Information Processing and Management 58 (2021) 102732

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

Demographics and topics impact on the co-spread of COVID-19 misinformation and fact-checks on Twitter

Grégoire Burel [*], Tracie Farrell, Harith Alani

Knowledge Media Institute, The Open University, United Kingdom

ARTICLE INFO

Keywords:
COVID-19
Misinformation
Fact-checking
Social media
Demographics

ABSTRACT

Correcting misconceptions and false beliefs are important for injecting reliable information about COVID-19 into public discourse, but what impact does this have on the continued proliferation of misinforming claims? Fact-checking organisations produce content with the aim of reducing misinformation spread, but our knowledge of its impact on misinformation for particular topics and demographics is limited. In this article, we explore the relation between misinformation and fact-checking spread during the COVID-19 pandemic for different topics, user demographics and attributes. We specifically follow misinformation and fact-checks emerging from December 2019 until the 4th of January 2021 on Twitter. Using a combination of spread variance analysis, impulse response modelling and causal analysis, we highlight the bidirectional, weak causation spread behaviour between misinformation and fact-checks. Although we observe that fact-checks about COVID-19 are appearing fairly quickly after misinformation is circulated, its ability to reduce overall misinformation spread appears to be limited. This is especially visible for misinformation about conspiracy theories and the causes of the virus.

1. Introduction

Misinformation spreads faster than true information by exploiting strong emotions like fear, surprise and disgust (Brennen, Simon, Howard, & Nielsen, 2020; Vosoughi, Roy, & Aral, 2018). It is not surprising, therefore, that misinformation about COVID-19 is spreading faster than we can contain it (Brennen et al., 2020; Cinelli, Quattrociocchi, Galeazzi, Valensise, Brugnoli, Schmidt, Zola, Zollo, & Scala, 2020; Kouzy et al., 2020). This is worrying because misinformation can influence beliefs (Roozenbeek et al., 2020) and behaviour (Kouzy et al., 2020) in several ways that are detrimental to solving the crisis. For example, exposure to misinformation about COVID-19 has been shown to impact mental health and trigger misconceptions, resulting in poorer overall knowledge about COVID-19 and fewer preventative behaviours (Lee et al., 2020).

However, the impact of corrective information and its diffusion through large networks has received much less attention until now (Almaliki, 2019). During COVID-19, access to trustworthy corrective information is correlated with engaging in more protective behaviours, such as following government guidelines on social distancing, hygiene, or even vaccine acceptance (Lazarus et al., 2020). Understanding how, when and how much to correct misinformation online can help us save lives.

In this work, we study how misinformation and fact-checks spread together, to give us insights on the impact of fact-checking in reducing misinformation spread about specific topics. This allows us to distinguish patterns and timings that are significant in the spread of misinformation and where fact-checking efforts should be focused. This knowledge can then be used to empower fact-checkers to target specific topics and user demographics with the potential to improve the effect of fact-checking on misinformation

[*] Corresponding author.
E-mail addresses: gregoire.burel@open.ac.uk (G. Burel), tracie.farrell@open.ac.uk (T. Farrell), harith.alani@open.ac.uk (H. Alani).

https://doi.org/10.1016/j.ipm.2021.102732
Received 31 January 2021; Received in revised form 19 July 2021; Accepted 12 August 2021
Available online 30 August 2021

spread. Although previous work has shown that fact-checking did appear to impact the spread of misinformation over time (Burel, Farrell, Mensio, Khare, & Alani, 2020), it remains unclear if such a behaviour generalises over a longer period and a larger amount of data.

In this article, we extend our previous work in understanding how fact-checks and misinformation spread during the COVID-19 pandemic (Burel et al., 2020). We analyse the co-spread of misinformation and fact-checks on a larger dataset that covers more than one year of the pandemic and related fact-checks. Our extended dataset is nearly 17 times larger than our previous dataset of 358,776 Twitter posts mentioning misinformation or fact-checks related to COVID-19. Our new analysis investigates how misinformation and fact-checks spread, according to different COVID-19-related topics, and user demographics and attributes.

Our comparison focuses on the diffusion of 7370 misinformation and 9151 fact-checking URLs about COVID-19 on Twitter from early December 2019 to early January 2021 in order to understand the spread of misinformation and fact-checks over time. We also use 6 of the 7 COVID-19-related topics identified by the Poynter Institute's Corona Virus Facts Alliance.[1] and extract demographics information from users (i.e., user type and gender) for increasing the granularity of our study.

In the first part of our work, we analyse misinformation and fact-checking spread by aligning individual URL spreads and analysing how specific misinformation spreads after their initial appearance (relative spread analysis). In the second part of our study, we analyse how fact-checking impacts misinformation spread by analysing the intrinsic relation between fact-checking and misinformation spread for COVID-19 topics.

In this paper, we aim to answer the following main research questions:

1. Do COVID-19 misinformation and fact-checks spread similarly on Twitter?
2. Do these sharing patterns differ with topics, demographics, and relative time?
3. How does the spread of fact-checks affects the diffusion of misinformation about COVID-19 for different topics?

2. Related work

In this section, we discuss some of the propositions that researchers have made regarding the spread of misinformation and fact-checks on social networks. We highlight the complexity of establishing the impact of fact-checks on misinformation sharing, to which our study contributes. We also discuss some of the approaches used for extracting user demographics on social media.

2.1. Misinformation spread analysis

WHO Director-General Tedros Adhanom Ghebreyesus referred to misinformation about COVID-19 as an "infodemic" that could "undermine" the global response to the pandemic.[2] This metaphor is already aligned with existing research on misinformation spread that uses epidemiological modelling to help represent the information ecosystem (and misinformation as a virus) (Almaliki, 2019; Jin, Dougherty, Saraf, Cao, & Ramakrishnan, 2013; Jin et al., 2014; Roozenbeek et al., 2020). Additional features, like weighted values for influential users (Tong & Du, 2019), the existence of debunkers and the dynamics of opinion evolution (Saxena et al., 2020) have added to the granularity of models. "Network epidemiology" also considers the adaptive features of social networks that may interfere with the spread of misinformation (Masuda & Holme, 2017). Previous work has shown that chains or groups of nodes may accelerate the spread of misinformation (Sarkar, Guo, & Shakarian, 2019), and that individuals can be exposed to and share misinformation across platforms (Allgaier & Svalastog, 2015; Xian, Yang, Pan, Wang, & Wang, 2019). During COVID-19, researchers have observed that misinformation diffusion patterns can be quite different across platforms, which may have to do with the cultural and demographic make-up of the platform, as well as its affordances (Cinelli et al., 2020).

Until now, most research on misinformation spread was centred on the United States, particularly the 2016 US Presidential Election (Roozenbeek et al., 2020). However, we now understand more about how subject matter (topology) impacts misinformation spread (Allgaier & Svalastog, 2015; Harman, 2020; Vaezi & Javanmard, 2020; Xie et al., 2020), with some topic/audience interdependencies increasing the spread of misinformation. This is, perhaps, related to cultural norms, experiences or values (Farrell, Piccolo, Perfumi, Alani, & Mensio, 2019). During COVID-19, we have revisited this challenge, with many more studies emerging that address some of the cognitive aspects of susceptibility to misinformation about the pandemic and its impacts (Roozenbeek et al., 2020). For example, trust in science has been correlated with belief in prevailing scientific explanations about COVID-19 and may provide a "defence" against misinforming alternative explanations (Agley & Xiao, 2021). Perhaps more crucially, accepting what you do not know (intellectual humility) is also associated with lower susceptibility to misinformation about COVID-19 (Koetke, Schumann, & Porter, 2021).

One possible explanation for this is that COVID-19 is an evolving, ambiguous event in which trust in science and intellectual humility are assets. Accessibility to quality information has been limited at times (Cuan-Baltazar, Muñoz Perez, Robledo-Vega, Pérez-Zepeda, & Soto-Vega, 2020; Kouzy et al., 2020), which can be distressing for the public. It is not surprising that the biggest impact of misinformation appears to happen within a short time span from the initial circulation (Starbird, Dailey, Mohamed, Lee, & Spiro, 2018). Spikes are observed during times of conflict and war (Lewandowsky, Stritzke, Freund, Oberauer, & Krueger, 2013), political events (Kuklinski, Quirk, Jerit, Schwieder, & Rich, 2000), and other types of crisis events (like COVID-19), when the public desperately needs information about where to go or what to do next (Starbird et al., 2018). Health crises, in general, are often

[1] Corona Virus Facts Alliance, https://www.poynter.org/coronavirusfactsalliance.

[2] WHO: Immunising the public against misinformation, https://www.who.int/news-room/feature-stories/detail/immunising-the-public-against-misinformation.

accompanied by misinformation (Vaezi & Javanmard, 2020; Xie et al., 2020). The scale of misinformation about COVID-19 has been tremendous, further motivating the necessity of early intervention (Kouzy et al., 2020).

The majority of works about misinformation spread before COVID-19 tended to focus on early intervention and removal of misinformation, with the assumption that exposure to misinformation will impact behaviour (Almaliki, 2019). The behaviours that misinformation could impact during COVID-19 include following government advice on social distancing and hygiene, self-isolating when necessary and (more recently) accepting a vaccine (Cinelli et al., 2020). These are global concerns because they impact how and when we will get the pandemic under control. What is relevant to behaviour right now appears to be access to trustworthy information. In a survey of 13,426 people living in 19 countries, researchers found that respondents reporting higher levels of trust in information from government sources were more likely to accept a vaccine (Lazarus et al., 2020). Understanding how to disseminate trustworthy information is, therefore, a timely contribution to fighting COVID-19.

In terms of what kinds of misinformation are being spread about COVID-19, topics include conspiracy theories, misinformation about the symptoms, causes or spread of COVID-19, misinformation about cures and misinformation about the authorities involved in managing the pandemic (Brennen et al., 2020). Often, misinformation is relatively easy to debunk. For example, a study of visuals accompanying misinformation about COVID-19, researchers found that mislabelled images were more common than images that were manipulated (Brennen, Simon, & Nielsen, 2021). Still, despite access to factual information that directly refutes a specific claim, misinformation about COVID-19 continues to proliferate. Our previous work indicated that the volume of misinformation may be simply too large to extinguish and that more work was needed to amplify and extend the impact of fact-checking (Burel et al., 2020). In addition, certain health beliefs, such as those underpinned by conspiracy theories may be difficult to break through (van der Linden, Roozenbeek, & Compton, 2020). Conspiracy theories are part of a sense-making activity, in which people want to explain significant events that do not, as of yet, have a satisfactory explanation (Douglas et al., 2019). The cause of a novel coronavirus and its future impacts will likely remain ambiguous. Understanding more about the interaction between misinformation and corrective information is crucial for getting misinformation under control, and managing misconceptions and beliefs during COVID-19.

Existing research on misinformation at scale has focused mostly on analysing misinformation spread alone without much focus on whether fact-checking information impacts the spread of specific claims. In this paper, we seek to analyse the co-spread of misinformation and fact-checking to compare and contrast their diffusion patterns during COVID-19, and further interrogate some of the topological patterns that can be observed in how certain types of claims or fact-checks spread.

2.2. *Fact-checking information spread analysis*

Before COVID-19, subjectivity in fact-checking assessment, overemphasis of fact-checking in the United States, and a lack of clarity around correcting beliefs were continued challenges (Nieminen & Rapeli, 2019). As COVID-19 is a global problem, we can expect our knowledge and practices to mature.

Researchers have investigated the usefulness of fact-checking from a variety of perspectives, including attitudes toward fact-checking (Nyhan & Reifler, 2015), impact on knowledge and behaviour (Barrera, Guriev, Henry, & Zhuravskaya, 2020), or perception of events and people (Swire, Berinsky, Lewandowsky, & Ecker, 2017). The number of corrections appears to be an important feature in outweighing the impact of misinformation (Aird, Ecker, Swire, Berinsky, & Lewandowsky, 2018; Bode, Vraga, & Tully, 2020; Starbird et al., 2018). Researchers have been able to establish tipping points, in terms of the number of fact-checks that might be theoretically necessary to extinguish a misinforming claim (Tambuscio, Ruffo, Flammini, & Menczer, 2015). Later, such models were transferred to real-world datasets, like Twitter and Weibo (Kim, Tabibian, Oh, Schölkopf, & Gomez-Rodriguez, 2018), to model how the network could be mobilised to spread corrective information effectively. Still, these models are meant to predict how future fact-checks may diffuse and not to estimate existing causal relationships.

Fact-checks assess claims for accuracy (Vlachos & Riedel, 2014), representing a new category of information to be observed (Jiang & Wilson, 2018). The different diffusion patterns of real and false information (Vosoughi et al., 2018) suggest that fact-checks will also diffuse differently in a network. In keeping with the metaphor of misinformation as a virus, fact-checking can be positioned as both a protective and therapeutic practice of “inoculation” against misinformation. With the former, the focus is on keeping well-informed those still not “infected” with misinformation. With the latter, individuals who already believe misinformation are targeted with fact-checking to help slowly build more trust and positive attitudes toward healthy behaviour (van der Linden et al., 2020). Understanding the different functions fact-checking can have in a network help us to question the “knowledge deficit model" as being the full explanation for why fact-checking is necessary. Not only does fact-checking communicate informative content, it also communicates values and intentions for accountability (Krause, Freiling, Beets, & Brossard, 2020).

There are also other types of beliefs and behaviours that govern how we accept authoritative intervention in our health, and fact-checking can help facilitate this. First, fact-checking about COVID-19 is often about explaining science and medicine. For example, in a study on participants in Turkey and in the UK, vaccine acceptance was correlated with belief in the current scientific explanation around the natural origin of COVID-19 (Salali & Uysal, 2020). While the two are not topically connected, they share an attribute of trust in science. Second, quality fact-checking can help the public to understand ambiguities in an evolving event like COVID-19. Studies have shown that highlighting uncertainty is important for building trust in fact-checking (Krause et al., 2020).

However, there are some concerns about how little individuals are exposed to fact-checks. In one cross-sectional online study spanning over 35 countries, fact-checking about COVID-19 was spreading much slower than misinformation, on a sublinear trend. More worryingly, misinformation was permeating more deeply into countries with a lower economic stability, who already appear to be most impacted by the pandemic (Cha et al., 2020). The link between offline and online experiences during COVID-19 may indicate that similar trust factors are involved. Once again, COVID-19 motivates the need for early intervention in misinformation as a matter of public health.

The best delivery of fact-checking, and by whom, is also a critical subject now, including how best to use high profile users like celebrities and politicians, or even just users with large and influential networks. During COVID-19, researchers found that platform users were not only more prolific in fact-checking than professionals, they were more efficient, tended to have greater reach and some used concrete, verifiable evidence. Unfortunately, evidence was not always complete, and often it was not information provided by professional fact-checkers, but rather other sources (Micallef, He, Kumar, Ahamad, & Memon, 2020). Motivations to share fact-checks are also unclear. Previous work from political contexts in the United States showed that motivations to share fact-checks could be linked to age, ideology, and political behaviours (Amazeen, Vargo, & Hopp, 2019), with more politically liberal users as the primary consumer of fact-checking materials (Robertson, Mourão, & Thorson, 2020). During COVID-19, concern for health and safety was correlated with a higher acceptance of fact-checking in the United States, with greater effect sizes for republicans (Rich, Milden, & Wagner, 2020). Still, there are many open questions about how to engage users and fact-checking organisations together to help fight misinforming claims as they emerge.

2.3. Demographics extraction and analysis

Understanding how misinformation and fact-checks spread for specific demographics, topics and user groups may be critical for the creation of targeted approaches for reducing online misinformation since some demographic indicators are correlated with misinformation and fact-check sharing (Amazeen et al., 2019; Bedard & Schoenthaler, 2018; Guess, Nagler, & Tucker, 2019; Rampersad & Althiyabi, 2020). As a result, having a reliable way for identifying social media demographics is key for understanding if specific spread patterns disproportionately affect particular user demographics.

Accessing the demographics behind particular social media accounts is a complex matter that may be hindered by platform privacy rules or the decision by account holders to withdraw information or keep it private. In addition, users may decide to represent themselves differently to whom they really are, leading to their inaccurate representation (Kendall, 1998). In this context, most studies about user demographics and misinformation spread have been limited to small surveys and user groups (Bedard & Schoenthaler, 2018; Rampersad & Althiyabi, 2020).

In order to deal with such issues, researchers have been investigating methods for automatically identifying various social media demographics and user account traits with various degrees of success and accuracy. Areas of research such as user age (Brandt et al., 2020; Giorgi, Lynn, Matz, Ungar, & Schwartz, 2019; Wang et al., 2019), gender (Brandt et al., 2020; Culotta & Cutler, 2016; Sloan et al., 2013; Wang et al., 2019; Yang, Al-Garadi, Love, Perrone, & Sarker, 2021), language (Sloan et al., 2013; Wang et al., 2019), account type (e.g., identifying if a particular account is an individual, organisation or bot) (Beskow & Carley, 2018; Gürlek, 2021; Rodríguez-Ruiz, Mata-Sánchez, Monroy, Loyola-Gonzalez, & López-Cuevas, 2020; Wang et al., 2019), income range (Giorgi et al., 2019; Wang et al., 2019) and education level (Giorgi et al., 2019) identification have all received a large amount of attention particularly for Twitter data.

The various works relating to the identification of the gender of individuals rely mostly on profile image analysis, statistical information about user first names usage, profile text and user language analysis (Brandt et al., 2020; Giorgi et al., 2019; Wang et al., 2019) as well as historical posts (Yang et al., 2021). The combination of these approaches tends to lead to highly accurate results ($F_1 > 90\%$), but they all consider gender as a purely binary classification problem. It is worth noting that, while demographic information on gender may allow for easier comparison of results across different studies, gender may not be a purely binary classification. Some marginalised groups may, therefore, "resist classification and capture" when studying the impact of gender on behaviour (Ruberg & Ruelos, 2020). There is also evidence, primarily from the domain of online gaming, that individuals do engage in the act of "gender switching", in which they wish to appear online as a gender with which they do not identify personally (Kendall, 1998). These limitations should be taken into consideration whenever interpreting online behaviour at scale from a gendered perspective. Standout accurate approaches for gender identification include: (1) the multi-task Deep Neural Network (DNN) M3 model proposed by Wang et al. (2019) that uses usernames, profile images, screen names and profile description with an F_1 of 91.8%, and; (2) the SVM-based meta-classifier M3 extension proposed by Yang et al. (2021) that improves over the M3 model ($F_1 = 94.7\%$) with the drawback of requiring user historical posts for inference rather than user profiles alone. This requirement makes the model of Yang et al. (2021) less useable in practice as it requires a large amount of additional data compared to the M3 model.

The identification of account types mostly derives from the need to identify fraudulent or non-human accounts (bots) (Beskow & Carley, 2018; Rodríguez-Ruiz et al., 2020) with some work targeting the differentiation between individual and organisation accounts (Gürlek, 2021; Wang et al., 2019). Most of these approaches have shown to be highly accurate with accuracy typically higher than 85%.

Results in identifying other types of user demographics on Twitter such as user age (Brandt et al., 2020; Giorgi et al., 2019; Wang et al., 2019), language (Sloan et al., 2013; Wang et al., 2019), income range (Giorgi et al., 2019; Wang et al., 2019) and education level (Giorgi et al., 2019) have been less accurate, meaning that their usage in quantitative user behaviour studies may lead to misleading conclusions. For example, although the user's language is usually relatively easy to identify based on user posts, the limited length of Twitter posts shows some important accuracy issues when typical approaches are applied (Sloan et al., 2013).

In this work, we decide to use only models with higher than 85% accuracy. As a result, our study only analyses user gender and account type. For extracting such information, we decide to use the models proposed by Wang et al. (2019) since they are readily available and provide high accuracy for binary gender classification ($F_1 = 91.5\%$) and account type identification ($F_1 = 89.8\%$). For the gender identification model, we decide to use the M3 model over the model proposed by Yang et al. (2021) since it is accurate enough, readily available[3] and does not require collecting historical posts for each user identified during our analyses.

3. Co-spread of COVID-19 misinformation and related fact-checks

Existing research shows a gap in understanding the relation and interaction between the corrective information propagated by fact-checking practices and misinformation spread. Although in our previous work, we already showed how misinformation and fact-checks influence each other in the context of the COVID-19 pandemic (Burel et al., 2020), it remains unclear if such behaviour varies for different demographics and topics. Understanding such relations is necessary for improving the effectiveness of fact-checking campaigns and for identifying vulnerable demographics or more sensitive topics. We conduct an analysis on the co-spread of fact-check information and misinformation on Twitter based on the sharing of misinforming URLs that were collected from Poynter Institute's International Fact-Checking Network (IFCN).[4] The data was collected up to the 4th of January 2021. Our approach for collecting and processing data for the purpose of our analysis can be summarised in Fig. 1. Our approach extends previous work on co-spread analysis (Burel et al., 2020). First, we collect Twitter data by looking for the appearance of misinforming URLs collected from the IFCN database. Then, the demographics (gender and account type) associated with the authors of each post are extracted using automatic methods. By combining the posts obtained from Twitter, metadata from the URLS obtained from the IFCN and user demographics data, we generate the two different datasets that we use for our analyses. Each dataset consists of the collected misinformation and fact-checking data aggregated for three different time periods from the initial emergence of a given misinforming URL or fact-check URL (relative level analysis). This approach allows for a better analysis of spread with the ability to observe spread variation for multiple topics and demographics.

For our work, we perform multiple analyses to investigate how fact-checks and misinformation spread behaviour differ for varying topics and user demographics. This analysis allows the identification of significant relations between misinformation spread, fact-check information, topics and demographics, which can be used for designing better methods for spreading fact-checking information on social media. Besides this analysis, weak causation and impulse response analyses are also performed between fact-checks and misinformation in order to identify if fact-check information diffusion impacts misinformation spread and if such behaviour varies for different topics.

3.1. Datasets and data collection

Improving upon our previous misinformation and fact-check co-spread analysis (Burel et al., 2020), we create a new dataset that contains misinformation and their fact-checks, associated topics, and demographics about who shares such information. We focus our work on Twitter due to its popularity and its accessibility. We rely on COVID-19-related reports from legitimate fact-checking websites that identify misinforming content by their URLs and search for the occurrences of these URLs in user posts on Twitter. We update our data collection method so that an increasing number of tweets is retrieved for each misinforming and fact-check URL compared to our previous data collection approach (Burel et al., 2020). As a result, we obtain a more complete representation of how misinforming and fact-check URLs spread on Twitter. The URL-based data collection approach has multiple advantages: first, it ensures that experts have assessed the accuracy of claims. Second, we can look specifically at pairs of misinformation and their corresponding fact-checks instead of disconnected true and false information as in most related literature. And finally, the appearance of a known misinforming URL in a Tweet is a more robust indicator of sharing than searching for textual claims. Although the inclusion of a misinforming URL in a Tweet does not necessarily reflect endorsement, it has been found to increase the proliferation of misinformation nevertheless. See Section 6 for a further discussion on this issue.

3.1.1. Fact-check URLs and topics dataset

The dataset of fact-checks and misinforming URLs comes from the Poynter Institute's International Fact-Checking Network (IFCN). The dataset is extracted from Poynter's COVID-19 specific fact-check alliance database[5] rather than keyword-filtered URLs from the Data Commons ClaimReview public feed.[6] This approach leads to more precise URLs and, therefore, to more precise results compared to our previous approach (Burel et al., 2020). The Poynter database aggregates fact-checking reviews from more than 100 verified fact-checking websites around the world about issues surrounding COVID-19, and consists of fact-checking URLs, the reviewed URLs, and the fact-checker rating (e.g., False, True).

The COVID-19 reviews aggregated by the Poynter Institute follow the standard `ClaimReview` schema,[7] which is defined specifically for the purpose of annotating reviews of claims analysed by fact-checking organisations. Each review generally consists of two URLs (the fact-checker article that review a claim and one or more URLs linking to the reviewed claim or content) and a rating indicating the validity of the claim. In addition to this information, the database also contains topical information about the reviewed claims (e.g., virus origins, cures, symptoms).

[3] At the time of writing, the link pointing to the models proposed by Yang et al. (2021) redirects to an empty code repository whereas the M3 model implementation is publicly available.

[4] IFCN, https://ifcncodeofprinciples.poynter.org.

[5] Corona Virus Facts Alliance, https://www.poynter.org/coronavirusfactsalliance.

[6] ClaimReview Public Feed, https://www.datacommons.org/factcheck/download.

[7] ClaimReview Schema, https://schema.org/ClaimReview.

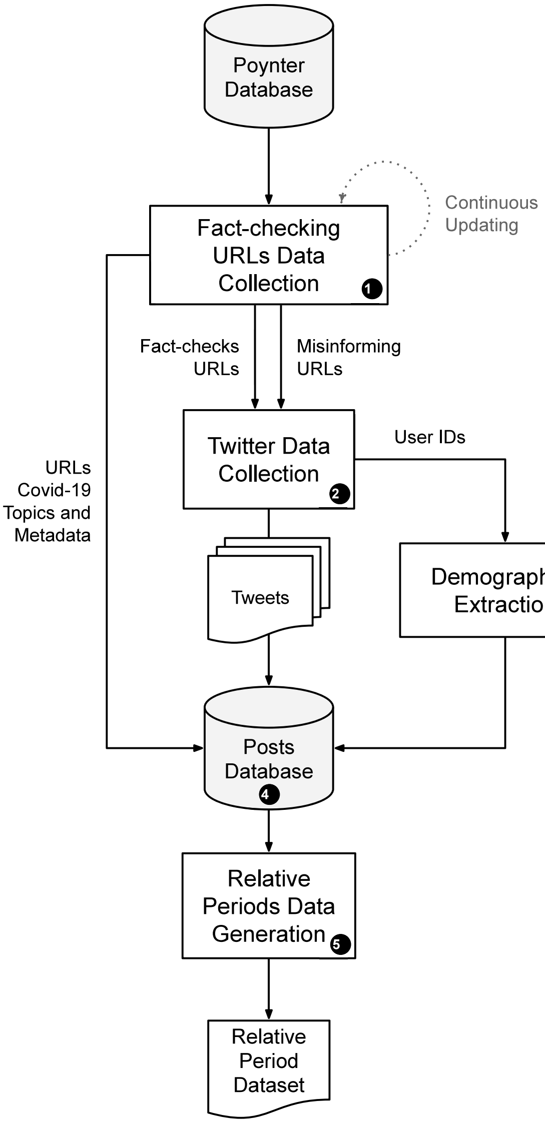

Fig. 1. Data collection and processing pipeline for generating the analysed datasets.

For the purpose of the COVID-19 specific fact-check alliance database, the following 7 topics are identified in relation to the pandemic[8]:

1. *Authorities*: Information relating to government or authorities communication and general involvement during the COVID-19 pandemic (e.g., crime, government, aid, lockdown).
2. *Causes*: Information about the virus causes and outbreaks (e.g., China, animals).
3. *Conspiracy theories*: COVID-19-related conspiracy theories (e.g., 5G, biological weapon).
4. *Cures*: Information about potential virus cures (e.g., vaccines, hydroxychloroquine, bleach).
5. *Spread*: Information relating to the spread of COVID-19 (e.g., travel, animals).
6. *Symptoms*: Information relating to symptoms and symptomatic treatments of COVID-19 (e.g., cough, sore throat).
7. *Other*: Any topic that does not fit the aforementioned categories directly. In this paper, we omit the *other* category since it does not denote any specific COVID-19 topics.

[8] The IFCN database does not give any explicit description of each category. The description of each category is derived based on the claims in each topic.

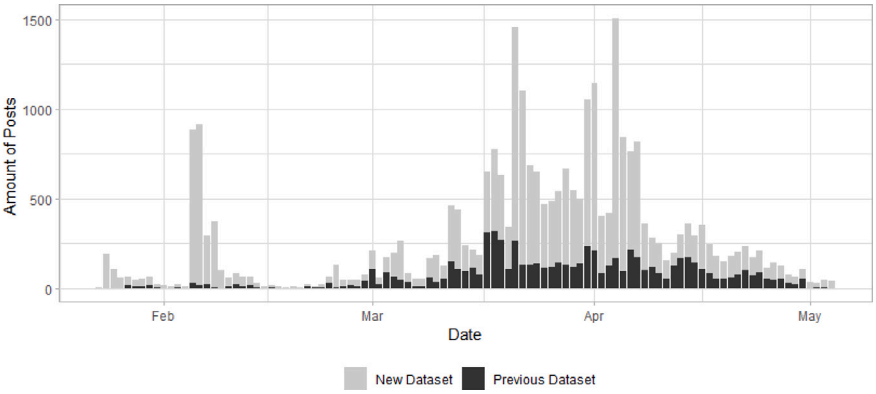

Fig. 2. Amount of collected Twitter posts for the URLs appearing in both the new and previous dataset (Burel et al., 2020) during the 1st December 2019 and the 4th of May 2020 period.

Given that different validity labels are used by different fact-checkers in their `ClaimReview` annotations, we normalise the data so that each review is associated to a common credibility score that is then used for distinguishing misinformation from trustworthy claims. The ratings are normalised between $[-1;+1]$ depending on their credibility using the rules proposed in Mensio and Alani (2019). Using these ratings, we select only misinforming URLs (ratings ≤ 0). Although different levels of misinformation exist (e.g., manipulation, misleading information, forgery), we focus our investigation on any type of misinformation in order to simplify the analysis.

The claims that cannot be normalised effectively, such as links to incomplete URLs or URL fields that contain free form text, are discarded from our final dataset.

The final URL dataset includes fact-checks published until the 4th of January 2021, with a total of 7370 distinct misinforming URLs and 9151 fact-check URLs. For the 7 topics described above, the distribution of URLs obtained are as follows: (1) *Authorities*: 1692; (2) *Causes*: 203; (3) *Conspiracy Theory*: 1280; (4) *Cure*: 1220; (5) *Other*: 2489; (6) *Spread*: 746, and; (7) *Symptoms*: 119.

3.1.2. Twitter dataset

Using the misinforming and fact-check URLs, we create the Twitter dataset by searching their occurrences on Twitter a posteriori using a crawler based on the TWINT Intelligence Tool.[9] This approach differs from our previous work, where we adapted an existing Twitter Hashtag crawler based on Twitter's mobile interface (Burel et al., 2020). As previously highlighted, the new approach yields a much higher number of tweets leading to more refined analysis results.

The difference in the amount of posts retrieved for the URLs appearing both in our previous work (Burel et al., 2020) and the updated dataset for the period between the 1st December 2019 and the 4th of May 2020 is displayed on Fig. 2. Overall, 67% more data was collected for that period compared to the data previously obtained. The higher amount of data can be explained by the higher amount of seed fact-check URLs used for the new analysis (75% more URLs), the way URLs were selected and the updated data collection approach.

Out of the total of 16,521 seed URLs, we found posts for only 14,706 distinct URLs, giving a total of 358,776 posts. On average, there are 46.3 posts for each URL ($\sigma = 574.47$, $min = 1$, $max = 29{,}141$).

Fig. 3 shows the cumulative amount of shared misinforming and fact-check URLs shared over time. The figure shows that misinformation and fact-checking URLs spread started increasing from mid-January 2020.

3.1.3. Demographics dataset

To understand the impact of misinformation and fact-checks on COVID-19-related misinformation, we consider some demographic information about the users who share the relevant information on Twitter. We are interested in understanding how gender and user type (i.e., organisations vs. individual users) correlate with different spreading patterns of misinformation and fact-checks. We only focus on those demographics since the accuracy of automatic demographic extraction methods is significantly lower for other types of demographics (Section 2.3).

Since accessing demographic data of user profiles is not supported by the Twitter API, we extract this information using the machine learning models developed by Wang et al. (2019). We use these models due to their readily available implementation and their high accuracy for gender and account type identification. These models also do not require historical user timelines for inferring demographics.

The models proposed by Wang et al. (2019) identify user gender as a binary classification task (male or female only), if a user account represents an individual or an organisation (e.g., a company, institution), their language, as well as user age group (i.e., '19–29', '30–39', '≤18', '≥40'). Their approach uses Twitter profile description and image to infer the aforementioned information.

[9] TWINT Intelligence Tool, https://github.com/twintproject/twint.

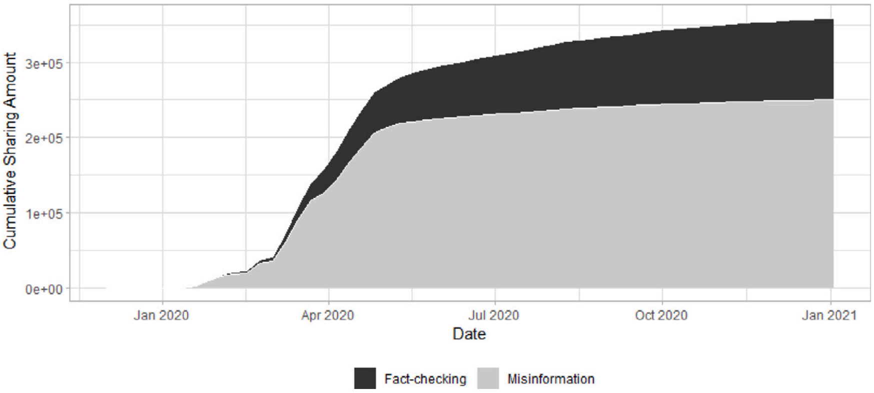

Fig. 3. Stacked cumulative amount of shared misinforming and corrective URLs over time.

Their model reports an F_1 of 91.5% for determining user gender and an average F_1 of 89.8% for identifying if an account is an organisation or an individual. The result for determining age shows a lower F_1 of around 42.5%. For languages, performance varies significantly by language from 28% F_1 for Bosnian to 73% for Slovenian and Welsh. As previously discussed (Section 2.3), we leave out language information as well as user age group due to their low accuracy. It is also important to note that Wang et al. (2019) model, like other existing automatic approaches, does not consider non-binary gender representation and may, therefore, misclassify marginalised user groups. Since we only use models with high precision for predicting user gender and account type, the demographic analysis of COVID-19-related misinformation and fact-check spread should not suffer significantly from the potential misclassification errors from these models. Nevertheless, we perform some error analysis in order to understand the potential impact of misclassified demographic indicators (Section 4.2).

3.2. *Relative periods generation*

To understand sharing behaviour independently from when each URL has been initially shared (relative period analysis), we normalise the dates for each analysed URL. First, we identify the first occurrence of each URL and then obtain the number of times it has been shared per day for each day following its initial appearance. Through this alignment method, we aim to study how misinformation and fact-checks spread over time independently from when they were posted.

In order to analyse how misinformation and fact-checks spread over time, we divide the data into three different time periods: *initial*, *early* and *late*. We decided to use three time periods in order to keep our analysis manageable and based on how misinformation typically spreads (Starbird et al., 2018).

The identification of the periods is done using segmented linear regression (Muggeo, 2003) on the daily aggregated curve containing all the shared URLs (i.e., misinforming and fact-checking URLs). This method applies an iterative procedure for identifying a fixed number of break points (in our case 2) by fitting multiple linear regressions to a curve (in our case the daily amount of shared misinformation and fact-checks URLs over time). Although, in principle, individual URLs may have different *initial*, *early* and *late* periods, we decided to aggregate the periods across all the analysed URLs so they can be compared more easily.

Using this method, the *initial* time period is specified as any URL shares happening within the first 3 days after its first occurrence. The *early* period corresponds to shares between day 4 and day 10. Finally, the *late* period is for any shares happening after 10 days. These periods are slightly different to the ones obtained in our previous work (Burel et al., 2020), where the *initial* period finished on day 2 and the *early* period finished at 14 days.

3.3. *Final relative period dataset*

As displayed in Fig. 1, we integrated multiple types of data in the final dataset used for generating the analysed periods. The final database used for our analysis consists of the integrated fact-checking and misinforming URLs, their topics, the Twitter posts mentioning them and the demographics of the users sharing such URLs tuned to the relative time periods described in the previous section.

4. Multivariate spread variance analysis

For the first part of the analysis, we identify the different patterns of appearance of misinformation and fact-check URLs over varying periods of time for all the shared URLs and for different topics and demographic groups. This analysis is done using the one-way Multivariate ANalysis Of VAriance (MANOVA) and the one way ANalysis Of VAriance (ANOVA) methods. Depending on the group analysed, Dunn post-hocs tests are also performed in order to better understand where spread patterns differ within different sub-groups. This approach allows us to determine if there are significant differences in information spread between the fact-checking information and misinformation groups in each *initial*, *early* and *late* period.

4.1. Experimental setup

MANOVA and ANOVA rely on the definition of independent variables and dependent variables. For the general analysis, the amount of information spread is the dependent variable whereas each information type (i.e., misinformation and fact-check information) is an independent variable. When dealing with specific demographic groups or topics, the amount of spread for misinformation or fact-checks is also the dependent variable, whereas each dimension of a demographic group (i.e., gender and account type) or topic is an independent variable. For instance, the relation between the target population gender (independent variable) and the amount of misinformation spread (dependent variables) for the *initial*, *early* and *late* periods can be analysed by grouping the information spread data according to the different gender groups (male and female).

Since our data does not follow all the assumptions required for the standard ANOVA and MANOVA methods (i.e., multicollinearity, normality and homogeneity), we use non-parametric versions of MANOVA and ANOVA for the analysis, using F-approximations permutation tests. The F-approximation of ANOVA's test, as well as Wilks' Lambda Type Statistic, are obtained with their p-value and the associated permutation test p-value.

Our analysis is divided into two or three different parts depending on the groups analysed: (1) First, a non-parametric MANOVA analysis is performed for identifying if there are differences in spread between the different periods and information types, then; (2) Non-parametric ANOVA analysis is then performed if the MANOVA results are significant for each individual time period for determining in which sub-period (i.e., *initial*, *early* and *late*) the spreading patterns differ, finally, when dealing with larger groups (e.g., individual topics); (3) A Bonferroni adjusted Dunn test is performed for each significant ANOVA result for identifying pairwise spread differences between sub-group (e.g., spread difference between individual topic groups in a particular time period).

For the non-parametric ANOVA analysis, the Kruskal–Wallis test is used and the p-values were adjusted using a Bonferroni correction (since multiple dependent variables are analysed). Significant results mean that the behaviour of corrective information and misinformation are significantly different, whereas a non-significant result means that the distribution of spread for each time period is non-significant. When more than two sub-groups were analysed (e.g., individual topics), a Bonferroni adjusted Dunn test is used for post-hoc analysis in order to determine where spread patterns differ within different sub-groups. A significant result means a difference in spread for a given period and sub-group pair (e.g., male and female for the *early* period).

4.2. Analysed groups

Besides analysing how misinformation and fact-check spread differs over time (general analysis), we also try to understand how misinformation and fact-check propagation differs for the 6 of the 7 topics defined in Section 3.1.1 (we leave out the *other* category since it is not well defined compared to the other COVID-19-related topics) and the various demographics (Section 3.1.3) over time (individual analysis).

When analysing the topic spread of the misinforming and fact-check URLs, we remove all the posts belonging to the *other* topic. As a result, 262,044 posts are kept when performing the topic analyses (74% of the full dataset is used). For the gender analyses, the posts belonging to the users accounts that are identified to be about organisations are removed. As a result, we are left with 323,655 posts when analysing gender (90% of the full dataset is used).

For the global analysis, we use misinformation and fact-checks as the independent variable, whereas both the type of information and misinformation, and the demographics or topics, are used together as the independent variable for the individual analysis.

Although the models we use for extracting gender and user account types are highly accurate (with and $F_1 = 91.5\%$ for determining user gender, and an average $F_1 = 89.8\%$ for identifying user account types), we check if their error rate can lead to different observations by changing randomly the classifications results obtained for the gender and account types classifiers according to their reported F_1 score. This means that for the gender classifier, we randomly sample 8.5% of both user groups and reverse their gender (i.e., 8.5% of the female users are changed to male and 8.5% of the male users are changed to female). The same approach is used for the account type, except that the sample rate is 10.2%. For each group, we perform the operation 10 times and report the mean of the significance scores for our analyses. This helps us to better understand the maximal potential impact (i.e., worst case scenario) of the classification errors for each demographic classifier. We use this approach as we do not have manually annotated user profiles with demographics in our dataset. Such data would be necessary for directly evaluating the accuracy and associated errors of the used pre-trained demographic models on our dataset.

The different analyses, groups and research questions can be summarised as follows:

- *Global analysis*:
    1. Does misinformation and fact-checks spread vary? (MANOVA)
    2. If there is a variation, what time periods differ? (ANOVAs)
- *Individual analyses*:
    - *Topics*:
        1. Does information spread vary depending on topic and information type? (MANOVA)
        2. If there is a variation, what are the topics that differ in spread? What topic spreads the most misinformation and fact-checks? (ANOVAs)
        3. Which sub-topics exhibit different spread patterns? (post-hoc)

- *Account type*:
  1. Does information spread vary depending on account type (i.e., *organisation* or *individual*) and information type? (MANOVA)
  2. If there is a variation, what time periods and information types differ? What account type spreads the most misinformation and fact-checks? (ANOVAs)
  3. Which account and information type exhibit different spread patterns? (post-hoc)
- *Gender*:
  1. Does information spread vary depending on gender (i.e, *male* or *female*) and information type? (MANOVA)
  2. If there is a variation, what time periods and information types differ? What gender spreads the most misinformation and fact-checks? (ANOVAs)
  3. Which gender and information types exhibit different spread patterns? (post-hoc)

In the following section, we only report important results for brevity. Supplementary material is provided at the following URL: https://github.com/evhart/fc-co-spread.

4.3. General results

The one-way MANOVA analysis comparison at the relative URL shares level for misinforming URLs and fact-check URLs shows a significant permuted p-value of 0. This means that at the global level, there are significant differences in how misinforming URLs and fact-checking URLs relatively spread and that the type of shared URLs influences the amount of spread at different relative time periods.

Following the significant result of the MANOVA analysis, a one-way ANOVA analysis is performed for each relative time period. The Bonferroni adjusted Kruskal–Wallis tests are only significant for the *initial* ($p = 1.27 \times 10^{-71}$) and *late* ($p = 1.07 \times 10^{-239}$) periods. This means that sharing behaviour during the *early* ($p = 1$) relative period does not differ during that period, whereas differences exist when looking at the *initial* and *late* periods.

The individual distributions of misinforming and fact-check URLs for each time period show that misinformation spreads more than fact-checks with a large difference during the initial period in particular (initial misinformation mean = 18.4 vs. initial fact-check mean = 1.41). Interestingly, the highest difference in terms of mean and standard deviation between the different URL types appears to be mostly during the initial phase with a more important standard deviation for the misinforming URLs ($\sigma = 216$ for misinforming content and $\sigma = 8.51$ for fact-checks).

4.4. Individual results

We perform the same analysis for different topics and demographics. For the topic analysis, we leave out the *other* category and focus on the remaining six topics since the *other* topic is not clearly defined, as mentioned earlier (Section 4.2). For the gender and age group, we perform the analysis only on accounts that are identified as individuals, since organisations are not normally associated with a particular age group and gender.

4.4.1. Topic results

Similar to the global analysis, the MANOVA result shows a significant permuted p-value of 0, which means that there are differences in how information types and topics spread. The ANOVA analysis show significant differences in all the individual periods (initial p = 6.72×10^{-24}; early p = 3.69×10^{-181}, and; late p = 7.08×10^{-14}).

The post-hoc Bonferroni adjusted Dunn tests show significant differences between the different information types and topics spread for 44 cases out of the 198 Dunn tests (Fig. 4). Most of the differences in behaviour happen during the initial period, with misinformation spreading generally differently compared to fact-checks.

Spread for misinformation tends to be similar across all topics. The same observation can be made for fact-check URLs, which also appear to propagate similarly across the various topics. Notable similarity in behaviour across topics is observed between misinformation about Symptoms, and other topics and information types. Misinformation about Causes also shows a similarity with other topics and information types, except for the spread of fact-checks about Cures (p = 0.012).

For the *early* period, differences are mostly linked to misinformation about Conspiracy Theories and COVID-19 Causes, whereas behaviour between misinformation about Causes and Conspiracy Theories is similar (p = 1), suggesting a strong behavioural link between what caused the virus and conspiracies.

As the spread of misinformation and fact-checks propagate from the *initial* to the *late* period, spread behaviour becomes more consistent and similar across the topics and information types during the later periods, except for a few topics (24 significant p-values for the *initial* periods, then 10 for the *early* period and 10 for the *late* phase). The results show that misinformation about COVID-19 Causes and Conspiracy Theories spread significantly differently compared to fact-checks about Authorities (p = 0.005 for COVID-19 Causes and p = 0.0005 for Conspiracy Theories) and COVID-19 Spread (p = 1.862×10^{-05} for COVID-19 Causes and p = 1.418×10^{-08} for Conspiracy Theories).

Spread of misinformation overall seems to be mostly focused on Authorities. This finding is consistent with previous findings that misinforming claims about the World Health Organisation and other international bodies like the UN make up a significant

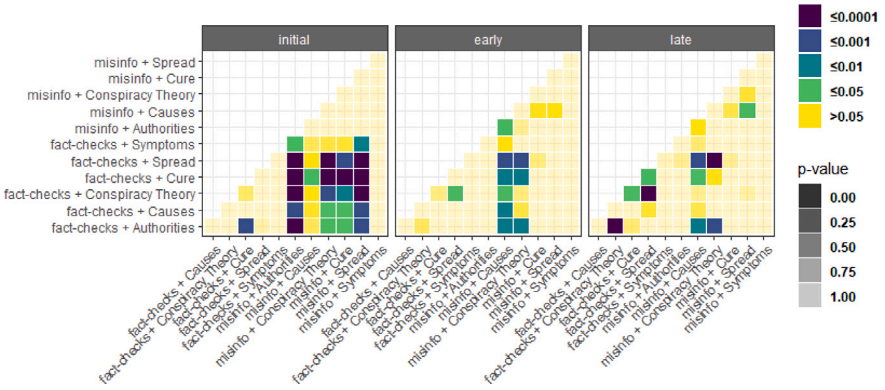

Fig. 4. Post-hoc Bonferroni adjusted Dunn tests results for different topics and information types for the *initial*, *early* and *late* periods. The transparency scale indicates *p*-value score, whereas the colour indicates the significance level.

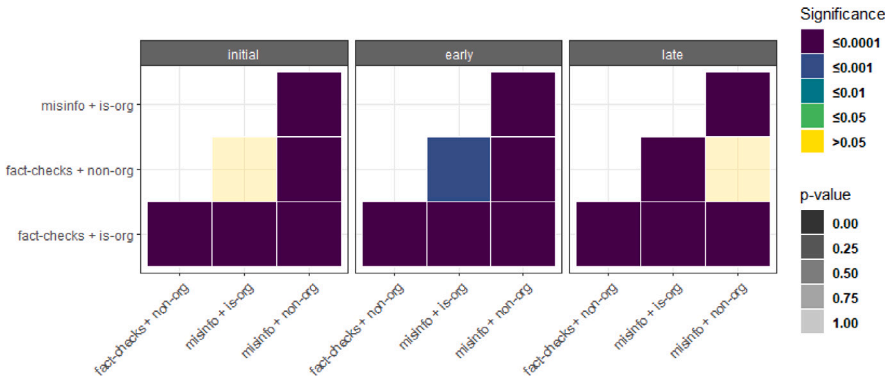

Fig. 5. Post-hoc Bonferroni adjusted Dunn tests results for different account and information types for the *initial*, *early* and *late* periods. Transparency scale indicates *p*-value score, whereas the colour indicates the significance level.

percentage of misinformation about COVID-19 (Brennen et al., 2020). Donald Trump has also been implicated as a significant figure in one study, mentioned in over 37% of all claims the authors examined (Evanega, Lynas, Adams, Smolenyak, & Insights, 2020). In our study, 63,934 posts identified as misinformation can be linked to authorities. This potentially shows how communication about the pandemic and appropriate responses may gather more misconceptions than, for example, than misinformation about Symptoms (1146 posts), which is harder to spread. When looking at the distribution of spread in the different periods, we can observe that deviations in spread are much higher for misinforming content for all the topics, with Causes showing the highest deviation overall and particularly for the *initial* period ($\sigma = 2675.2$). This observation may be linked to the uncertainties of such a topic, particularly during the beginning of the COVID-19 pandemic, where most attention was centred on what caused the pandemic amidst much confusion amongst people on why the pandemic was happening.

4.4.2. Account type results

The MANOVA analysis for the type of accounts sharing misinformation and fact-checks shows a significant *p*-value. This indicates that there are variations overall with regard to how different account and information types share posts over time. This result is confirmed with the ANOVA analysis for each of the *initial* (adjusted $p = 1.84 \times 10^{-159}$), *early* (adjusted $p = 9.39 \times 10^{-121}$) and *late* (adjusted $p = 2.72 \times 10^{-262}$) time periods. These observations are stable when considering the potential maximal account classification errors globally and for the *initial* (mean adjusted $p = 1.25 \times 10^{-138}$), *early* (mean adjusted $p = 4.31 \times 10^{-75}$) and *late* (mean adjusted $p = 8.78 \times 10^{-135}$) time periods.

The post-hoc Bonferroni adjusted Dunn tests (Fig. 5) show that, overall, there are significant differences for each of the analysed pairs, except for 2 of the 18 analysed pairs during the *initial* and *late* phases. A small variation is observed when accounting for potential maximal classification errors with two behaviour similarities observed during the *early* and *late* phases.

In the *initial* phase, it appears that the spread of fact-checks by individuals exhibits a similar behaviour to the spread of misinformation from organisations (adjusted p = 1). This reverse association during the initial period may be due to the fact that, in our data, individuals spread much more misinformation overall (n = 234,222 for individuals and n = 16,145 for organisations) and that most information spread happens during the first phase. During the *late* period, it appears that individual sharing behaviour remains constant, independently from the type of information shared (adjusted p = 1). This result may be linked to the fact that

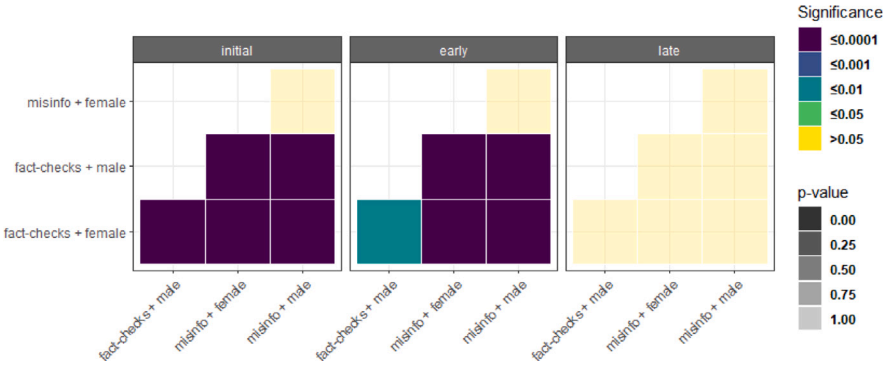

Fig. 6. Post-hoc Bonferroni adjusted Dunn tests results for gender and information types for the *initial*, *early* and *late* periods.

the amount of shared misinformation and fact-checks is relatively similar, overall, in the *late* phase for individuals (n = 35,407 for fact-checks and n = 44,648 for misinformation), compared to the other periods.

The maximum error analysis shows that misinformation and fact-checks spread differently during the *initial* period, whereas a convergence is observed for the *early* period between organisations sharing misinformation and individuals sharing fact-checks (mean adjusted p = 1). This association also appears less significant for the standard analysis (adjusted $p > 0.001$). As with the general *initial* phase observation, the result difference may be due to the small amount of data relating to organisations in our dataset, compared to individual accounts and how the classification errors may propagate in such small populations. In any case the observations remain the same for all the other cases.

The distribution of misinformation spread for the account types shows that 93% of the shared posts are from individuals rather than organisations and that 82% of the shared fact-check posts are from individuals. For organisations, the relative percentage of shared fact-checks is much higher with more than half the posts being fact-check URLs (54% instead of 27%). Similarly to our previous observations, most misinformation and fact-check spread happens during the *initial* and *early* phases with misinformation spreading in larger amounts and exhibiting high spreading variations, compared to fact-checks during the *initial* period ($\sigma = 66.3$ and $\sigma = 978$ for misinforming content for organisations and individuals compared to $\sigma = 2.63$ and $\sigma = 29.7$ for fact-checks).

4.4.3. Gender results

The MANOVA analysis for different genders sharing misinformation and fact-checks shows a significant p-value, indicating a significant difference in sharing behaviour overall (p = 0). Compared to the previous groups, we only observe significant differences in behaviour in the *initial* ($p = 1.11\times10^{-184}$) and *early* ($p = 1.78\times10^{-19}$) phases, which means that **gender is not linked with different spreading behaviour for misinforming content and fact-checks in the long-term**. All these observations are confirmed when accounting for potential gender classification errors globally and for the *initial* (mean adjusted $p = 8.04\times10^{-169}$), *early* (mean adjusted $p = 7.16 \times 10^{-17}$) and *late* (mean adjusted p = 1) time periods.

The post-hoc Bonferroni adjusted Dunn tests results for the *initial* and *early* periods are very similar (Fig. 6) with differences in spread behaviour for all the possible gender and information types spread except for misinformation spread for male and females (p = 1 for the *initial* and early *periods*). This result shows that misinformation spreading behaviour appears to be independent of gender, but that there are differences in how fact-checks are spread between genders (p = 0.009 for the *initial* period and $p = 6.07\times10^{-55}$ for the *early* period). The fact that spread behaviour seems to be similar across gender and information type in the late periods seems to indicate that the impact of gender on COVID-related information spread is only short term (day 1–10).

The results obtained when considering the potential gender classification errors produce the same results overall with spreading behaviour similarities in the *late* period for each gender and information type and in the *initial* (mean adjusted p = 0.935) and early phases (mean adjusted p = 1) for misinforming URLs for each gender. The only difference appears during the *early* phase, where fact-check URLs spread behaviour now appears similar across gender (mean adjusted p = 0.057). This behaviour was previously identified as slightly significant (adjusted p > 0.001). This result confirms that male and female users, in the binary classification, spread fact-checking URLs differently, mostly during the *initial* spreading period.

The distribution of misinformation spread is mostly propagated by those classified as male (68%), as are fact-checks (68%). This result can be directly compared with the distribution of genders on Twitter where 70% of users are believed to be male.[10]

As with the observation of the previous groups, we see again that most misinformation and fact-checks spread happens during the *initial* and *early* phases with misinformation spreading in larger amounts and exhibiting high spreading variations compared to fact-checks. This is particularly true during the *initial* period ($\sigma = 690$ and $\sigma = 372$ for misinforming content for those classified as male and those classified as female compared to $\sigma = 19.2$ and $\sigma = 14.5$ for fact-checks).

[10] Distribution of Twitter users worldwide (Statista), https://www.statista.com/statistics/828092/distribution-of-users-on-twitter-worldwide-gender.

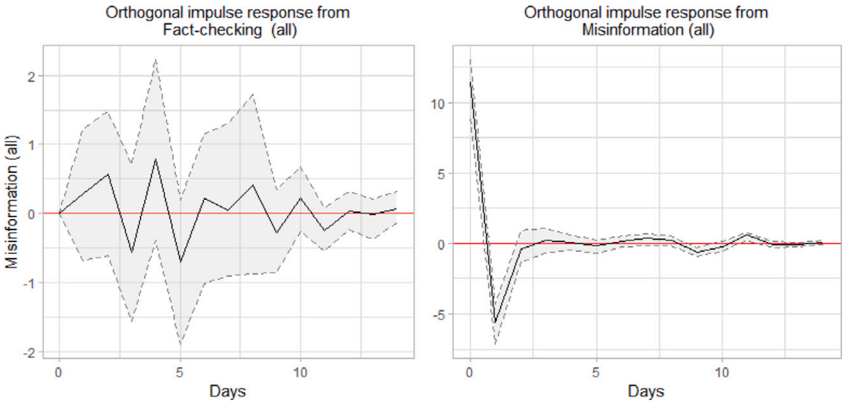

Fig. 7. Bootstrapped relative-level orthogonal impulse response for misinformation information (95% confidence interval).

5. Fact-checking and misinformation impact analysis

For the second part of our analysis, we study how misinformation and fact-checks impact each other globally and identify if topic-specific relations between the spread of misinformation and fact-check URLs exist. We investigate if the spread of fact-check information for different topics has a beneficial impact in reducing the diffusion of misinformation. For this analysis, we model the spread of URLs as a Vector AutoRegression (VAR) model using the misinformation and fact-check URLs as endogenous variables. We perform this analysis at the relative level (i.e., the relative number of days since the first appearance of a URL related to a particular misinformation) and determines if weak causation relations between each information type exists. We perform such analysis between individual topics as well as globally.

5.1. Experimental setup

Similarly to our previous work (Burel et al., 2020), we estimate if the spread of a given information type can be used to predict the spread of another information type using a Granger causality test. In order to compute the Granger causality test, we first build a Vector AutoRegression (VAR) model using the combined misinformation spread and fact-check information both globally and for the analysed topic (e.g., Causes, Authorities). Since we are interested in the relation between misinformation and fact-checks, we only select posts where related misinformation and fact-check are shared (i.e., misinforming URLs that spread with their corresponding fact-check URL counterparts). Since our data is non-stationary, we first integrate each analysed information type and analysed group value so that the spread amount for each day is represented as the difference between the current day value and the previous day value.

Depending on the analysed group (topics or global spread), a different order is used for the VAR model based on Akaike's information criterion. Using the VAR(n) model, we perform a bootstrapped Granger causality test for determining if misinformation spread can be associated with fact-check URL spread or if fact-check spread can be inferred from misinformation spread for the global analysis. For the topic analysis, we perform two different tests in order to understand the relation that topics have on the spread of fact-checks and misinformation. For example, when analysing the *Causes* topic, we perform two analyses in order to answer the following questions: (1) Does misinformation spreads about the *Symptoms* topic causes fact-check spread for the *Symptoms* topic? (2) Does fact-check spreads about the *Symptoms* topic causes misinformation spread for the *Symptoms* topic?

In order to understand the dynamics that relate fact-checks and misinformation globally and for the topics, impulse response analysis is performed as well as Forecast Error Variance Decomposition (FEVD). For the impulse response analysis, we use orthogonal impulse responses in order to evaluate the spread response of the different types of URLs for 14-day periods. This approach enables us to determine how a particular sharing behaviour may affect other types of URL shares in the future. We run the FEVD with the same 14-day periods in order to obtain the contribution importance of each information type on the spread of both misinforming URLs and fact-check URLs.

5.2. Global results

Using the Akaike's information criterion, we obtain a lag value of 10 days for the global analysis and create a VAR(10) model for performing the Granger causation tests. We obtain a significant value for the Granger causation between misinformation and fact-checks ($p = 0.03$) as well as between fact-checks and misinformation ($p = 0.01$). This bidirectional result suggests that changes in fact-check information spread may cause a change in misinformation spread and that misinformation spread may also cause a change in fact-check information spread. This suggests that **fact-checks have an impact on misinformation spread but that misinformation spread also impacts fact-check spread**.

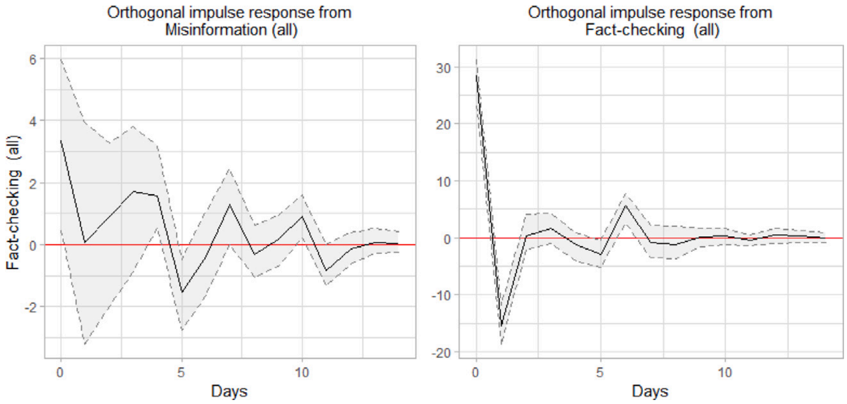

Fig. 8. Bootstrapped relative-level orthogonal impulse response for misinformation content (95% confidence interval).

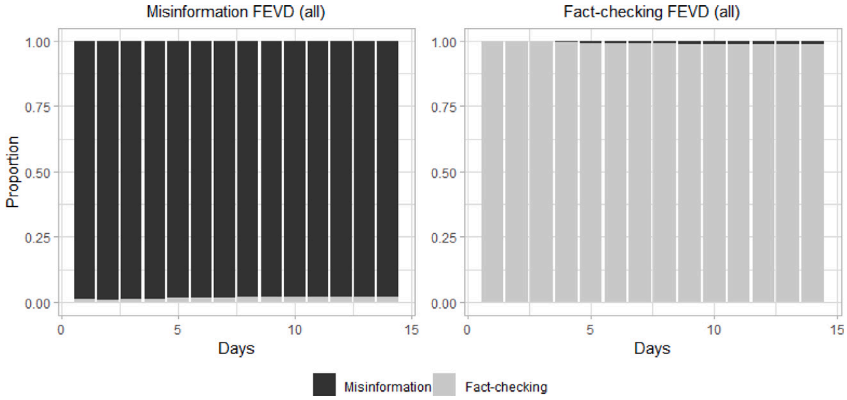

Fig. 9. Forecast Error Variance Decomposition (FEVD) for misinformation and fact-checking spread.

The impulse response for the orthogonal shock in the amount of shared fact-checks (Fig. 7) does not appear to lead to a clear drop in misinformation sharing even though, over time, misinformation seems to decrease. However, this behaviour may be simply linked to how information spread in general decreases over time. Contrary to our previous work (Burel et al., 2020), we do not observe an initial decrease in misinformation, suggesting that the impact of fact-checks on misinformation is not as strong as previously thought. Nevertheless, as observed in the Granger analysis, fact-checks spread does impact misinformation spread, but not necessarily in the way we wish it would. A sharp increase in misinformation leads to a sudden drop in misinformation spread. This confirms previous observation that misinformation spreads mostly after its initial appearance and decreases quickly in the following days (Burel et al., 2020).

The impulse response for the orthogonal shock in the amount of shared misinforming URLs (Fig. 8) shows a quick initial response in increase of fact-check URLs spread. This result suggests that fact-check spread seems to respond quickly to a rise in misinformation as fact-checking articles are created as a response to misinformation and shared on Twitter. Contrary to our previous investigation where we observed a delayed response (Burel et al., 2020), it appears that fact-checks are almost shared as fast as misinformation is created when fact-checks are available. As with the misinformation sharing behaviour, we observe a sharp decrease in fact-check sharing behaviour after the initial shock as initial sharing behaviour reduces.

The FEVD results as shown in Fig. 9 show that both misinformation spread and fact-checks spread predictions over time are only slightly affected by the spread of their counterpart. Although both misinformation and fact-checking FEVD are much less impacted than previously observed (Burel et al., 2020), we can see that misinformation prediction getting more affected by fact-check spread as time goes by and fact-checks predictions getting more affected by misinformation over time. Interestingly, it appears that misinformation spread prediction is always affected by fact-check spread even though this relation is very minimal, whereas fact-check spread prediction is only affected by misinformation spread after the first day. This result suggests that fact-checking impacts misinformation spread from day one, even though it appears that **misinformation spread does not slow down when fact-check spread increases**, as we observed in the impact analysis.

Table 1
Granger causality test results for different topics between misinformation and fact-checking URLs spread.

Topic	Model	Granger Relation			*p*-value
Authorities	VAR(14)	*Fact-checks*	⇒	*Misinformation*	0.36
		Misinformation	⇒	*Fact-checks*	0.61
Spread	VAR(8)	*Fact-checks*	⇒	*Misinformation*	0.16
		Misinformation	⇒	*Fact-checks*	0.02
Cure	VAR(11)	*Fact-checks*	⇒	*Misinformation*	0.59
		Misinformation	⇒	*Fact-checks*	0.06
Conspiracy Theory	VAR(14)	*Fact-checks*	⇒	*Misinformation*	0.43
		Misinformation	⇒	*Fact-checks*	0.06
Causes	VAR(4)	*Fact-checks*	⇒	*Misinformation*	0.13
		Misinformation	⇒	*Fact-checks*	0.32
Symptoms	VAR(14)	*Fact-checks*	⇒	*Misinformation*	0.14
		Misinformation	⇒	*Fact-checks*	0.01

5.3. *Topic results*

The Granger causation tests and lag numbers identified by the Akaike's information criterion for creating the VAR(n) models are described in Table 1. In general, we can observe that most results show no weak causation relation between misinformation and fact-checks spread for individual topics with only two exceptions. First, it appears that misinformation spread about the virus *Spread* Granger causes fact-check URLs spread about the virus *Spread* (p = 0.02). Second, we see that misinformation spread about the virus *Symptoms* Granger causes fact-check URLs spread about the virus *Symptoms* (p = 0.01). Interestingly, we do not observe any weak causation from fact-check to misinformation spread for all the topics.

The Granger causation relating to the *Spread* topic may be linked to the number of shares the *Spread* topic has, compared to the other analysed topics (11% of the total information spread), with around 71% of the *Spread* topic shares linked to misinformation spread. When looking at the distribution of fact-checks over time, for the *Spread* topic we see that fact-checking URLs diffusion seems to overtake the spread of misinforming URLs in the *late* period (misinformation goes from 18,498 shares in the *initial* period to 745 in the *late* period, whereas we observe a change from 4400 to 2791 for the fact-check URLs). This result suggests that the rise in misinformation about *Spread* induces a rise in fact-check URL shares concerning the same topic, meaning that corrective information to particular misinformation about the virus *Spread* diffuses well compared to other topics.

The Granger causation relating to the *Symptoms* topic appears to have a similar dynamic to the *Spread* topic. Similarly to the *Spread* topic, the *Symptoms* topic spread is much smaller than any other topic spread (1% of the total information spread) with around 43% of the *Symptoms* topic shares linked to misinformation spread. This result is sensibly different to the other topics as fact-check information appears to spread more than misinformation overall for this topic. This general behaviour may be due to the scope of the *Symptoms* topic compared to potentially more controversial topics like *Cures*, *Causes* and *Authorities* that are more susceptible to misinformation spread. As with the *Spread* topic, the *Symptoms* topic misinformation spread decrease faster than fact-checks over time and suggests that corrective information to particular misinformation about the virus *Symptoms* diffuses well compared to other topics.

The impulse response and FEVD analyses for each topic generally highlight very little impact between misinformation URLs and fact-check URLs spread. Similarly, most of the FEVD results also show **minimal predictive dependencies between misinformation and fact-check spread**. Standout impulse responses can be observed for the *Conspiracy theory* and *Causes* topics which have been identified during the post-hoc analysis as the topics that are the most likely to behave differently than others during the *late* period (Section 4.4.1).

The impulse response for the orthogonal shock in the amount of shared fact-checks about conspiracies (Fig. 10) leads to an initial erratic reduction in misinformation that seems to oscillate over time. This observation suggests that fact-checking does not have a clearly defined long-term impact on misinformation sharing. As with previous general observations (Fig. 10), we observe a sharp decrease in misinformation sharing behaviour after an initial misinformation increase as initial sharing behaviour reduces.

The impulse response for the orthogonal shock in the amount of shared misinforming URLs for the *Conspiracy theory* topic (Fig. 11) shows a quick initial response in the increase of fact-check URL spread. However, this increase quickly reduces after the first day. This result suggests that new misinformation about conspiracy theories quickly generates a fact-checking response from the public and organisations, but that this response is relatively short-lived. For the response of fact-check spread, following an increase of fact-checking information shares, we observe a quick decrease in fact-check spread that is similar to previous observations.

FEVD analysis for the *Conspiracy theory* topic (Fig. 12) highlight clear dependencies between misinformation and fact-check spread for both predicting misinformation and fact-check spread about conspiracies over time. As time increases, we can observe that the dependencies increase. Although misinformation spread about conspiracy theories appears to be always affected by fact-check spread, it appears that fact-checking spread is not initially impacted by misinformation spread. This observation suggests that the spread of misinformation is directly affected by how fact-check about conspiracies spreads, even though no Granger causation relation was observed for this topic. For the fact-checks, the impact of misinformation spread seems to be delayed, meaning that fact-checking does not depend on how misinformation is shared initially.

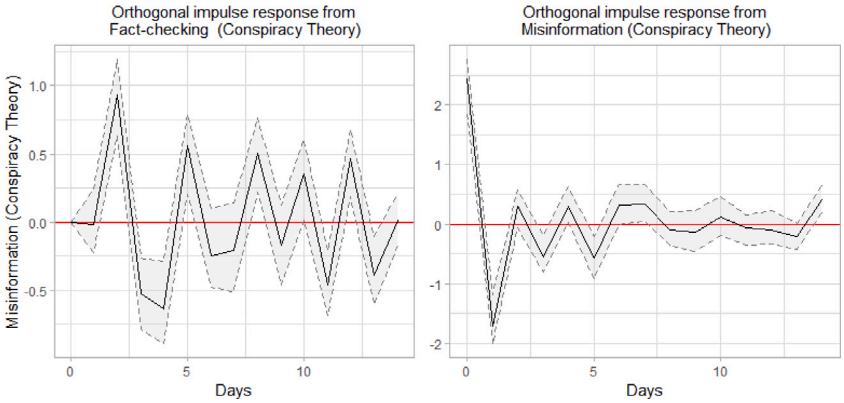

Fig. 10. Bootstrapped relative-level orthogonal impulse response for misinformation and the *Conspiracy theory* topic (95% confidence interval).

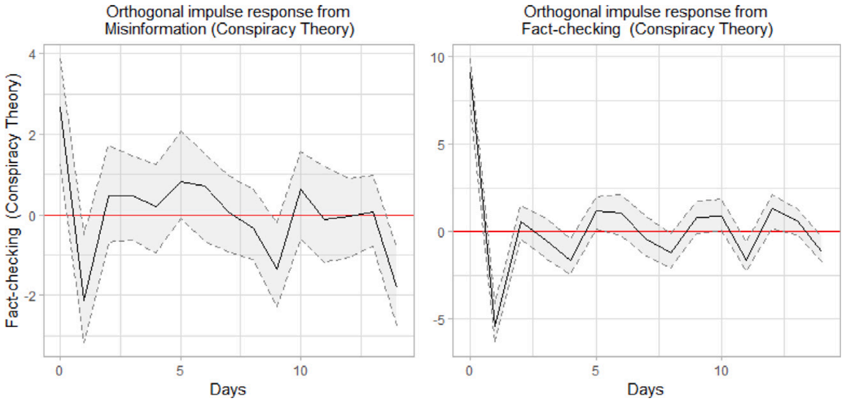

Fig. 11. Bootstrapped relative-level orthogonal impulse response for fact-checking information and the *Conspiracy theory* topic (95% confidence interval).

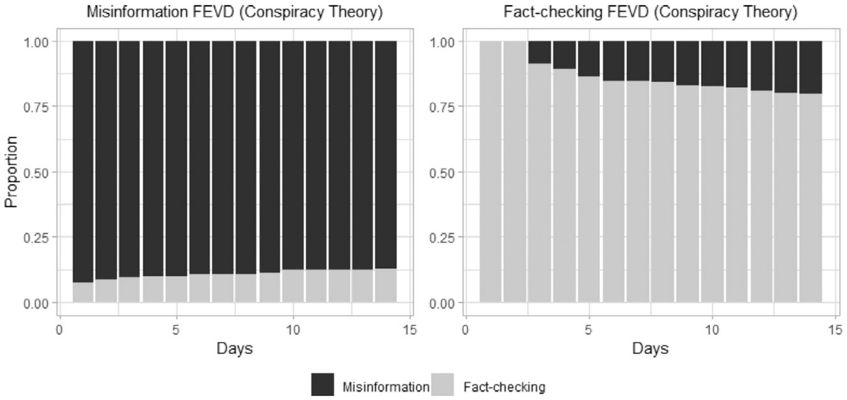

Fig. 12. Forecast Error Variance Decomposition (FEVD) for misinformation and fact-checking spread for the *Conspiracy theory* topic.

The impulse response for the orthogonal shock in the amount of shared fact-checks about the virus causes (Fig. 13) initially creates a general decrease in misinformation sharing, meaning that fact-checking appears to reduce misinformation spread about them virus causes even though we do not observe any Granger causation relation for this topic. For the fact-checking spread, we observe the familiar decrease in fact-checking spread after the initial impulse.

The impulse response for the orthogonal shock in the amount of shared misinforming URLs for the *Causes* topic (Fig. 14) creates an important rise in fact-checking spread that is sustained over a few days. Similarly to the global analysis, this appears to show that misinformation about virus causes create an important public response compared to other topics and shows that fact-check spread seems to respond quickly to a rise in misinformation as fact-checking articles are created as a response to misinformation and

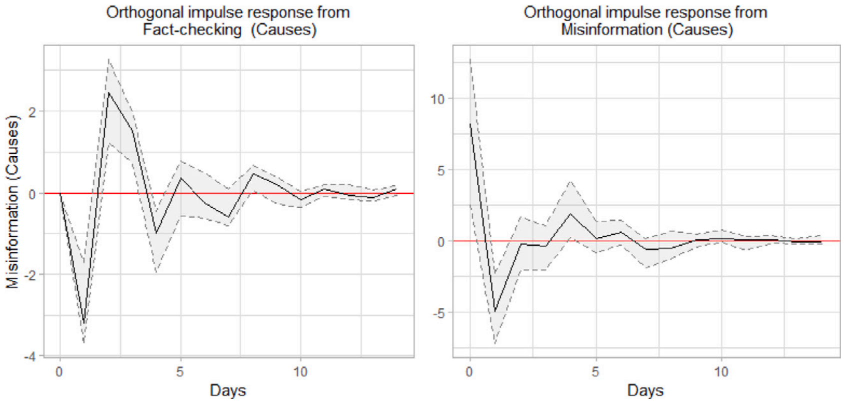

Fig. 13. Bootstrapped relative-level orthogonal impulse response for misinformation and the *Causes* topic (95% confidence interval).

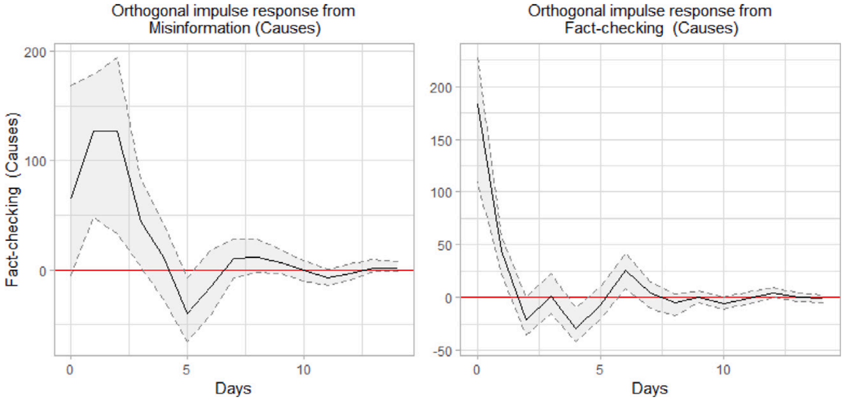

Fig. 14. Bootstrapped relative-level orthogonal impulse response for fact-checking information and the *Causes* topic (95% confidence interval).

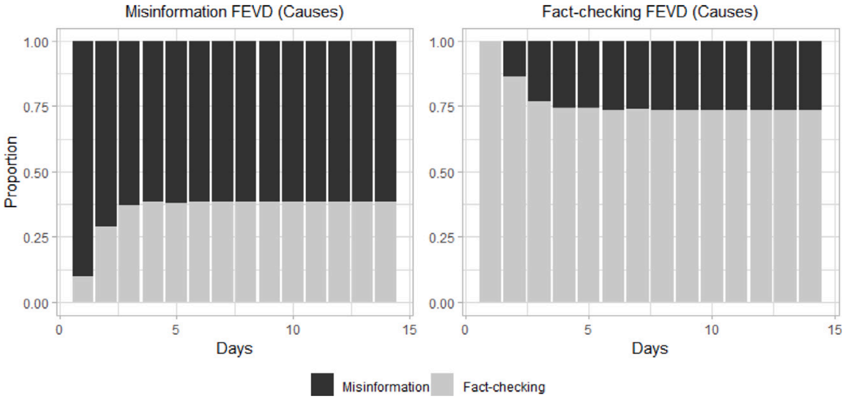

Fig. 15. Forecast Error Variance Decomposition (FEVD) for misinformation and fact-checking spread for the *Causes* topic.

shared on Twitter. It seems that fact-checks about the virus causes are shared as fast as misinformation is created when fact-checks are available. For the misinformation sharing behaviour for the virus causes, we observe a sharp decrease in fact-check sharing behaviour after the initial shock as initial sharing behaviour reduces, confirming the previous general observations.

The FEVD results for the *Causes* topic (Fig. 15) shows much higher dependencies between misinformation and fact-checks spread than our previous results, with fact-checking spread affecting around 30% of the misinformation prediction model over time. This result confirms that fact-checks have an impact on misinformation spread. For the spread of fact-checking URLs, it seems that misinformation does not play an initial role in its spread but then account for 25% of the information that is used for predicting the spreading behaviour of fact-checks. In general it seems that for the *Causes* topic in particular, the spread of misinformation and fact-check highly depends on the spread of each other, meaning that it affecting the spread of misinformation or fact-check for this topic leads to a direct change in its counterpart diffusion behaviour.

6. Discussion

In this section, we discuss a small selection of the most relevant findings related to the temporal, topological and demographic features of misinformation and fact-checking spread during COVID-19. We also highlight the differences between our previous work results (Burel et al., 2020) and the novel observations made in this study. The scope of this investigation was particularly large, so we are not able to present all results. Further insights and analyses can be viewed at the following address: https://github.com/evhart/fc-co-spread.

6.1. Fact-checks and misinformation spread behaviour

Concerning the temporal aspects of misinformation spread, the timing of misinformation and its persistence are always key issues (Starbird et al., 2018). Our study confirms previous findings that misinformation spreads mostly in the initial phases after circulation (Starbird et al., 2018). We were also able to confirm that misinformation spreads much more than fact-checking information (Vosoughi et al., 2018). In our study, misinformation accounted for 70% of all spread. What our current study illuminated, however, is that fact-checking is following a similar pattern during the COVID-19 pandemic. This is promising, as previous work has shown that early intervention is crucial (Cuan-Baltazar et al., 2020; Kouzy et al., 2020), and our previous work indicated a delay in fact-checking response (Burel et al., 2020). This should continue to be monitored, as health crises are often accompanied by misinformation (Vaezi & Javanmard, 2020; Xie et al., 2020) and early indications show that COVID-19 continues this trend (Kouzy et al., 2020).

In addition to this discovery, we were also able to add to the description of misinformation and fact-checking spread in several ways: First, we found that topic behaviour differs across all the periods, but that the most interesting behaviour is linked to misinformation spread about authorities and causes, showing links between these two topics. Earlier studies have also indicated that authorities and causes gather a particular level of misinforming claims around them (Brennen et al., 2020). We explained this finding with uncertainty about the pandemic in Section 4.4, as well as the reliance on trustworthy information in an ambiguous setting. Trust in science (Agley & Xiao, 2021), trust in information (Lazarus et al., 2020) and trust in the scientific process (through intellectual humility) (Koetke et al., 2021) appear to be important during COVID-19 for avoiding *infection* with misinformation. Perhaps this explains the different behaviour at the beginning of a serious health crisis, when authorities may not have the information the public needs (Cuan-Baltazar et al., 2020; Kouzy et al., 2020). Conspiracy theories are part of how people make sense of events that do not yet have a satisfactory explanation (Douglas et al., 2019).

Second, we find that different spread behaviour of misinformation about authorities and causes continues over the early and late periods. This may be linked to the difficulty in eradicating conspiracy theories (Douglas et al., 2019). Our previous work indicated that the volume of misinformation may be simply too large to extinguish and that more work was needed to amplify and extend the impact of fact-checking (Burel et al., 2020). Perhaps this effect is greater for misinformation about topics that attract conspiracy thinking. In addition, certain health beliefs, if underpinned by conspiracy theories, may be difficult to break through (van der Linden et al., 2020). The combination of these factors may create more space for conspiracy thinking to flourish. Finally, groups and individuals that are committed to certain political or ideological positions may exploit the pandemic uncertainty to revive previous grievances or issues (Pyszczynski, Lockett, Greenberg, & Solomon, 2020). These explanations cover a range of altruistic, practical and more sinister reasons for why we observed this finding, but future work will be necessary to explore in more detail the category of misinformation about causes of COVID-19 and why misinformation thrives in particular contexts.

Third, across topics, spread deviations are much more pronounced for misinformation than for fact-checks, both globally and for individual groups. In addition, for all the different groups (except account type), misinformation spreads similarly across demographics and topics. We cannot observe the same for fact-checks. For misinformation, this makes sense as different misinforming claims may target different groups of users or emotions. For fact-checking, this finding is potentially important because it may show how fact-checks are targeting users and communicating in more generic ways, which may limit their reach. Our previous work indicated that those sharing fact-checks may not be the same as those sharing misinformation (Burel et al., 2020), strengthening this proposition.

Globally, behaviour differs between misinformation and fact-check spread except for the early phase (days 3–10). These results may be linked to the general behaviour of information spread and how misinformation can sometimes re-emerge over time. During the early period, misinformation reduces as well as fact-checking spread. However, as misinformation is spread again later, fact-checks are shared again as a response. Once again, this may have to do with the persistence of certain types of misinformation, like conspiracy theories (Douglas et al., 2019).

Fourth, in terms of demographic analysis, our account type analysis showed that individuals were sharing much more misinformation than organisations (93%). Individuals also share more fact-check URLs, which confirms earlier work that individuals are contributing significantly to the process of debunking online (Micallef et al., 2020). Understanding and leveraging the role of individuals in amplifying the work of professional fact-checkers will be crucial for future efforts to fight misinformation effectively. However, other demographic analyses did not yield particularly promising results. Behaviour based on classifications of users by gender shows different gendered behaviour for fact-checks initially. However, this completely converges in the late period, meaning that gender-based spreading behaviour is more temporary than other types of analysed demographics. As Twitter is believed to have a gender disparity among users in favour of men, this could explain part of this difference. A closer examination will be necessary to understand more about the different gendered motivations for sharing fact-checking information. One early study indicated that motivations for sharing misinformation about COVID-19 may follow gendered patterns with regard to cyberchondria or sharing without verifying, for example Laato, Islam, Islam, and Whelan (2020).

6.2. Impact of fact-checking on misinformation spread

The second part of our work examined the impact of fact-checking spread on misinformation spread. As mentioned above, one promising result is that our study showed that an increase in misinformation clearly generated a instantaneous increase in fact-check spreads. This tapered off quickly (1–6 days). However, compared to our previous results, where we observed a delay (Burel et al., 2020), it seems that fact-checkers' response to new misinformation spread may be less delayed than initially thought.

General topic spread for misinformation and fact-checks naturally decreases over time. This reflects more general expectations of how information spreads. What is perhaps more interesting is that we observed a bidirectional, weak causation overall suggesting that misinformation and fact-check have a co-dependence in how they spread. One explanation for this is that the appearance of fact-checks may motivate the exploration of the original piece of misinformation being debunked. Whether or not this has any impact on belief, or is just a cross-checking exercise, however, would still be unknown.

Unfortunately, while our previous work showed a slight decrease in misinformation as a result of the presence of fact-checking information (Burel et al., 2020), the current study did not confirm this. Globally, impulse response showed that misinformation did not decrease faster as a response of an increase in fact-check spreads. In addition, both misinformation spread and fact-checking spread predictions over time are only slightly affected by the spread of their counterpart, whereas our previous work indicated a higher relation (Burel et al., 2020). Interestingly, it appears that misinformation spread prediction is always affected by fact-checking spread even though this relation is very minimal. These observation differences may be linked to the difference in our data collection method and longer period of study.

Our work showed only two weak causation relations for the topics, suggesting that the causation relations are more global than topic specific. The two causation relations we did observe are only for *misinformation weakly causing fact-checks*, meaning that fact-checks may not affect how misinformation spreads directly. This may provide evidence of the more prophylactic rather than therapeutic impact of fact-checking (Krause et al., 2020), but again, these causal relationships require much more analysis. However, for two cases we observe that fact-check spreads as misinformation spreads for the *Spread of COVID-19* topic and the *Symptoms of COVID-19* topic, which are the two topics least susceptible to misinformation (with the smallest information shares overall). In particular, *Symptoms* are more associated with fact-checks overall. This may be why misinformation appears to generate fact-check spread.

While these results may seem disheartening, we know from extensive cognitive research that correcting misinformation about COVID-19 helps (Bode et al., 2020). Our results indicate a need to look more closely into the relationship between fact-checking, belief in misinformation and motivations for sharing information during COVID-19, in order to trigger a more targeted, effective impact.

6.3. Comparison to our previous findings

Our analysis can be seen as a major extension to our previous co-spread analysis study (Burel et al., 2020) both in term of how data is collected and how it is analysed. As discussed in Section 3.1, we updated both the way the data is collected and how fact-checks and misinformation URLs are selected. The updated methodology and longer period analysed means that we get a more precise and more complete picture of how COVID-19-related misinformation and fact-checks spread. As previously discussed (Section 3.1.2), our new dataset is much larger than the one previously studied. For the period between the 1st December 2019 and the 4th of May 2020, we collected 67% more data and 75% more fact-checks seed URLs. Additionally, the new data also covers more than a year of misinformation and fact-checks spread, whereas the previous analysis only covered the first six months of the COVID-19 pandemic.

Globally, the new analysis confirms the general finding of our previous work: misinformation spreads more than fact-checks and they spread significantly differently overall. However, we start to observe differences when looking at their spread for each of the individual periods.[11] In our previous work, we observed that spread behaviour was similar in the *initial* period only, whereas we only observe a similar spread behaviour between fact-checks and misinforming URLs during the *early* period in this study. This result may be simply explained by the longer period and larger amount of data analysed. Since misinformation spreads more than fact-checks URLs (Vosoughi et al., 2018) and mostly in the initial phases after circulation (Starbird et al., 2018), the larger amount of data reinforce these observations and appears to shift the spread behaviour similarity between each URL types from the *initial* period to the *early*. The *late* period spread behaviour observation remains unchanged and, as observed in this article, may be due to how specific misinforming topics are likely to re-emerge over longer periods.

The impact analysis also shows some significant differences that may be also explained by the additional amount of data collected and the longer time period analysed. In particular, we previously observed a unidirectional weak causation relation that suggested that fact-checks had some positive impact in reducing how misinformation spread (Burel et al., 2020), whereas our additional data

[11] Note that the segmented linear regression method created slightly different periods thresholds for each dataset (see Section 3.2).

suggests a bi-directional relation between how fact-check and misinformation spread relate. Unfortunately, this means that the impact of fact-checking spread is less clear than we thought. However, the observed bi-directional weak causation relation suggests that fact-checks spread directly as a response to how misinformation propagates. This is confirmed by the difference in impulse response graphs. This result is more representative to how fact-checkers work as they try to push corrective information as new misinforming claims appear.

Thanks to the new topic and demographic analyses, our extended study gives important insights concerning how specific COVID-19 topics spread and why spreading behaviour differs in the *late* period (Burel et al., 2020). It appears that misinformation related to topics like the virus causes remain active over long time period compared to other topics and may be linked to the difficulty in dealing with topics related to conspiracy theories (Douglas et al., 2019). Our new study also investigates demographics and confirms that misinformation and fact-checks are mostly shared by individuals (Micallef et al., 2020). We also observe that gender sharing behaviour appears to be only different when dealing with fact-checking URLs.

7. Limitations and future work

Although the method we used for identifying both misinforming and fact-checking content on Twitter has the advantage of not depending on automatic tools, the data collection approach is also limited by the number of identified URLs from fact-checking organisations. As a result, our data covers only a small amount of misinformation and does not contain variations of the same posts. Similarly, the amount of collected posts is limited by the data collection method. This can lead to some differences in what is observed (Section 6.3). Coupling our data collection method with automatic misinformation identification methods could increase our dataset dramatically. However, that approach may add some false-positives to the collected content.

In our work, we have also used some automatic methods for extracting different user demographics in order to extend our knowledge on how misinformation and fact-check spread. Although we use automatic models that provide good accuracy, automatic tools are always susceptible to misclassification errors. The model we use for identifying gender is also unable to identify non-binary gender. Therefore, the gender of non-binary users cannot be identified correctly, leading to misgendering issues and potentially a misunderstanding about gendered sharing behaviour. The recent extension to the gender M3 model used in our analyses (Wang et al., 2019) proposed by Yang et al. (2021) could potentially lead to more accurate results. However, the model implementation remains currently unavailable publicly. Future work could investigate how this model impacts and compares to the presented results. It is worth mentioning that this model requires collecting the historical posts of the users present in the analysed dataset and does not fix the mentioned potential misgendering issues.

Investigating additional demographics such as user age and language is key for better understanding how fact-checks and misinformation spread. Although models for Twitter exist for predicting such demographics, at the moment, existing models seem to report accuracy that is too low to be used in our analysis (Section 2.3). This is the main reason why we did not use such models in our work (Section 3.1.3).

In this work, we have proposed to evaluate the potential impact of classification errors based on a bootstrapping method. Although this approach gives a good idea of the potential impact of misclassification, a more precise approach would be to manually annotate some user profiles to check the actual error rate of the pre-trained classifiers on our dataset. In this context, future work should consider the manual annotation of user profiles and even the potential fine-tuning of the pre-existing models to the dataset for reducing the error rate of the classifiers. Nevertheless, it is important to note that the proposed approach used in this article presents a worst-case scenario and that in practice the observed errors are likely to be lower. This would be particularly true after fine-tuning the models to manually annotated profiles in the dataset.

For this analysis, we did not consider the demographics of Twitter in general when estimating the importance of specific demographics group. In future work, we hope to integrate this information in order to have a better understanding of the impact of specific information types at the level of Twitter.

The topics used for our analysis may also need to be extended for increasing the granularity of our understanding of COVID-19-related misinformation and fact-check spread. We should investigate how the *other* topic groups can be divided into smaller categories, such as vaccines, lockdown or virus mutations.

Another source of limitations can be linked to the social network analysed. Future work should investigate additional social networks, since misinformation and fact-checks also circulate on other platforms besides Twitter.

8. Outcomes and suggestions

The demographics and topic analysis results highlight some potential issues concerning how fact-checks impact COVID-19 misinformation spread, and suggest some directions where Twitter and fact-checkers could adapt how they handle misinformation-related content.

Improving the reach of fact-checks is key for reducing misinformation since misinforming content spreads significantly more than fact-checks (Vosoughi et al., 2018). This behaviour may simply be due to the intrinsic factual nature of fact-checks that makes them less shareable than highly viral misinforming content (Vosoughi et al., 2018). This observation suggests that fact-checkers may want to adopt publication methods that are more likely to increase the spread of their content on social media by making them more attractive to social media users (Berger & Milkman, 2012).

An important observed issue is that topics linked to COVID-19 conspiracy theories, such as the causes of the virus, are hard to deal with using conventional fact-checking methods. This may be due to missing satisfactory explanations (Douglas et al., 2019). This observation means that fact-checkers may need new approaches for tackling this type of misinformation.

Misinformation reappearance is also another important issue (Section 6.1). Unfortunately, it seems that when this happens, the associated fact-checking content is not always shared again. Multiple approaches may help in dealing with such issues. First, fact-checkers may want to monitor the re-sharing of the misinformation linked to their own fact-checks so they can republish the corresponding fact-checks and re-stimulate the spread of their content. Existing tools may help them dealing with such issues, such as the Fact-checking Observatory.[12] Second, social media platforms may want to provide tools that identify old conversations linked to the same content so that previous fact-check conversations about specific misinforming posts are not completely lost.

Although Twitter has made some recent improvements concerning how it deals with misinformation,[13] additional work is still necessary so that misinforming content is identified faster. In order to do this, better communication between fact-checking organisations and Twitter seems key for the prompt flagging of fact-checked content.

An important observation that confirms previous work (Micallef et al., 2020) is that individual users are more likely to share misinformation but also fact-checks. In order to increase the sharing of fact-checks by individuals, fact-checking organisations may want to create more personalised content so that individuals are more likely to spread it and find methods for leveraging the role of individuals in amplifying the spread of their work.

9. Conclusion

We have significantly extended our previous analysis about the co-spread of misinformation and fact-checks on Twitter (Burel et al., 2020) by covering a larger period of COVID-19-related information spread and investigating their diffusion for different topics and demographics. Compared to our previous results, it seems that fact-checking may not be as successful as expected in reducing misinformation spread on Twitter. However, we know from previous work that exposure to corrective information does work on the cognitive level, if there is enough corrective information and if it comes early enough in the information cycle (Bode et al., 2020; Starbird et al., 2018). We also understand that fact-checking can have other prophylactic advantages, such as building trust in the information environment (Krause et al., 2020). Our work suggests that fact-checking organisations are better than previously thought at responding to misinformation diffusion during COVID-19 (Burel et al., 2020). Nevertheless, fact-checks seem to not impact misinformation spread as well as they should. The analysis for different topics and demographic groups shows that misinformation spread may be independent from particular groups, whereas fact-checks may be more susceptible to group effects. In order to overcome these issues, it appears that understanding why fact-checking works better in some contexts (e.g., particular topics and demographics) and not in others is key to creating better interaction bridges between fact-checking and misinformation spreaders for reducing the spread of misinformation effectively.

CRediT authorship contribution statement

Grégoire Burel: Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing - original draft, Writing - review & editing, Visualization. **Tracie Farrell:** Writing - original draft, Writing - review & editing. **Harith Alani:** Writing - review & editing, Funding acquisition.

Acknowledgements

This work has received support from the European Union's Horizon 2020 research and innovation programme under grants agreement No. 770302 (Co-Inform) and No. 101003606 (HERoS).

References

Agley, J., & Xiao, Y. (2021). Misinformation about COVID-19: evidence for differential latent profiles and a strong association with trust in science. *BMC Public Health*, *21*(1), 1–12.

Aird, M. J., Ecker, U. K., Swire, B., Berinsky, A. J., & Lewandowsky, S. (2018). Does truth matter to voters? The effects of correcting political misinformation in an Australian sample. *Royal Society Open Science*, *5*(12), Article 180593.

Allgaier, J., & Svalastog, A. L. (2015). The communication aspects of the ebola virus disease outbreak in Western Africa–do we need to counter one, two, or many epidemics? *Croatian Medical Journal*, *56*(5), 496.

Almaliki, M. (2019). Online misinformation spread: A systematic literature map. In *Proceedings of the 2019 3rd international conference on information system and data mining*. (pp. 171–178).

Amazeen, M. A., Vargo, C. J., & Hopp, T. (2019). Reinforcing attitudes in a gatewatching news era: Individual-level antecedents to sharing fact-checks on social media. *Communication Monographs*, *86*(1), 112–132.

Barrera, O., Guriev, S., Henry, E., & Zhuravskaya, E. (2020). Facts, alternative facts, and fact checking in times of post-truth politics. *Journal of Public Economics*, *182*, Article 104123.

Bedard, M., & Schoenthaler, C. (2018). Satire or fake news: Social media consumers' socio-demographics decide. In *Companion proceedings of the the web conference 2018* (pp. 613–619).

Berger, J., & Milkman, K. L. (2012). What makes online content viral? *Journal of Marketing Research*, *49*(2), 192–205.

Beskow, D. M., & Carley, K. M. (2018). Bot conversations are different: leveraging network metrics for bot detection in twitter. In *2018 IEEE/ACM international conference on advances in social networks analysis and mining* (pp. 825–832). IEEE.

[12] Fact-checking Observatory, https://fcobservatory.org.

[13] Twitter: Updating our approach to misleading information, https://blog.twitter.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information.html.

Bode, L., Vraga, E. K., & Tully, M. (2020). Do the right thing: tone may not affect correction of misinformation on social media. *Harvard Kennedy School Misinformation Review*.

Brandt, J., Buckingham, K., Buntain, C., Anderson, W., Ray, S., Pool, J.-R., et al. (2020). Identifying social media user demographics and topic diversity with computational social science: a case study of a major international policy forum. *Journal of Computational Social Science*, 1–22.

Brennen, J. S., Simon, F., Howard, P. N., & Nielsen, R. K. (2020). *Types, sources, and claims of COVID-19 misinformation*. Reuters Institute.

Brennen, J. S., Simon, F. M., & Nielsen, R. K. (2021). Beyond (mis) representation: Visuals in COVID-19 misinformation. *The International Journal of Press/Politics*, *26*(1), 277–299.

Burel, G., Farrell, T., Mensio, M., Khare, P., & Alani, H. (2020). Co-spread of misinformation and fact-checking content during the Covid-19 pandemic. In S. Aref, K. Bontcheva, M. Braghieri, F. Dignum, F. Giannotti, F. Grisolia, & D. Pedreschi (Eds.), *Social Informatics* (pp. 28–42). Cham: Springer International Publishing.

Cha, M., Cha, C., Singh, K., Lima, G., Ahn, Y.-Y., Kulshrestha, J., et al. (2020). COVID-19 infodemic prevalence over 35 countries. *JMIR Human Factors*.

Cinelli, M., Quattrociocchi, W., Galeazzi, A., Valensise, C. M., Brugnoli, E., Schmidt, A. L., et al. (2020). The covid-19 social media infodemic. *Scientific Reports*, *10*(1), 1–10.

Cuan-Baltazar, J. Y., Muñoz Perez, M. J., Robledo-Vega, C., Pérez-Zepeda, M. F., & Soto-Vega, E. (2020). Misinformation of COVID-19 on the internet: infodemiology study. *JMIR Public Health and Surveillance*, *6*(2), Article e18444.

Culotta, A., & Cutler, J. (2016). Mining brand perceptions from twitter social networks. *Marketing Science*, *35*(3), 343–362.

Douglas, K. M., Uscinski, J. E., Sutton, R. M., Cichocka, A., Nefes, T., Ang, C. S., et al. (2019). Understanding conspiracy theories. *Political Psychology*, *40*, 3–35.

Evanega, S., Lynas, M., Adams, J., Smolenyak, K., & Insights, C. G. (2020). Coronavirus misinformation: quantifying sources and themes in the COVID-19 'infodemic'. JMIR Preprints.

Farrell, T., Piccolo, L., Perfumi, S. C., Alani, H., & Mensio, M. (2019). Understanding the role of human values in the spread of misinformation. In *Conference for truth and trust online*.

Giorgi, S., Lynn, V., Matz, S., Ungar, L., & Schwartz, H. A. (2019). Correcting sociodemographic selection biases for accurate population prediction from social media. arXiv preprint arXiv:1911.03855.

Guess, A., Nagler, J., & Tucker, J. (2019). Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances*, *5*(1), eaau4586.

Gürlek, M. (2021). *Determining user types from twitter account contentand structure* (Master's thesis), Middle East Technical University.

Harman, S. (2020). The danger of stories in global health. *The Lancet*, *395*(10226), 776–777.

Jiang, S., & Wilson, C. (2018). Linguistic signals under misinformation and fact-checking: Evidence from user comments on social media. *Proceedings of the ACM on Human-Computer Interaction*, *2*(CSCW), 1–23.

Jin, F., Dougherty, E., Saraf, P., Cao, Y., & Ramakrishnan, N. (2013). Epidemiological modeling of news and rumors on twitter. In *Proceedings of the 7th workshop on social network mining and analysis* (pp. 1–9).

Jin, F., Wang, W., Zhao, L., Dougherty, E., Cao, Y., Lu, C.-T., et al. (2014). Misinformation propagation in the age of twitter. *Computer*, *47*(12), 90–94.

Kendall, L. (1998). Meaning and identity in "cyberspace": The performance of gender, class, and race online. *Symbolic Interaction*, *21*(2), 129–153.

Kim, J., Tabibian, B., Oh, A., Schölkopf, B., & Gomez-Rodriguez, M. (2018). Leveraging the crowd to detect and reduce the spread of fake news and misinformation. In *Proceedings of the eleventh ACM international conference on web search and data mining* (pp. 324–332).

Koetke, J., Schumann, K., & Porter, T. (2021). Intellectual humility predicts scrutiny of COVID-19 misinformation. *Social Psychological and Personality Science*, 1948550620988242.

Kouzy, R., Abi Jaoude, J., Kraitem, A., El Alam, M. B., Karam, B., Adib, E., et al. (2020). Coronavirus goes viral: quantifying the COVID-19 misinformation epidemic on Twitter. *Cureus*, *12*(3).

Krause, N. M., Freiling, I., Beets, B., & Brossard, D. (2020). Fact-checking as risk communication: the multi-layered risk of misinformation in times of COVID-19. *Journal of Risk Research*, *23*(7–8), 1052–1059.

Kuklinski, J. H., Quirk, P. J., Jerit, J., Schwieder, D., & Rich, R. F. (2000). Misinformation and the currency of democratic citizenship. *Journal of Politics*, *62*(3), 790–816.

Laato, S., Islam, A. N., Islam, M. N., & Whelan, E. (2020). What drives unverified information sharing and cyberchondria during the COVID-19 pandemic? *European Journal of Information Systems*, *29*(3), 288–305.

Lazarus, J. V., Ratzan, S. C., Palayew, A., Gostin, L. O., Larson, H. J., Rabin, K., et al. (2020). A global survey of potential acceptance of a COVID-19 vaccine. *Nature Medicine*, 1–4.

Lee, J. J., Kang, K.-A., Wang, M. P., Zhao, S. Z., Wong, J. Y. H., O'Connor, S., et al. (2020). Associations between COVID-19 misinformation exposure and belief with COVID-19 knowledge and preventive behaviors: cross-sectional online study. *Journal of Medical Internet Research*, *22*(11), Article e22205.

Lewandowsky, S., Stritzke, W. G., Freund, A. M., Oberauer, K., & Krueger, J. I. (2013). Misinformation, disinformation, and violent conflict: From Iraq and the "War on Terror" to future threats to peace. *American Psychologist*, *68*(7), 487.

van der Linden, S., Roozenbeek, J., & Compton, J. (2020). Inoculating against fake news about COVID-19. *Frontiers in Psychology*, *11*, 2928.

Masuda, N., & Holme, P. (2017). *Temporal network epidemiology*. Springer.

Mensio, M., & Alani, H. (2019). News source credibility in the eyes of different assessors. In *Truth and trust conference*.

Micallef, N., He, B., Kumar, S., Ahamad, M., & Memon, N. (2020). The role of the crowd in countering misinformation: a case study of the covid-19 infodemic. In *2020 IEEE International Conference on Big Data (Big Data)* (pp. 748–757).

Muggeo, V. M. (2003). Estimating regression models with unknown break-points. *Statistics in Medicine*, *22*(19), 3055–3071.

Nieminen, S., & Rapeli, L. (2019). Fighting misperceptions and doubting journalists' objectivity: A review of fact-checking literature. *Political Studies Review*, *17*(3), 296–309.

Nyhan, B., & Reifler, J. (2015). *Estimating fact-checking's effects*. Arlington, VA: American Press Institute.

Pyszczynski, T., Lockett, M., Greenberg, J., & Solomon, S. (2020). Terror management theory and the COVID-19 pandemic. *Journal of Humanistic Psychology*, 0022167820959488.

Rampersad, G., & Althiyabi, T. (2020). Fake news: Acceptance by demographics and culture on social media. *Journal of Information Technology & Politics*, *17*(1), 1–11.

Rich, T. S., Milden, I., & Wagner, M. T. (2020). Research note: Does the public support fact-checking social media? It depends who and how you ask. *The Harvard Kennedy School Misinformation Review*.

Robertson, C. T., Mourão, R. R., & Thorson, E. (2020). Who uses fact-checking sites? The impact of demographics, political antecedents, and media use on fact-checking site awareness, attitudes, and behavior. *The International Journal of Press/Politics*, *25*(2), 217–237.

Rodríguez-Ruiz, J., Mata-Sánchez, J. I., Monroy, R., Loyola-Gonzalez, O., & López-Cuevas, A. (2020). A one-class classification approach for bot detection on twitter. *Computers & Security*, *91*, Article 101715.

Roozenbeek, J., Schneider, C. R., Dryhurst, S., Kerr, J., Freeman, A. L., Recchia, G., et al. (2020). Susceptibility to misinformation about COVID-19 around the world. *Royal Society Open Science*, *7*(10), Article 201199.

Ruberg, B., & Ruelos, S. (2020). Data for queer lives: How LGBTQ gender and sexuality identities challenge norms of demographics. *Big Data & Society*, *7*(1), Article 2053951720933286.

Salali, G. D., & Uysal, M. S. (2020). COVID-19 vaccine hesitancy is associated with beliefs on the origin of the novel coronavirus in the UK and Turkey. *Psychological Medicine*, 1–3.

Sarkar, S., Guo, R., & Shakarian, P. (2019). Using network motifs to characterize temporal network evolution leading to diffusion inhibition. *Social Network Analysis and Mining*, *9*(1), 14.

Saxena, A., Hsu, W., Lee, M. L., Leong Chieu, H., Ng, L., & Teow, L. N. (2020). Mitigating misinformation in online social network with top-k debunkers and evolving user opinions. In *Companion proceedings of the web conference 2020* (pp. 363–370).

Sloan, L., Morgan, J., Housley, W., Williams, M., Edwards, A., Burnap, P., et al. (2013). Knowing the tweeters: Deriving sociologically relevant demographics from Twitter. *Sociological Research Online*, *18*(3), 74–84.

Starbird, K., Dailey, D., Mohamed, O., Lee, G., & Spiro, E. S. (2018). Engage early, correct more: How journalists participate in false rumors online during crisis events. In *Proceedings of the 2018 CHI conference on human factors in computing systems* (pp. 1–12).

Swire, B., Berinsky, A. J., Lewandowsky, S., & Ecker, U. K. (2017). Processing political misinformation: comprehending the Trump phenomenon. *Royal Society Open Science*, *4*(3), Article 160802.

Tambuscio, M., Ruffo, G., Flammini, A., & Menczer, F. (2015). Fact-checking effect on viral hoaxes: A model of misinformation spread in social networks. In *Proceedings of the 24th international conference on world wide web* (pp. 977–982).

Tong, G. A., & Du, D.-Z. (2019). Beyond uniform reverse sampling: A hybrid sampling technique for misinformation prevention. In *IEEE INFOCOM 2019-IEEE conference on computer communications* (pp. 1711–1719). IEEE.

Vaezi, A., & Javanmard, S. H. (2020). Infodemic and risk communication in the era of CoV-19. *Advanced Biomedical Research*, *9*.

Vlachos, A., & Riedel, S. (2014). Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 workshop on language technologies and computational social science* (pp. 18–22).

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151.

Wang, Z., Hale, S., Adelani, D. I., Grabowicz, P., Hartman, T., Flöck, F., et al. (2019). Demographic inference and representative population estimates from multilingual social media data. In *The world wide web conference* (pp. 2056–2067). New York, NY, USA: Association for Computing Machinery.

Xian, J., Yang, D., Pan, L., Wang, W., & Wang, Z. (2019). Misinformation spreading on correlated multiplex networks. *Chaos. An Interdisciplinary Journal of Nonlinear Science*, *29*(11), Article 113123.

Xie, B., He, D., Mercer, T., Wang, Y., Wu, D., Fleischmann, K. R., et al. (2020). Global health crises are also information crises: A call to action. *Journal of the Association for Information Science and Technology*.

Yang, Y.-C., Al-Garadi, M. A., Love, J. S., Perrone, J., & Sarker, A. (2021). Automatic gender detection in Twitter profiles for health-related cohort studies. *JAMIA Open*, *4*(2), ooab042.