Rapid Review on Publicly Available Datasets for Health Misinformation Detection

Zhenni Ni¹ᵇ¹, Cédric BOUSQUETᵇ, Pascal VAILLANTᵇ⁰ and Marie-Christine JAULENTᵇ⁰

¹School of Information Management, Wuhan University, Wuhan, China
²Sorbonne 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_ni@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:

\[ \text{ALL=(health or medical) AND ALL=(dataset* or Repositor* or Data set*) AND ALL=(misinformation or disinformation or 'fake news' or rumor* or infodemic)} \]

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 supplement².

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.

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/1E4lb6cybF7QkcSb8UbTihYqXhnP5ZumXy70kBjt2g/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.

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