in the Finnish-speaking Twitter community and their potential roles in the discourse.

Regarding our first research question, Fig. 2 shows the timeline of vaccination-related tweets from THL and Twitter engagements (retweets and mentions) targeting the THL account during the COVID-19 pandemic, using a logarithmic scale to account for the varying magnitude of interactions. THL's tweets can be retweeted multiple times, and the account can be mentioned in various discussions about COVID-19. The subplots in the figure feature a red regression line indicating the

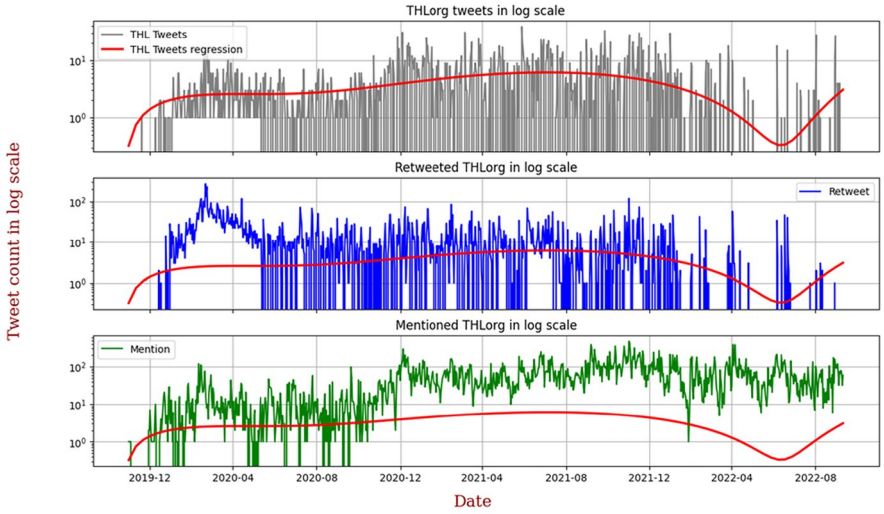

Fig. 2 Overall engagements with THL tweets in log scale. Engagements include all retweets and mentions targeting the THL account

trend of THL's daily tweet count during the pandemic. In the initial stage, THL tweeted daily[3] from February 2020 to mid-June 2020, and during this period, the daily volume of retweets of THL tweets (n = 35) was over 10 times higher, and the volume of mentions of the THL account (n = 13) was over 4 times higher than that of THL tweets (n = 3). The maximum number of retweets of THL tweets also occurred during this period.

After the initial months, both THL tweets and retweets of THL happened less frequently, and the daily volume of retweets of THL tweets (n = 8) decreased to an average of over 2 times higher than that of THL tweets (n = 3) (Fig. 4). In contrast, mentions of THL became more frequent, and there was at least one mention of a THL account from mid-October 2020 onwards. After mid-October 2020, the volume of daily mentions of THL (n = 71) was almost 20 times higher than that of THL tweets (n = 4) on average. (See Fig. 1 in Online Appendix for more details.) Put differently, while posts from THL were widely circulated by retweets at the onset of the pandemic, it lost its initial appeal, and the organization subsequently became the primary focus for information scrutiny, as evidenced by an increase in mentions.

Responding to the second research question, we analyzed the fluctuations of misinformation and negative vaccine stance targeting THL. We defined MINVS engagement as any retweets or mentions that contain either misinformation or negative vaccine sentiment targeting the THL account from all account types. In the context of MINVS engagement, we observed, on average, 23.77 daily malicious engagements

[3] THL produces multiple tweets on various subjects each day, however, the findings presented here correspond to the outcomes that meet our search query criteria.

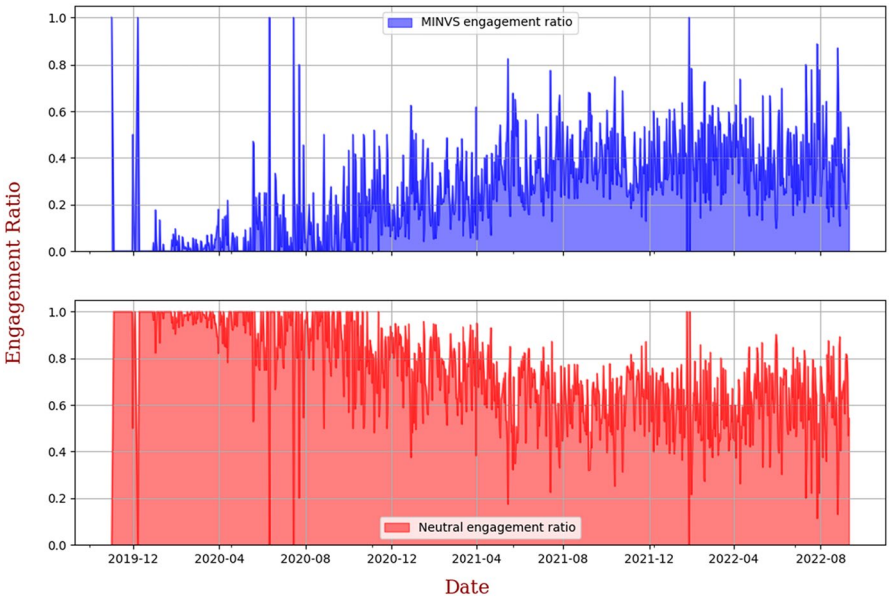

Fig. 3 Comparative ratios of MINVS and neutral engagements in THL tweets. Engagements include all retweets and mentions targeting THL

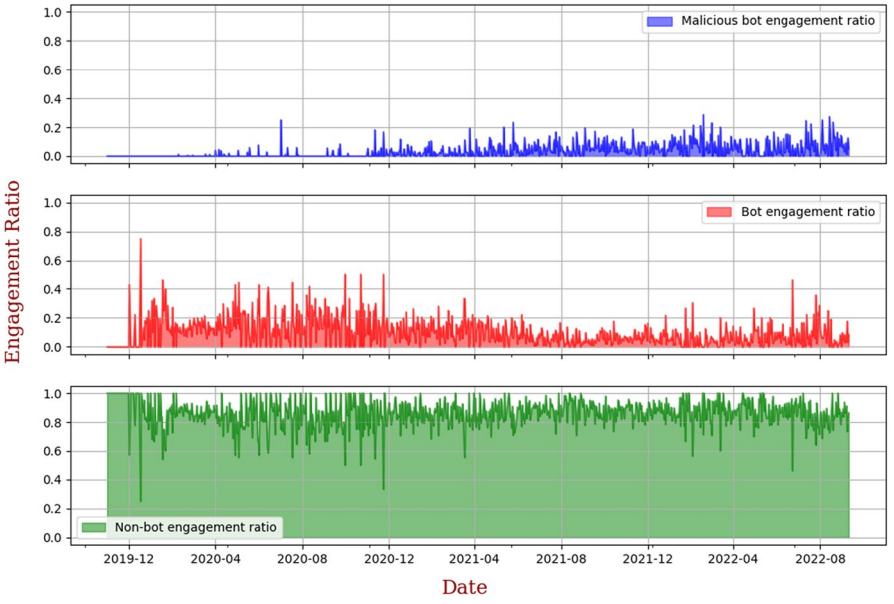

Fig. 4 Proportional engagement by bot types targeting THL. Engagements include all retweets and mentions targeting THL

(retweeting or mentioning) targeting the THL Twitter handle. Figure 3 illustrates the MINVS and neutral engagement ratio in relation to total engagement on the THL account regarding the COVID-19 vaccine topic.

Between the 1st of December 2019 and the 1st of October 2020, the median[4] MINVS engagement ratio is 6.8%. From the 1st of October 2020 onwards, this number increased to 33.3%. On the other hand, the median neutral engagement ratio was 98.8% before October 2020 and reduced to 67.2% afterward. Put differently, during the pandemic, there was a notable escalation in the frequency of engagements with THL content related to MINVS. Concurrently, the ratio of neutral engagements experienced a gradual decline over the same period. This trend suggests that accounts exhibiting skepticism toward THL's posts, or aligned with MINVS, are increasingly likely to interact with THL.

Responding to the third research question, we classified accounts into three types: non-bot (or actual users), bot, and malicious bot, based on our malicious bot classification. Figure 4 shows the ratio of engagements made by each account type over the total engagement with THL. Non-bot engagements accounted for 86% of all engagements with THL, followed by bot engagements at 12% and malicious bot engagements at 6%. This suggests that real users were more likely to interact with THL content than bots. It is also worth noting that while non-bot and bot accounts consistently mentioned and retweeted THL during the whole pandemic, malicious bot accounts only started to interact frequently with THL accounts from November 2020.

Figure 5 illustrates the retweet patterns[5] among non-bot users (human accounts), excluding the THL account. Two categories of retweeted sources are considered: malicious bots, representing unreliable sources, and the THL account, symbolizing reliable sources. Retweet ratios are computed by determining the proportion of retweets originating from non-bot users, where the original authors are either the malicious bots (subplot 1) or THL (subplot 2), relative to the total number of retweets observed during the same time frame within the non-bot retweet activity data. The figure indicates that during the initial year of the pandemic, particularly between January and October 2020, non-bot users predominantly retweeted content from THL, with an average retweet from THL ratio of 19%. Subsequently, this trend exhibited a noteworthy decline, with an average ratio of 0.8%.

On the other hand, the retweet ratio originating from malicious bots consistently remains below 5%, reflecting the infrequent tendency of regular users to retweet content from such sources. This trend is particularly evident in the early stages of the pandemic, during which the average retweet from malicious bot ratio is 0.23%. However, from June 2021, there is an observable rise in this ratio (from 0.23 to 3%), indicating some regular users are influenced by the content from malicious bots.

[4] The median value is reported because there are few outliners (dates when only one interaction is recorded, and it is a malicious interaction) in the data resulting in a higher mean value.

[5] Similar figure for the mention-engagement is included as Fig. 1 in Appendix.

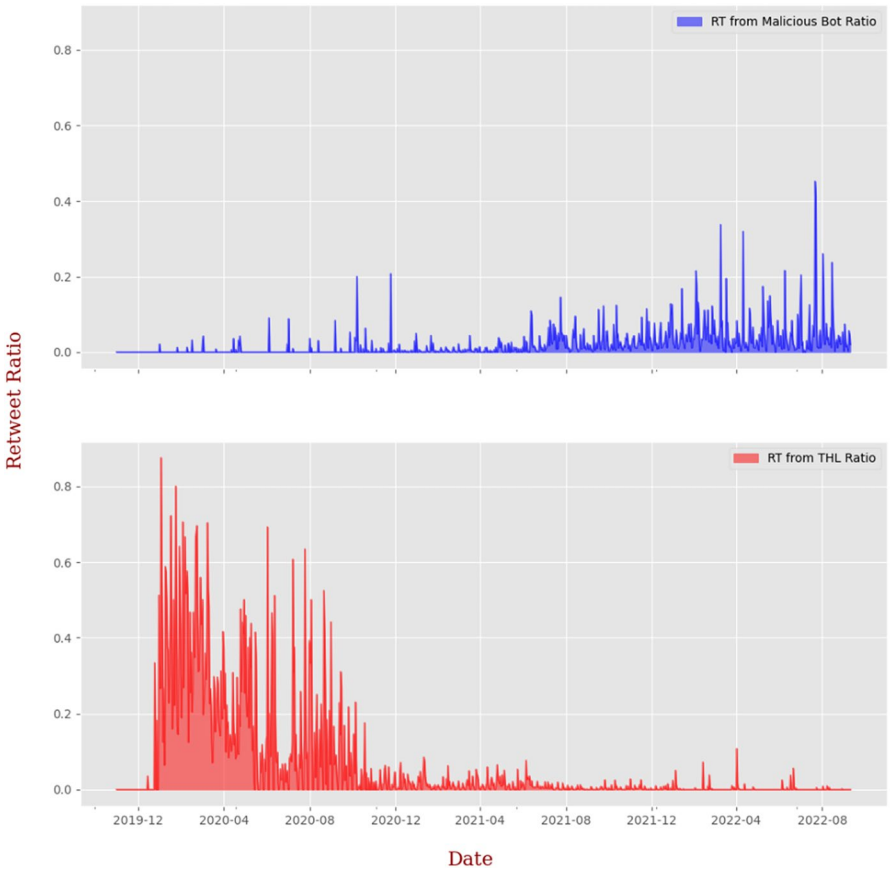

Fig. 5 Temporal trends in retweet ratios by non-bot accounts: malicious bots and THL. The retweet ratios are determined by dividing the number of retweets from non-bot accounts whose original authors are either Malicious Bots (subplot 1) or THL (subplot 2) by the total number of retweets within the same time frame in our dataset of non-bot retweet activity

Main networks

Our final research questions delve into the distinctive characteristics of networks that propagate misinformation compared to those expressing a negative stance towards vaccines, with a particular emphasis on the scale of these networks within the Finnish-speaking Twitter community discussing COVID-19. In investigating this, we examined unique connectivity patterns and bot-driven activity in networks identified by mentions and retweets.

At the onset of the pandemic, THL was centrally positioned within a relatively sparse mention network, primarily characterized by neutral communication. However, by the end of the pandemic, while THL remained at the heart of the network, it increasingly became the target of misinformation posts (For additional details, refer to Figs. 4 and 5 in the Online Appendix). Figure 6 below specifically illustrates the

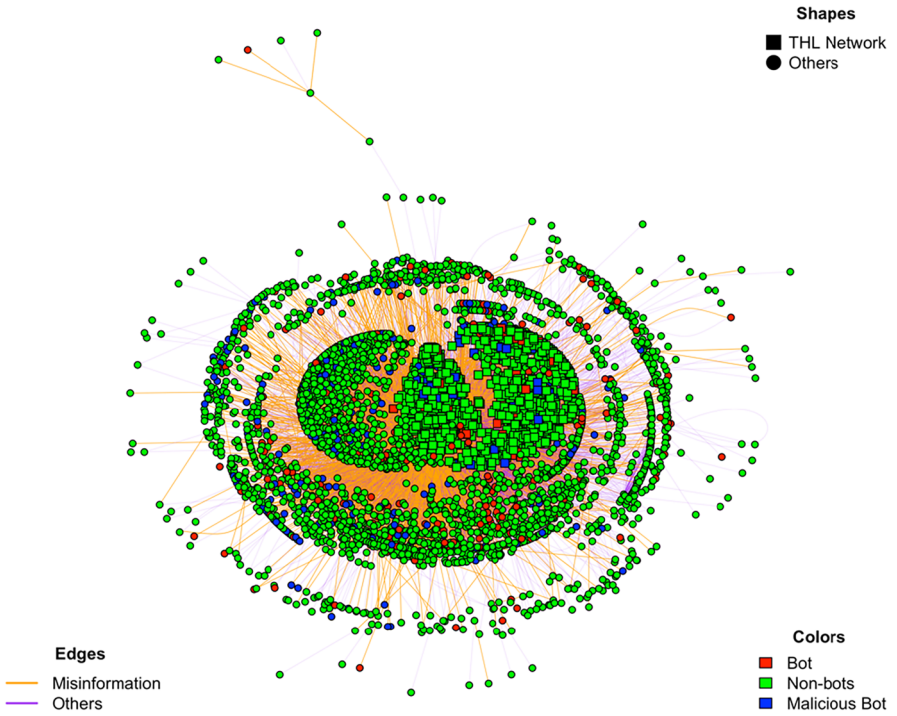

Fig. 6 All mentions network and THL-connected accounts in the network. The depicted network includes only those accounts that have been mentioned more than 10 times, and it showcases the connectivity amongst these accounts. The term 'THL network' is used to describe the direct account interactions involving the Finnish Institute for Health and Welfare's Twitter account (@THLorg). The vertices in the network are differentiated by both color and shape, representing the distinct types of accounts involved. Edge colors represent post types and thickness shows the frequency (normalized) of mentions between vertices

overall pattern of the mention activity, confirming THL's role as a central hub in COVID-19 communication. Although most accounts are authentic, malicious bots are present, seeking to disrupt and distort communication. These bots do not form clusters but are dispersed throughout the network. The presence of such malicious bots highlights the risks of potential threats to the communication network's integrity. The orange lines in Fig. 6 indicate that most active accounts were disseminating misinformation through mentions.

On the other hand, retweet analysis identified two primary clusters. The first cluster is characterized by a network that disseminates reliable and trustworthy information regarding COVID-19. In contrast, the second cluster is associated with a network promoting anti-vaccination narratives and disseminating misinformation (Fig. 7). This pattern becomes more evident in the network structures observed both at the onset and the end of the pandemic (See Figs. 2 and 3 in Online Appendix for more details). Initially, the THL served as the primary hub for information dissemination, and the network featured minimal misinformation-related retweet activity.

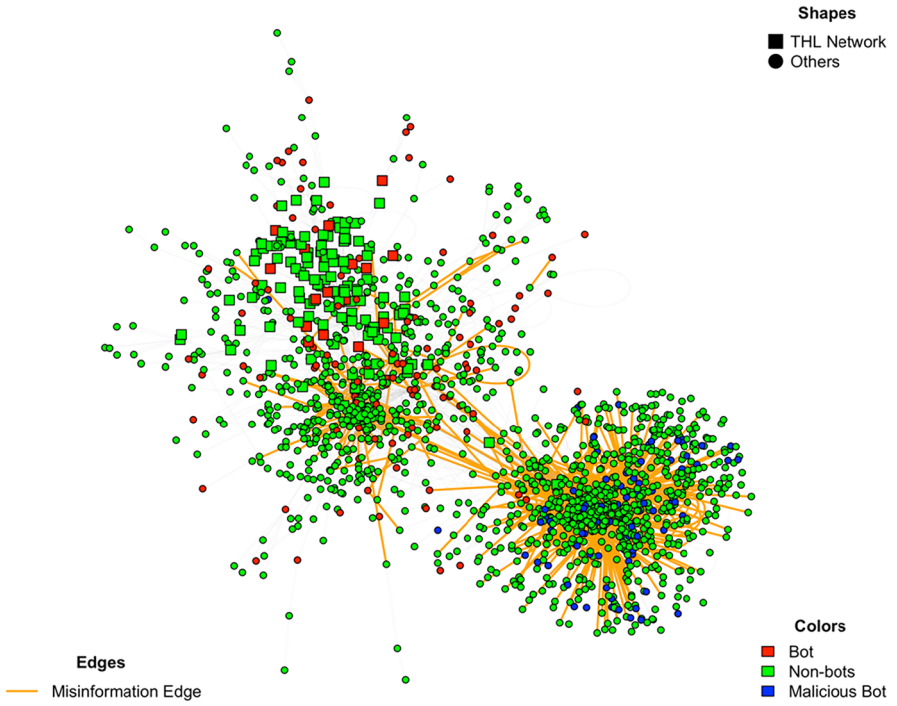

Fig. 7 All retweets network and THL-connected accounts in the network. The depicted network includes only those accounts that have been retweeted more than 10 times, and it showcases the connectivity amongst these accounts. The term 'THL network' is used to describe the direct account interactions involving the Finnish Institute for Health and Welfare's Twitter account (@THLorg). The vertices in the network are differentiated by both color and shape, representing the distinct types of accounts involved. Edge color represent misinformation and thickness shows the frequency (normalized) of retweets between vertices

However, by the end of the pandemic, a distinct network had emerged that exclusively circulated misinformation. Moreover, misinformation circulation was also observable within networks affiliated with the THL.

Moreover, there is a noticeable divergence in the retweet activity of various types of bots; malicious bots tend to retweet accounts different from THL-related ones, while regular bots tend to align with THL's retweeting behavior (red colors in Fig. 7 below). This finding implies that malicious bots alleged their retweeted activities with a network promoting MINVS. In contrast, regular bots may have a positive impact on the COVID-19 discussion network in Finland by disseminating accurate information and endorsing the messages of relevant PHA such as the THL.

The characteristics of the networks are summarized in Online Appendix, Table 3. Our results show that, Twitter users in Finland who tweet more about MINVS exhibit distinct interaction patterns. The network of all mentions demonstrates higher assortativity (− 0.1102) relative to all retweets (− 0.1881), implying that users who mention a particular topic tend to connect with others who possess

similar levels of activity and influence. This may indicate a significant degree of awareness and skepticism among users concerning COVID-19. Nevertheless, specific subgroups of users are more prone to spreading or engaging MINVS.

The study found several differences between retweets and mentions in the network. Retweets were found to be less common, with lower density (0.00006) than all mentions (0.00023), and they may take longer to spread than mentions, as indicated by the larger diameter (134) of all retweets compared to all mentions (46). However, mentions may have a higher prevalence of misinformation, as shown by the highest density (0.00035) of the network of mentions in misinformation among all groups.

In terms of reciprocity, all retweets (0.0244) were less likely to be reciprocated than all mentions (0.1626). Negative vaccine stance may be more likely to be reciprocated in mentions, as suggested by the higher reciprocity (0.0963) of the network of mentions in negative stance than that of all mentions.

Additionally, nodes may be less inclined to retweet content from nodes with similar characteristics, as indicated by the lower assortativity of all retweets (− 0.1881) than all mentions (− 0.1102). However, negative vaccine stance may be less assortative in retweets (− 0.1941) than in mentions (− 0.1376).

The study found that users who tweet about MINVS have different interaction patterns. The network of misinformation—mentions showed the highest density (0.0003485) and reciprocity (0.1176), indicating a strong echo chamber effect where users reinforce each other's beliefs without being exposed to alternative views or facts. The network of misinformation—retweets had the lowest assortativity (− 0.1963), indicating a high level of vulnerability or susceptibility to influence from more active or influential users who spread misinformation. The network of negative vaccine stance-retweets had a low density (0.0001877) and reciprocity (0.0201), but a high diameter (53), suggesting a low level of engagement or commitment among these users, who may simply follow or share what they see without much reflection or discussion. The network of negative vaccine stance—mentions had a moderate density (0.0003272) and reciprocity (0.0963), but also a high diameter (56), indicating a mixed level of engagement or commitment among these users, who may express negative vaccine stance without necessarily agreeing or disagreeing with each other. Overall, these findings suggest that users who tweet about MINVS may be more likely to interact with like-minded individuals and may be vulnerable to misinformation and influence from more active or influential users.

Overall, the study highlights different patterns of interaction among Finnish Twitter users who tweet about MINVS. While some subgroups of users are more susceptible to spreading or engaging with MINVS, the results also suggest a high level of awareness and skepticism among Finnish Twitter users.

THL network

The THL network was analyzed at two levels; mentions and retweets. Our analysis revealed that THL was retweeted by 2898 different accounts for a total of 18,538 times. This represents a relatively small portion of the network's interactions that account for approximately 5% of all retweet interactions. On the other hand, only

0.6% of the malicious bot accounts were identified in the THL retweet network, indicating that THL's content did not align with their goals or objectives.

The relatively small fraction of accounts (3.7%, or 2898 out of 76,815 distinct accounts) that shared THL's content suggests that THL's messages may not have been disseminated and received as widely as other types of content circulating within the network. Although THL's content did not generate high levels of direct engagement or interaction from its retweeters, the centrality measures suggest that it still played a significant role in shaping the conversation around COVID-19 on Twitter, particularly at the beginning of the pandemic as discussed above. THL was identified as the most important account to be retweeted (in degree) and had the highest PageRank score, which suggests that it was viewed as a reliable and relevant source of information by a broad range of Twitter users. According to the betweenness score, THL served as a critical link between various accounts in the network, ranking ninth among all accounts in this regard. However, THL's status as an authoritative source of information makes it a valuable resource for PHA and other stakeholders looking to disseminate accurate and timely information about COVID-19 to the public. It is also important to note that, two government accounts (@valtioneuvosto [State news agency] and @Fimea [Safety and development center for the pharmaceutical industry]) come after THL in the Page Rank list, which shows that when sharing COVID-19-related tweets, the public relies on government accounts.

A total of 8,060 unique accounts mentioned THL, which constitutes approximately 15% of all 54,754 unique accounts (referring to connected accounts by mentions) that were mentioned in the network. The number of accounts involved in mentions engagement is 3.5 times greater than the number of accounts in the retweet network. Likewise, among all mentions (700,753), THL constitutes 45% (317,582) of the interactions. Moreover, among the accounts that mentioned THL, 8% were identified as malicious bots. Among the top 100 accounts that mentioned other accounts, only one was identified as a malicious bot. However, among the top 100 accounts that mentioned THL directly, 24 were identified as malicious bots. This indicates that malicious bot accounts particularly targeted THL by mentioning specific topics, such as MINVS. In Fig. 8, we combined all the accounts that mentioned THL with MINVS. The results show that non-bots had the highest frequency of mentions (outdegree) among all the accounts. On the other hand, despite posting less frequently than some non-bot accounts, the THL account was specifically targeted by a high number of malicious bots.

The study found that THL was a highly relevant and influential source of information about COVID-19 in Finland, with a large number of direct and indirect connections to other nodes in the network. THL was the top account according to degree centrality, betweenness score, and PageRank score, suggesting that it played a central role in both retweet and mention networks. However, the low closeness score of THL indicates that it was relatively far from most other nodes in the network, which may have limited its direct access to information or influence over other nodes compared to more central accounts. This could potentially hinder THL's ability to effectively disseminate information and engage with other accounts in the network. Moreover, the presence of malicious bots in the network raises concerns about

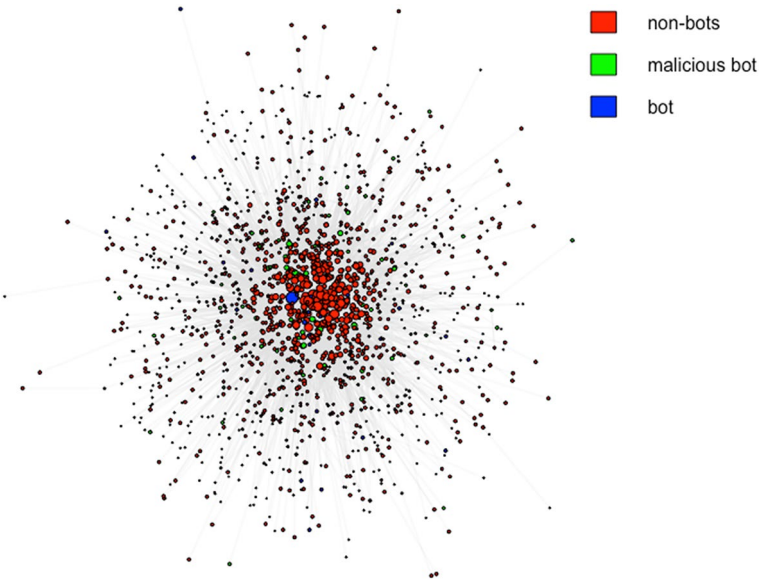

Fig. 8 Networks of Twitter accounts that have mentioned THL with both misinformation and negative vaccine stance. The depicted network includes only those accounts that have been mentioned more than 5 times, and it showcases the connectivity amongst these accounts. The vertices in the network are differentiated by color, representing the distinct types of accounts involved. The size of vertices represents the log scale out-degree score

the risks and challenges of social media for public health communication, as these bots may have attempted to spread misinformation and undermine trust in THL's messages despite being mentioned by many authentic and interested accounts.

Discussions

THL effectiveness on Twitter discussion

Given the paucity of authoritative information available at that time, individuals demonstrated heightened engagement with THL. This engagement was manifested predominantly through the sharing of THL's posts, thereby serving as conduits for disseminating reliable information to their respective followers. Malicious bots participated in the discussion at a later stage. Although the study design makes it challenging to establish a causal relationship between malicious bots and non-bot accounts, we observed that active Twitter users were more likely to engage with content related to MINVS during the pandemic.

Our findings indicate that THL's role transitioned from being a contributor (or steering the discussions) to becoming the subject of discussion in the latter phase of the pandemic. To put it another way, despite maintaining a consistent volume of tweets, THL experienced a decline in retweets and instead became a focal point

for discussions related to MINVS. As the Twitter discourse increasingly gravitated toward MINVS-related topics, THL's posts garnered less attention and were met with greater interactions containing elements of MINVS.

According to Fig. 6 (as well as in Figs. 2 and 3 in Appendix), the retweet network linked to THL aligns more closely with an accurate and trustworthy information network, displaying minimal malicious bot activity. These findings indicate that THL successfully established a reliable and trustworthy network for COVID-19 information in Finland, although its influence waned somewhat compared to the initial stages of the pandemic. However, the content shared by THL appeared to be misaligned with the expectations of active Twitter users following COVID-19-related discussions, compromising THL's influence and prompting the organization to make the strategic decision to suspend its Twitter account.

The decision of THL to suspend their Twitter can be interpreted as a proactive measure to limit the dissemination of MINVS and maintain trust in the health system. We found that while malicious bots have a restricted network and limited overall reach, they can use targeted attacks on authoritative accounts to amplify their message and potentially cause harm. By suspending their Twitter activity, THL may have prevented further amplification of MINVS content and prevented the spread of potentially harmful misinformation.

However, it is crucial to acknowledge that further investigations are required to comprehensively examine the long-term effects of the suspension of THL's Twitter account. Additionally, it is important to emphasize that Twitter does not represent the sole platform or medium through which public perceptions are expressed and shaped. In fact, Twitter usage in Finland is limited to a specific segment of society and cannot be considered fully representative of the entire population.

Structure of networks

Our analysis uncovered clear differences in COVID-related communication patterns within retweet and mention networks, as well as across networks associated with misinformation and negative vaccine stances. These findings address our research questions and provide insights into the diverse dynamics within these distinct networks.

We found retweet networks involved more participants, but lower engagement compared to mention networks across all conversations (See Table 1–3 in Online Appendix). This suggests retweets disseminate information more broadly while mentions elicit greater user engagement. Additionally, retweet networks had larger diameters, indicating they enable wider propagation of information. Importantly, authoritative sources like THL were predominantly retweeted, permeating expansive networks. In contrast, despite malicious bots, misinformation, and negative vaccine stance networks had smaller diameters, exhibiting selective sharing.

In terms of the negative vaccine stance networks, mentions networks are wider, denser, more reciprocating, and less assertive than retweet networks. This suggests users opposed to vaccination may be more likely to engage each other directly to build stronger relationships, rather than simply sharing content.

Comparatively, the misinformation network's mentions are denser, smaller, more reciprocating, and less assertive than retweets. This implies users spreading COVID-19 misinformation are likely to engage a smaller group directly to reinforce misinformation homogeneity by connecting like-minded users, rather than attracting diverse perspectives.

Negative vaccine stance is more reciprocated in mentions than retweets possibly because users empathize with others' frustration, anger, or fear about the pandemic or government response. Users may also be less likely to retweet similar content than mention creators to diversify information sources and perspectives.

Our analysis reveals that at the onset of the pandemic, there were no discernible clusters of misinformation or negative sentiment towards vaccines; only a few active accounts were identified. However, by the end of the pandemic, robust and dense networks had emerged for both misinformation and negative vaccine stances. Several factors contribute to this phenomenon. In the initial stages, a lack of clear information and prevailing uncertainty may have led to cautious information-sharing behaviors. However, as vaccines became more accessible, there appeared to be a polarization of public opinion, giving rise to well-defined social groupings. The enduring nature of the pandemic likely intensified pandemic fatigue as described above, making individuals more susceptible to misinformation and less trusting of vaccines. Additionally, the availability of vaccines redirected the public discourse from broad pandemic concerns to vaccine-specific debates, offering a targeted subject for misinformation and skepticism to thrive. The appearance of these highly interconnected networks poses a significant challenge to public health initiatives, as they not only obstruct the flow of accurate information but also undermine the impact of vaccination efforts, thus extending the duration of the public health crisis.

Influential actors in networks

Results of the centrality measures show that THL emerges as the preeminent actor in both retweet and mention networks. Although there are several other influencers in the retweet network, our analysis indicates that THL still wields a significant and influential role in shaping the overall network. Moreover, nearly half of all mentioned interactions are directed toward THL. Two possible explanations may account for these findings.

The public's perception of COVID-19 risks and the acceptable norms of conduct in relation to it was influenced and challenged by many tweets that expressed responses and reactions to the social and political actions of others [44]. These tweets engaged in a dynamic and ongoing dialogue that shaped and contested the boundaries and rules of COVID-19 risk management. At the pre-problem stage, which is marked by a gradual rise in media and public attention as the COVID-19 crisis develops, we contend that we would observe minimal amplification regardless, mainly delivering THL posts by retweeting.

However, during the final stage, particularly at the point of gradual decline, there may be indications of COVID-19 fatigue that leads to information avoidance [45] and a diminished risk perception among the public. This could in turn lead to a

reduced willingness to engage in or follow related discussions. It is also possible that those who are particularly concerned about COVID-19 vaccines or are active in anti-vaccination campaigns remain engaged in Twitter discussions. These aspects of online engagement among specific groups of motivated individuals warrant further exploration in future research.

The degree to which we trust information and its source has a significant impact on our perception and interpretation of it, a phenomenon known as the "trust heuristic". This heuristic allows us to make judgments and decisions in uncertain environments without demanding too many resources from the decisionmaker [37]. During the early stages of the pandemic when information was scarce, our research reveals that the messages disseminated by THL were more prone to being retweeted. This finding suggests that THL's communication held a higher level of credibility and authority within the network during a time when reliable information sources were limited. The increased likelihood of retweets indicates that THL played a crucial role in shaping the information landscape and influencing the spread of information during the initial phase of the pandemic. As more vaccination-related information circulated online, people were more likely to engage in direct interactions. However, as the perceived risk decreased, people tended to passively follow updates without engaging actively.

This is also confirmed by previous studies that only a small percentage of individuals participate actively in online discussions, with the majority merely observing the debates. Focusing solely on the vocal minority (the "loud" 10%) who may hold extreme views may lead to the formation of a "trust deficit model" and distort our perception of the majority's attitudes toward risks and risk events [23, 65]. Our analysis revealed that the vocal 10% of online users were predominantly anti-vaccination campaigners and malicious bots. In fact, over 82% of the MINVS contents from our dataset are shared by the vocal minority.

However, it is important to note that as the pandemic progressed and more vaccination-related information circulated online, people's engagement with THL shifted from retweeting to more direct interactions, as illustrated in Figs. 1, 3, and 4. From Fig. 3, we observed that non-bot users account for most of the interactions with THL; and Fig. 1 reveals the shift of interacting type from retweeting to mentioning. Furthermore, as the perceived risk diminishes and individuals adopt a more complacent stance toward staying informed, the role of THL in the retweet network becomes normalized.

Influence of malicious bot accounts

Regarding the characteristics of malicious bots, despite being noticeable participants in online public health discourse, they have a relatively restricted network and are limited to a certain level of users. While they are known to influence discussions on vaccination and can cause distortions [12], their overall reach is found to be comparatively small in our study.

Our results show that malicious bot accounts were organized, specifically targeting THL and other health authorities. The malicious bots aimed to use the THL

account as a social station to disseminate MINVS to a much wider audience in a manner unavailable by their reach or network [47]. Our study found that while malicious accounts exhibit specialized patterns (mentions) that enable them to initiate information scrutiny more rapidly than active users, there is no conclusive evidence to suggest that such bots have significantly impacted the vaccine discourse on Twitter. Despite this, our research suggests that bots can amplify a small subset of accounts, but their overall influence is limited, consistent with previous studies [66].

Contribution to SARF

Our research contributes to the SARF by operationalizing several key indicators: tweet content types (misinformation and negative vaccine stances), interaction types (retweets and mentions), network structures, and interaction volumes. This multidimensional approach enabled us to explore the nuanced dynamics between THL and the Twitter audience during different stages of the COVID-19 pandemic [46].

Initially, THL's role as a social station involved effectively signaling risks and clarifying uncertainties, which garnered a largely positive response from Twitter users who subsequently disseminated this information. However, as the pandemic progressed, the appearance of malicious bots and other accounts signaling vaccine and management risks led to a diversification of public responses, encompassing behavioral, economic, and symbolic dimensions [38, 45].

Our analytical framework further revealed that risk perception, information transmission, and response mechanisms are influenced not just by the severity of the risk but also by the emotional content of the information [38, 41–43]. For example, tweets containing negative emotions had a propensity for rapid dissemination, leading to the emergence of more organized networks. This observation underscores the evolving nature of risk communication and the necessity for timely interventions, particularly when misinformation becomes potent enough to halt organizational activity, as was the case with THL [67].

Therefore, our work refines the SARF by not only offering a method for risk measurement on social media but also by providing insights into the changing dynamics of public interactions and network structures during a crisis, thereby informing strategies for timely and effective interventions.

Policy implications

Our research identifies critical factors influencing the spread of MINVS on social media platforms, particularly Twitter. These findings offer actionable insights for PHA and organizations like THL, emphasizing the need for dynamic communication strategies [38].

Firstly, as discussed above, when the COVID-19-related risks decreased and vaccines were introduced, concern and anxiety about the virus were replaced by anti-vaccination-related discourses on Twitter. The THL shared facts and guidelines

throughout the pandemic, and the general Twitter audience followed and disseminated these posts. However, when online discussions turned to misinformation and anti-vaccination-related topics, only those who were interested in these topics remained engaged. In the information transmission stage [38], focusing solely on health promotion is not enough, PHA should also prioritize counter-arguments in the response mechanism. There is very little evidence to suggest that providing facts alone will stop the spread of misinformation. More effective ways to combat misinformation and negative vaccine stance include debunking and prebunking [68].

Secondly, PHA must monitor and engage with key actors to maximize their influence [40]. Keeping abreast of the rhetoric and logic behind these actors' claims will enable targeted communication interventions, in terms of debunking and rebunking techniques, thereby mitigating the spread of misinformation [44]. Although THL forged partnerships with influential individuals and organizations, such as the Prime Minister's office, our results show that they are not enough to steer the Twitter discussions in the last period.

Furthermore, the constrained nature of anti-vaccine and misinformation networks suggests limited openness to contrary perspectives. This observation aligns with the SARF's conceptualization of how information selectively propagates, emphasizing the need for nuanced communication tailored to subgroups within these networks [38].

Sentiment analysis can provide additional granularity, revealing the emotional dimensions that shape risk perceptions and guide public behavior [42]. This understanding can inform more effective messaging, particularly in the context of an evolving pandemic where public sentiment and engagement patterns fluctuate [45].

Lastly, our study underscores the importance of transparency and timely communication by health authorities to maintain public trust, especially when making significant decisions like withdrawing from a platform [23]. Continuing coordination between different health organizations can optimize message reach and impact, especially in combating the efforts of bots and trolls [66].

In summary, PHA can employ these insights to develop comprehensive communication strategies that counter MINVS, thereby fostering an informed public discourse and enhancing the effectiveness of public health initiatives.

Limitations

The SARF was designed for empirical research on risk communication complexity in real-world situations. While useful for understanding how "expected" risks can affect beyond one risk, the framework has limitations in comprehending changes in content data over time [41, 69]. Establishing an agreed-upon degree of risk to analyze changes in content data regarding risk can be challenging and not useful.

Twitter data may not represent the general population's opinions or experiences since Twitter users are not a representative sample. Thus, only tentative conclusions can be drawn from this study. Additionally, not Twitter users' posts are accessible to

researchers, and Twitter data can be subject to biases and manipulation, such as the use of bots and fake accounts to amplify certain messages or distort public opinion. The limitations of Twitter data may affect the accuracy and generalizability of research findings related to COVID-19. Social network analysis can identify patterns of communication and information flow, but it may be challenging to measure the impact of specific messages or interventions on public health outcomes.

Conclusions

This study offers novel insights into the dynamics of COVID-19 vaccine discourse on Twitter, elucidating the evolving interplay between the THL and various online stakeholders. Our multi-dimensional analytical framework reveals key differences in retweet and mention networks across different stages of the pandemic. Initially, THL served as a major information source, disseminating critical updates that Twitter users readily amplified through retweets. However, THL's influence on Twitter was compromised as the discourse evolved, with increased visibility of vocal anti-vaccine subgroups and malicious bots challenging THL's credibility.

The emergence of highly interconnected misinformation and anti-vaccine networks towards the pandemic's latter stages poses significant challenges for public health communication. This polarization reveals that simply providing facts is insufficient to counter misinformation. In contrast, PHA should proactively present counterarguments and develop nuanced communication strategies tailored to distinct audience segments. Granular sentiment analysis can uncover the emotional dimensions shaping online vaccine discourse and risk perceptions.

While malicious bots exhibit specialized behaviors amplifying targeted accounts, their overall reach remains limited. This underscores the need for PHA to forge expansive partnerships, coordinate messaging, and exercise transparency in social media communications activities. Adapting communication strategies based on the evolving dynamics revealed in this study will empower PHA to mitigate misinformation and enhance the effectiveness of their health promotion efforts.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s42001-024-00257-8.

Acknowledgements The authors wish to acknowledge the scientific computing support provided by the team at the Aalto University Research Software Engineer (RSE).

Funding Open Access funding provided by Finnish Institute for Health and Welfare. This article was funded by the Research Council of Finland (Grant no. 339931).

Data availability The data that support the findings of this study are available from Aalto University, but restrictions apply to the availability of these data, which were used under license for the current study and so are not publicly available.

Declarations

Conflict of interest None declared.

References

1. Kim, H. K., Ahn, J., Atkinson, L., & Kahlor, L. A. (2020). Effects of COVID-19 misinformation on information seeking, avoidance, and processing: A multicountry comparative study. *Science Communication, 42*(5), 586–615. https://doi.org/10.1177/1075547020959670
2. Brüssow, H., & Timmis, K. (2021). COVID-19: Long Covid and its societal consequences. *Environmental Microbiology, 23*(8), 4077–4091. https://doi.org/10.1111/1462-2920.15634
3. Van Huijstee, D., Vermeulen, I., Kerkhof, P., & Droog, E. (2022). Continued influence of misinformation in times of COVID-19. *International Journal of Psychology, 57*(1), 136–145. https://doi.org/10.1002/ijop.12805
4. WHO. (2021). *WHO public health research agenda for managing infodemics.* World Health Organization. https://www.who.int/publications/i/item/9789240019508
5. Méndiz-Noguero, A., Wennberg-Capellades, L., Regadera-González, E., & Goni-Fuste, B. (2023). Public health communication and the Covid-19: A review of the literature during the first wave. *El Profesional de La Información*. https://doi.org/10.3145/epi.2023.may.13
6. Alizadeh, H., Sharifi, A., Damanbagh, S., Nazarnia, H., & Nazarnia, M. (2023). Impacts of the COVID-19 pandemic on the social sphere and lessons for crisis management: A literature review. *Natural Hazards, 117*(3), 2139–2164. https://doi.org/10.1007/s11069-023-05959-2
7. Iberszer, K., Litwiniuk, M., Zaniuk, M., Hurkała, K., Antonik, D., Denys, B., Góra, K., Zdziennicki, W., Zimnicki, P., & Lato, M. (2023). Influence of social media on the fight against COVID-19 pandemic—Literature review. *Journal of Education, Health and Sport, 39*(1), 17–28. https://doi.org/10.12775/JEHS.2023.39.01.002
8. Etta, G., Galeazzi, A., Hutchings, J. R., James Smith, C. S., Conti, M., Quattrociocchi, W., & Riva, G. V. D. (2022). COVID-19 infodemic on Facebook and containment measures in Italy, United Kingdom and New Zealand. *PLoS ONE, 17*(5), e0267022. https://doi.org/10.1371/journal.pone.0267022
9. Chen, M., Yu, W., & Cao, X. (2023). Experience pandemic fatigue? social media use may play a role: Testing a model of pandemic fatigue development from a social media perspective. *Health Communication, 38*(14), 3346–3356. https://doi.org/10.1080/10410236.2022.2149095
10. Al-Rawi, A., & Shukla, V. (2020). Bots as active news promoters: A digital analysis of COVID-19 tweets. *Information*. https://doi.org/10.3390/info11100461
11. Xu, W., & Sasahara, K. (2022). Characterizing the roles of bots on Twitter during the COVID-19 infodemic. *Journal of Computational Social Science, 5*(1), 591–609. https://doi.org/10.1007/s42001-021-00139-3
12. Broniatowski, D. A., Jamison, A. M., Qi, S., AlKulaib, L., Chen, T., Benton, A., Quinn, S. C., & Dredze, M. (2018). Weaponized health communication: Twitter Bots and Russian trolls amplify the vaccine debate. *American Journal of Public Health, 108*(10), 1378–1384. https://doi.org/10.2105/AJPH.2018.304567
13. Gilani, Z., Farahbakhsh, R., Tyson, G., & Crowcroft, J. (2019). A large-scale behavioural analysis of bots and humans on Twitter. *ACM Transactions on the Web, 13*(1), 7:1-7:23. https://doi.org/10.1145/3298789
14. Chang, H.-C.H., & Ferrara, E. (2022). Comparative analysis of social bots and humans during the COVID-19 pandemic. *Journal of Computational Social Science, 5*(2), 1409–1425. https://doi.org/10.1007/s42001-022-00173-9

15. Bruns, H., Dessart, F. J., & Pantazi, M. (2022). *Covid-19 misinformation: Preparing for future crises: An overview of the early behavioural sciences literature.* Publications Office of the European Union. https://doi.org/10.2760/41905
16. Seara-Morais, G. J., Avelino-Silva, T. J., Couto, M., & Avelino-Silva, V. I. (2023). The pervasive association between political ideology and COVID-19 vaccine uptake in Brazil: An ecologic study. *BMC Public Health, 23*(1), 1606. https://doi.org/10.1186/s12889-023-16409-w
17. Jemielniak, D., & Krempovych, Y. (2021). An analysis of AstraZeneca COVID-19 vaccine misinformation and fear mongering on Twitter. *Public Health, 200*, 4–6. https://doi.org/10.1016/j.puhe.2021.08.019
18. Pierri, F., DeVerna, M. R., Yang, K.-C., Axelrod, D., Bryden, J., & Menczer, F. (2023). One Year of COVID-19 vaccine misinformation on Twitter: Longitudinal study. *Journal of Medical Internet Research, 25*, e42227. https://doi.org/10.2196/42227
19. Sufi, F. K., Razzak, I., & Khalil, I. (2022). Tracking anti-vax social movement using AI-based social media monitoring. *IEEE Transactions on Technology and Society, 3*(4), 290–299. https://doi.org/10.1109/TTS.2022.3192757
20. Larson, H. J., Lin, L., & Goble, R. (2022). Vaccines and the social amplification of risk. *Risk Analysis, 42*(7), 1409–1422. https://doi.org/10.1111/risa.13942
21. Muric, G., Wu, Y., & Ferrara, E. (2021). COVID-19 vaccine hesitancy on social media: Building a public twitter data set of antivaccine content, vaccine misinformation, and conspiracies. *JMIR Public Health and Surveillance, 7*(11), e30642. https://doi.org/10.2196/30642
22. Hwang, J., Su, M.-H., Jiang, X., Lian, R., Tveleneva, A., & Shah, D. (2022). Vaccine discourse during the onset of the COVID-19 pandemic: Topical structure and source patterns informing efforts to combat vaccine hesitancy. *PLoS ONE, 17*(7), e0271394. https://doi.org/10.1371/journal.pone.0271394
23. Bearth, A., & Siegrist, M. (2022). The social amplification of risk framework: A normative perspective on trust? *Risk Analysis, 42*(7), 1381–1392. https://doi.org/10.1111/risa.13757
24. Sutton, J. (2018). Health communication trolls and bots versus public health agencies' trusted voices. *American Journal of Public Health, 108*(10), 1281–1282. https://doi.org/10.2105/AJPH.2018.304661
25. Shahsavari, S., Holur, P., Wang, T., Tangherlini, T. R., & Roychowdhury, V. (2020). Conspiracy in the time of corona: Automatic detection of emerging COVID-19 conspiracy theories in social media and the news. *Journal of Computational Social Science, 3*(2), 279–317. https://doi.org/10.1007/s42001-020-00086-5
26. Khan, F. S., Ullah, A., Khan, O. J., Sehar, B., Alsubaie, A. S. R., Asmat, S., & Zeb, F. (2022). Comparable public health responses to COVID-19 pandemic. *The Open Public Health Journal, 15*(1), e187494452207290. https://doi.org/10.2174/18749445-v15-e2207290
27. Zhai, S., Li, Y. J., & Chi, M. (2022). The impact of government social media information quality on public panic during the infodemic. *Frontiers in Psychology, 13*, 908213. https://doi.org/10.3389/fpsyg.2022.908213
28. James, L., McPhail, H., Foisey, L., Donelle, L., Bauer, M., & Kothari, A. (2023). Exploring communication by public health leaders and organizations during the pandemic: A content analysis of COVID-related tweets. *Canadian Journal of Public Health, 114*(4), 563–583. https://doi.org/10.17269/s41997-023-00783-4
29. Tagliacozzo, S., Albrecht, F., & Ganapati, N. E. (2023). Public agencies tweeting the COVID-19 pandemic: Cross-country comparison of must have and forgotten communication topics. *Frontiers in Communication, 8*, 1062241. https://doi.org/10.3389/fcomm.2023.1062241
30. Catalan-Matamoros, D., Prieto-Sanchez, I., & Langbecker, A. (2023). Crisis communication during COVID-19: English, French, Portuguese, and Spanish discourse of AstraZeneca vaccine and omicron variant on social media. *Vaccines, 11*(6), 1100. https://doi.org/10.3390/vaccines11061100
31. Finnish Institute For Health and Welfare. (2023). *Vaccines and coronavirus*. Finnish Institute For Health and Welfare. https://thl.fi/en/web/infectious-diseases-and-vaccinations/what-s-new/coronavirus-covid-19-latest-updates/vaccines-and-coronavirus
32. Browne, E. (2021). Fact Check: Have Finland, Sweden, Norway and Iceland "Banned" Moderna Vaccine? *Newsweek*. https://www.newsweek.com/fact-check-has-sweden-denmark-norway-iceland-banned-moderna-vaccine-1638563
33. Yle News. (2023). *THL takes Twitter break over disinformation concerns*. https://yle.fi/a/74-20013022

34. Chou, W.-Y.S., & Budenz, A. (2020). Considering emotion in COVID-19 vaccine communication: Addressing vaccine hesitancy and fostering vaccine confidence. *Health Communication, 35*(14), 1718–1722. https://doi.org/10.1080/10410236.2020.1838096
35. Hudson, A., & Montelpare, W. J. (2021). Predictors of vaccine hesitancy: Implications for COVID-19 public health messaging. *International Journal of Environmental Research and Public Health*. https://doi.org/10.3390/ijerph18158054
36. Tangcharoensathien, V., Calleja, N., Nguyen, T., Purnat, T., D'Agostino, M., Garcia-Saiso, S., Landry, M., Rashidian, A., Hamilton, C., AbdAllah, A., Ghiga, I., Hill, A., Hougendobler, D., Van Andel, J., Nunn, M., Brooks, I., Sacco, P. L., De Domenico, M., Mai, P., et al. (2020). Framework for managing the COVID-19 infodemic: Methods and results of an online, crowdsourced WHO technical consultation. *Journal of Medical Internet Research, 22*(6), e19659. https://doi.org/10.2196/19659
37. Siegrist, M. (2021). Trust and risk perception: A critical review of the literature. *Risk Analysis, 41*(3), 480–490. https://doi.org/10.1111/risa.13325
38. Kasperson, R. E., Renn, O., Slovic, P., Brown, H. S., Emel, J., Goble, R., Kasperson, J. X., & Ratick, S. (1988). The social amplification of risk: A conceptual framework. *Risk Analysis, 8*(2), 177–187. https://doi.org/10.1111/j.1539-6924.1988.tb01168.x
39. Kasperson, R. E., Webler, T., Ram, B., & Sutton, J. (2022). The social amplification of risk framework: New perspectives. *Risk Analysis, 42*(7), 1367–1380. https://doi.org/10.1111/risa.13926
40. Chong, M., & Choy, M. (2018). The social amplification of haze-related risks on the Internet. *Health Communication, 33*(1), 14–21. https://doi.org/10.1080/10410236.2016.1242031
41. Wirz, C. D., Xenos, M. A., Brossard, D., Scheufele, D., Chung, J. H., & Massarani, L. (2018). Rethinking social amplification of risk: Social media and Zika in three languages. *Risk Analysis, 38*(12), 2599–2624. https://doi.org/10.1111/risa.13228
42. Zhang, X. A., & Cozma, R. (2022). Risk sharing on Twitter: Social amplification and attenuation of risk in the early stages of the COVID-19 pandemic. *Computers in Human Behavior, 126*, 106983. https://doi.org/10.1016/j.chb.2021.106983
43. Strekalova, Y. A., & Krieger, J. L. (2017). Beyond words: Amplification of cancer risk communication on social media. *Journal of Health Communication, 22*(10), 849–857. https://doi.org/10.1080/10810730.2017.1367336
44. Hopfer, S., Fields, E. J., Lu, Y., Ramakrishnan, G., Grover, T., Bai, Q., Huang, Y., Li, C., & Mark, G. (2021). The social amplification and attenuation of COVID-19 risk perception shaping mask wearing behavior: A longitudinal twitter analysis. *PLoS ONE, 16*(9), e0257428.
45. Lee, E. W. J., Zheng, H., Goh, D.H.-L., Lee, C. S., & Theng, Y.-L. (2023). Examining COVID-19 Tweet diffusion using an integrated social amplification of risk and issue-attention cycle framework. *Health Communication*. https://doi.org/10.1080/10410236.2023.2170201
46. Kasperson, J. X., Kasperson, R. E., Pidgeon, N., & Slovic, P. (2013). The social amplification of risk: Assessing fifteen years of research and theory. In N. Pidgeon, J. E. Kasperson, & P. Slovic (Eds.), *The social amplification of risk* (pp. 13–47). Cambridge: Cambridge University Press.
47. Chung, I. J. (2011). Social amplification of risk in the Internet environment. *Risk Analysis, 31*(12), 1883–1896. https://doi.org/10.1111/j.1539-6924.2011.01623.x
48. Brown, A. (2021). Understanding the technical and societal relationship between shadowbanning and algorithmic bias. *Forbest*. https://www.forbes.com/sites/anniebrown/2021/10/27/understanding-the-technical-and-societal-relationship-between-shadowbanning-and-algorithmic-bias/?sh=184ad12d6296
49. Crockett, M. J. (2017). Moral outrage in the digital age. *Nature Human Behaviour, 1*(11), 769–771. https://doi.org/10.1038/s41562-017-0213-3
50. Barrie, C., & Ho, J. C. (2021). academictwitteR: An R package to access the Twitter Academic Research Product Track v2 API endpoint. *Journal of Open Source Software, 6*(62), 3272. https://doi.org/10.21105/joss.03272
51. Du, J., Xu, J., Song, H., Liu, X., & Tao, C. (2017). Optimization on machine learning based approaches for sentiment analysis on HPV vaccines related tweets. *Journal of Biomedical Semantics, 8*(1), 9. https://doi.org/10.1186/s13326-017-0120-6
52. Lindelöf, G., Aledavood, T., & Keller, B. (2022). Vaccine discourse on Twitter during the COVID-19 pandemic. arXiv Preprint arXiv:2207.11521.
53. Memon, S. A., & Carley, K. M. (2020). Characterizing COVID-19 misinformation communities using a novel Twitter dataset. CoRR, abs/2008.00791. https://arxiv.org/abs/2008.00791

54. Moffitt, J. D., King, C., & Carley, K. M. (2021). Hunting conspiracy theories during the COVID-19 pandemic. *Social Media + Society, 7*(3), 20563051211043212. https://doi.org/10.1177/20563051211043212
55. Hughes, B., Miller-Idriss, C., Piltch-Loeb, R., Goldberg, B., White, K., Criezis, M., & Savoia, E. (2021). Development of a codebook of online anti-vaccination rhetoric to manage COVID-19 vaccine misinformation. *International Journal of Environmental Research and Public Health*. https://doi.org/10.3390/ijerph18147556
56. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. CoRR, abs/1810.04805. http://arxiv.org/abs/1810.04805
57. Virtanen, A., Kanerva, J., Ilo, R., Luoma, J., Luotolahti, J., Salakoski, T., Ginter, F., & Pyysalo, S. (2019). *Multilingual is not enough: BERT for Finnish* (arXiv:1912.07076). arXiv. http://arxiv.org/abs/1912.07076
58. Yang, K.-C., Varol, O., Davis, C. A., Ferrara, E., Flammini, A., & Menczer, F. (2019). Arming the public with artificial intelligence to counter social bots. *Human Behavior and Emerging Technologies, 1*(1), 48–61. https://doi.org/10.1002/hbe2.115
59. Sayyadiharikandeh, M., Varol, O., Yang, K.-C., Flammini, A., & Menczer, F. (2020). Detection of novel social bots by ensembles of specialized classifiers. In *Proceedings of the 29th ACM international conference on information & knowledge management* (pp. 2725–2732). https://doi.org/10.1145/3340531.3412698
60. Yang, K.-C., Ferrara, E., & Menczer, F. (2022). Botometer 101: Social bot practicum for computational social scientists. *Journal of Computational Social Science, 5*(2), 1511–1528. https://doi.org/10.1007/s42001-022-00177-5
61. Unlu, A., Lac, T., Sawhney, N., & Tammi, T. *Unveiling the veiled threat: the impact of bots on COVID-19 health communication* (Under review)
62. Al-Taie, M. Z., & Kadry, S. (2017). *Python for graph and network analysis*. Berlin: Springer.
63. Kolaczyk, E. D., & Csárdi, G. (2020). *Statistical analysis of network data with R* (2nd ed.). Cham: Springer. https://doi.org/10.1007/978-3-030-44129-6
64. Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal*, *Complex Systems*, 1695. https://igraph.org
65. Hargittai, E., & Walejko, G. (2008). THE PARTICIPATION DIVIDE: Content creation and sharing in the digital age. *Information, Communication and Society, 11*(2), 239–256. https://doi.org/10.1080/13691180801946150
66. Bastos, M., & Mercea, D. (2018). The public accountability of social platforms: Lessons from a study on bots and trolls in the Brexit campaign. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 376*(2128), 20180003. https://doi.org/10.1098/rsta.2018.0003
67. Zhao, S., & Wu, X. (2021). From information exposure to protective behaviors: Investigating the underlying mechanism in COVID-19 outbreak using social amplification theory and extended parallel process model. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2021.631116
68. Helfers, A., & Ebersbach, M. (2023). The differential effects of a governmental debunking campaign concerning COVID-19 vaccination misinformation. *Journal of Communication in Healthcare, 16*(1), 113–121. https://doi.org/10.1080/17538068.2022.2047497
69. Bakir, V. (2005). Greenpeace v. Shell: Media exploitation and the Social Amplification of Risk Framework (SARF). *Journal of Risk Research, 8*(7–8), 679–691. https://doi.org/10.1080/13669870500166898