Humanities & Social Sciences Communications

ARTICLE

https://doi.org/10.1057/s41599-023-01838-0 OPEN

Systematic meta-analysis of research on AI tools to deal with misinformation on social media during natural and anthropogenic hazards and disasters

Rosa Vicari[1]✉ & Nadejda Komendatova[1]

The spread of misinformation on social media has led to the development of artificial intelligence (AI) tools to deal with this phenomenon. These tools are particularly needed when misinformation relates to natural or anthropogenic disasters such as the COVID-19 pandemic. The major research question of our work was as follows: what kind of gatekeepers (i.e. news moderators) do we wish social media algorithms and users to be when misinformation on hazards and disasters is being dealt with? To address this question, we carried out a meta-analysis of studies published in Scopus and Web of Science. We extracted 668 papers that contained keyterms related to the topic of "AI tools to deal with misinformation on social media during hazards and disasters." The methodology included several steps. First, we selected 13 review papers to identify relevant variables and refine the scope of our meta-analysis. Then we screened the rest of the papers and identified 266 publications as being significant for our research goals. For each eligible paper, we analyzed its objective, sponsor's location, year of publication, research area, type of hazard, and related topics. As methods of analysis, we applied: descriptive statistics, network representation of keyword co-occurrences, and flow representation of research rationale. Our results show that few studies come from the social sciences (5.8%) and humanities (3.5%), and that most of those papers are dedicated to the COVID-19 risk (92%). Most of the studies deal with the question of detecting misinformation (68%). Few countries are major funders of the development of the topic. These results allow some inferences. Social sciences and humanities seem underrepresented for a topic that is strongly connected to human reasoning. A reflection on the optimum balance between algorithm recommendations and user choices seems to be missing. Research results on the pandemic could be exploited to enhance research advances on other risks.

[1] Cooperation and Transformative Governance Research Group Advancing Systems Analysis, International Institute for Applied Systems Analysis (IIASA), Laxenburg, Austria. ✉email: vicari@iiasa.ac.at

Introduction

Fake news is old news: different forms of misinformation have occurred repeatedly throughout history (Novaes and de Ridder, 2021). Nevertheless, the impact of different communication technologies on content, production, distribution, and consumption of misinformation—and hence, its dissemination—has changed over history (Posetti and Matthews, 2018). In the age of social media, misinformation spreads at a fast pace: the pervasive nature of misinformation in the digital age is being reinforced by both technical and socio-psychological factors (Dallo et al., 2022).

The speed of diffusion influences the amplitude of negative impacts which, because of their cascading effects, can be exponential in the context of disasters (McGee et al., 2016). The COVID-19 "infodemic," as the World Health Organization (2022, p. 1) defines it, had various negative impacts; these include psychological consequences (such as anxiety, depression, or post-traumatic stress disorder), reduced trust in public authorities and health institutions, adoption of inadequate protective measures by the population, and increased purchases of medical supplies and other products which stresses on the market (Pian et al., 2021).

Misinformation can also occur in relation to natural disasters; for instance, during and after the Hurricane Irma disaster which hit the Caribbean in September 2017, several rumors spread, among others, fake news concerning the number of deaths on the French territory of Saint Martin. According to the rumors, there were between over 100 and over 1000 dead, while the real death toll was 11. This fake news continued to circulate for more than a year, negatively impacting the territory's social cohesion and post-hurricane reconstruction (Moatty et al., 2019).

In the last two years, numerous studies have been carried out to develop artificial intelligence (AI) tools that can deal with misinformation in risk management contexts that require a very fast response. The very recent and fast development of research efforts on this topic necessitates a timely and efficient review of the current research trends. In this study, we conduct a meta-analysis of the literature to identify the main research gaps and to answer the following question: what kind of gatekeepers (i.e., news moderators) do we wish social media algorithms and users to be in terms of dealing with misinformation on hazards and disasters? This meta-analysis will contribute to developing a communication model based on social media moderation and recommendation practices that are aligned with human rights and journalism ethics.

Background

The spread of misinformation on social media. Misinformation has ancient origins. According to Kaminska (2017), one of the earliest records of misinformation goes back to 30 BCE, to the time of hostilities between Mark Antony and Octavian over the leadership of the Roman world. Across history, the impact of misinformation has changed with the evolution of technology: for instance, the invention of the printing press led to the first large-scale news hoax (Thornton, 2000). In the digital age, the dissemination of misinformation is so greatly amplified that, since 2018, several governments have started to introduce regulatory measures at the national level to combat fake news (Posetti and Matthews, 2018).

Various definitions are proposed in the literature to define different kinds of information disorders. Lazer et al. (2018, p. 2) define fake news as "fabricated information that mimics news media content in form but not in organizational process or intent [and] overlaps with other information disorders, such as misinformation (false or misleading information) and disinformation (false information that is purposely spread to deceive people)." Ireton and Posetti (2018, p. 43) recommend using the terms "misinformation" and "disinformation" to indicate information disorders, and to avoid using the term "fake news" as it has been "politicized and deployed as a weapon against the news industry, as a way of undermining reporting that people in power do not like." In this study, we refer exclusively to misinformation, as this is frequently used as a general term in the scientific literature to include different information disorders, such as disinformation, rumors, misinformation, and hoaxes.

Misinformation on social media and risk management. Social media contribute to the social representation of hazards and disasters (Sarrica et al., 2018); in other words, they shape the population's perception and attitude regarding hazards and disasters. Ng et al. (2018) compare traditional media with social media and highlight that the latter has a stronger effect in terms of increasing their readers' risk perception. Tsoy et al. (2021) suggest that social media can shape hazard experience in two ways: either by amplifying risk perception or reducing it.

In this context, misinformation can strongly affect risk management. One example is the spread of rumors and hoaxes on social media that followed the 2017 Manchester Arena Bombing (Qiu, 2017). In particular, the news that unaccompanied children had been sheltered in hotels was a false rumor; this illustrates how such misinformation can misdirect the affected population and cause confusion and chaos (Hunt et al., 2020). Obviously, other examples can be taken from the COVID-19 pandemic. The significant impact of misinformation during the pandemic led the United Nations to urge countries to take action to combat the "infodemic," defined as "too much information including false or misleading information in digital and physical environments during a disease outbreak. It causes confusion and risk-taking behaviors that can harm health" (World Health Organization, 2022, p. 1).

This research addresses both hazards and disasters—concepts that are related, but distinct. According to the United Nations Office for Disaster Risk Reduction (United Nations Office for Disaster Risk Reduction (UNDRR), 2023) and the Federal Emergency Management Agency FEMA (2023) of the United States, a hazard represents a potential threat, while a disaster is the actual damage caused by a hazard. Misinformation can impact both hazards and disasters and can hinder efforts to prevent and reduce risks associated with these events.

The need for AI tools to deal with misinformation. In the last decade, a wide variety of data mining tools have been developed to gauge public opinion by exploring big digital communication datasets. It has become possible to automatically detect misinformation thanks to natural language processing, machine learning, and deep learning (Ayo et al., 2020; Hossein and Miller, 2018; Murfi et al., 2019). Machine learning and deep learning algorithms (Fig. 1)—two subsets of the broader category of artificial intelligence—are two of the most common approaches to automating the process of classifying unreliable or reliable news (Varma et al., 2021).

Over a decade ago, Grzywińska and Borden (2012) stated that social media were replacing traditional media as the preeminent information source and the main player in public agenda-setting. This trend has strengthened to such an extent that newspaper headlines frequently quote social media items as sources of information (Paulussen and Harder, 2014). Along with the increasing importance of social media, AI tools are playing a key role in our society to deal with the rapid spread of misinformation, while guaranteeing the right of access to information

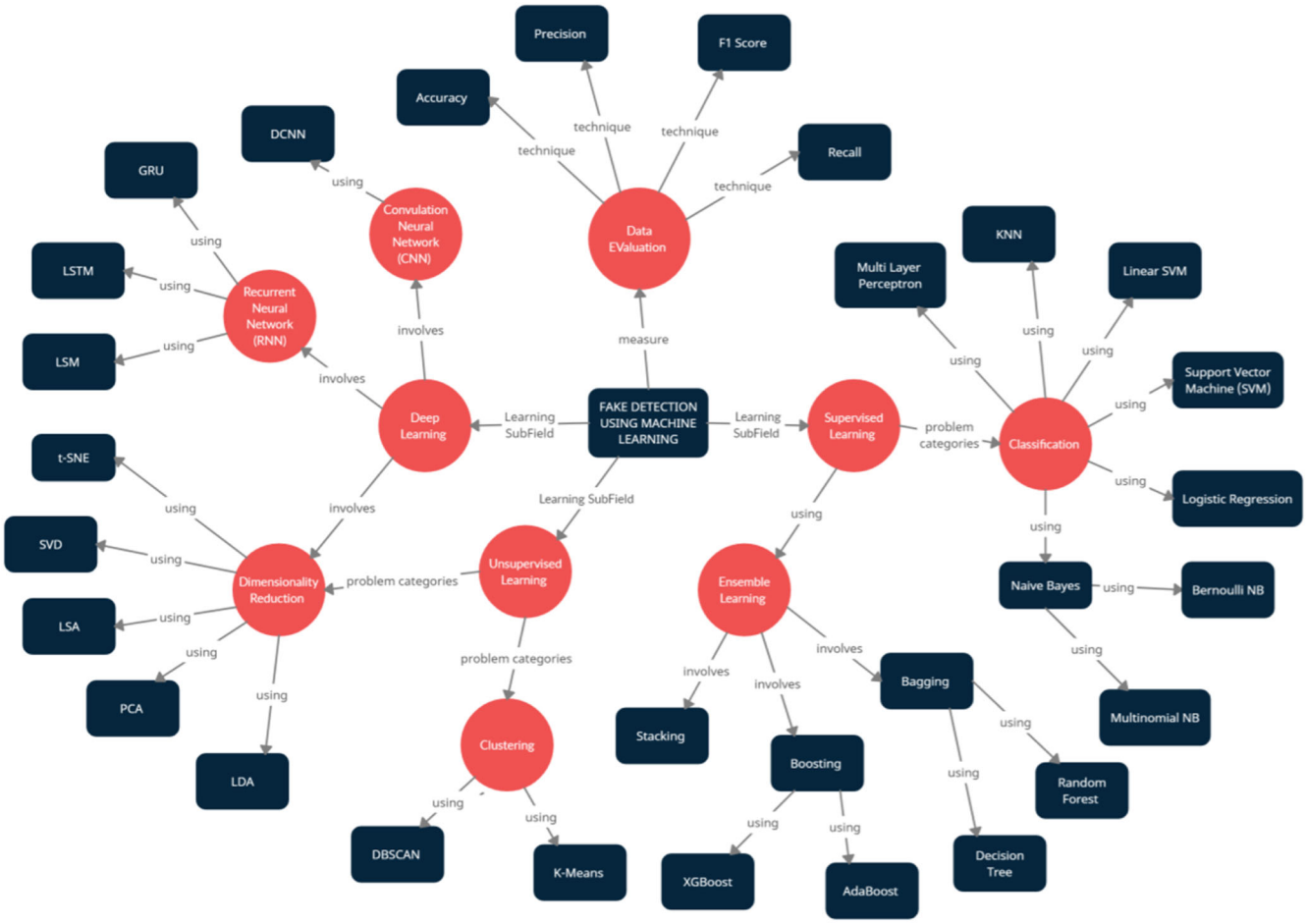

Fig. 1 Concept map outlining the various techniques used for detecting fake news detection, as proposed by Varma et al. (2021). The authors focused their study on Machine Learning (ML) and Deep Learning (DL), which are two of the most commonly used approaches for classifying misinformation. ML algorithms can be divided into two subfields: supervised learning and unsupervised learning. The former includes techniques such as Naïve Bayes, support vector machine, and logistic regression, while the latter includes *K*-Means and DB-Scan among others. Ensemble Learning is a type of Supervised Learning that is both accurate and innovative, according to the authors. It encompasses techniques such as AdaBoost, XGBoost, decision tree, and random forest. DL algorithms are also effective in detecting misinformation. The most popular DL techniques include convolution neural network (CNN) and recurrent neural networks such as long short-term memory. LSTM is a widely used technique.

mentioned in Article 19 of the Universal Declaration of Human Rights (United Nations, 1948).

Toward a new communication model. The traditional role of the press as "gatekeepers" has weakened, as journalists no longer hold exclusive rights to select, extract, and disseminate the news in the digital age (Canter, 2014, p. 102). Within the context of social media, content recommendation algorithms and individual media users have taken on a "gatekeeping function" (Napoli, 2015, p. 755).

Within the expanding and diverse area of studies dedicated to AI tools to deal with misinformation, scientists cannot discard the following question: what kind of gatekeepers do we wish moderation and recommendation algorithms—and also social media users—to be? This question addresses fundamental human rights and journalism ethics, such as freedom of expression, the right of the public to be informed, accuracy, and differentiation between fact and opinion, privacy protection, hate and discrimination prevention, and plagiarism prevention. For instance, Jørgense and Zuleta (2020, p. 52) point out that social media constitute a public space where freedom of expression is vulnerable, given that these platforms are "governed by private actors operating outside the direct reach of human rights law." According to the two authors, European Union policies encourage content moderation by private companies, but their "over-removal" practices are rarely aligned with the principles of "strictly necessary" and "proportionate" limitation that are mentioned in the international human rights law (Jørgensen and Zuleta, 2020, p. 59).

Methodology

Data for analysis. We extracted our corpus of abstracts from Web of Science and Scopus, two platforms that provide access to abstract, reference, and citation data from academic journals. The abstracts were all studies published up until 1 July 2022. We extracted the corpus of studies on the basis of the following key terms related to the research area of AI tools to deal with misinformation on social media during hazards and disasters.

Corpus selection based on keyterms included in abstracts (Web of Science & Scopus)

Abstract =(disaster) OR (emergenc) OR (hazard) OR (disaster) OR (flood) OR (earthquake) OR (industrial accident) OR (terrorist attack*) OR (COVID) OR (pandemic) OR (wildfire) OR (Coronavirus)*

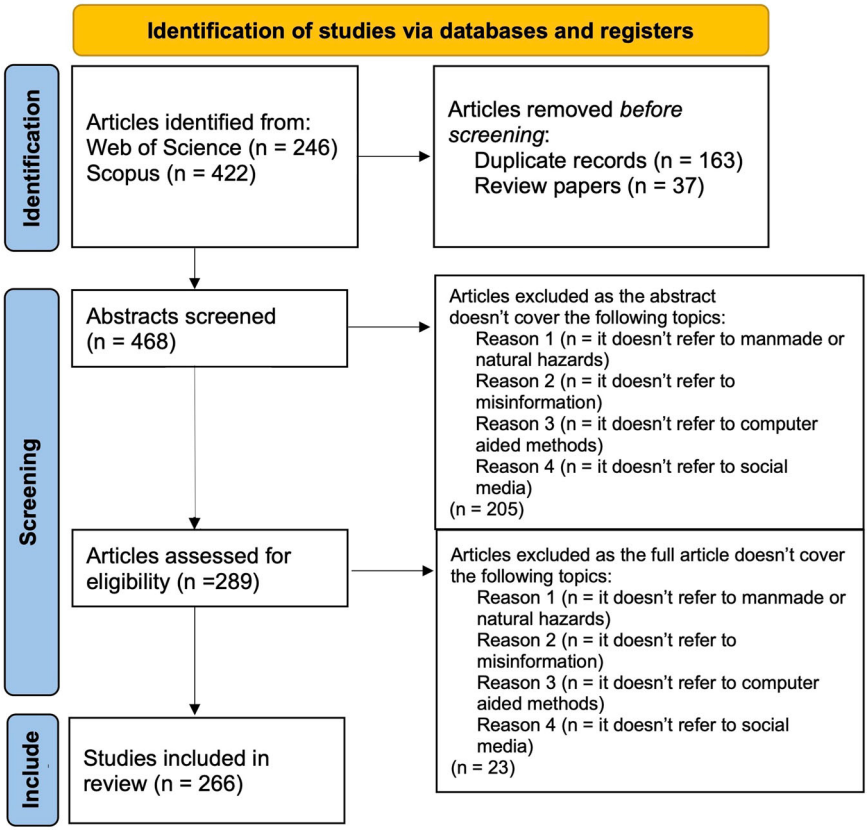

Fig. 2 Our data selection process, guided by the PRISMA 2020 flow diagram, which outlines the steps involved in conducting a meta-analysis and the corresponding information flow. This diagram serves as a useful tool for documenting the number of documents that were selected, assessed, deemed eligible or ineligible, as well as the reasons for exclusion (Page et al., 2021).

AND

Abstract =(social media) OR (Twitter) OR (WhatsApp) OR (Facebook) OR (Instagram) OR (YouTube)

AND

Abstract =(detect) OR (monitor) OR (prevent) OR (screen) OR (AI) OR (artificial intelligence)

AND

Abstract =(fake news) OR (misinformation)

The Boolean operator "OR" means that the selected abstracts must include one of the keyterms. The Boolean operator "AND" means that the selected abstracts must combine two search queries. A search query always starts with "Abstract" = in order to search for keyterms included in the abstracts.

The search keyterms include different anthropogenic and natural hazards. As discussed in section "Introduction", misinformation can affect both anthropogenic and natural hazards and disasters. Furthermore, the Horizon 2020 CORE project (European Union's Horizon 2020 research and innovation program, 2023) highlights that the practitioners (e.g., civil protection), who have to cope with different types of hazards and sometimes have to face multiple overlapping risks, are requesting a dedicated strategy to deal with risk misinformation in different contexts.

The search keyterms include "social media" and the names of popular platforms: "Twitter," "WhatsApp," "Facebook," "Instagram." We used these keyterms in order to include studies with a focus on one of these platforms.

We then used the PRISMA 2020 flow diagram (Fig. 2) (Page et al., 2021) to report which papers were selected and included in our study. After extracting 246 abstracts from Web of Science and 422 abstracts from Scopus, we removed 163 duplicate records and 37 review papers. We then manually screened the remaining 468 abstracts and excluded 205 of them, as they did not refer to one of the following topics as central subjects of the study presented in the corresponding paper: anthropogenic or natural disasters and hazards, misinformation, social media, and AI methods. Finally, we manually assessed for eligibility 289 articles, excluding 23 papers that did not refer to the above-mentioned topics as key topics for the study. As a result, 266 studies were included in the meta-analysis.

Literature review. Our initial corpus of studies on AI tools to deal with misinformation on social media during hazards and disasters included various review papers. These papers were examined to establish the current state of the art for our research topic. The corpus of studies, presented in the section "Data for analysis", includes 37 review papers, of which 24 were excluded for not being central to our literature review. The 24 review papers in question did not indeed consider the following themes to be central to *their* literature review: anthropogenic or natural hazards and disasters, misinformation, social media, and computer-aided methods. This left us with 13 review papers that were directly relevant to our research topic.

We analyzed these 13 review papers to achieve two goals:

1. Verify if they covered all of our research themes, namely all the search key terms used to select our paper corpus or only a part of them;
2. Identify any research variables that were not addressed in these papers.

To begin, we identified all the research themes and variables proposed in the 13 review papers. These are listed in the two tables (Tables 1 and 2).

Different research themes are addressed in each of the 13 papers to select the corpus of documents for the review. Table 1 summarizes and compares these research themes. The authors of the reviews also propose different research variables. We present and compare the main research variables in Table 2.

We used the research themes and research variables identified in Tables 1 and 2 to elicit the following observations, which are an important step for our research:

1. All the review papers focus on the COVID-19 crisis or other disease outbreaks. None of the 13 review articles cover other types of natural or anthropogenic hazards. Hence, using a meta-analysis of a corpus covering both anthropogenic and natural hazards, we opened up the scope of our study compared to other reviews. Indeed, there does not seem to be a precedent for a meta-analysis covering such research themes.
2. With regard to the research variables, four review papers (Ansar and Goswami, 2021; Gabarron et al., 2021; Himelein-Wachowiak et al., 2021; Varma et al., 2021) focus on reviewing the main AI tools used to deal with misinformation. Varma et al. (2021) compare different AI methods and two different publication periods (before and after the pandemic). Ansar and Gaswami (2021) compare the AI methods, misinformation origins, and contents. Gabarron et al. (2021) compare misinformation contents and impacts. Himelein-Wachowiak et al. (2021) specifically focus on bots and compare their origins, topics, and dissemination patterns. What appears to be missing as a research variable, however, is a wider reflection on the different research objectives in the literature on the topic of "tools to deal with misinformation on social media related to hazards and disasters."
3. As well as the research variable *objectives of the study*, three other research variables seem to be missing:

- The *research areas* covered by this topic;
- The natural and anthropogenic *hazards* covered;
- The *location of the funding sponsors*,

These research variables define the scope of our meta-analysis.

Methods of analysis. The proposed meta-analysis aims to explore the 266 studies according to the research themes and research variables identified above: research area, type of hazard, research objective, and location of the funding sponsor. We also considered the "publication year" in order to comprehend how relevant the recent increase in publications is and if it is correlated with other trends. For each research question we applied different methods of analysis: descriptive statistics (for the year of publication, research area, type of hazard, sponsor's location), network representation of keyword co-occurrences (for the type of hazard and the related topics), and flow representation of research rationale (for the objective of the study).

Year of publication. Scopus and Web of Science automatically provide the information on the year of publication in a separate column of the abstract dataset (in CVS format) that can be exported from both websites. The number of publications per

Table 1 Research themes described in the review papers.

	COVID	Disease outbreak	Vaccine	Social media	Digital media	Detect	Hesitancy	Fake news	Bots	Health literacy	Computer-aided methods	Other methods
Varma et al. (2021)	×			×		×		×			×	
Himelein-Wachowiak et al. (2021)	×							×	×		×	
Ansar and Goswami (2021)	×							×			×	
Gabarron et al. (2021)	×			×				×			×	×
Alamoodi et al. (2021)	×		×				×				×	
Garett and Young (2021)	×		×		×		×	×				
Joseph et al. (2022)	×			×				×				
Chowdhury et al. (2021)		×						×				
Bin Naeem and Kamel Boulos (2021)	×			×				×				
Liu and Xiao (2021)		×								×	×	
Tsao et al. (2021)	×			×								
Ivarez-Galvez et al. (2021)		×		×	×			×				
Salehinejad et al. (2021)		×				×		×				

Each row corresponds to a review paper (the references are indicated in the first column on the left of the table) and each column corresponds to a different theme.

Table 2 Research variables proposed in the review papers.

	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.
Varma et al. (2021)		x			x				x			x	
Himelein-Wachowiak et al. (2021)		x	x	x									
Ansar and Goswami (2021)		x	x						x				
Gabarron et al. (2021)			x										x
Alamoodi et al. (2021)			x	x		x	x	x					x
Garett and Young (2021)								x	x				
Joseph et al. (2022)		x					x	x					x
Chowdhury et al. (2021)			x	x	x	x							
bin Naeem and Kamel Boulos (2021)							x	x	x	x			
Liu and Xiao (2021)							x						
Tsao et al. (2021)							x	x					x
Alvarez-Galvez et al. (2021)	x	x				x			x	x		x	
Salehinejad et al. (2021)	x	x							x			x	

Each row corresponds to a review paper (the references are shown in the first column on the left of the table) and each column corresponds to a different research variable. Research variables: 1. Which diseases are the subject of misinformation? 2. What are the sources of misinformation? 3. What kind of content is subject to misinformation? 4. What are the patterns of misinformation dissemination? 5. What are the channels of misinformation dissemination? 6. What factors are correlated with misinformation? 7. What solutions can help to detect misinformation? 8. What solutions can help to combat misinformation? 9. What are the main computer-aided methods able to deal with misinformation? 10. What are other quantitative and qualitative methods of dealing with misinformation described in the literature? 11. How can health literacy be developed? 12. What geographical area/years of publication are covered by existing research papers? 13. What are the impacts of misinformation?

year can easily be extracted and visualized with a bar chart (available in Microsoft Excel for descriptive statistics).

Research areas. Web of Science automatically provides the research area of each paper as part of the abstract dataset, in the column entitled "WoS categories." Scopus does not include information on the research area in the exportable abstract dataset. The list of research areas is, however, available on the search result webpage of Scopus as part of the filter tool entitled "Subject Area." This search filter makes it possible to organize and extract the abstract from the dataset in different subsets corresponding to different research areas.

The next step was to refine the research area classifications proposed by Web of Science and Scopus. The list of research areas is indeed rich, but it is not uniform in Scopus and Web of Science, and an important number of the studies are associated with more than one research area. We thus simplified the list of research areas by merging neighboring disciplines and synonymous terms. We obtained a single simplified list of 12 research areas.

We used the following scoring system to calculate the portion of studies that refers to each research area. We assigned 12 points to a research area when a study referred to it as its sole research area; 6 points when a study referred to it plus a second research area; 4 points when a study referred to it plus two other research areas; and 3 points when a study referred to it plus three other research areas.

We summed the points assigned to each research area. We then converted the total scores, corresponding to each research area, to a percentage. To illustrate the distribution of studies across research areas, we created a bar chart with Excel.

Type of hazard and related topics. We manually screened the abstracts and articles to identify which hazard each study refers to. We identified six types of hazards (multiple hazards, disease outbreaks, COVID-19, floods, earthquakes, and hurricanes) and we counted the number of articles referring to each type of hazard. We finally calculated the percentage of studies referring to each type of hazard.

To further develop the analysis, we tried to explore other topics covered in each study and their relation to the type of hazard that we had previously identified. We followed different steps to explore these topics. First, we extracted the author keywords and the index keywords associated with each article (which were provided by Scopus and Web of Science in two dedicated columns of the abstract dataset). We merged the synonyms presented in Table 3.

Table 3 After extracting the author keywords and the index keywords listed in each article of our corpus, we merged the synonyms presented in this table.

Keyword	Replaced by
machine-learning	machine learning
coronavirus disease 2019	covid-19
Lstm	long short-term memory
Coronaviruses	coronavirus
fake detection	fake news detection
social media platforms	social media
social networks	social network
social networking	social network
social networking (online)	social network
natural language processing sy	natural language processing
natural language processing systems	natural language processing
machine learning models	machine learning
machine learning processing	machine learning
Nlp	natural language processing
Pandemics	pandemic
Humans	human

Except for the first line, each line in the table specifies a label (in the "keyword" column) and an alternative label (in the "replaced by" column), meaning that the label was replaced by the alternative label.

In the next step, we produced a network representation with VOSviewer (Centre for Science and Technology Studies, 2022) based on the list of keywords and their co-occurrence in each paper. In the resulting network, each node corresponds to a keyword. The size of the node depends on how many papers refer to the keyword: the bigger a node, the greater the number of papers that cite it. A link between two nodes (i.e., between two keywords) appears if two keywords co-occur in the same paper. A color code is used to identify different node clusters. The only nodes and clusters that appear in the network representation are nodes with at least 5 co-occurrences and clusters with at least 5 nodes.

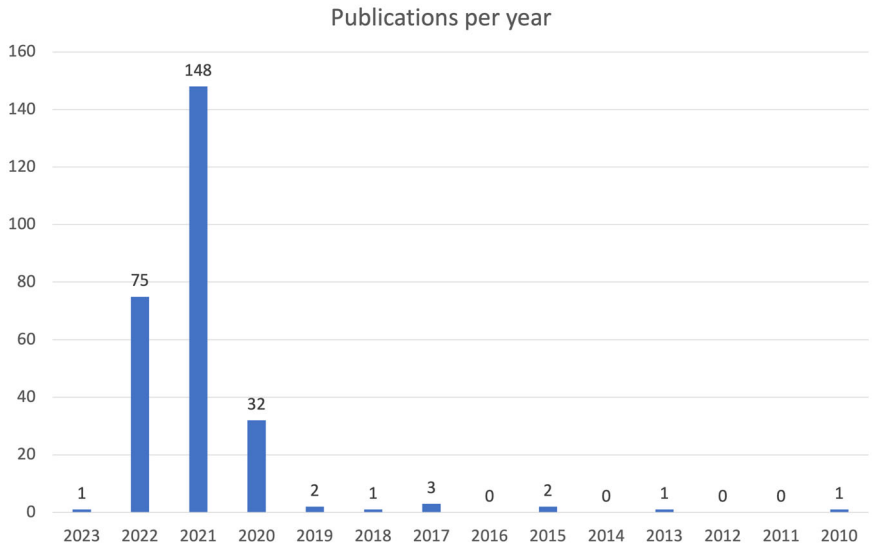

Fig. 3 Growth in research on "AI tools to deal with misinformation on social media during hazards and disasters". The number of publications per year began to increase in 2020, with 32 articles and experienced a significant peak in 2021 with 148 publications.

The objective of the study. Manual screening made it possible to identify 5 general research objectives and 21 research sub-objectives. Given that several studies in our sample refer to more than one general objective or sub-objective, we used the following scoring system to calculate the portion of articles that covers each general objective or sub-objective. We assigned 1 point to a general objective/sub-objective referred to in a study as its sole general objective/sub-objective. We assigned 0.5 points to a general objective/sub-objective referred to it in a study as a general objective/sub-objective plus a second general objective/sub-objective. We summed up the points assigned to each general objective and sub-objective. We then converted the total scores assigned to each general objective and sub-objective to a percentage. We created a Sankey plot, a flow diagram, with SankeyMATIC (Bogart, 2022) to illustrate the distribution of studies across the 5 general objectives and the 21 sub-objectives.

The geographical location of the sponsor. When an article refers to the organization that funded the study, Scopus and Web of Science provides this information in a dedicated column of the abstract dataset. Of the 266 papers that constitute our corpus, 90 refer to a funding organization. We labeled each of the 90 papers manually with the location of the funding organization. The location corresponds to a country or a region (in the case of studies funded by the European Union). We identified 33 different countries. We counted the number of studies associated with each sponsor's location and identified 7 ranges of values: (i) 25 papers; (ii)at least 14–16 papers; (iii) at least 12–13 papers; (iv) 11 papers; (v) 6 papers; (vi) at least 3–4 papers; and (vii) at least 1–2 papers. We defined a color code, with 7 different colors corresponding to the different ranges of values, so that we could use a map to illustrate which countries are associated with which value ranges.

We then wanted to verify if the research efforts vary in each country because of varying degrees of COVID-19 impact. Hence, we compared the number of publications per country with the local number of deaths due to COVID-19 (number of deaths per 1 million population reported by Worldometer (2023)).

Results and discussion

This study consists of a meta-analysis of 266 eligible studies on the topic of AI tools to deal with misinformation on social media during hazards and disasters. As described in the previous section, for each eligible paper we analyzed its objective, the sponsor's location, the year of publication, the research area, the type of hazard, and its related topics according to different methods. In this section, we present the results obtained for each of our five research variables.

Year of publication. Figure 3 illustrates how many studies in our sample have been published each year since 2010. We can observe that the number of publications per year starts to increase in 2020 with 32 articles, and an important peak follows in 2021 with 148 publications. Given that the papers were collected up until 1 July 2022, we cannot observe the results for the full current year. One study is dated 2023 because it was published before the journal revision process was finalized. These results confirm that since 2020 there has been a fast and significant development in the number of studies dedicated to AI tools to deal with misinformation on social media during hazards and disasters. We can infer that this trend is due to the COVID-19 pandemic. Indeed, this research topic is strongly connected to the information disorders that occurred during the pandemic.

Research area. Figure 4 shows the per research area distribution of the studies included in our sample. The chart highlights that the largest portion of papers (50.3%) concerns studies in the field of "Computer Science"; "Engineering" (12.8%) and "Medicine" (12.4%) appear as the second and third most relevant research areas in our corpus. We can also observe that "Social Sciences" (5.8%), "Humanities and Communication" (3.5%), "Business, Management and Decision Sciences" (3%), and "Psychology and Neuroscience" (1%) seem underrepresented, given that these last four research areas are strongly connected to human reasoning. This result could be explained by the use of different terminology in different scientific fields: it is possible that the research areas that are underrepresented rarely use terms such as "detect," "monitor," "prevent," "screen," "AI," "artificial intelligence," namely the search keyterms that we used to select our corpus of studies.

Type of hazard and related topics. Figure 5 illustrates the per hazard type distribution of the studies included in our sample. The chart clearly shows that a striking majority of studies concern COVID-19. This result seems to confirm our hypothesis that the context of the pandemic strongly contributed to the rapid

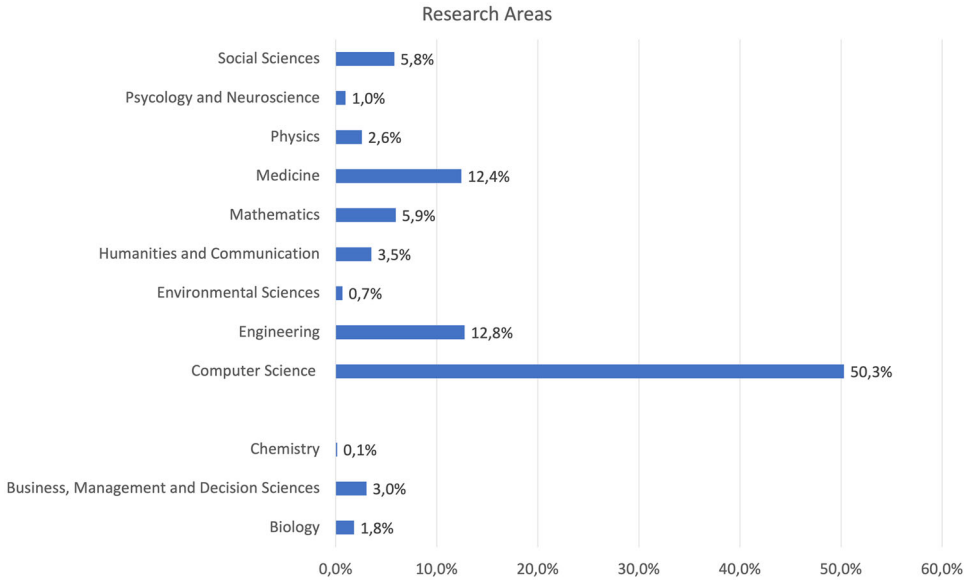

Fig. 4 Chart showing the distribution of research areas in the corpus of studies. It is clear from the chart that the largest share of papers (50,3%) pertains to studies in the field of "Computer Science".

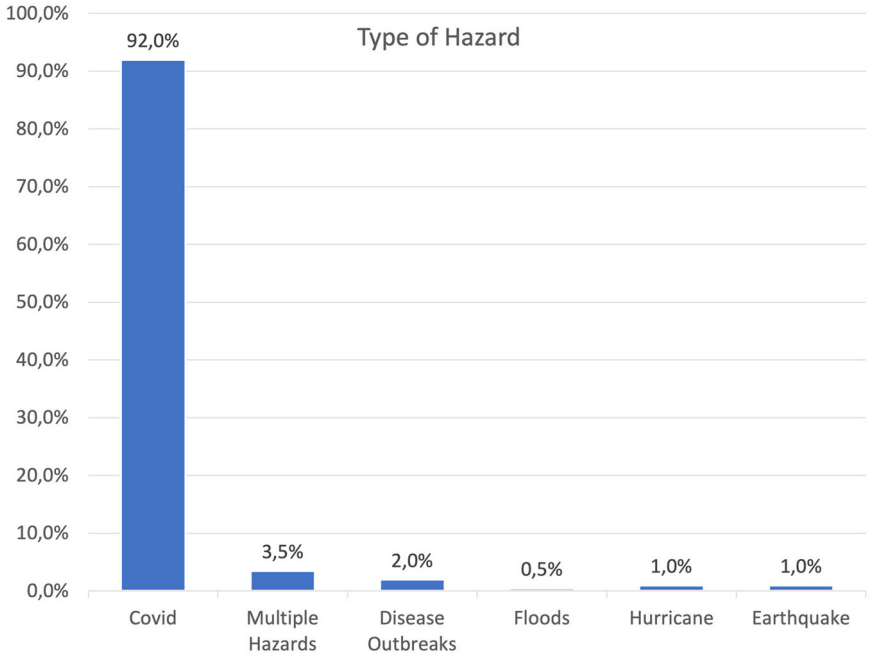

Fig. 5 Distribution of hazard types in the corpus of studies. The corpus covers various types of hazards, and the figure displays the percentage of papers pertaining to each hazard. Notably, the vast majority of studies focus on COVID-19 (92%).

increase in publications. On the other hand, studies that concern other types of hazards (i.e., floods, earthquakes, hurricanes, multiple hazards, and disease outbreaks other than COVID-19) seem underrepresented in our corpus.

Figure 6 shows a network representation with 71 nodes and four clusters of nodes. The network links highlight if two keywords (the nodes of the network) are cited in the same paper. One of the biggest nodes corresponds to the keyword "COVID-19", which means that many papers in the corpus refer to this keyword. This finding confirms the result presented in Fig. 6 (i.e., the majority of the studies in our sample concern COVID-19). This is also confirmed by 16 smaller nodes related to the topic of COVID-19, while other types of hazards are not covered by the keywords that appear in the network. We can also observe that another relevant node is "social media." This result is because "social media" is among those key terms that we used to select our corpus; the vast majority of papers in our sample thus include this key term.

We can observe four clusters of nodes: each cluster brings together the keywords that frequently co-occur. Most of the keywords included in the red cluster (with 21 nodes) and the blue cluster (with 15 nodes) refer to the research scope. For instance, the blue cluster includes keywords such as "coronavirus," "infodemic," "public health," and "information dissemination,"

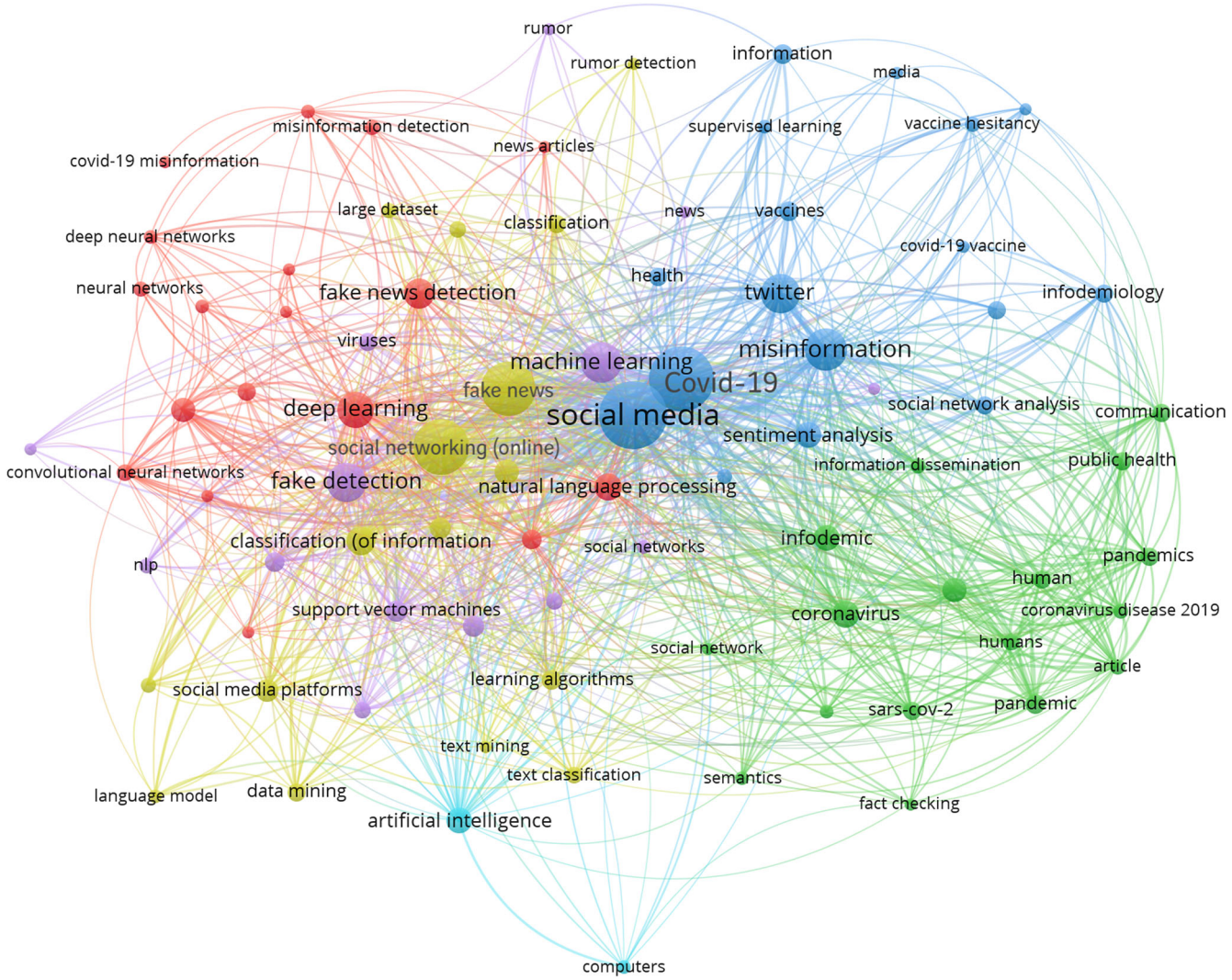

Fig. 6 Visualization of the abstract dataset as a network. The author keywords and the index keywords are depicted as nodes, and their co-occurrence in each publication is shown as links between two nodes. One of the most prominent nodes in the network corresponds to the keyword "COVID-19". The network can be further divided into four clusters, where each cluster groups together keywords that frequently appear together. The red and blue clusters, with 21 and 15 nodes respectively, consist mainly of keywords related to the research scope. In contrast, the yellow and green clusters, with 15 and 20 nodes respectively, contain mostly keywords related to analytical methods, especially AI techniques. These 25 keywords are part of the specialized language of "Computer Science".

and the red cluster includes keywords such as "rumor detection," "vaccine hesitancy," and "Twitter." Both clusters include keywords referring to COVID-19 (6 keywords in the red cluster and 8 keywords in the blue cluster), while other hazards do not appear among the keywords. The red cluster includes four keywords that refer to AI methods ("machine learning," "supervised learning," "topic modeling," and "sentiment analysis"), but the number of keywords of this type is small in comparison to those in yellow cluster and the green cluster.

Indeed, the yellow cluster (with 15 nodes) and the green cluster (with 20 nodes) include a majority of keywords referring to methods of analysis and more specifically to AI techniques. These keywords (25) come from the jargon of "Computer Science." This result is consistent with the data in Fig. 4 which highlight that "Computer Science" is the most prolific research area on the topic of AI tools to deal with misinformation on social media during hazards and disasters.

We can also observe that the yellow cluster does not include any keyword referring to hazards, while the green cluster includes only two words related to COVID-19.

The cluster structure highlights that part of our sample studies, through the keywords selected by the authors and the editors, is identified as a contribution to the research on COVID-19 information. Another part of the study is identified for its contribution to the development of new or improved AI methods.

The objective of the study. Figure 7 is a Sankey plot that illustrates, on the left, the research general objectives and, on the right, the corresponding sub-objectives, which are covered by the studies included in our corpus. We can see that a huge variety of general objectives are covered: from the detection of misinformation, impact assessment, and content analysis to the identification of the causes of misinformation and combating misinformation. The sub-objectives are also very diverse: from multilingual detection or bot-debunking to dissemination pattern monitoring or the analysis of the heuristic process, etc.

The plot highlights that most studies in the corpus refer to "detecting misinformation" (68%) as a general objective and to "classification" solutions (52%) as a sub-objective. These studies provide solutions to identify unreliable information but do not

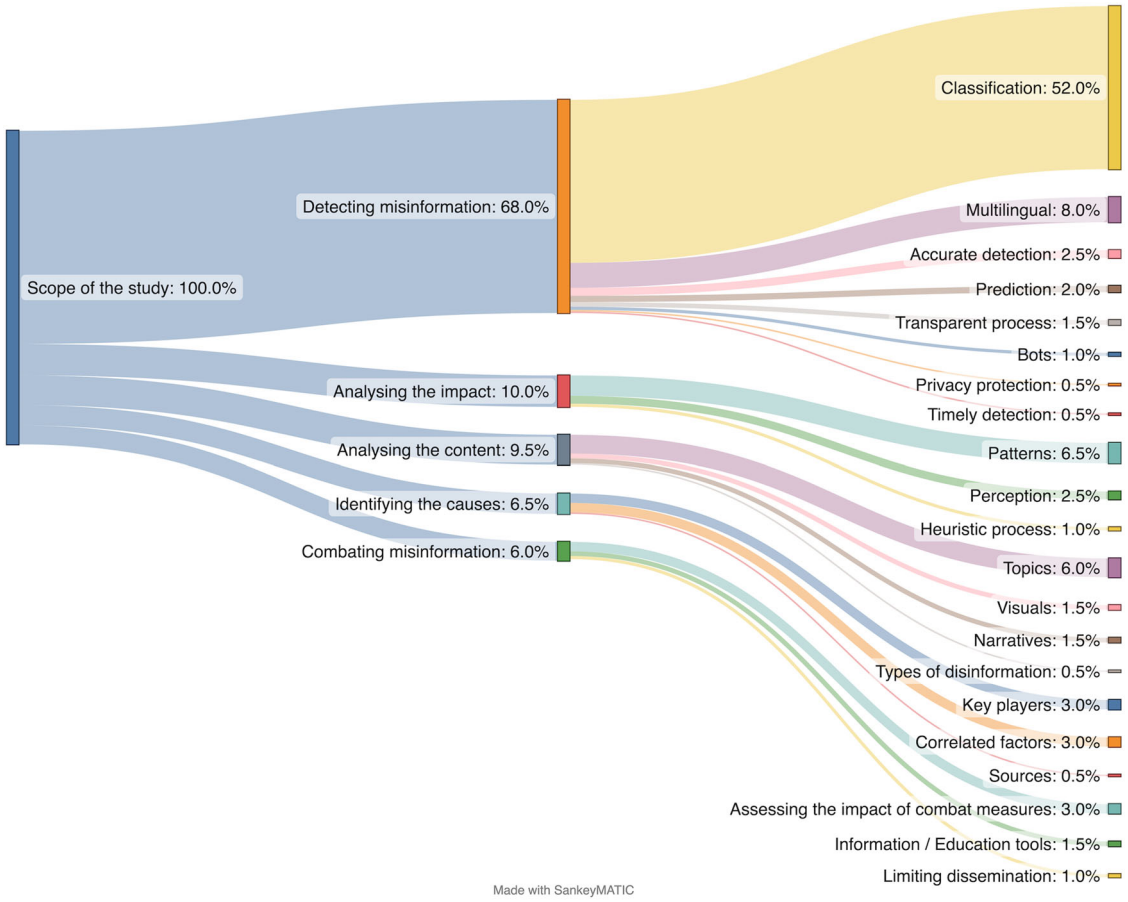

Fig. 7 Sankey plot displaying the distribution of studies across general and sub-objectives. The plot was created using sankeymatic.com (Bogart, 2022): it provides a summary of the diverse range of general objectives and sub-objectives covered in our corpus. Additionally, it highlights that the majority of studies focus on "detecting misinformation" (68%) and "classification" solutions (52%) as their general objectives.

directly deal with the question of "combating misinformation," a general objective that only 6% of the studies have.

Location of the sponsoring organization. According to Fig. 8, few countries are major funders of research on the topic of AI tools to deal with misinformation on social media during hazards and disasters. The United States is the most frequent funder (with 25 papers), followed by China, Spain, and Italy (with between 14 and 16 papers) in second position. The countries that have between 11 and 13 papers are all located in the European Union and can therefore access the programs funded by the European Commission. The number of sponsored papers per country is also presented in Fig. 9 and compared with the number of deaths due to COVID-19 per million population in each country (Worldometer, 2023). As we can see in Fig. 9, three countries (United States, Italy, and Spain) with the highest number of publications (between 14 and 25) are among the countries with the highest number of deaths due to COVID-19 per million population (between 2500 and 3500), leaving aside China which undercounts COVID-19 deaths according to the World Health Organization (Wang and Qi, 2023). Nevertheless, we can also notice that other countries with a very high number of deaths due to COVID-19 per million population (such as Brazil, Mexico, and Slovakia) have a very limited number of publications (between 1 and 4).

Conclusions and perspectives

This study aims to provide new insight into the research gaps that need to be filled on the topic of AI tools to deal with misinformation on social media during hazards and disasters. Such a meta-analysis will contribute to developing a communication model based on social media moderation and recommendation algorithms that are aligned with human rights and journalism ethics.

The results confirm that after the COVID-19 pandemic, there was a marked acceleration in the number of scientific publications per year on the topic of AI tools to deal with misinformation on social media related to hazards and disasters. This trend mainly concerns papers on COVID-19, while other risks are covered by a minor share of publications. We suggest that results developed in the framework of research on the COVID-19 pandemic could be exploited to enhance research advances on other risks. On the other hand, caution should be taken when interpreting the results. The trends we describe below characterize the studies on COVID-19 that are dominant in the sample we examined, and they cannot be generalized to the studies on other risks, as these are underrepresented in the sample.

The results suggest that research in the fields of social science, decision science, psychology, humanities, and communication is underrepresented if we consider that the topic, "AI tools to deal with misinformation on social media during hazards and disasters," is strongly connected to human reasoning. This result may be because social scientists rarely refer to detection, monitoring, prevention, screening, or artificial intelligence. This trend may be indicative of the limited involvement of social scientists in the design of AI detection tools.

There is a gap to be filled by supporting these research areas that are essential to enhancing the protection of human rights and

Sponsor's Location	N° of Papers
United States	25
China, Spain, Italy	14-16
Bulgaria, Germany, Ireland, Poland, Sweden	12-13
Austria, Hungary, Croatia, Greece, Malta, Luxembourg, Romania, Portugal, Latvia, Lithuania, Slovenia, Denmark, France, Belgium, Finland, Estonia, Cyprus	11
Saudi Arabia, United Arab Emirates	6
Brazil, India, South Korea, Malaysia, Qatar	3-4
Australia, Bangladesh, Canada, Egypt, Hong Kong, Iran, Israel, Japan, Mexico, Norway, Slovakia, South Africa, Switzerland, Taiwan, Thailand, United Kingdom	1-2

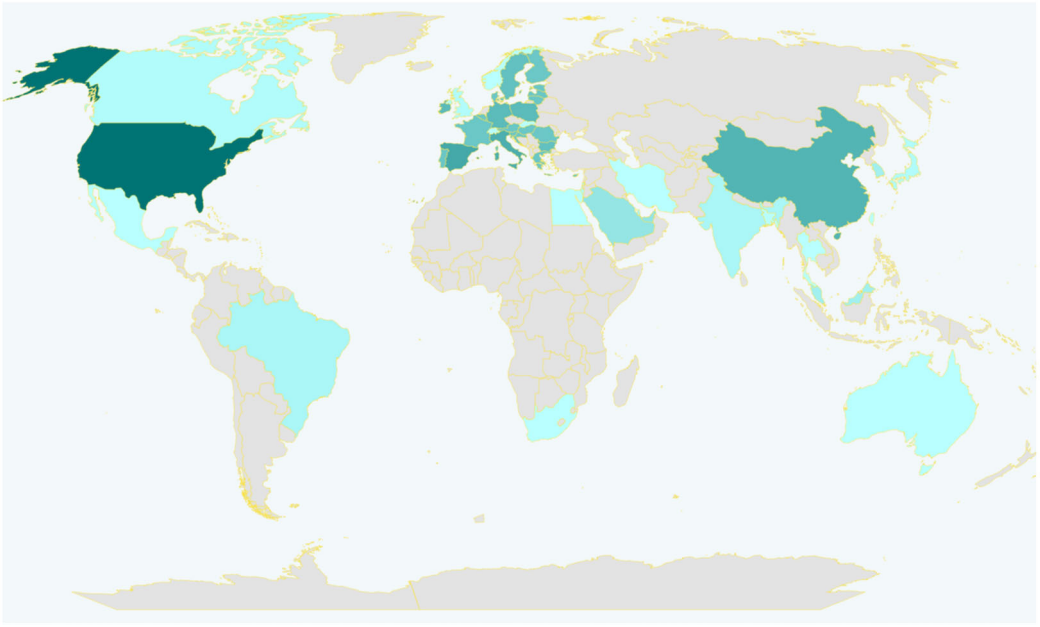

Fig. 8 Geographic distribution of sponsoring organizations and number of sponsored papers per country. The color code indicates which countries sponsor the highest number of publications (shown in dark green) and which ones sponsor only a few publications (shown in light green). Only a few countries are major funders of research on "AI tools to deal with misinformation on social media during hazards and disasters.".

journalism ethics. For instance, these research areas contribute to reflections on the regulatory, digital, or educational solutions that support digital inclusion and critical thinking.

The results also highlighted that most of the studies are dealing with the issue of detecting misinformation. This remark opens up new research questions: is the decision to filter the news left to the discretion of individual users? Are the individual user's considered active actors in the attempt to combat misinformation? Do researchers and practitioners have the same vision? A reflection on the optimum balance between algorithm recommendations and user choices seems to be missing.

Finally, the results section shows that there are a few countries, the main one being the United States, that fund research on the topic of "AI tools to deal with misinformation on social media during hazards and disasters." We can suppose that the high impact of COVID-19 contributed to increasing the research efforts on the topic. Nevertheless, this was not the only factor that determined the number of publications per country, as not all the countries that have been strongly affected by COVID-19 also have a high number of publications.

In the future, it would be interesting to compare these results with other data on digitalization trends at the national level—for instance, in the industry or education sectors—to verify if this trend is correlated with the leadership of a few countries in the field of digitalization.

The major research question of our work was about the kind of gatekeepers (i.e., news moderators) we wish social media algorithms and users to be when we are dealing with misinformation on hazards and disasters. In our view, gatekeeping should be based on communication standards that are aligned with international human rights and journalism ethics. These general principles need to be translated into operational guidelines that are tailored to the context of social media and its rapid evolution. Here, future research can play a key role by providing the knowledge required to develop and implement these operational guidelines. However, several research gaps must be filled on this topic, as we highlight in our study.

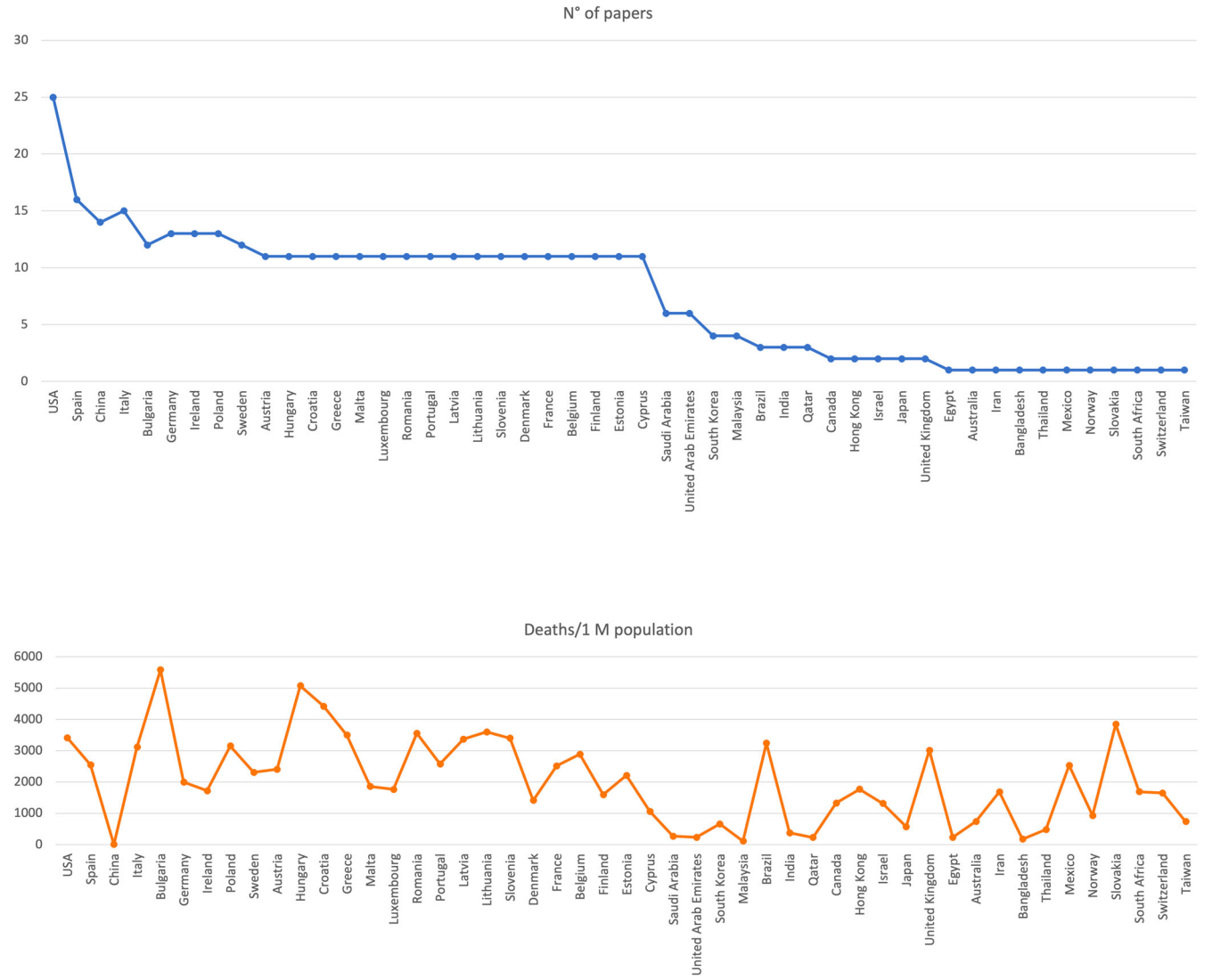

Fig. 9 The line chart displays the number of papers sponsored in each country, compared to the number of COVID-19 deaths per million population in each country. The *x*-axis lists the countries and the y-axis at the top indicates the number of publications, while the *y*-axis at the bottom indicates the number of Covid-19 deaths per million population. Three countries (United States, Italy, and Spain) with the highest number of publications (between 14 and 25) are among the countries with the highest number of deaths due to COVID-19 per million population (between 2500 and 3500), leaving aside China which undercounts COVID-19 deaths according to the World Health Organization (Wang and Qi, 2023). Nevertheless, we can also notice that other countries with a very high number of deaths due to COVID-19 per million population (such as Brazil, Mexico, and Slovakia) have a very limited number of publications (between 1 and 4).

Given these considerations, it seems to us essential that policies and programs encourage research on the topic of AI tools to deal with misinformation on social media: 1) about risks other than COVID-19 2) in the fields of social science, decision science, psychology, and humanities 3) with particular attention to the complementary role played by algorithms and users in gatekeeping, 4) as well as to the less digitally competitive countries. This policy framework would be essential to develop a communication model based on social media moderation and recommendation practices that are aligned to human rights and journalism ethics.

Data availability

The data that support the findings of this study are available from Scopus and Web of Science but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of Scopus and Web of Science.

Received: 22 November 2022; Accepted: 6 June 2023;
Published online: 17 June 2023

References

Alamoodi AH, Zaidan BB, Al-Masawa M, Taresh SM, Noman S, Ahmaro IYY, Garfan S, Chen J, Ahmed MA, Zaidan AA, Albahri OS, Aickelin U, Thamir NN, Fadhil JA, Salahaldin A (2021) Multi-perspectives systematic review on the applications of sentiment analysis for vaccine hesitancy. Comput Biol Med 139:104957. https://doi.org/10.1016/j.compbiomed.2021.104957

Alvarez-Galvez J, Suarez-Lledo V, Rojas-Garcia A (2021) Determinants of infodemics during disease outbreaks: a systematic review. Front Public Health 9. https://doi.org/10.3389/fpubh.2021.603603

Ansar W, Goswami S (2021) Combating the menace: a survey on characterization and detection of fake news from a data science perspective. Int J Inf Manag Data Insights 1(2):100052. https://doi.org/10.1016/j.jjimei.2021.100052
Ayo FE, Folorunso O, Ibharalu FT, Osinuga IA (2020) Hate speech detection in Twitter using hybrid embeddings and improved cuckoo search-based neural networks. Int J Intell Comput Cybern 13(4):485–525. https://doi.org/10.1108/IJICC-06-2020-0061
bin Naeem S, Kamel Boulos MN (2021) COVID-19 misinformation online and health literacy: a brief overview. Int J Environ Res Public Health 18(15):8091. https://doi.org/10.3390/ijerph18158091
Bogart S (2022) SankeyMATIC. https://sankeymatic.com/
Canter L (2014) From traditional gatekeeper to professional verifier: how local newspaper journalists are adapting to change. Journalism Educ 3–1:102–119. https://journalism-education.org/2014/05/dazed-and-confused/
Centre for Science and Technology Studies (2022) VOSviewer. https://www.vosviewer.com/
Chowdhury N, Khalid A, Turin TC (2021) Understanding misinformation infodemic during public health emergencies due to large-scale disease outbreaks: a rapid review. J Public Health. https://doi.org/10.1007/s10389-021-01565-3
Dallo I, Corradini M, Fallou L, Marti M (2022) How to fight misinformation about earthquakes? A communication guide. Swiss Seismological Service at ETH Zurich. https://www.research-collection.ethz.ch/bitstream/handle/20.500.11850/530319/CommunicationGuide_FightingEarthquakeMisinformation_Dallo_Corradini_Fallou_Marti.pdf?sequence=3&isAllowed=y
European Union's Horizon 2020 research and innovation program (2023) CORE (sCience& human factOr for Resilient sociEty). https://www.euproject-core.eu/
Federal Emergency Management Agency (FEMA) (2023) Natural hazards. https://hazards.fema.gov/nri/natural-hazards
Gabarron E, Oyeyemi SO, Wynn R (2021) COVID-19-related misinformation on social media: a systematic review. Bull World Health Organ 99(6):455–463A. https://doi.org/10.2471/BLT.20.276782
Garett R, Young SD (2021) Online misinformation and vaccine hesitancy. Transl Behav Med 11(12):2194–2199. https://doi.org/10.1093/tbm/ibab128
Grzywińska I, Borden J (2012) The impact of social media on traditional media agenda setting theory. The case study of Occupy Wall Street Movement in USA. In: Lodzki B, Wanta W, Dobek-Ostrowska B (eds) Agenda setting: old and new problems in old and new media. Wydawnictwo Uniwersytetu Wrocławskiego, pp. 133–155
Himelein-Wachowiak M, Giorgi S, Devoto A, Rahman M, Ungar L, Schwartz HA, Epstein DH, Leggio L, Curtis B (2021) Bots and misinformation spread on social media: implications for COVID-19. J Med Internet Res 23(5):e26933. https://doi.org/10.2196/26933
Hossein N, Miller DW (2018) Predicting motion picture box office performance using temporal tweet patterns. Int J Intell Comput Cybern 11(1):64–80. https://doi.org/10.1108/IJICC-04-2017-0033
Hunt K, Wang B, Zhuang J (2020) Misinformation debunking and cross-platform information sharing through Twitter during Hurricanes Harvey and Irma: a case study on shelters and ID checks. Nat Hazards 103(1):861–883. https://doi.org/10.1007/s11069-020-04016-6
Ireton C, Posetti J (2018) Journalism, 'fake news' & disinformation. United Nations Educational, Scientific and Cultural Organization. https://en.unesco.org/sites/default/files/journalism_fake_news_disinformation_print_friendly_0_0.pdf
Jørgensen RF, Zuleta L (2020) Private governance of freedom of expression on social media platforms: EU content regulation through the lens of human rights standards. Nordicom Rev 41(1):51–67. https://doi.org/10.2478/nor-2020-0003
Joseph AM, Fernandez V, Kritzman S, Eaddy I, Cook OM, Lambros S, Jara Silva CE, Arguelles D, Abraham C, Dorgham N, Gilbert ZA, Chacko L, Hirpara RJ, Mayi BS, Jacobs RJ (2022) COVID-19 misinformation on social media: a scoping review. Cureus. https://doi.org/10.7759/cureus.24601
Kaminska, I (2017) A lesson in fake news from the info-wars of ancient Rome. Financ Times. https://www.ft.com/content/aaf2bb08-dca2-11e6-86ac-f253db7791c6
Lazer DMJ, Baum MA, Benkler Y, Berinsky AJ, Greenhill KM, Menczer F, Metzger MJ, Nyhan B, Pennycook G, Rothschild D, Schudson M, Sloman SA, Sunstein CR, Thorson EA, Watts DJ, Zittrain JL (2018) The science of fake news. Science 359(6380):1094–1096. https://doi.org/10.1126/science.aao2998
Liu T, Xiao X (2021) A framework of AI-based approaches to improving eHealth Literacy and combating infodemic. Front Public Health 9. https://doi.org/10.3389/fpubh.2021.755808
McGee S, Frittman J, Ahn SJ, Murray S (2016) Implications of cascading effects for the Hyogo Framework. Int J Disaster Resil Built Environ 7(2):144–157. https://doi.org/10.1108/IJDRBE-03-2015-0012
Moatty A, Grancher D, Virmoux C, Cavero J (2019) Bilan humain de l'ouragan Irma à Saint-Martin: la rumeur post-catastrophe comme révélateur des disparités socio-territoriales. Géocarrefour 93(93). https://doi.org/10.4000/geocarrefour.12918
Murfi H, Siagian FL, Satria Y (2019) Topic features for machine learning-based sentiment analysis in Indonesian tweets. Int J Intell Comput Cybern 12(1):70–81. https://doi.org/10.1108/IJICC-04-2018-0057
Napoli PM (2015) Social media and the public interest: governance of news platforms in the realm of individual and algorithmic gatekeepers. Telecommun Policy 39(9):751–760. https://doi.org/10.1016/j.telpol.2014.12.003
Ng YJ, Yang ZJ, Vishwanath A (2018) To fear or not to fear? Applying the social amplification of risk framework on two environmental health risks in Singapore. J Risk Res 21(12):1487–1501. https://doi.org/10.1080/13669877.2017.1313762
Novaes CD, de Ridder J (2021) Is fake news old news? In: The epistemology of fake news. Oxford University Press, Oxford, pp. 156–179
Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, Shamseer L, Tetzlaff JM, Akl EA, Brennan SE, Chou R, Glanville J, Grimshaw JM, Hróbjartsson A, Lalu MM, Li T, Loder EW, Mayo-Wilson E, McDonald S, … Moher D (2021) The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 71. https://doi.org/10.1136/bmj.n71
Paulussen S, Harder RA (2014) Social media references in newspapers. Journalism Pract 8(5):542–551. https://doi.org/10.1080/17512786.2014.894327
Pian W, Chi J, Ma F (2021) The causes, impacts and countermeasures of COVID-19 "Infodemic": a systematic review using narrative synthesis. Inf Process Manag 58(6):102713. https://doi.org/10.1016/j.ipm.2021.102713
Posetti J, Matthews A (2018) A short guide to the history of 'fake news' and disinformation. International Center for Journalists. https://www.icfj.org/sites/default/files/2018-07/A Short Guide to History of Fake News and Disinformation_ICFJ Final.pdf
Qiu L (2017) Fact check: Manchester bombing rumours and hoaxes. N Y Times. https://www.nytimes.com/2017/05/24/world/europe/fact-check-manchester-bombing-rumors-and-hoaxes.html
Salehinejad S, Jangipour Afshar P, Borhaninejad V (2021) Rumor surveillance methods in outbreaks: a systematic literature review. Health Promot Perspect 11(1):12–19. https://doi.org/10.34172/hpp.2021.03
Sarrica M, Farinosi M, Comunello F, Brondi S, Parisi L, Fortunati L (2018) Shaken and stirred: Social representations, social media, and community empowerment in emergency contexts Semiotica 2018(222):321–346. https://doi.org/10.1515/sem-2016-0208
Thornton B (2000) The Moon Hoax: debates about ethics in 1835 New York newspapers. J Mass Media Eth 15(2):89–100. https://doi.org/10.1207/S15327728JMME1502_3
Tsao S-F, Chen H, Tisseverasinghe T, Yang Y, Li L, Butt ZA (2021) What social media told us in the time of COVID-19: a scoping review. Lancet Digit Health 3(3):e175–e194. https://doi.org/10.1016/S2589-7500(20)30315-0
Tsoy D, Tirasawasdichai T, Ivanovich Kurpayanidi K (2021) Role of social media in shaping public risk perception during COVID-19 pandemic: a theoretical review. Int J Manag Sci Bus Adm 7(2):35–41. https://doi.org/10.18775/ijmsba.1849-5664-5419.2014.72.1005
United Nations (1948) Universal declaration of human rights. United Nations https://www.ohchr.org/en/universal-declaration-of-human-rights
United Nations Office for Disaster Risk Reduction (UNDRR) (2023) UNDRR terminology. https://www.undrr.org/terminology/
Varma R, Verma Y, Vijayvargiya P, Churi PP (2021) A systematic survey on deep learning and machine learning approaches of fake news detection in the pre- and post-COVID-19 pandemic. Int J Intell Comput Cybern 14(4):617–646. https://doi.org/10.1108/IJICC-04-2021-0069
Wang J, Qi L (2023) WHO says China is undercounting Covid deaths, asks for more reliable data. Wall Str J https://www.wsj.com/articles/who-prodschina-to-release-reliable-covid-19-data-11672862046
World Health Organisation (2022) Infodemic. World Health Organisation. https://www.who.int/health-topics/infodemic#tab=tab_1
Worldometer (2023) COVID-19 coronavirus pandemic. Worldometer. https://www.worldometers.info/coronavirus/#countries

Acknowledgements

This study is part of the broader Horizon 2020 CORE project (sCience & Human factOr for Resilient sociEty). One of the work packages of this research project aims to define and apply a suitable methodology for the efficient use of social media in disaster situations. Specifically, this study contributes to the recommendations for the co-development of tools to deal with misinformation, which is one of the actions aimed at achieving the overall goals of CORE. The authors disclosed receipt of the following financial support for the research, authorship and publication of this article: this work was funded by the European Union's Horizon 2020 research and innovation program under grant agreement No. 101021746, CORE (science and human factor for resilient society).

Author contributions

These authors contributed equally to this work: RV and NK.

Competing interests

The authors declare no competing interests.

Ethical approval

This article does not contain any studies with human participants performed by any of the authors.

Informed consent

This article does not contain any studies with human participants performed by any of the authors.

Additional information

Correspondence and requests for materials should be addressed to Rosa Vicari.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.