Healthcare Transformation with Informatics and Artificial Intelligence
J. Mantas et al. (Eds.)

doi:10.3233/SHTI230439

Rapid Review on Publicly Available Datasets for Health Misinformation Detection

Zhenni NI[a b,1], Cédric BOUSQUET[b], Pascal VAILLANT[b]and Marie-Christine JAULENT[b]
[a]*School of Information Management, Wuhan University, Wuhan, China*
[b]*Sorbonne Université, UMR_S 1142, LIMICS, Paris, France*

Abstract. The proliferation of health misinformation in recent years has prompted the development of various methods for detecting and combatting this issue. This review aims to provide an overview of the implementation strategies and characteristics of publicly available datasets that can be used for health misinformation detection. Since 2020, a large number of such datasets have emerged, half of which are focused on COVID-19. Most of the datasets are based on fact-checkable websites, while only a few are annotated by experts. Furthermore, some datasets provide additional information such as social engagement and explanations, which can be utilized to study the spread of misinformation. Overall, these datasets offer a valuable resource for researchers working to combat the spread and consequences of health misinformation.

Keywords. Health misinformation, datasets, infodemic

1. Introduction

The internet has grown in popularity as a source of health information. However, individuals are prone to be misled given the abundance of health misinformation online. The broad definition of “misinformation” is a catch-all concept for related concepts like disinformation, rumors, fake news [1]. The narrow definition of “misinformation” emphasizes that misinformation is created and spread in an unintentional space, in contrast to disinformation, which refers specifically to false information operated deliberately [2]. Misinformation detection aims to employ technologies and methods to identify false information online. In the health domain, the implementation of annotated misinformation datasets requires expert knowledge, which can be time-consuming and labor-intensive. Therefore, this paper reports on a rapid review of the publicly available health misinformation datasets. The review process focused on finding two kinds of information: (i) How do these datasets define misinformation? How are these annotated misinformation datasets constructed? (ii) What are the characteristics of these datasets (publication year, topic, language, data type, labels, time span, data amount, and evaluation results)? What additional information do these datasets provide for further research?

[1] Corresponding Author: Zhenni Ni, E-mail: Jennie_n@whu.edu.cn.

2. Methods

We searched the Web of Science (WoS) Core Collection and arXiv for journal articles and conference papers written in English. The search query is as follows:

“ALL= (health or medical) AND ALL= (dataset* or Repositor* or Data set*) AND ALL= (misinformation or disinformation or 'fake news' or rumor* or infodemic)”

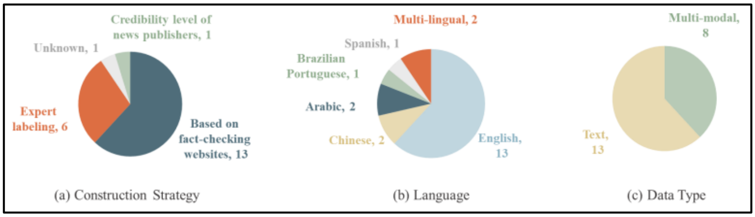
The search query was conducted on January 11, 2023. We defined the following three inclusion criteria: (i) a new dataset is proposed in the study; (ii) the data is health-related; and (iii) fact-checking labels are available for misinformation detection tasks. We then checked each dataset to see if it was public and available.

3. Results

WOS and arXiv returned respectively 457 and 83 studies. A total of 49 publications were obtained after removing duplicates. Among them, 21 publications provided publicly available datasets. These 21 public datasets were released between January 2020 and September 2022. Further details of datasets are provided in an online data supplement[2].

3.1. Concept definition and implementation strategy

The majority of the datasets lack a clear definition of detection targets. The most common concepts are "misinformation" and "fake news," which are used interchangeably in publications. Any false information, whether operated intentionally or not, is included in the category of misinformation. Figure 1(a) displays the implementation strategies for the publicly available datasets. In more than half of the datasets, misinformation was collected through fact-checking websites, and facts through reliable websites [3-15]. The credibility of source was used as an important criterion to judge the authenticity of the information [16]. Another implementation strategy is expert labeling [17-22]. It requires more than one expert to annotate a piece of information based on their knowledge. This strategy can reveal misinformation that is not included in fact-checking websites.



(a) Construction Strategy (b) Language (c) Data Type

Figure 1. Implementation strategy, language, and data type of misinformation datasets.

3.2. Characteristics of datasets

As shown in Figure 1 (b), datasets for health misinformation detection have been made available in different languages. Some of the datasets provide multimodal data, which

[2] https://docs.google.com/spreadsheets/d/1E4lh6cybF7QkcSb8UbTihYqxhnP5ZunxXy70kBjtt2g/edit?usp=sharing

means that multimedia files such as images and videos can be applied to the detection task, as shown in Figure 1 (c). Figure 2 shows the topics and data amount of the misinformation datasets. The topic keywords are extracted from the title or abstract of publications. More than half of the datasets focused on COVID-19. There is a lack of large-scale datasets. Two datasets provide large-scale predicted labels through automatic pipelines [5, 22], which can be used as weak labels for misinformation detection tasks. A few datasets manually labelled contained over 5,000 misinformation items [10, 17, 19]. Different datasets have different label sets. Most datasets classify data as binary (True or False). Some datasets consider undecidable cases by adding a third category [4, 7, 8, 18, 19, 22], such as "Not Enough Information," or "Irrelevant". Several datasets provide a more refined classification of information [5, 15, 21]. The level of annotation also differs across datasets. Most datasets are annotated at the article or tweet level [3, 5, 8-14, 16-18, 20-23]. Some datasets are annotated at the claim [4-6, 15] or sentence [19] level. A dataset for Spanish fake news detection [7] considers more fine grained items in newspaper articles (Who, What, When, Where, Why, and How).

In addition, many datasets provide additional features to specify misinformation, such as social engagement [5, 8, 10, 12-14, 16] and explanations [4, 5, 7, 12, 14, 15, 18]. Social engagement refers to how the public is involved in the spread of misinformation. Data on social engagement can contribute not only to the task of health misinformation checking but also to the study of network science and communication. Explanation refers to the evidence or reason for judging a piece of information to be false, which provides the possibility of explainable misinformation detection. Explainable detection models can help users understand why an article or tweet is false, which is especially important in the health domain.

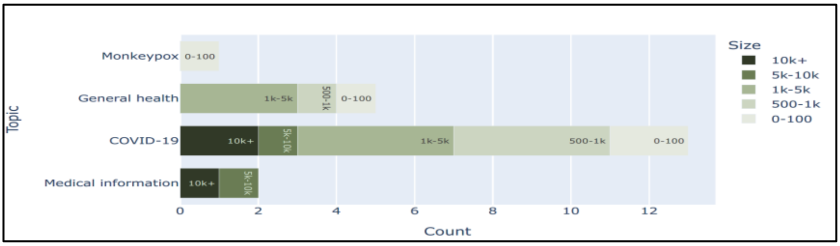


Figure 2. The distribution of topics and data count. The horizontal coordinate indicates the number of public datasets for different topics. The color indicates the amount of misinformation.

4. Conclusion

This review aims to summarize the implementation strategies and characteristics of publicly available datasets for health misinformation detection. A large number of publicly available datasets that can be used for misinformation detection have emerged since 2020, half of which are on COVID-19. Some datasets provide multilingual and multimodal data that can meet the requirements of different detection tasks. Some datasets also provide additional information, such as social engagement and explanations, which can be further used to study the spread of misinformation. The use of publicly available datasets is time-saving and necessary, which allows the comparison of different detection algorithms. Given the limited search sources and terms, current work can be considered as starting point for more comprehensive in-depth research.

Acknowledgment

This study is funded by LIMICS and China Scholarship Council (No. 202206270136).

References

[1] Cacciatore MA. Misinformation and public opinion of science and health: Approaches, findings, and future directions. Proceedings of the National Academy of Sciences. 2021;118(15):e1912437117

[2] Lewandowsky S, Stritzke WG, Freund AM, et al. Misinformation, disinformation, and violent conflict: From Iraq and the "War on Terror" to future threats to peace. American psychologist. 2013;68(7):487

[3] Endo PT, Santos GL, Xavier MED, et al. Illusion of Truth: Analysing and Classifying COVID-19 Fake News in Brazilian Portuguese Language. Big Data and Cognitive Computing. 2022;6(2).

[4] Hu XM, Guo ZJ, Wu GY, et al. CHEF: A Pilot Chinese Dataset for Evidence-Based Fact-Checking. Conference of the North-American-Chapter-of-the-Association-for-Computational-Linguistics (NAAACL) - Human Language Technologies; 2022 Jul 10-15; Seattle, WA2022.

[5] Srba I, Pecher B, Tomlein M, et al. Monant Medical Misinformation Dataset: Mapping Articles to Fact-Checked Claims. 45th International ACM SIGIR Conference on Research and Development in Information Retrieval; 2022 Jul 11-15; Madrid, SPAIN2022.

[6] Kolluri A, Vinton K, Murthy D. PoxVerifi: An Information Verification System to Combat Monkeypox Misinformation. arXiv preprint arXiv:220909300. 2022.

[7] Bonet-Jover A, Piad-Morffis A, Saquete E, et al. Exploiting discourse structure of traditional digital media to enhance automatic fake news detection. Expert Systems with Applications. 2021;169.

[8] Haouari F, Hasanain M, Suwaileh R, et al. ArCOV19-rumors: Arabic COVID-19 twitter dataset for misinformation detection. arXiv preprint arXiv:201008768. 2021.

[9] Du JS, Dou YT, Xia CY, et al. Cross-lingual COVID-19 Fake News Detection. 21st IEEE International Conference on Data Mining (IEEE ICDM); 2021 Dec 07-10; Electr Network2021.

[10] Paka WS, Bansal R, Kaushik A, et al. Cross-SEAN: A cross-stitch semi-supervised neural attention model for COVID-19 fake news detection. Applied Soft Computing. 2021;107.

[11] Cui LM, Seo H, Tabar M, et al. DETERRENT: Knowledge Guided Graph Attention Network for Detecting Healthcare Misinformation. 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD); 2020 Aug 23-27; Electr Network2020.

[12] Li YCA, Jiang BH, Shu K, et al. Toward A Multilingual and Multimodal Data Repository for COVID-19 Disinformation. 8th IEEE International Conference on Big Data; 2020 Dec 10-132020.

[13] Cui L, Lee D. Coaid: Covid-19 healthcare misinformation dataset. arXiv preprint arXiv:200600885. 2020.

[14] Dai E, Sun Y, Wang S. Ginger cannot cure cancer: Battling fake health news with a comprehensive data repository. Proceedings of the International AAAI Conference on Web and Social Media; 2020.

[15] Kotonya N, Toni F, Assoc Computat L. Explainable Automated Fact-Checking for Public Health Claims. Conference on Empirical Methods in Natural Language Processing; 2020 Nov 16-202020.

[16] Zhou XY, Mulay A, Ferrara E, et al. ReCOVery: A Multimodal Repository for COVID-19 News Credibility Research. 29th ACM International Conference on Information and Knowledge Management (CIKM); 2020 Oct 19-23; Electr Network2020.

[17] Hayawi K, Shahriar S, Serhani MA, et al. ANTi-Vax: a novel Twitter dataset for COVID-19 vaccine misinformation detection. Public Health. 2022;203:23-30.

[18] Mohr I, Wührl A, Klinger R. Covert: A corpus of fact-checked biomedical covid-19 tweets. arXiv preprint arXiv:220412164. 2022.

[19] Nabozny A, Balcerzak B, Wierzbicki A, et al. Active Annotation in Evaluating the Credibility of Web-Based Medical Information: Guidelines for Creating Training Data Sets for Machine Learning. JMIR MEDICAL INFORMATICS. 2021;9(11).

[20] Mahlous AR, Al-Laith A. Fake News Detection in Arabic Tweets during the COVID-19 Pandemic. International Journal of Advanced Computer Science and Applications. 2021;12(6):776-85.

[21] Alam F, Shaar S, Dalvi F, et al. Fighting the COVID-19 infodemic: modeling the perspective of journalists, fact-checkers, social media platforms, policy makers, and the society. arXiv preprint arXiv:200500033. 2021.

[22] Micallef N, He B, Kumar S, et al. The Role of the Crowd in Countering Misinformation: A Case Study of the COVID-19 Infodemic. 8th IEEE International Conference on Big Data; 2020.

[23] Khan S, Hakak S, Deepa N, et al. Detecting COVID-19-Related Fake News Using Feature Extraction. Frontiers in Public Health. 2022;9.