Review Article

DIGITAL HEALTH
Volume 9: 1–33
© The Author(s) 2023
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/20552076231212296
journals.sagepub.com/home/dhj

Evaluating online health information quality using machine learning and deep learning: A systematic literature review

Yousef Khamis Ahmed Baqraf[1], **Pantea Keikhosrokiani**[1,2,3] **and Manal Al-Rawashdeh**[1]

Abstract

Background: Due to the large volume of online health information, while quality remains dubious, understanding the usage of artificial intelligence to evaluate health information and surpass human-level performance is crucial. However, the existing studies still need a comprehensive review highlighting the vital machine, and Deep learning techniques for the automatic health information evaluation process.

Objective: Therefore, this study outlines the most recent developments and the current state of the art regarding evaluating the quality of online health information on web pages and specifies the direction of future research.

Methods: In this article, a systematic literature is conducted according to the PRISMA statement in eight online databases PubMed, Science Direct, Scopus, ACM, Springer Link, Wiley Online Library, Emerald Insight, and Web of Science to identify all empirical studies that use machine and deep learning models for evaluating the online health information quality. Furthermore, the selected techniques are compared based on their characteristics, such as health quality criteria, quality measurement tools, algorithm type, and achieved performance.

Results: The included papers evaluate health information on web pages using over 100 quality criteria. The results show no universal quality dimensions used by health professionals and machine or deep learning practitioners while evaluating health information quality. In addition, the metrics used to assess the model performance are not the same as those used to evaluate human performance.

Conclusions: This systemic review offers a novel perspective in approaching the health information quality in web pages that can be used by machine and deep learning practitioners to tackle the problem more effectively.

Keywords

Machine learning, deep learning, quality metrics, online health information, quality assessment

Submission date: 3 February 2023; Acceptance date: 12 October 2023

[1]School of Computer Sciences, Universiti Sains Malaysia, Minden, Penang, Malaysia
[2]Faculty of Information Technology and Electrical Engineering, University of Oulu, Oulun Yliopisto, PL, Finland
[3]Faculty of Medicine, University of Oulu, Oulun Yliopisto, PL, Finland

Corresponding authors:
Yousef Khamis Ahmed Baqraf, School of Computer Sciences, Universiti Sains Malaysia, Minden, Penang 11800, Malaysia.
Email: baqraf.cs@student.usm.my

Pantea Keikhosrokiani, School of Computer Sciences, Universiti Sains Malaysia, Minden, Penang 11800, Malaysia; Faculty of Information Technology and Electrical Engineering, University of Oulu, PL 8000, Oulun yliopisto, Finland; Faculty of Medicine, University of Oulu, PL 8000, Oulun yliopisto, Finland.
Email: pantea.keikhosrokiani@oulu.fi, pantea@usm.my

Introduction

The internet has evolved into a leading source of health-related information, with an increasing number of users of this information.[1,2] Furthermore, patients often use search engines and online health information before or as a replacement for talking to a health professional.[3,4] The quality of online health formation remains questionable or misleading in some situations.[2]

Consequently, health misinformation can significantly impact public health, considering that one in three American adults has relied on online searches for health-related information regarding their potential medical conditions.[5] The information they acquire from these online sources plays a crucial role in shaping and influencing their health beliefs, behaviors, and healthcare decisions.[6,7,3,8] As a result, evaluating the quality of such information becomes crucial, as it saves time and contributes to preserving the well-being and lives of consumers of health information. In response to the problem of health information quality, the researchers find three methods to examine the quality of online health information. First, and perhaps the most widely used in the literature,[9–12] is the manual method that uses available guidelines like Journal of the American Medical Association (JAMA) score,[13] Health On the Net Foundation (HON) code,[14] and Discern Criteria[15] to evaluate the quality. These guidelines contain a set of standards to be met for the web application or the website content to be considered of high quality. However, the researchers show that users of online health information lack the motivation and knowledge to evaluate the quality of online health information.[3,16] Coupled with the previous results, other studies find that online health information consumers do not take the time to evaluate the quality of the information they obtain,[17] which puts them in danger of health misinformation. Health misinformation is defined by researchers as a health-related claim of fact that is presently untrue because it lacks supporting scientific evidence.[18] On the other hand, in their study,[19] the concept of data quality is introduced, which is characterized as the extent to which data meets the needs and requirements of its users or consumers. Essentially, it refers to the suitability of the information for a specific use case, and this definition is the most widely used in recent studies.[20,21] More than one systematic review examines how information quality is defined and assessed using the manual method.[1,17,22,23] The second method uses a centralized database that provides certification for websites that pass certain conditions, for instance, the HON certification, which includes eight conditions to be fulfilled. The organization manually checks the web pages for health information quality in this method. Because of the difficulty of the first and second methods for quality assessment (QA), both approaches need a lot of time and effort to carry out. In order to achieve promising outcomes through manual methods, patients and healthcare professionals must possess a deep understanding of the guidelines, acquire the necessary skills to analyze the criteria, and dedicate substantial time to applying these principles to every encountered website.[24] Additionally, using a centralized database proves to be impractical when faced with the exponential expansion of online health information. Not only new web pages must be evaluated, but the assessment also extends to previously existing evaluated web pages to be reevaluated, which makes the process unscalable.[24,25] The most promising approach to provide the general public with scalable resources for evaluating the quality of online health information is through an automated evaluation process.[24] Therefore, recent research has focused on a third method: developing an automatic evaluation process using machine learning (ML) and deep learning (DL) algorithms.[26,24,27] DL language models have demonstrated remarkable potential across diverse domains, encompassing translation, question answering, and the assessment of health misinformation in social media content.[28,29] These models have yielded exceptionally promising outcomes in these areas and beyond.[30–33]

Building on this potential, there are three significant types of studies on automating the evaluation of the quality of online health information. First, the studies that employ ML to address health information quality, such as.[26,27,34,35] Despite the difficulties of manual feature engineering in ML, these studies have shown encouraging results, with *F*-score performance ranging from 51 to 91. Continuing with the types of studies, the second type of studies, including,[36,37,24] employ both ML and DL techniques to evaluate health information quality. While these studies face challenges like language-specific models and scientific medical text word embeddings, the studies achieved *F*-score performance from 79 to 95, providing valuable insights into the understanding of health information QA. Lastly, the studies[38,39] utilized only DL techniques by using the web2vec framework for identifying health misinformation. The studies employed a hybrid CNN-BiLSTM model to capture web page features and classify them accordingly. The limitation of this framework is that it is a language-specific framework built for the English language. In summary, these studies offer insights into evaluating health information quality using various ML and DL approaches, addressing some of the challenges and opportunities in this field.

emphasizing the limited studies on health information quality on web pages, there is a need for more studies about the automatic evaluation of these web pages. For example, Figure 1 shows that from 3073 articles in the systemic review, only nine discuss health information quality on the web pages. The main aim of this systemic review is to summarize the current state of the art regarding the research on the automatic evaluation of online health information quality. Therefore, this study will cover the architectures and models used to evaluate health information and make the comparison between their performance with human performance in assessing the same information.

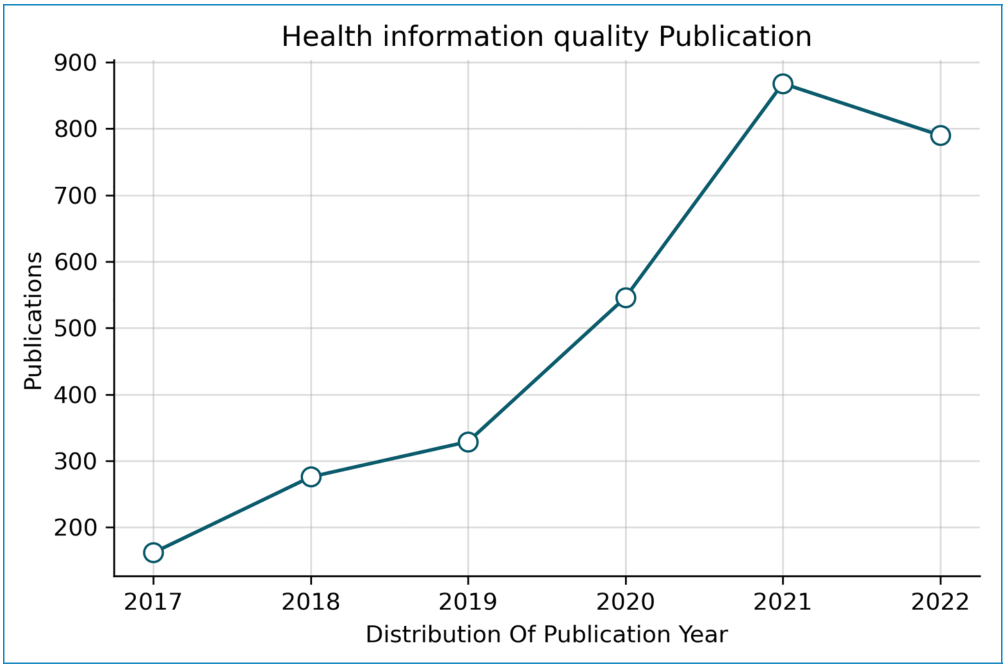

Figure 1. Health information quality publication for the last six years.

Specifically, this systemic review will cover all the empirical studies that have used predefined criteria and ML or DL in evaluating health information quality on the web. The following question with the associated hypothesis is proposed to achieve this goal.

To what extent do ML and DL language models augment the precision of evaluating health information quality on web pages?

Alternative Hypothesis (Ha): Implementing ML and DL language models to assess online health information quality in web pages significantly improves accuracy, surpassing human evaluative capabilities.

This article is organized as follows. The "Introduction" section introduces the current health information quality situation and the problem that needs to be covered in this systematic review. , The "Method" section sets the stages of PRISMA guidelines used to conduct this study. , The "Results" section presents extended results of conducting the systematic review. , The "Discussion" section discusses the review results in the context of other studies and what makes them new. , The "Summary and conclusion" section draws upon the main findings and provides future research recommendations to extend this work.

Method

This study uses PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement[40] for carrying out the systematic review. The researchers search the following eight databases:

1. ACM (Association for Computing Machinery) Digital Library
2. Science Direct
3. Scopus
4. Web of Science
5. Springer link
6. PubMed
7. Emerald Insight
8. Wiley Online Library

The search on the databases was performed on 3 August 2022. The retrieved records, the corresponding search string, fields, and filters used are specified in Table 1.

Search strategy

To build a search string that uncovers the most relevant papers in the literature, the researchers write it in a way that covers the three main concepts:

1. Quality
2. Health information
3. DL or ML

The search is performed by inserting more general search strings covering all possible synonyms for each central concept to more specific ones covering only a few possible alternative words for the main ideas. The strategies to identify the relevant keywords to include in the search are as follows:

A. Scan primary search results using quality, health information, DL, and ML.

Table 1. Database names, search strings, and retrieved records.

No.	Database name	Retrieved records	Search string	Fields and filters
1	ACM (Association for Computing Machinery) Digital Library	180	B	2017–2022
2	Science Direct	346	C	Research Articles, Conference Abstracts
				2017–2022
3	Scopus	2212	A	English, Research Articles, Conference Papers
				2017–2022
4	Web of Science	42	A	2017–2022
5	Springer link	228	A	English, Discipline: Medicine & Public Health,
				Subdiscipline: Health Informatics, Articles
				2017–2022
6	Wiley Online Library	41	B	Journals, computer science
				2017–2022
7	PubMed	128	A	2017–2022
8	Emerald Insight	101	B	2017–2022
9	Citation Tracking	1		2017–2022
	The total retrieved records		3073	

B. To comprehensively analyze the relevant articles, we reviewed the literature and incorporated keywords used by the authors in their previous studies or the main keywords emphasized by these authors.[23]

C. Check the validity of the search string, sort results from the databases by relevance and review the first 10 papers.

In addition, the search string is organized using PICO (population, intervention, comparison, and output) structure suggested by Kitchenham et al.[41] To summarize the previous steps, they are put into four categories:

1. Population: Articles related to online or web-based health information (“health information” OR “health document”) AND (online OR Internet OR web-based)
2. Intervention: QA or evaluation (quality OR credibility OR reliability) AND (evaluate OR assessment).
3. Comparison: Search strategy is compared with the systematic review of articles about manual evaluation.[23]
4. Output: Criteria, tools, model architecture, and the achieved model performance (“Deep learning” OR “Neural networks” OR “Natural language processing” OR “text classification” OR “Machine learning”).

Moreover, these concepts are connected using “AND” and “OR” operators, and three search strings are obtained:

1. The first search string (A): (quality OR credibility OR reliability OR accuracy OR standards OR “content credibility” OR “quality standards”) AND (evaluate* OR assessment) AND (health information OR health information seeking OR consumer health informatics OR online health information OR misinformation OR “health information quality assessment” OR “information quality metrics”) AND (online OR Internet OR

web-based) AND (criteria OR criterion OR metrics) AND ("Deep learning" OR "Neural networks" OR "Natural language processing" OR "text classification" OR "Machine learning").

2. The second search string (B): (quality OR credibility OR reliability) AND (evaluate OR assessment) AND ("health information" OR "health document") AND (online OR web-based) AND criteria AND ("Deep learning" OR "Machine learning").
3. The third search string (C): Quality AND (evaluate OR assessment) AND "health information" AND (online OR web-based) AND criteria AND ("Deep learning" OR "Machine learning").

Table 1 presents a concise overview of different databases, the search parameters and criteria used, and the number of records obtained during the systematic review search.

Search parameters and criteria

The fields and filters column in Table 1 provides additional information about the search Parameters and criteria used for each database. It includes details such as the publication date range (e.g. "2017–2022"), specific study types (e.g. "Research Articles," "Conference"), language preferences (e.g. "English"), and more. It helps specify the scope of the search within each database.

Screening

The researchers used Rayyan for the screening process, which involved several steps. Initially, one researcher imported the full citations of the retrieved articles into the Rayyan web application.[42] Subsequently, duplicates were removed using Rayyan's features. To ensure objectivity, the blind feature in Rayyan was activated, and the first 100 papers were independently coded by two researchers. Following this, the remaining articles were coded by one reviewer. After completing this initial phase, an additional 100 papers were randomly selected from the retrieved dataset to assess inter-rater reliability.

To determine inter-rater reliability, the coding results from the two researchers during both the initial and subsequent phases were compared, and Cohen's Kappa was employed as a metric. The obtained Kappa values were 0.9 for the first stage and 0.49 for the second stage. These values indicate a very good agreement in the first stage and a moderate agreement in the second, as defined by Brennan and Silman.[43] Any discrepancies in coding decisions were thoroughly discussed and resolved through consensus.

Given the moderate agreement observed during the second stage, the three researchers independently coded another set of 100 randomly selected records. This additional round of coding yielded an intercoder agreement score of 0.86, which is considered a very good level of agreement. Any discrepancies that emerged during this process were similarly addressed and resolved through discussion. The detailed results of this entire procedure are presented in Figures 2 in the "Results" section.

The following criteria are used to perform the screening process:

- The article's primary focus is on the automatic assessment of health information using ML or DL.
- Peer-reviewed journal articles and conference papers are accessible on the web.
- The language of the journal articles or conference papers is English.
- Published year: 2017–2022.
- Study setting: Web-based health information (general health information, treatment description, and medical advice) in the form of web pages with the target audience being health information consumers. In addition, the review does not cover social media or other platforms.

Extraction of data

To address the extraction of the data, we first created a data extraction sheet in Excel and conducted a pilot test using five randomly selected articles from the pool of included articles. The pilot test involved all three reviewers. Following this initial test, we gathered feedback and made necessary improvements to the extraction sheet. Any unknown or uncodable elements in the initial stage were left uncoded, and a subsequent review was conducted by one of the researchers to identify and apply the appropriate coding. Subsequently, one of the reviewers carried out the actual data extraction for all the included articles, while the second and third reviewers verified and cross-checked the extraction process. Any discrepancies that emerged during this extraction process were resolved through discussions among the reviewers. You can find additional details about the extracted data in Table 2.

Quality assessment

QA of the research papers can be defined as the evaluation process of the overall quality of the selected papers.[44] Consequently, the QA of ML and DL papers requires a complete and accurate evaluation, and it includes considering multiple aspects of the paper. First, its relevance to the research question and the learning algorithms used to solve the problem (supervised, unsupervised, and reinforcement learning). Second, the methodologies used for data collection, data preprocessing, and handling of data imbalance. Third, the model development considerations, such as hyperparameters (learning rate, regularization parameters,

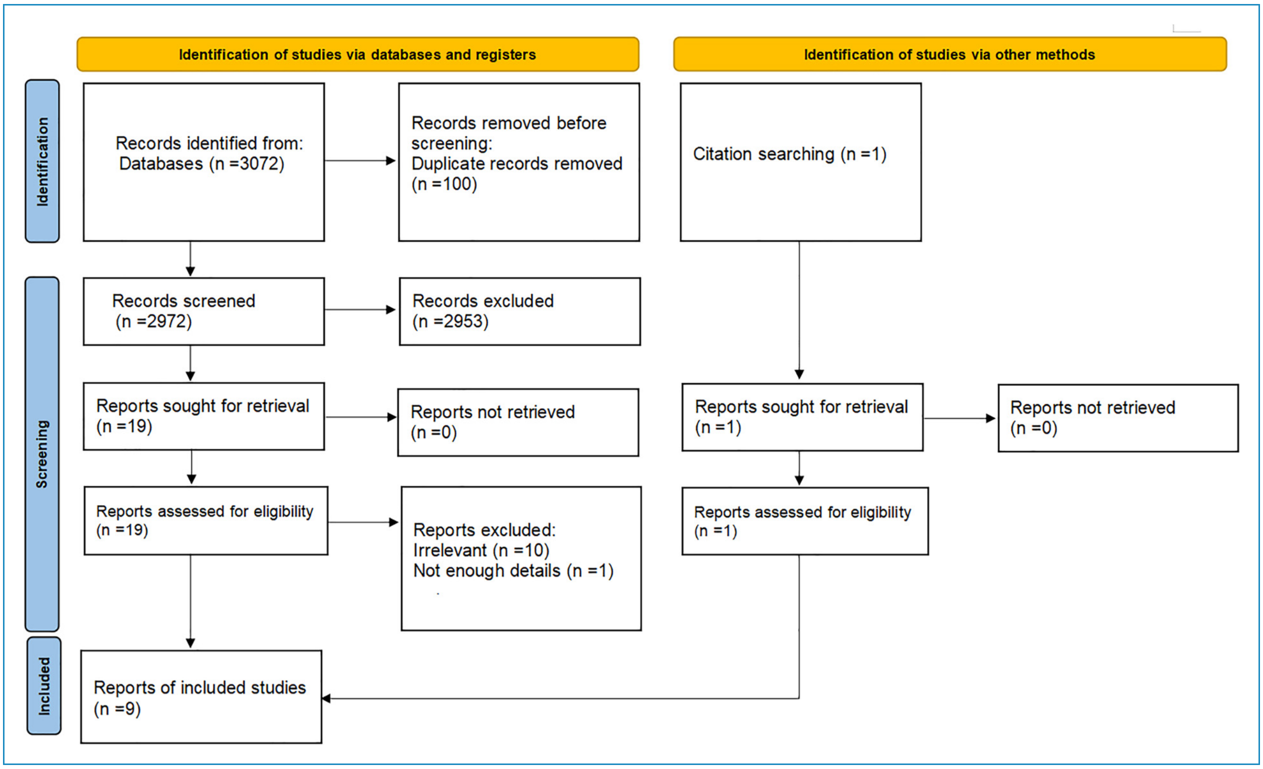
Identification of studies via databases and registers
Identification of studies via other methods
Identification
Screening
Included
Records identified from: Databases (n =3072)
Records removed before screening: Duplicate records removed (n =100)
Citation searching (n =1)
Records screened (n =2972)
Records excluded (n =2953)
Reports sought for retrieval (n =19)
Reports not retrieved (n =0)
Reports sought for retrieval (n =1)
Reports not retrieved (n =0)
Reports assessed for eligibility (n =19)
Reports excluded: Irrelevant (n =10) Not enough details (n =1)
Reports assessed for eligibility (n =1)
Reports of included studies (n =9)

Figure 2. Study selection process.

batch size, or network architecture). Finally, the evaluation metrics used are also assessed to check the effectiveness and appropriateness of the chosen measures for reporting the model performance. Conducting this comprehensive evaluation allows us to determine the overall quality of the included papers accurately. For this purpose, to evaluate the papers in our systematic review, we propose the following six quality criteria in the form of questions to be assessed:

1. Does the paper clearly focus on evaluating health information quality on web pages using ML and DL language models?
2. Are there any biases in the data collection and preprocessing phase that could impact the model's performance or generalizability?
3. Does the paper introduce any basis in the model selection, including model hyperparameters (learning rate, regularization parameters, batch size, or network architecture)?
4. Does the paper introduce any basis in the evaluation metrics selection?
5. Does the paper introduce any biases related to the language representation used in the model and its suitability for diverse web page content?
6. Are there any potential conflicts of interest, funding sources, or affiliations that could influence the study's design, conduct, or reporting of results?

Summary of the criteria

The purpose of these criteria is to assess bias and guarantee the excellence of the papers that have been included. Below is a brief overview of the intent behind each criterion:

1. Focus on the relevance to the research question: The first criterion ensures that the selected papers directly align with the research question, evaluating health information quality using DL and ML. By including only studies that meet this criterion, the potential bias of the inclusion of irrelevant papers is minimized.
2. Identifying data biases: The second criterion evaluates data biases in the data collection, cleaning, and preparation process before the model training. Identifying any bias at this stage is crucial as it will impact the final results and the generalizability of ML or DL models.
3. The model selection: This criterion investigates the model selection process. It ensures that models are chosen based on their suitability for the task rather than relying only on their previous performance, as reported in the literature.
4. Appropriate evaluation metrics: The fourth criterion emphasizes selecting suitable evaluation metrics.

Table 2. Essential elements of the extracted data.

Category	Item/description
Basic article information	Title
	Year
	Authors
Study focus	Credibility
	Health misinformation
	Reliability
	Quality
Types of algorithm used	Machine learning algorithms
	Deep learning algorithms
	Machine and deep learning
The name of the algorithm	The name of neural network architecture
	The name of the machine learing algorithm
Data preprocessing techniques	Identification of preprocessing techniques
	used such as normalization, etc.
Measurement tools	None or undertermined
	HON code
	JAMA
	DISCERN
	Mix
	Other
Model perfromnce	*F*1_score
	AUC
	Precision
	Recall
Human perfromnce	Inter-rater reliability
Dataset context	General health information
	Specific disease
	Model language
Sample data	Web pages
	Whole website
Sample size	Number of examples in dataset
Search engine	Google
	Yahoo
	Bing
Data type	New collected data
	Archival data
Sampling method	Procedure used to collect the data
Study design	Supervised learning
	Unsupervised learning
	Reinforcement learning

JAMA: American Medical Association; HON: Health On the Net Foundation; AUC: area under the curve.

Ensuring the absence of bias in this step is vital to obtaining a comprehensive view of the model's effectiveness.

5. Language representation bias: The fifth criterion addresses possible biases related to the language representation and selection used, which could impact the model's generalizability to various web page content.
6. Accounting for conflicts of interest: This criterion improves the assessment process by considering possible sources of bias related to conflicts of interest, funding, or affiliations that could impact the study's outcomes.

The evaluation of included papers was conducted by the three researchers using a quality score categorized into three levels: "low," "medium," and "high."[44,45] Each research paper was assigned a high-quality score of 1 if it fully met the corresponding requirements for the criteria, 0.5 if it partially met the requirements, and 0 if it failed to meet the quality requirements. Regarding the overall

Table 3. Quality evaluation scores for the review's included papers.

No.	Title	Q1	Q2	Q3	Q4	Q5	Q6	Scores	Quality
1	Hybrid machine learning approach for Arabic medical web page credibility assessment	1	0.5	1	1	0.5	1	5	High
2	Health misinformation detection in web content A structural-, content-based, and context-aware approach based on Web2Vec information detection	1	0.5	1	1	0.5	1	5	High
3	Reliable or not? An automated classification of web pages about early childhood vaccination using supervised machine learning	1	0	0	1	0.5	1	3.5	Medium
4	Predicting the quality of health web documents using their characteristics	1	0	1	0.5	0.5	1	4	High
5	AutoDiscern: rating the quality of online health information with hierarchical encoder attention-based neural networks	1	0.5	1	1	0.5	1	5	Medium
6	Health misinformation detection in the social web: an overview and a data science approach	1	0	0	1	0.5	1	3.5	Medium
7	Using machine learning for automatic identification of evidence-based health information on the web	1	0.5	1	1	0.5	1	5	High
8	Intelligent and effectively aligned evaluation of online health information for older adults	1	0.5	0	1	0.5	1	4	Medium
9	Vec4Cred: a model for health misinformation detection in web pages	1	0.5	1	1	0.5	1	5	High

score, the maximum score was 6, defined as achieving the six quality criteria, and a minimum score of 0 if it failed to achieve any criteria. Therefore, papers that scored below 3 were considered low quality, those with a score from 3 up to 4 were considered medium quality, and those with a score above 4 to 6 were considered high-quality papers. The evaluation result is provided in Table 3. Furthermore, any inconsistencies that emerged during the manual assessment process, which was based on the researchers' subjective evaluations of the papers, were resolved through discussions among the researchers. It is important to note that no papers were excluded from the evaluation, as they were all categorized as being of medium or high quality.

Results

Key aspects of the included papers

Figure 2 shows the selection process of the articles in this study. There is a total of 3072 records found from searching the databases. After removing duplicates, the screen records a total of 2972, from which 19 full-text documents are reviewed and finally included eight papers. Afterwards, we investigated both backward and forward citations. For the backward citations, we examined the references cited in the included papers and included any papers that met our criteria. Similarly, we used Google Scholar to track the forward citations of the included papers. However, only one article that fulfills the inclusion criteria is found in these searches.

After a careful study selection process based on Figure 2, only nine papers are included for the systematic literature review in this study; six are journal papers, and three are conference papers. Furthermore, the distribution of article publications by year is illustrated in Figure 3, and Table 4 provides details information about each publication. Furthermore, Figure 4 illustrates the distribution of citations for the included paper by year. Finally, Tables 5 and 6 provides the critical features of included papers, including the types of algorithms, measurement tools, model performance, and human performance. Likewise, for more information about the measurement tools and criteria, refer to Tables 7 and 8. Most of the studies use the Google search engine to collect the data. Furthermore, the maximum sample size used is 12,245, as provided by Goeuriot et al.,[46] and the minimum sample size is 50 provided by Robillard and Feng.[47]

The review revealed several significant findings across different areas, including publications information, data preprocessing, frequently used algorithms, model performance, measurement tools and quality dimensions, generalization and boundaries of the developed models, including (the evaluated health information and the evaluated language in the studies), and language representation. Detailed information about each topic is in the following sections:

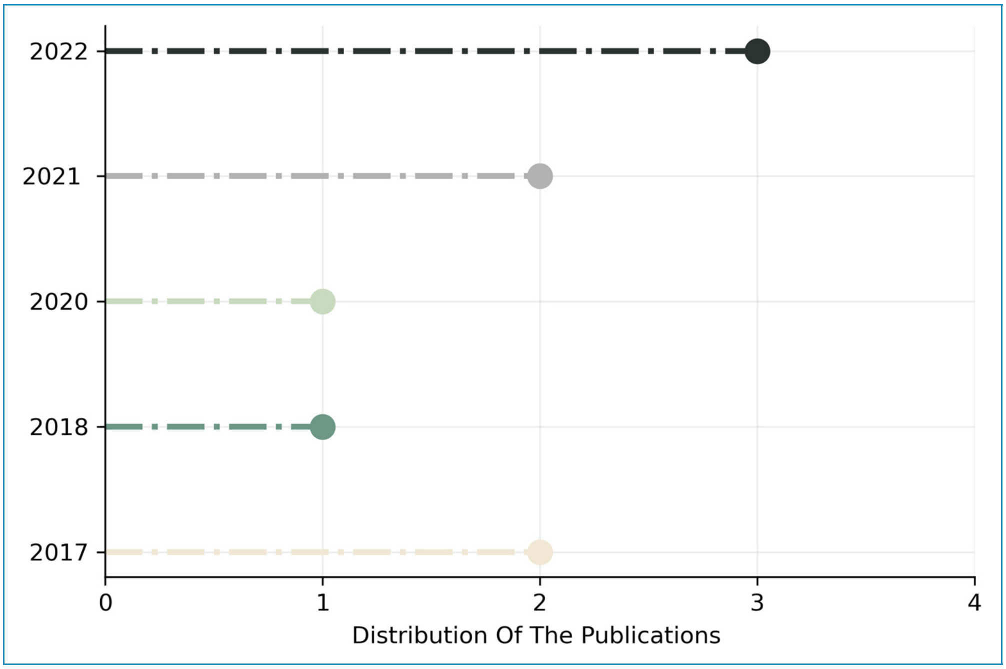

Figure 3. The publications distribution of the included paper by year.

Publications information

Table 4 shows information about the nine publications included in this review on various topics, including credibility assessment of Arabic medical web pages, health misinformation detection, automated classification of web pages about childhood vaccination, predicting the quality of health web documents, rating the quality of online health information, health misinformation detection in the social web, automatic identification of evidence-based health information, and assessment of digital health information targeting older adults.

Additionally, the table mentions the journals or conferences associated with these publications, such as Health Informatics Journal, Conference on Information Technology for Social Good, Patient Education and Counseling, Online Information Review, BMC Medical Informatics and Decision Making, Environmental Research and Public Health, International Conference on Digital Health, Multimedia Tools and Applications, and Association for the Advancement of Artificial Intelligence (AAAI) Conference. Furthermore, the table provides additional information about citations, journal impact factor, and year of publication. Figure 4 illustrates the distribution of citations for the included paper by year.

Data preprocessing

Preprocessing could be defined as converting raw data into a format suitable for further analysis or model training in DL and ML. Preprocessing entails several stages: cleansing, normalization, feature extraction, and transformation of the raw data. Preprocessing is essential for ensuring data quality, removing noise, dealing with missing values, standardizing variables, and performing the required transformations. This section will provide a full overview of the preprocessing techniques or software employed in the studies covered in this review.

In the study, Al-Jefri et al.[26] followed more than one step to organize the corpus into separate files and their corresponding labels. Firstly, they imported the saved web pages into Sketch Engine,[48] a tool that helped remove HTML tags, web scripts, and irrelevant content such as advertisements. Secondly, punctuation marks were eliminated, and commonly occurring unigram feature stopwords like “a,” “the,” and “is” were excluded from the text before applying feature extraction methods.

In these studies, Upadhyay et al.[38,39] followed three steps in the preprocessing phase to extract the critical feature. The first step, the data parsing process of Document Object Model (DOM) structure, examines the target web page’s DOM structure, content, and URLs to extract useful features. The DOM structure parsing extracts a list of HTML tags in a particular order, from the top-level tags, such as The HEAD tag, to child elements, such as the IMG tag. These HTML tags are all used to express the DOM structure and are thus considered word-level corpus in the following data representation phase.

In the second step, web page content parsing, links and tags are ignored, and only unstructured text content is

Table 4. Publications information of the included paper.

Study references	Citation	Title	Journal/conference	Impact factor	Year
Alasmari et al.[36]	4	Hybrid machine learning approach for Arabic medical web page credibility assessment	Health Informatics Journal	3.092	2022
Upadhyay et al.[38]	14	Health misinformation detection in web content A structural-, content-based, and context-aware approach based on Web2Vec information detection	Conference on Information Technology for Social Good	N/A	2021
Meppelink et al.[27]	9	Reliable or not? An automated classification of web pages about early childhood vaccination using supervised machine learning	Patient Education and Counseling	3.467	2021
Oroszlányová et al.[35]	11	Predicting the quality of health web documents using their characteristics	Online Information Review	2.325	2018
Kinkead et al.[24]	21	AutoDiscern: rating the quality of online health information with hierarchical encoder attention-based neural networks	BMC Medical Informatics and Decision Making	3.894	2020
Di Sotto and Viviani.[37]	26	Health misinformation detection in the social web: an overview and a data science approach	Environmental Research and Public Health	4.799	2022
Al-Jefri et al.[26]	12	Using machine learning for automatic identification of evidence-based health information on the web	International Conference on Digital Health	N/A	2017
Robillard et al.[34]	4	Intelligent and effectively aligned evaluation of online health information for older adults	Association for the Advancement of Artificial Intelligence(AAAI) Conference	N/A	2017
Upadhyay et al.[39]	7	Vec4Cred: a model for health misinformation detection in web pages	Multimedia Tools and Applications	2.396	2022

included when parsing web pages. Afterwards, two types of corpus are generated. First, the sentence-level corpus is formed by identifying word sequences separated by periods. Second, a word-level corpus is created by capturing each word on the page. Additionally, the web page's POS (parts of speech)-level corpus is extracted because POS tags might offer linguistic insights[49–52] useful for tasks like detecting fake news.[53]

In the third step, the domain names of the URLs for the target page and any related pages found inside are collected. The list of domain names functions as a word-level corpus because it effectively helps to detect misinformation.[54–56] Furthermore, a statistic-based algorithm, YAKE,[57] is utilized to extract keywords from the content of related pages automatically. YAKE generates a list of the top 20 keywords for each included page. These keywords represent the word-level corpus for future examination. The study also found that YEAK greatly outperformed TextRank,[58] another word extractor, in terms of speed and output quality.

In the research, Alasmari et al.[36] reported the removal of Arabic stop words, diacritics, non-Arabic words, and special characters using the PyArabic library.[59] Moreover, words including three letters or fewer were also eliminated. Additionally, the DISCERN instrument ratings, originally on a scale of 1-5, were reclassified in the study[24] as a binary classification: scores of 3-5 considered a pass, while scores of 1-2 indicated failure based on the established criteria. The text extraction and cleansing from HTML tags were achieved using the Beautiful Soup library.[60]

Finally, according to the study,[34] the papers were cleaned by removing stop words such as "and" and "the," and all content was normalized to lowercase letters. Furthermore, three of the included studies lack information on the preprocessing stage.

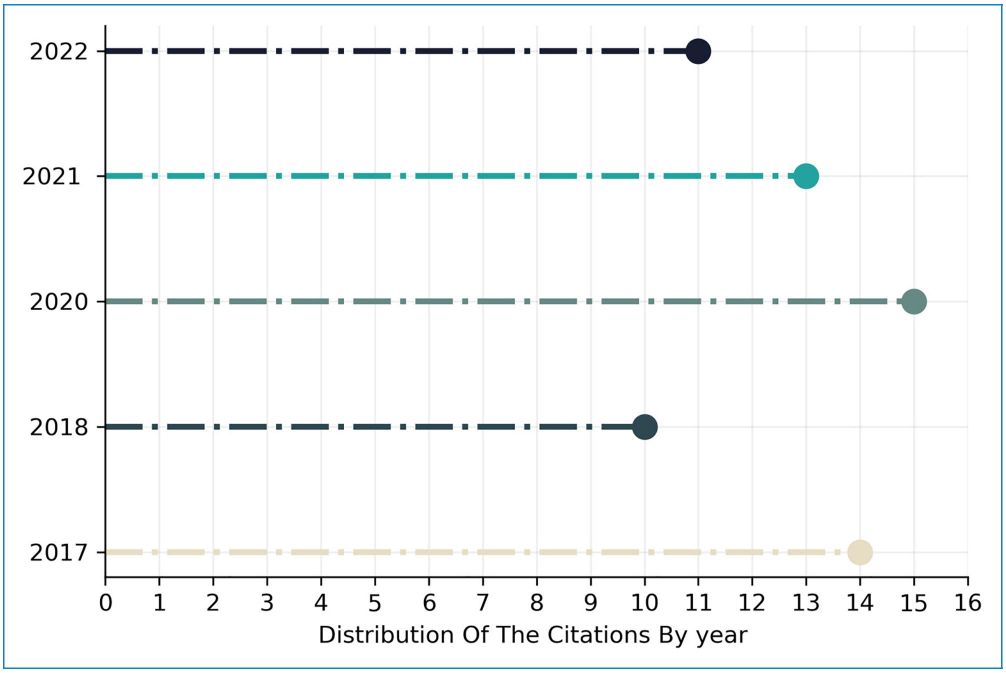

Figure 4. The citation distribution of the included paper by year.

Frequently used algorithms

Different ML and DL algorithms are used in the included studies to automatically evaluate health information quality. Of the nine included papers, two (22.2%) use DL, and three (33.3%) use mixed method DL and ML. Finally, four papers (44.4%) use only ML. See Table 5 for more details about each study. The DL algorithms used in the included studies are bidirectional long short-term memory (LSTM), convolutional neural network (CNN) architecture, hierarchical encoder attention (HEA)-based neural network, and a feed-forward neural network with one layer to perform the classification. In contrast, the ML algorithms used are logistic regression (LR), Naïve-Bayes, gradient boosting, support vector machine (SVM), decision tree (DT), and random forest. Figure 5 shows the number of times each algorithm is used.

Model performance

The included studies report several metrics for measuring the model performance, including accuracy, recall, precision, and f_score. The maximum achieved f_score is 95.3%, and only a few studies report human performance. The highest score is 96% see Table 5 for individual study f_score or accuracy. Figure 6 shows the accuracy, recall, precision, f_score, and human performance of studies Alasmari et al.[36], Meppelink et al.[27], and Kinkead et al.[24] in Table 5. Furthermore, in terms of performance, the first study's best algorithm is LSTM, and the third and fourth are Naïve Bayes, HEA-based neural network with Bidirectional Encoder Representations from Transformers (BERT) for word embedding, respectively. Below is a comprehensive explanation of each evaluation matrix and the mathematical equation used to calculate each matrix. Additionally, the strengths and limitations of each matrix are also examined:

Accuracy

The accuracy metric demonstrates the general correctness of the model. In other words, it indicates the proportion of accurate predictions made by the model in relation to the total number of predictions.

$$\textbf{Accuracy} = \frac{\textit{True Positives} + \textit{True Negatives}}{\textit{Total Predictions}}$$

Strengths

Accuracy provides a comprehensive evaluation score. The accuracy score comprehensively assesses the overall performance of the classification model by analyzing correct and incorrect predictions across all classes. It provides a detailed understanding of the model's performance.

Shortcomings

The accuracy matrix can provide inaccurate information when dealing with imbalanced datasets if the number of instances in positive and negative classes is unequal.

Recall

Recall, also known as sensitivity, measures the model's capability to accurately anticipate the positive class by

Table 5. The critical features of the included studies.

Study references	Types of algorithm	Measurement tools	Model performance	Human performance	Study language	Study type
Alasmari et al.[36]	Deep learning, machine learning	Textual and nontextual features	*F*1-score: 79%	28%	Arabic	Experimental study
Upadhyay et al.[38]	Deep learning	textual and nontextual features	*F*1-score: 94.17	N/A	English	Experimental study
Meppelink et al.[27]	Machine learning	Guidelines provided by Dutch National Institute	Unreliable information F1_score: 0.54–0.86 Reliable information: F1_score: 0.82–0.91.	96%	Dutch	Analytical study
Kinkead et al.[24]	Deep learning, machine learning	DISCERN instrument	*F*-score: 86%	94%	English	Experimental study
Di Sotto and Viviani[37]	Deep learning, machine learning	Textual and nontextual features	F_means: 95.3	N/A	English	Analytical study
Al-Jefri et al.[26]	Machine learning	Evidence-based	F1_score: 89.61		English	Experimental study
Robillard et al.[34]	Machine learning	Quest	F1_score: 90%	93%	English	Analytical study
Oroszlányová et al.[35]	Machine learning	Health On the Net Foundation (HON) code as ground truth	Accuracy: 89	N/A	English	Experimental study
Upadhyay et al.[39]	Deep learning	Textual and nontextual features	*F*1: 94.21	N/A	English	Experimental study

normalizing correct prediction by the total amount of times the model predicts it as a negative class.

$$\textbf{Recall} = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

Strengths

The recall score helps to discover the percentage of true positives by showing the model basis in predicting the negative class very often. A high recall value suggests the model successfully identifies and captures many positive instances within the dataset.

Shortcomings

The recall score is informative regarding the model's overall performance in correctly identifying positive class instances. However, it does not offer any insights into the percentage of true negatives, representing the model's ability to accurately identify instances of the negative class.

Precision

Precision measures the model's capability to accurately anticipate the positive class by normalizing correct prediction by the total amount of times the model predicts the negative class as positive.

$$\textbf{Precision} = \frac{TP}{TP + FP}$$

Strengths

The precision score helps to discover the proportion of true positive predictions by showing if the model has a

Table 6. Comparative table of the included studies.

Study references	Focus	Dataset context	Study design	Sampling method	Sample size	Language representation
Alasmari et al.[36]	Credibility	General health information	Supervised learning	Convenience sampling	500 web pages	TF-IDF representation, Neural word embedding word2vec (AraVec).
Upadhyay et al.[38]	Health misinformation	Different domains: health, finance, politics, general health information	Supervised learning	Convenience sampling	Microsoft Credibility Dataset: 1000 web pages, Medical Web Reliability Corpus: 360 web pages, CLEF eHealth: 5509 credible, 6736 non-credible web pages	PubMed word2vec
Meppelink et al.[27]	Reliability	Early childhood vaccination	Supervised learning	Convenience sampling	468 web pages	Bag-of-words, TF-IDF
Kinkead et al.[24]	Quality	Breast cancer, arthritis, depression	Supervised learning	Convenience sampling	269 web pages	BERT and BioBERT
Di Sotto and Viviani[37]	Health misinformation	Covid-19, different health topics	Supervised learning	Convenience sampling	CoAID: 3555 web pages, Recovery, Fake Health: 2029 web pages	GLOVE
Al-Jefri et al.[26]	Evidence-based property	Shingles, flu, migraine	Supervised learning	Convenience sampling	276 web pages	Bag-of-words
Robillard et al.[34]	Quality	Alzheimer	Supervised learning	Convenience sampling	50 web pages	Bag-of-words
Oroszlányová et al.[35]	Quality	General health information	Supervised learning	Convenience sampling	734 web pages	Bag-of-words
Upadhyay et al.[39]	Health misinformation	Different domains: health, finance, politics,General health information	Supervised learning	Convenience sampling	Microsoft Credibility Dataset: 1000 web pages, Medical Web Reliability Corpus: 360 web pages, CLEF eHealth: 5509 credible, 6736 non-credible web pages	PubMed word2vec

Table 7. The language representation techniques with a details description (1).

Quality criteria/ tools	Details	Study references
Textual representation features		
TF-IDF	Term Frequency-inverse document frequency technique counts the importance of the word across the documents, so infrequent words get highly weighted	Alasmari et al.[36], Meppelink et al.[27]
AraVec[67]	The Arabic version of the word2vec is a neural embedding model pre-trained in general Arabic content. As a result, the model can capture many syntactic and semantic relations among words in the document	Alasmari et al.[36]
PubMed (Word2vec[68])	The word2vec neural model pre-trained on PubMed	Upadhyay et al.[38], Upadhyay et al.[39]
BERT,[69] BioBERT[62]	Two distinct versions of this neural embedding model for text representation were used in the included studies. BERT pre-trained on the English Wikipedia and books corpus, while BioBERT pre-trained on both general and scientific texts	Kinkead et al.[24]
Bag-of-the-words	With this technique, every word in the document becomes a feature by counting how many times it occurs. Unlike TF-IDF, frequent words are weighted highly	Meppelink et al.[27], Al-Jefri et al.[26], Robillard et al.[34]
GloVe[70]	By using this method, each word or phrase in the document is converted into a dense vector representation known as a word embedding that captures the semantic meaning and relationships between words	Di Sotto and Viviani.[37]

TF-IDF: term frequency-inverse document frequency; BERT: Bidirectional Encoder Representations from Transformers.

basis in predicting the positive class very often. A high precision value indicates that the model has a low rate of false positives.

Shortcomings

The precision score informs us about the model's effectiveness in identifying positive class instances. Still, it does not offer any insights into the model's accuracy in identifying instances of the negative class, which is represented by the proportion of true negatives.

*F*1-score

It is finding the harmonic mean between recall and precision.

$$\mathbf{F1} = \frac{2 * TP}{2 * TP + FP + FN}$$

Strengths

The $F1$-score offers a balanced evaluation of model performance, particularly valuable when dealing with imbalanced datasets, as it considers both precision and recall. Furthermore, it integrates the model's performance into a single metric, facilitating the comparison of multiple models and helping in the selection process.

Shortcomings

The $F1$-score fails to provide detailed information about the specific errors made by the classifier or the distribution of errors among different classes. Consequently, relying only on the $F1$-score may not be suitable for comprehensively assessing classifier performance.

Human performance

Cohen's Kappa coefficient[43] is a statistical measure that evaluates the agreement between two raters or more when categorizing items into discrete categories. In this study's case, it measures the agreement between the two health professionals about the health information on the web pages, whether it is of high or low quality. Consequently, Cohen's Kappa is an appropriate measure for evaluating the performance of a DL model and comparing its output to judgements or annotations made by human experts. Table 9 shows the suggested interpretation of the agreement for different values of K.

$$\mathbf{K} = \frac{p_o - p_e}{1 - P_e}$$

Table 8. The criteria and tools used in included studies with a details description (2).

Other representation features			
Quality criteria /tools		Details	Study references
QUEST tool	Complementary	It is tested for if health document support or replaces the doctor-patient relationship	Robillard et al.[34]
Brief DISCERN criteria	References	The sources of information used to write the health information	Kinkead et al.[24]
	Date	Include the date that the information was released	Kinkead et al.[24]
	How treatment work	Provide an explanation of the mechanisms underlying each treatment	Kinkead et al.[24]
	Treatment benefit	Outline the advantages associated with each treatment type offered	Kinkead et al.[24]
	Treatment risk	Outline the potential risks associated with each treatment option	Kinkead et al.[24]
HON code	Authoritative	include information regarding the credentials and qualifications	Alasmari et al.[36]
	Attribution	provide the source and date of publication for the information cited	Alasmari et al.[36]
	Advertising policy	Precisely differentiates between promotional material and main content	Alasmari et al.[36]
	Transparency	Show a straightforward way of contacting, such as email	Alasmari et al.[36]
Linguistic-stylistic features		It captures the stylistic features of the text	
	Strong modals	might, could, can, would, may	Di Sotto and Viviani[37]
	Weak models	should, ought, need, shall, will	Di Sotto and Viviani[37]
	Negations	If	Di Sotto and Viviani[37]
	To be form	To form be, am, is are, was, were, been	Di Sotto and Viviani[37]
	Conclusive conjunctions	until, despite, in spite, though	Di Sotto and Viviani[37]
	Following conjunctions	but, however, otherwise, yet	Di Sotto and Viviani[37]
	Definite determiners	the this, that, those, these	Di Sotto and Viviani[37]
	Personal pronouns	I, you	Di Sotto and Viviani[37]
	First person	I, we, me, my, mine, us, our	Di Sotto and Viviani[37]

(continued)

Table 8. Continued.

Other representation features		
Quality criteria /tools	Details	Study references
Second person	you, your, yours	Di Sotto and Viviani[37]
Third person	he, she, him, her, his, it, its	Di Sotto and Viviani[37]
Question particles	why, what, when, which, who	Di Sotto and Viviani[37]
Adjectives	correct, extreme, long, visible	Di Sotto and Viviani[37]
Adverbs	maybe, about, probably, much	Di Sotto and Viviani[37]
Proper nouns	names of places, things, etc.	Di Sotto and Viviani[37]
Other nouns	other nouns	Di Sotto and Viviani[37]
To have form	have, has, had, having	Di Sotto and Viviani[37]
Past tense verb	past tense verb	Di Sotto and Viviani[37]
Gerund	gerund	Di Sotto and Viviani[37]
Participle verb	past or present participle verb	Di Sotto and Viviani[37]
Superlatives	superlative adjectives or adverbs	Di Sotto and Viviani[37]
Exclamation	exclamation mark	Di Sotto and Viviani[37]
Linguistic-emotional Features	These features capture the emotions expressed in the text	Di Sotto and Viviani[37]
Linguistic-medical features	Capture the statics of medical terms in the text that affect the health information quality	Di Sotto and Viviani[37]
Normalized count of medical terms	is counting the number of times the medical term occurs, normalized by the number of words. Extracting this feature requires the use of Named-entity recognition (NER)	Di Sotto and Viviani[37]

(continued)

Table 8. Continued.

Other representation features		
Quality criteria /tools	Details	Study references
Normalized count of unique medical terms	A unique count of the medical term is considered	Di Sotto and Viviani[37]
Hyperlink count	This count the proportion number of external links as a factor of the total number of links	Di Sotto and Viviani[37]
Normalized count of commercial terms	It counts the number of commercial terms. The more commercial terms in the document, the less credible the source information is	Di Sotto and Viviani[37]
Feedback	The health information provider allows for feedback from the user	Alasmari et al.[36]
Source reliability	Website rank (Alexa and page rank)	Alasmari et al.[36]
Search engine	The information source includes built-in search capabilities	Alasmari et al.[36]
Certification	The information source owns an official certification such as HON certification	Alasmari et al.[36]
Childhood vaccines guidelines	Provide by Dutch National Institute for Public Health and the Environment (RIVM)	Meppelink et al.[27]
Evidence-based	Capture whether medical authorities like the US Food and Drug Administration agency and the National Institute for Clinical Excellence in the UK licensed the treatment	Al-Jefri et al.[26]

HON: Health On the Net Foundation.

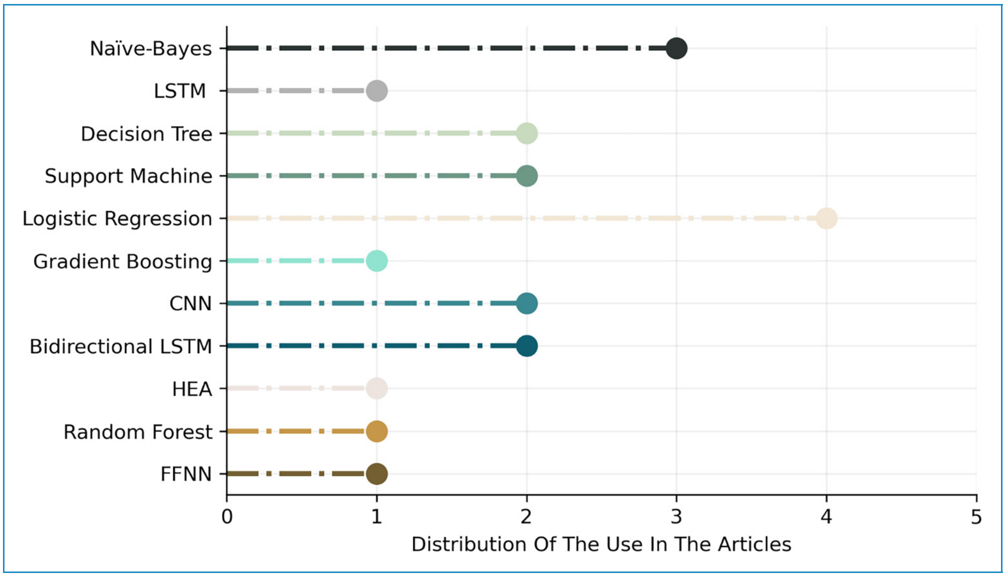

Figure 5. A number of times different algorithms have been used in the articles.

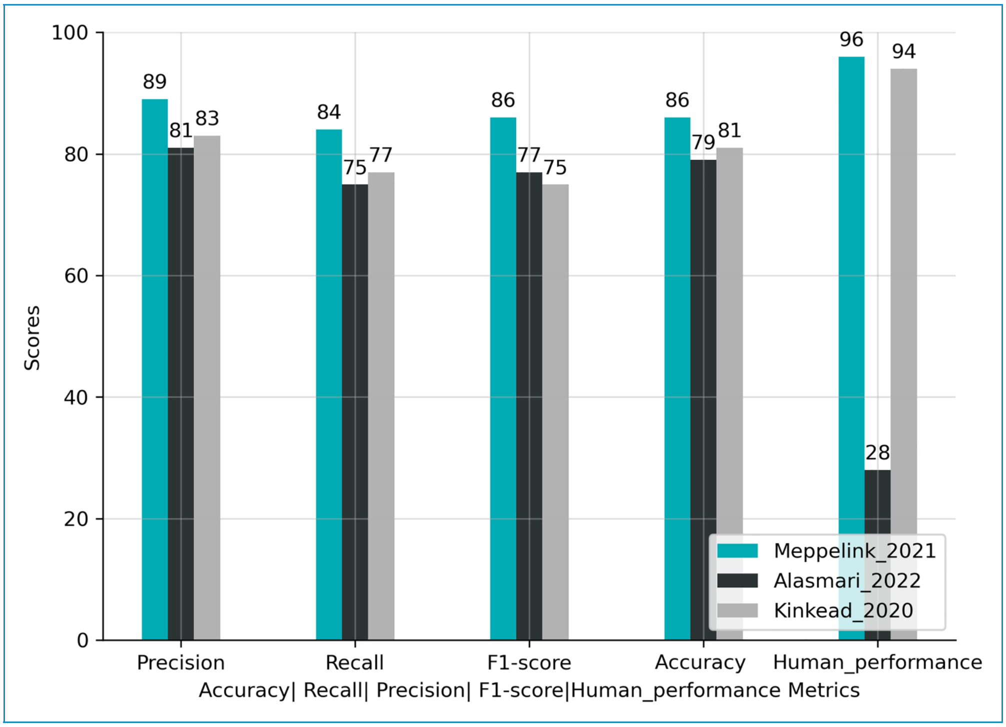

Figure 6. The performance of the three best algorithms in each study with human performance.

Strengths

Cohen's Kappa coefficient is valuable because it considers chance agreement, manages imbalanced data, applies to multiple raters, and offers a consistent scale for straightforward interpretation and study comparison.

Shortcomings

Although Cohen's Kappa coefficient is considered a very good score, it has limitations uniquely when there is a kind of data bias, which can result in incorrect results, mainly when dealing with skewed distributions or intentional biases among raters.

Table 9. *K* values with associated interceptions.

K values	Strength of agreement
<0.20	Poor
<0.21–0.40	Fair
<0.41–0.60	Moderate
<0.61–0.81	Good
<0.81–1	Very good

Measurement tools and quality dimensions

The included studies used multiple instruments to evaluate health information quality. First, JAMA benchmarks consist of four standards for determining the trustworthiness of a health information source: authorship, attribution, disclosure, and currency, which give information about the author, sources, funding, and the date the health information was written. Second, the HON code extends beyond trustworthiness and contains five additional criteria: complementarity, privacy, justification, transparency, and advertising policy. These criteria require health information providers to enhance doctor–patient relationships, protect personal information, support claims with scientific evidence, maintain direct communication channels, and distinguish between main content and advertising.

Third, the brief DISCERN instrument further expands the evaluation of health information quality by providing guidelines for treatment choices. It includes six criteria: treatment works, treatment benefits, treatment risks, effect on quality of life, side effects of no treatment, and areas of uncertainty. By considering these criteria, health information providers can ensure high-quality and reliable information delivery to consumers, enabling informed decision-making. Table 8 shows the cite every criterion and tool used in the included studies to evaluate health information quality.

Generalization and boundaries of the developed models

DL and ML models' success heavily rely on their generalization ability. It refers to a model's ability to perform

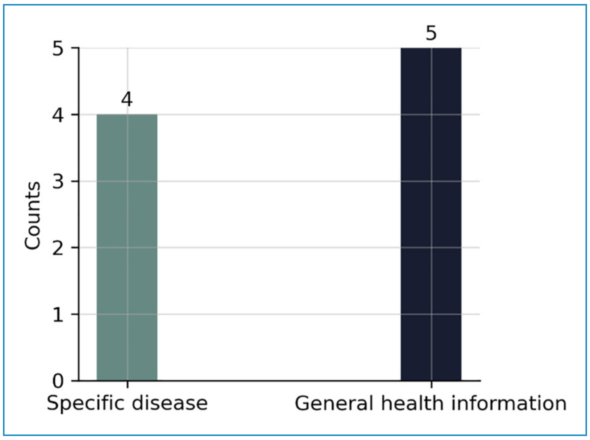

Figure 7. General versus specific health information.

effectively on new data it hasn't seen during training. Models have boundaries, which means they are created to apply to particular problems or datasets. The range of scenarios a model should perform well is defined by generalization boundaries. The results of this review show that the proposed models in the included studies have two substantial generalizability constraints:

1. The evaluated health information of all the included papers, only two (22.2%) collect new data for training and testing the models, six papers (66.67%) use archival, and one paper uses both new and archival. For example, Figure 7 shows that five papers (55.56%) of the included studies use general health information. In contrast, the other four studies (44.4%) use health information about specific diseases such as early childhood vaccination, depression, and Alzheimer's.
2. The evaluated language in the studies Figure 8 shows out of nine included studies, 77.8% $(n = 7)$ corresponds to research conducted to evaluate health information quality in English. On the other hand, 11.1% $(n = 1)$ and 11.1% $(n = 1)$ of studies correspond to research conducted in Arabic and Dutch.

Language representation

Regarding the textual representation, among the included papers, five (55.5%) used neural network embeddings trained on either general or scientific health-related text, such as AraVec, PubMed (word2vec), GloVe, BERT, and BioBERT. Detailed information about each type and the studies that employed them can be found in Table 7. One paper (11.11%) used both term frequency-inverse document frequency (TF-IDF) and neural network embeddings, while another paper (11.11%) employed the bag of words approach alongside TF-IDF. Additionally, one paper (11.11%) relied exclusively on the bag of words technique. For a more comprehensive overview, refer to Figure 9.

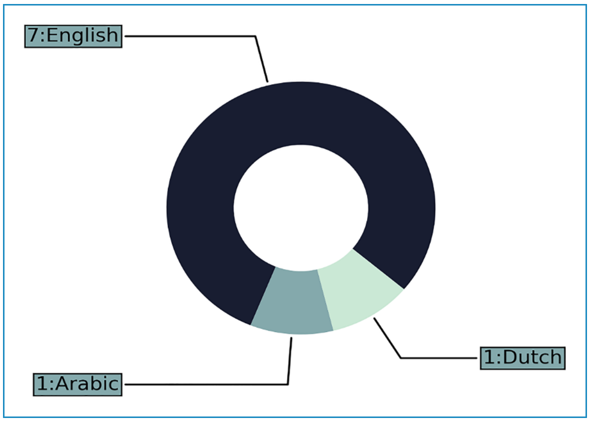

Figure 8. The language of the web pages.

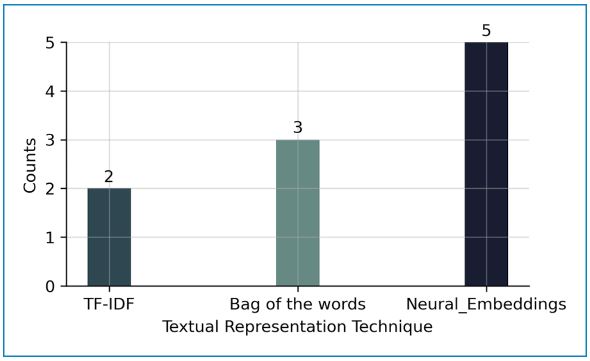

Figure 9. Shows how often each technique was used in the included studies.

Comparative information for the included studies

The included studies covered various aspects of health information quality, including credibility, misinformation, reliability, and evidence-based properties. Precisely, 3 (33.3%) of the included studies focus on the misinformation aspect of health information, the studies seeking to detect misleading information in health, finance, politics, and general health information. These studies used a supervised learning paradigm for training and testing the modes and convenience sampling for the data collection. They used datasets with different sizes ranging from small sample sizes of 360 web pages to larger sample sizes of 5509 credible and 6736 non-credible web pages. Moreover, the studies used PubMed, word2vec, and GloVe for language representations.

One (11.1%) of the included studies focuses on the credibility of general health information in Arabic. The study used a supervised learning paradigm and convenience sampling for the model training and data collection. The study used for language representation TF-IDF and neural word embedding word2vec (AraVec) with sample sizes of 500 web pages. Precisely, 3 (33.3%) of the included studies focus on health information quality, especially in the context of specific health conditions such as breast cancer, arthritis, depression, Alzheimer's disease, and general health information. These studies used a supervised learning paradigm and convenience sampling, and for language representations, they used bag-of-words or BERT. The sample sizes range from 50 to 734 web pages.

One (11%) of the included studies focuses on evidence-based properties of health information, particularly exploring the reliability of information related to shingles, flu, and migraine diseases. The study used a supervised learning paradigm, convenience sampling, and bag-of-words for language representation. The sample size was 276 web pages. Lastly, one (11%) of the included studies focuses on the reliability of early childhood vaccination health information. The study used a supervised learning paradigm and convenience sampling, and for language representations, the study used the bag-of-words and TF-IDF techniques. The sample size was 468 web pages. Table 6 shows detailed information about each study.

Discussion

Summary of contributions

This article discusses the significance of applying ML and DL methods to evaluate health information quality. Specifically, it investigates whether these approaches can enhance the accuracy of health information evaluation and surpass human performance. However, the results of a systematic review reveal a lack of universally accepted quality dimensions among health professionals and practitioners of ML and DL when evaluating health information. Furthermore, the metrics used to evaluate model performance differ from those used to assess human performance, making a comparison unfeasible. Nevertheless, this study contributes valuable insights into the automated evaluation of health information on web pages, as outlined in the following sections:

Tables 4 and 5 present details regarding the incorporated papers, which will be elaborated upon in the subsequent sections. The discussion will be divided into three parts. The first part, "Publications information and categorizing the included papers based on algorithm types," this section divides the papers into three distinct categories: papers that used ML, papers that integrated both ML and DL, and papers that focused solely on DL. The second part, "Health Information Quality Assessment: Tools, Algorithms, Dataset and Critical Analysis," looks into critical aspects related to the QA of health information, including measurement tools and quality dimensions, generalization and boundaries of developed models, model and human performance comparison. Frequently used algorithms and measurement tools, benchmark datasets, and comparative analysis. These topics will be examined for critical analyses. This analysis aims to offer valuable perspectives into the strengths and limitations of the included studies. Furthermore, this analysis will enhance our understanding of health information QA with the use of ML and DL approaches and specify the direction of future studies. The last part includes implications, limitations, directions for future studies, and a summary and conclusion.

Publications information and categorizing the included papers based on algorithm types

Publications information

The total number of citations for the included publications is 63; one of the main reasons for the low number of included studies and citations is the lack of a benchmark dataset to train and assess the built model performance.[36,37] In addition to this, assessing the online health information quality is a challenging task, as it involves over two dozen dimensions that users can assess subjectively.[38,39] The papers included can be generally categorized into three groups based on the algorithms used:

The included paper that used ML

The included papers that used ML to tackle the problem of health information quality are Al-Jefri et al.[26], Meppelink et al.[27], Alasmari et al.[34], and Di Sotto and Viviani[35] One major problem with ML is hand-craft features; in this case, most of the traditional ML algorithms follow two steps:[61]

1. The initial stage involves deriving manually designed features from the documents (or any equivalent units of text).
2. Then, features are fed into a classifier in the second step to produce the prediction.

The problem with this two-step procedure depends on manually designed features, which require much time for feature designing and analysis to achieve acceptable performance. Furthermore, because the technique relies on domain expertize to produce features, it can be challenging to generalize to new tasks. Despite this limitation, these studies have good performance ranging from a minimum *F*-score of 51 to a maximum of 91.

For checking the evidence-based property of the text, Al-Jefri et al.[26] suggested six types of ML algorithms named multinomial Naïve-Bayes, K-nearest neighbor (KNN), SVMs, stochastic gradient descent (SGD), LR, and multilayer perceptron (MLP). In addition, the study uses domain-specific criteria related to JAMA criteria. Finally, after training, the algorithm tries to classify the document according to the preceding features, whether evidence-based or not. In other words, medical authorities like the US Food and Drug Administration agency and the National Institute for Clinical Excellence in the UK licensed the treatment. The best algorithm of the study for evaluating the information regarding the three diseases migraine, flu, and shingles treatments, is LR with an *F*-score 81.7.

The study by Meppelink et al.[27] proposes two ML algorithms, Naïve-Bayes, and LR, to specify the reliability of webpages content. For evaluation purpose, the study uses a set of guidelines provided by the Dutch National Institute for Public Health and the Environmen (RIVM) about early childhood vaccination. The best algorithm of the study is Naïve-Bayes, with an *F*-score of 86.

The study by Oroszlányová et al.[35] explores the potential of predicting document quality based on its attributes. The study uses LR to achieve the goal with the HON code as the ground truth. The model achieves an accuracy of 89 in predicting the document's quality without using HON code characteristics. The most important features discovered by the model are split content, type, the process of revision, and place of treatment.

The study by Robillard et al.[34] investigates the quality of the health document and the alignment of the information provided regarding the patient and physical relationship. Therefore, the study proposes a DT classifier to tackle the problem. Concerning the evaluation process, the study uses complimentary quality criteria of the QUEST tool (authorship, attribution, conflict of interest, currency, and complimentary tone), and the achieved *F*-score is 90% in evaluating the Alzheimer's disease.

The included papers that used ML and DL

The papers by Alasmari et al.[36], Di Sotto and Viviani,[37]and Kinkead et al.[24] used different ML algorithms to tackle the issue and use ML for comparison purposes. Still, two significant problems of these studies are using word embeddings trained in general or scientific medial text, which could yield poor performance,[62,63] and these models oriented toward evaluating health information in a specific language making them unimodal language modal.[64,65] Nonetheless, these studies achieved a lowest *F*-score of 79 and a highest of 95.

The study by Alasmari et al.[36] proposes the hybrid features method to investigate the credibility of web pages; both textual and non-textual features. For the textual features, the study uses the term frequency-inverse document frequency technique (TF-IDF) and the Arabic version of the word2vec, a neural embedding model pre-trained in general Arabic content. In contrast, regarding the non-textual features, the study uses four criteria from HON code, including authoritative, attribution, ads policy, and transparency, along with other features. For the evaluation process, the study uses LR, SVM, DT, and LSTM. The best-performing model is LSTM, with word embeddings achieving an *F*1-score of 75 with textual features and an *F*1-score of 77 with hybrid features. The study by Di Sotto and Viviani[37] classifies health information using binary classification to distinguish health information from misinformation. The study uses over 119 features and five classical ML algorithms: gradient boosting, LR, Naïve Bayes, and random forests. In addition, the study uses a word embedding technique and a pre-trained neural model Golve to have a more complex representation of the text. Finally, the study uses two DL architectures: CNNs and bidirectional LSTM. The best-performing classifier of the study is random forests, with 88 for AUC (the area under the curve score).

In the study by Kinkead et al.[24] the authors proposed a method for automatically implementing DISCERN tool. First, the four hierarchy encoder attention (HEA)-based architectures (hierarchy encoder with BERT, hierarchy encoder with BioBERT, hierarchy encoder attention with BERT, and hierarchy encoder attention with BioBERT and random forest are trained on articles about breast cancer, arthritis, and depression (all related to the five DISCERN criteria). Then, a BERT layer converts words, phrases, and documents to dense vector representations. Finally, a SoftMax layer is used to classify the articles. This model is specifically relevant for articles that discuss various choices for treatment. Therefore, its usability is restricted to this particular category of articles. Despite limitations regarding the type of the article, the model's performance is quite good, with a lowest f-macro score of 74 and a highest f-macro score of 75. The best model is hierarchy encoder with attention (HEA) BERT, achieving an *F*1-score of 75.

The included papers that used DL

The research conducted by Upadhyay et al.[38,39] employs the web2vec algorithm introduced by Feng et al.[66] This algorithm is designed to identify fraudulent web pages using an approach that generates an embedded version of web pages incorporating their URL, content, and DOM structure. Subsequently, the hybrid CNN-BiLSTM model leverages this condensed representation to capture both local and global characteristics of the web pages. The study by Upadhyay et al.[38] uses the same proposed algorithm, web2vec, for detecting misinformation by taking into consideration the context the study suggested using PubMed word2vec word embeddings trained in health-related text from PubMed. Rather than concentrating on attributes associated with the webpage's URL, the study focuses on the URL links inside the webpage because it is a better indicator of the reliability of the webpage (e.g. the presence of a lot of commercial weblinks). With the preceding information, the suggested solution includes the subsequent elements:

(i) The data parsing stage: The page links, content, and DOM is parsed from the HTML document to extract the required information, which will be used in the subsequent steps.
(ii) Data representation stage: In this stage of web page processing, the web page's content is represented in the form of word and sentence-level representation while considering the other features, page links, and DOM.
(iii) Feature extraction stage: The hybrid model of CNN-BiLSTM is used to extract the local and global features.
(iv) The final stage: The web page classification was performed using a fully connected layer that utilized a sigmoid function. This categorization aimed to distinguish between credible and incredible web pages. The top-performing model in the study was Cred-W2V, which was trained using the web page's content, DOM, and link embeddings. Moreover, it used PubMed's word2vec as a starting point for its weights. The paper of Upadhyay et al.[39] is considered an extension of the previous article with an additional semantic aspect: the part of the speech extracted from the web pages and the grammatical aspects of the web pages. The "Comparative and critical analysis" section offers additional details and suggestions for future research. It shows a comprehensive examination and evaluation of the included study's methods and limitations, aiming to provide a deeper understanding and potential avenues for further exploration.

Health information quality assessment: Tools, algorithms, dataset, and critical analysis

Measurement tools and quality dimensions

The results of this study are consistent with prior research into measurement instruments and quality dimensions.[23] The concept of health information quality lacks a universal definition, leading different ML and DL practitioners to use different criteria and indicators. This highlights the nuanced and complex nature of health information quality.[38,39] The study also reveals the crucial role quality criteria and indicators play in comparing ML and DL model results and helping in selecting the model that covers most aspects of information quality.

For instance, Table 8 demonstrates significant limitations and inconsistencies in the approach of included studies when assessing health information quality. Some studies use only one quality criterion,[26] while others, like,[37] employ 119 different criteria, the highest number among all included studies. Surprisingly, only two studies[36,24] employ commonly used tools by health professionals, such as the HON code[14] and discern criteria.[15] Furthermore, none of the included studies uses the JAMA score,[13] which is widely used in recent studies of manual evaluations of health information.[11]

While the current review suggests careful consideration in using quality criteria to prefer one study model over another or comparing it with human performance due to the mentioned limitations, it emphasizes the importance of using various quality criteria to capture the multidimensional nature of health information quality.[71] Each criterion focuses on specific aspects, such as accuracy, reliability, and credibility. The selection of criteria depends on the research context and objectives. While all criteria contribute to the overall quality, their relative importance may vary depending on the context, making it challenging to recommend specific criteria for universal use.

Building upon the findings of the current review, it becomes clear that using various quality criteria is essential to completely capture the multidimensional nature of health information quality.[71] It should be emphasized that not all quality criteria offer the same advantages, as each measure serves a distinct purpose. The JAMA benchmarks focus on the trustworthiness of health information sources by assessing authorship, attribution, disclosure, and currency to ensure that the information comes from reliable sources, is up-to-date and includes essential details about funding and authorship. This will help consumers trust the information and make informed decisions about their health conditions.

The HON code goes beyond trustworthiness and includes additional criteria like complementarity, privacy, justifiability, transparency, and advertising policy. By considering these criteria, health information providers can enhance doctor–patient relationships, protect personal information, support claims with scientific evidence, maintain direct communication channels, and differentiate between main content and advertising. This will enhance the overall quality and credibility of health information.

The DISCERN instrument focuses on evaluating health information quality related to treatment choices by providing guidelines in treatment description like treatment works, treatment benefits, treatment risks, effect on quality of life, side effects of no treatment, and areas of uncertainty; it enables health information providers to provide comprehensive and reliable information to consumers about the treatment. This will enable consumers to make well-informed choices regarding their treatment options.

In summary, the benefits of these tools vary, and it's essential to consider the study's objective. If the main focus is on assessing the trustworthiness of health information sources, then the JAMA benchmarks would be the ideal choice. On the other hand, if the study aims to offer a comprehensive evaluation, including verifying claims and supporting them with scientific evidence, complementarity, privacy, justifiability, transparency, and advertising policy, the HON code would be more suitable because it goes beyond trustworthiness. Finally, the DISCERN instrument emerges as the best option for studies focusing on the quality of treatment choices. Its guidelines related to treatment descriptions, such as treatment works, treatment benefits, treatment risks, etc., enable a thorough evaluation of health information about treatment options and help make informed decision-making.

To address these challenges, we propose in the future research section combining multiple tools, including the practical experience and common sense of health specialists, to create a more comprehensive tool that covers various aspects of health information quality. Furthermore, these criteria should be grouped based on their functions and ranked according to their importance to health information quality, as perceived by health specialists. The "Directions for future studies" section offers further insights into this method.

Generalization and boundaries of the developed models

The findings show that the suggested models suffer from three significant limitations concerning generalizability. The first limitation is the language limitation. In this case, the results converge with previous findings,[36] where most studies focus on English while other languages are neglected. A total of seven studies from the included paper focus on English.[26,37,24,35,34,38,39] On the other hand, only one study focuses on Arabic,[36] and one in Dutch.[27] The second limitation is the health information used to train the models. There are two types of health information used to develop the models. The first type is general health information used by the subsequent studies,[36,38,37,35,39] and the second type is for a specific disease used by the subsequent studies.[27,24,26,47] These limitations make it hard to generalize the model to a particular illness or language and only work within the boundaries [language and specific disease information]. In other words, the datasets used for training and testing the ML and DL models may not accurately reflect the actual distribution of the health information quality problem, and this could happen when the datasets fail to cover all possible variations and data quality complexities present in real-world health information.

Coupled with the findings of the previous two limitations, the third limitation that affected the generalizability was that the textual representation features in the included studies were quite limited. To give an illustration, studies of Meppelink et al.[27] and Alasmari et al.[36] used TF-IDF term frequency-inverse document frequency technique, and study of Meppelink et al.[27] also uses a bag-of-words method; both approaches are limited and do not capture many of the syntactic and semantic relations among words in the document. The studies of Al-Jefri et al.[26] and Robillard and Feng[47] used only the bag of words technique, which is the least effective technique for weighting features where the most frequent features are weighted highly. Even though some of the included studies use neural embedding techniques, which can capture many of the syntactic and semantic relations among words in the document, they still have some limitations. Firstly, studies of Kinkead et al.[24] and Alasmari et al.[36] used words2vec and BERT, BioBERT, respectively. Study of Alasmari et al.[36] uses AraVec, which word2vec pre-trained word embedding trained in general Arabic text from Tweets, World Wide Web pages, and Wikipedia articles, making the models imprecise for health information evaluation or detection.

Similarly, study of Kinkead et al.[24] uses BERT pre-trained on English Wikipedia and books corpus and BioBERT pre-trained on scientific health-related text as well as a general text. Nevertheless, studies of Upadhyay et al.[38,39] used the PubMed version of the word2vec, which is pre-trained on PubMed and is considered as scientific health-related text and may not cover general health information beyond the biomedical field. In conclusion, the findings show that regarding textual representation features, there is clear evidence of omitting the importance of contextualized embedding as supported by recent studies.[62,63] More precisely, they show that applying word embeddings directly to context-specific natural language problems yields unsatisfactory performance.

The model and human performance comparison

Contrary to the attempt by Kinkead et al.[24] to compare human and model performance in their study, the results of this review show that the comparison is not possible due to the different metrics used to measure the model and human performance, as shown in the result section. More precisely, the included studies use recall, precision, f_score, and accuracy for model performance and Cohen's Kappa for human performance. Firstly, we will discuss the importance of comparing model and human performance. Secondly, we will examine the validity of the comparisons made in the included studies between model performance and human performance.

Firstly, with numerous ML systems striving to automate tasks that humans do well, three important benefits can be obtained by integrating human-level performance into future research on evaluating health information quality. These benefits include the following aspects:[72] first, it is easy to get label data from health professionals.

Second, error analysis is based on health specialist intuition and cognitive abilities. Contrary to the general health information user, who often struggles to assess the health information quality due to their limited cognitive ability[73] and low health literacy,[74] health specialists have two qualities that make them different from DL and ML language models: first, health literacy: health specialists have a unique ability to perceive linguistic nuances that influence health information quality using their knowledge and experience. Second, personal judgement: assessing health information quality most of the time needs subjective judgement[38,39] (a.k.a Personal Judgement) because it contains aspects that cannot be easily quantified, so incorporating the health specialist in the evaluation process of health information quality will include the subjective evaluation criteria.

Third, use human-level error as a proxy of Bayes error to estimate the optimal error and performance rates. For instance,[24] reports a manual accuracy of 94%, where the optimal error rate will be 6%, and the optimal model performance will be 94% for that specific dataset. In other words, If a model's error rate is close to the estimated Bayes error (i.e. the human-level error in this case), it indicates that the model is performing at or near the best possible level on that dataset. However, suppose the model's error rate is significantly higher than the estimated Bayes error. In that case, it suggests that there is room for improvement and the model is not performing as well as it could be. It should be emphasized that Bayes error is an essential concept in ML and DL and represents an idealized lower bound on error. In practice, it's not always achievable due to limitations in data quality, model complexity, and other factors. Nonetheless, it serves as a useful benchmark for assessing model performance.[72,75]

Secondly, regarding the validity of the comparisons made in the included studies, the review findings indicate that the comparison between human and model performance is invalid. This conclusion is drawn from analyzing each matrix's definitions and mathematical calculations. As a result, it is essential to use the same matrix for evaluating the model and human performance to ensure a fair and accurate comparison.

Guidance from optimal error rate

Based on the previous example,[24] an "optimal" classifier, which is a health expert, can achieve an error rate of ~ 6%. When assessing a health document, a health expert should be able to distinguish whether the information it contains is of high quality or not, with a potential error rate of 6%. It is reasonable to expect that a machine could perform just as well.[72] For instance, let's say the error rate on the training set is 7% and on the development set is 12%. This means that the training set performance is almost at the optimal error rate of 6%. Therefore, there isn't much potential for improvement in the training set performance. However, the model is not generalizing well to the development set, which indicates that there is significant room for improvement in reducing errors in the development set. In sum, From these instances, it becomes evident that having knowledge of the ideal error rate serves as valuable guidance for determining our subsequent actions.

Frequently used algorithms and measurement tools

As observed in the preceding sections, most of the recent research examined in this systematic review uses DL to tackle the problem of health information quality.[36–38] In addition, some of these studies use ML algorithms to compare the language models built using ML versus DL. The findings of this study converge with previous results about using language models produced by DL algorithms. These models have demonstrated enormous success in translation and question-answering by capturing features and nuances in language that has been considered problematic for a long time.[30,31] In addition, concerning health misinformation, the recent interest of researchers has turned to DL after achieving massive success, especially in social media.[76,28] Therefore, the findings suggest that DL algorithms are the most promising avenue in artificial intelligence to tackle the problem of health information quality.

Regarding the frequently used quality dimensions and measurement tools, The findings of this review converge with the previous systemic review[77,23] that the quality is evaluated using different dimensions and indicators in each study. These dimensions and indicators will be examined in further detail in the subsequent section. In addition, the present study is the first to investigate the automatic evaluation process of health information quality on web pages using ML and DL.

Benchmark dataset

With respect to the dataset used to evaluate the developed models, this study's findings converge with previous results;[36,37] there is a lack of an existing benchmark dataset to assess the performance of the developed model. Therefore, some researchers collect new data to test the model's performance. Three of the included studies collect new data. Firstly, Alasmari et al.[36] collected general health information in Arabic using the keywords "medical advice," "medical information," and "health information" within the total dataset size of 500 examples. Secondly, Meppelink et al.[27] collected data in Dutch about early childhood vaccination using the keywords "vaccinations safe," "vaccinations unsafe," "vaccinations good," and "vaccinations bad" within the total dataset size of 468 examples. Thirdly, Al-Jefri et al.[26] collected new data in English about shingles using the keyword "shingles treatment." In addition, he used archival data about the flu[78] and migraine[79] within the total dataset size of 276 examples.

On the other hand, six papers in the systematic review use archival data[38] and the more recent version of the same study[39] uses three archival data. First, Microsoft Credibility Dataset[80] consists of 1000 examples and covers health, finance, and political topics. Second, Medical Web Reliability Corpus[81] collected uses the HON code. The authors consider all websites with HON certification as reliable websites. For unreliable examples, they search the web using keywords such as the disease and "miracle cure." The dataset consists of 360 web pages; 180 reliable, and 180 unreliable. Third, CLEF eHealth[46] medical content has 5509 credible and 6736 non-credible web pages.

Robillard et al.[34] used 50 random samples from a total of 380 web pages about Alzheimer's provided by Robillard and Feng.[47] Similarly,[24] use a dataset consisting of 269 web pages about three diseases: breast cancer, arthritis, and depression. The dataset is collected using Google and Yahoo search engines, and data provided in the paper.[82] The study of Di Sotto and Viviani[37] uses three datasets in their study. First, CoAID is a collection of news and claims written in English about Covid-19 provided by Cui and Lee[83] consisting of 3555 examples. Second, recovery is a collection of news about Covid-19 written in English collected from 60 websites provided by Zhou et al.[84] Third, Fake Health is a collection of 2029 examples of reviews from health experts about different health topics, such as medical intervention, wellness, etc., provided by Dai et al.[85] Finally, the study[35] used a dataset consisting of 734 web pages that covered a wide range of topics, including treatment options and diseases provided by Lopes and Ribeiro.[86]

In sum, the findings indicate that there is no dataset collected in a way that covers all quality dimensions and indicators considered by health professionals in the manual evaluation, which can be used as a benchmark dataset to evaluate the model's performance.

Comparative and critical analysis

The essential analysis of the included studies points to several similarities and differences in their methodologies, metrics, and other factors. While all studies strive to handle diverse characteristics of health information quality, such as credibility, misinformation, and reliability, they use different approaches to achieve their objectives. Detailed information is shown in Table 6.

First, all the studies included in the review shared a common similarity in their method for data collection for training and testing the models. They all used a convenience sampling method. Nonetheless, this approach has one limitation that could affect the final conclusion. This limitation is the possibility of introducing bias,[87,88] as convenience sampling may not accurately represent the entire population of health information web pages.

Additionally, the models developed using this sampling method may not generalize well to other health information sources or types. For instance, The research study conducted by Meppelink et al.[27] gathered data on early childhood vaccination information using convenience sampling. However, an important aspect to emphasize is that the results may not accurately represent the complete range of early childhood vaccination information accessible online.

Second, another common similarity among the included studies is the use of supervised learning paradigms. Nevertheless, it is essential to highlight the limitation of supervised learning, the reliance on labeled data. For the supervised learning models to achieve accurate predictions and high classification performance, the labeled data must be of high quality and effectively represent the real-world distribution of the target problem. In sum, to overcome the first and second limitations, future studies need to use more representative methods for data collection, such as probability sampling methods. Moreover, regarding the learning paradigms, future studies could use semi-supervised, unsupervised, or self-supervised learning paradigms to provide alternative methods to overcome the limitations of the supervised learning paradigm.

Third, a wide-ranging of variations could be found when studying the information regarding the different types of algorithms used, metrics, or the Dataset context. Some studies used a mixed method approach, combining different ML algorithms with DL algorithms, while others focused on DL algorithms or traditional ML algorithms, as seen in the preceding section. The choice of metrics also varies, with studies reporting accuracy, $F1$-score, precision, recall, and AUC score, depending on their specific research objective. Furthermore, the study's datasets targeted different contexts, including general health information, specific diseases, and other domains such as finance and politics.

These variations have several implications associated with them as follows: First, the studies that only used ML models, these models may not capture the complex semantic nuances in textual data as efficiently as DL algorithms.[69,24] As a result, using these models may lead to less accurate predictions and lower classification performance. Second, the utilization of diverse measurements for assessing model performance in various studies complicates the comparison of their outcomes. Finally, analyzing the dataset context is crucial in assessing the generalizability of the results obtained in the included studies, considering that each study used a dataset with a different context. As an illustration, the dataset context of study[36] was focused on general health information in Arabic. Although the dataset context provides a proper understanding of credibility assessment in this particular context, the results may not apply to other languages or specific medical topics; more information about each study dataset context is provided in Table 6.

Fourth, the final crucial aspects to consider in the comparison between the included studies are the sample size and language representation. These two aspects significantly impact specifying the generalizability of the results.

Firstly, the sample sizes there were deviations among the included studies, with the smallest sample containing 50 web pages and the largest sample containing 5509 credible and 6736 non-credible web pages. The studies Robillard et al.[24], Di Sotto and Viviani[27], Alasmari et al.[34], and Kinkead et al.[36] employed a relatively small sample size, which could affect the reliability and validity of the final results of the models. Furthermore, with a small sample size, the models may not fully capture the dataset's complete variability and will impact the generalizability.

The study of Oroszlányová et al.[35] has a sample size of 734 web pages, which is somewhat larger than the previous studies. Nonetheless, it may still not be complete enough to capture the full population of general health information web pages. Lastly, the studies[38,39,37] used various datasets with different sample sizes ranging from small sample sizes of 360 web pages to large sample sizes of 5509 credible and 6736 non-credible web pages. Yet, the variation in sample sizes across different datasets can present an inconsistency in the results, subsequently compromising their reliability and validity.

Secondly, the choice of language representation varied among the included studies, with each study using a different technique to create a word or sentence representation. The studies used a range of techniques for language representation, ranging from simpler techniques like TF-IDF and bag-of-words to more sophisticated techniques such as BERT or Bio-BERT. These techniques vary in their ability to capture health-related text's semantic meaning and contextual nuances.

The use of simple methods such as TF-IDF or bag-of-words alone limits the model's ability to understand the meaning and nuances of health text data, and it may impact classification accuracy. In contrast, using a relatively complex neural word embedding such as word2vec, and Glove, which relies on the distributional hypothesis[89,90] to understand the meaning of the words within specific contexts, could provide useful understandings of the semantic relationships of the words. Nevertheless, these techniques are incapable to capture the contextual meaning of the health text data. Finally, although sophisticated techniques like BERT effectively capture the contextual meaning of words or sentences, retraining them on online general health text data is essential to improve their performance,[62,63] which requires sufficient time and computation resources. For additional information, the researchers may refer to the following recent reviews.[89–91]

In summary, the included studies provide valuable insights into health information QA. However, the limitations of sample size and language representations may impact the generalizability of the results. Future research should consider a large sample size with diverse examples of classes and advanced language representations to enhance generalizability.

Implications

This systematic review provides valuable insights for addressing health information quality issues on web pages, with implications for both practice and research.

Implications for practice

The findings of this review offer practical applications for improving health information QA. Practitioners can use the insights gained from this review to develop a set of quality criteria. These criteria will serve as a tool for automating the evaluation process, enhancing efficiency, and enabling comparisons between human evaluations and model performance.

Implications for research

The review also lays the groundwork for future research in this domain, suggesting several potential research directions:

1. Explore the development of pre-trained models specifically tailored to health-related text available on the internet. Incorporating domain knowledge into word representations can significantly enhance the performance of DL algorithms in evaluating health information quality.
2. Investigate the creation of a multilingual model capable of evaluating health information quality in various languages. Current models are limited by their language specificity, and developing a multilingual approach can expand the scope and applicability of health information assessment research.
3. Consider further research into benchmark datasets and evaluation metrics that align with the quality criteria examined in this review. This will contribute to a standardized framework for assessing health information quality.

In conclusion, this systematic review not only provides practical guidance for practitioners but also lays the foundation for future research efforts in the realm of health information QA on the web.

Limitations

There are three limitations to consider in this study. Firstly, it focuses exclusively on assessing health information quality on web pages, overlooking the assessment of social media platforms and other communication mediums. Secondly, the evaluation of the included papers may be restricted in terms of human and model performance perspectives. Thirdly, this study is limited to research that uses predefined textual or non-textual criteria, such as the HON or JAMA score for the non-textual criteria and BERT or Word2vec textual criteria. Additionally, it is essential to note that this review suffers from a limited number of studies because the automatic evaluation of health information on web pages is relatively new and in its infancy.

Directions for future studies

Several studies have addressed various aspects of health information quality, as discussed in this review. Still, several methodological shortcomings have been identified, requiring further investigation and refinement. This future research section aims to address two primary areas of improvement: establishing universal quality dimensions and using DL techniques. The first area of concern revolves around the need for universal quality dimensions which will provide the following benefits:

1. It will enhance the effectiveness of comparing the models' performance of different studies and selecting the best one.
2. It will make human and model performance comparison possible by using Cohen's Kappa coefficient or *f*-score for measuring the model performance and human performance. Similarly, the *d* prime[92] metric from the single detection theory could be used for measuring human and model performance.
3. It will make collecting benchmark datasets that all studies on health information quality can use more simplified. Furthermore, collecting benchmark datasets promotes collaboration, accelerates research progress, and enhances the quality of studies in the field of health information.

Steps of developing the universal quality dimensions

Figure 10 shows the proposed solution for a universal quality dimension that health professionals, ML, and DL practitioners could use, which consists of three steps for creating quality dimensions.[93]

First, the researcher should focus on the existing literature using the ontological approach (a.k.a theoretical approach) by identifying all the quality criteria used by health professionals in recent studies such asJasem et al.[94], Leung et al.[95], and Ölçer and Taşkaya Temizel[96] and DL and ML practitioners such as Kinkead et al.[24] and Upadhyay et al.[38,39] Second, a questionnaire can be developed using the empirical method to determine the most relevant criteria for assessing health information quality. This questionnaire will include quality criteria identified in the first step and factors related to search behaviors and personal information, such as profession, major, country, search terms, and much more, which can influence perception. Third, modify the health criteria using a health professional's common sense and experience if required.

Finally, this questionnaire could be promoted through social media platforms such as Facebook, WhatsApp, and Email to collect participants' perceptions, mainly health specialists. Researchers will use the questionnaire to gather data on various aspects of health information quality, enabling them to efficiently determine the essential criteria for evaluating health information quality. To conduct a robust statistical analysis of the collected data, it is recommended to use an analysis of variance test. In this test, the independent variable would consist of the type of profession (caregivers or doctors), education level, academic field, and country of residence. On the other hand, the dependent variable would encompass the quality criteria. The primary goal of this research will be to explore how different factors within the independent variables, as well as their associated levels, influence the perception of health information quality criteria. Ultimately, the researchers then use the perception of the health specialist to rank criteria important to the health information quality and select the most important one.

Deep learning for health information quality

The second area of improvement focuses on integrating DL techniques to address the challenges associated with health information quality. DL has demonstrated promising results in various domains[30–33] and has theoretical support through the universal approximation theorem (UAT). UAT is one of the essential theoretical underpinnings of neural networks. The theorem states that a sufficiently broad or deep network can approximate any continuous function as long as the problem can be set in input–output pairs.[97] However, It is essential to acknowledge that the theorem does not ensure that every sufficiently broad or deep network can approximate every possible function. While UAT provides valuable insights into shallow networks' capabilities, deep neural networks' power and complexity go beyond what is covered by the UAT. By using DL, a solution for the following issues can be provided: The theorem states that a sufficiently broad or deep network can approximate any continuous function as long as the problem can be set in input–output pairs

1. Reduce the training data by using self-supervised learning to generate pre-trained models consisting of language representation of general health-related text available for the average online health information consumer.
2. It will help create a multilingual model to evaluate health information quality in multiple languages.

Figure 11 outlines the steps for developing DL language models to enhance health information QA. The proposed framework is based on the following three steps: 1. Prepare pre-trained models: The researchers must first develop pre-trained models using a self-supervised learning paradigm with pretext tasks (either by predicting the next word in a sequence or by predicting a complete sentence) explicitly customized for online general health information. Self-supervised learning enables DL models to learn from larger volumes of data, an essential aspect for discerning and understanding patterns present in more nuanced and infrequent representations of the world.[98] By using online health information during the training and

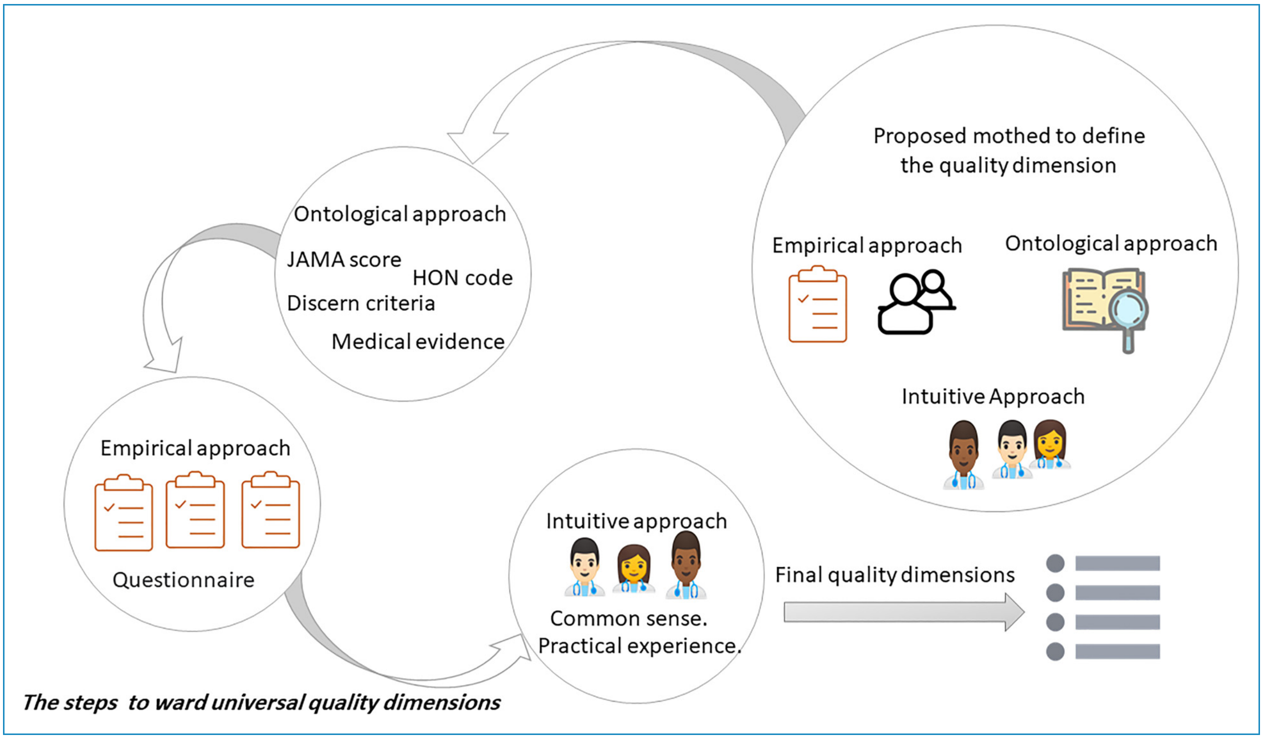

Figure 10. Three approaches to defining quality dimensions.

development phases, the pre-trained models effectively capture domain-specific knowledge related to health information (a.k.a contextual understanding of words and sentences). This contextual understanding of words and sentences in the health information domain can lead to more informative and specialized word embeddings.[62,63]

After successfully developing pre-trained models through self-supervised learning with pretext tasks for online general health information, the next step is to use these models for generating word embeddings. Word embeddings are dense vector representations of words that capture semantic relationships between them.[90,91] By leveraging the learned representations from self-supervised learning pretext tasks, the pre-trained models can transform individual words into high-dimensional continuous vectors in a way that encodes their contextual meanings within the domain of health information.

These pre-trained models will impact research in natural language processing about online health information quality for any downstream tasks, such as text classification, question answering, etc. Several studies have shown the effectiveness of the pre-trained language models in improving various natural language processing tasks.[98–102] Nevertheless, expanding the pre-trained such as BERT or word2vec models, to cover health information requires a large amount of data and a substantial amount of computation time.

Our study suggests a good approach to minimizing data requirements and computation time is by focusing on expanding the capabilities of existing pre-trained models. Regarding Arabic health data, researchers could consider expanding AraVec[67] or AraBERT[103] to contain a wider range of language representations by including Arabic health data. Additionally, they could explore the possibility of extending unilingual BERT, or multilingual BERT (mBERT)[69] and the updated version[104] to include the health data targeting the general consumers of health information of single or multiple languages. These extensions will enhance the efficiency of the existing models' capabilities in a medical context. For further insights, the researchers could refer to the recent systematic reviews in word embedding[105] and pre-trained models.[106]

2. Labeling the health data: The second step involves the health professionals labeling the health data into high and low quality within the three most common classes(general health information, treatment description, and medical advice) for the downstream task.

3. Developing a proposed model: The third step concerns addressing the requirement for a proposed novel framework to improve the assessment of health information quality. During this review, we thoroughly investigated the various suggested models, including their limitations and drawbacks, including Kinkead et al.[24], Alasmari et al.[36], Di Sotto and Viviani,[37] and Upadhyay et al.[38,39] Consequently, future studies need to suggest a novel framework to tackle the issue of assessing health information quality more effectively.

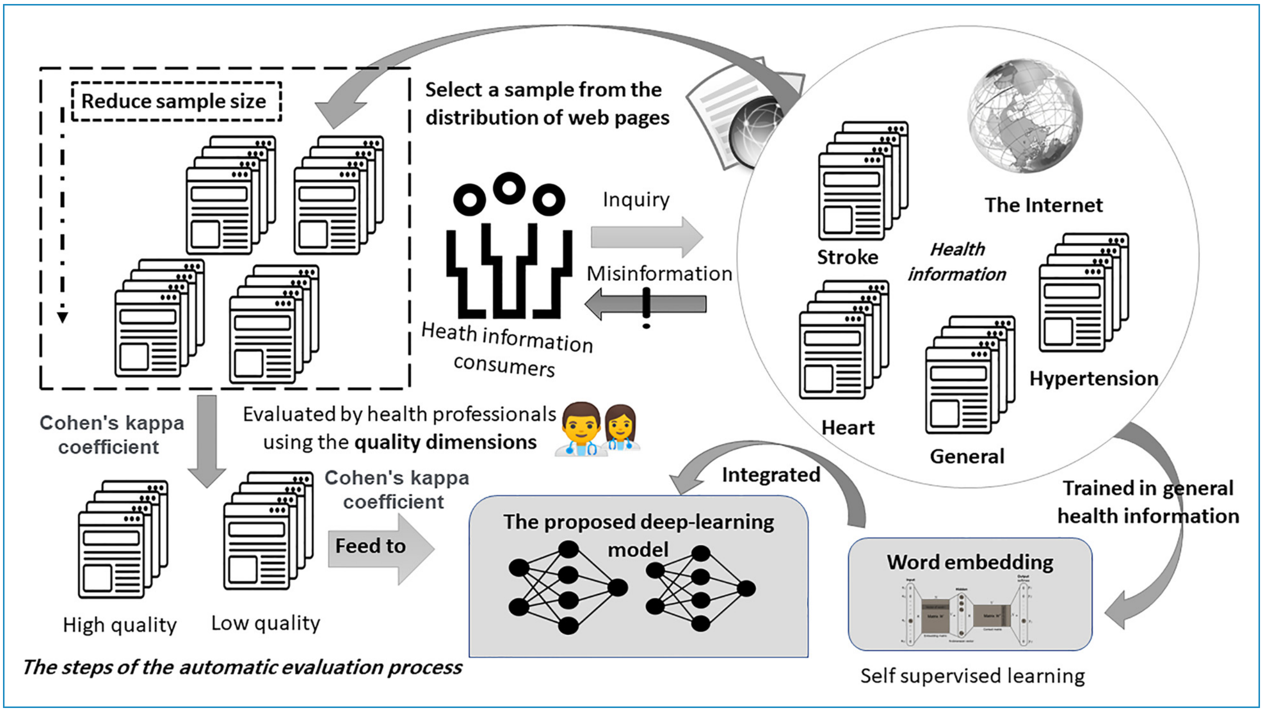
Reduce sample size
Select a sample from the distribution of web pages
Inquiry
Misinformation
Heath information consumers
The Internet
Health information
Stroke
Heart
General
Hypertension
Cohen's kappa coefficient
Evaluated by health professionals using the quality dimensions
Cohen's kappa coefficient
Feed to
High quality
Low quality
Integrated
The proposed deep-learning model
Trained in general health information
Word embedding
Self supervised learning
The steps of the automatic evaluation process

Figure 11. Proposed steps for developing deep learning language models for the health information quality.

Summary and conclusion

In this review, we investigate the extent to which ML and DL language models improve the precision of evaluating health information quality on web pages. Our systematic review aimed to bridge the research gap by comprehensively summarizing the current state of automatic health information quality evaluation. We focused on ML and DL models and found that they hold significant promise in surpassing human evaluative capabilities. This systematic review offers a new perspective on the automatic evaluation process of online health information quality and identifies several shortcomings. Future research should focus on the following areas to address the identified shortcomings and advance the understanding of this field.

Defining universal quality dimensions

To facilitate consistent comparisons between human and model performance, researchers should prioritize the development of universally accepted quality dimensions and indicators. This entails conducting a comprehensive literature review to identify existing quality criteria used by health professionals and ML practitioners. Subsequently, an empirical questionnaire should be developed, and a statistical analysis should be conducted to prioritize the most critical criteria based on the input from health specialists.

Standardizing evaluation metrics

To ensure fair and accurate comparisons, it is crucial to align the metrics for measuring both model and human performance. The adoption of specific metrics, such as Cohen's Kappa coefficient, should be encouraged, enhancing the reliability of evaluation results.

Specialized embeddings for health information

Overcoming the limitations of pre-trained models necessitates the development of specialized embeddings customized for evaluating health information quality. Future research should focus on self-supervised learning with health-specific pretext tasks and training on online health information to capture domain knowledge. Additionally, extending existing models like AraVec or AraBERT for health data evaluation is an avenue worth exploring.

Multilingual models for enhanced generalizability

The development of multilingual models is vital to extend evaluation across diverse languages. Researchers can leverage cross-lingual transfer learning techniques and fine-tune pre-trained models such as mBERT or XLM-RoBERTa on health-related content, which have already been trained on extensive text in various languages.

Creating a benchmark dataset

The absence of a benchmark dataset is a significant challenge. Establishing a standardized benchmark dataset specifically for evaluating online health information quality is crucial. This process should involve defining the dataset's goals and scope, collecting health-related web pages in various languages, and establishing precise quality criteria. An expert team should then annotate the web pages based on these defined criteria. Expanding the evaluation to include content from social media and other platforms would offer a more comprehensive perspective. By addressing these areas, future research can improve the reliability, generalizability, and practicality of evaluating online health information quality, benefiting healthcare consumers and professionals.

Acknowledgements: The authors are thankful to the School of Computer Science USM for all the support they provide and to the Hadhramout Establishment for Human Development Foundation for funding this work.

Contributorship: Not applicable.

Consent statement: Our study is a systematic review and does not involve human participation. Therefore, the need for a consent statement from patients is not applicable to our study.

Declarations of conflicting interest: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Ethical approval: Not applicable.

Funding: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research is funded by the Hadhramout Establishment for Human Development Foundation, Yemen, Hadhramout.

Guarantor: Not applicable.

ORCID iDs: Yousef Khamis Ahmed Baqraf https://orcid.org/0000-0002-8741-4622
Pantea Keikhosrokiani https://orcid.org/0000-0003-4705-2732

References

1. Daraz L, Morrow AS, Ponce OJ et al. Can patients trust online health information? A meta-narrative systematic review addressing the quality of health information on the internet. *J Gen Intern Med* 2019; 34: 1884–1891.
2. Zhang Y and Kim Y. Consumers' evaluation of web-based health information quality: meta-analysis. *J Med Internet Res* 2022; 24: e36463.
3. Hesse BW, Nelson DE, Kreps GL et al. Trust and sources of health information: the impact of the internet and its implications for health care providers: findings from the first health information national trends survey. *Arch Intern Med* 2005; 165: 2618–2624.
4. Popovac M and Roomaney R. Measuring online health-seeking behaviour: construction and initial validation of a new scale. *Br J Health Psychol* 2022; 27: 756–776.
5. Fox S and Duggan M. Pew Research Center. Washington, DC: Pew Internet & American Life Project; 2013. https://www.pewresearch.org/internet/2013/01/15/health-online-2013/, 2013.
6. Kealey E and Berkman CS. The relationship between health information sources and mental models of cancer: findings from the 2005 health information national trends survey. *J Health Commun* 2010; 15: 236–251.
7. Fox S. Pew Research Center. The social life of health information, 2014.
8. McKinley CJ and Wright PJ. Informational social support and online health information seeking: examining the association between factors contributing to healthy eating behavior. *Comput Human Behav* 2014; 37: 107–116.
9. Alhajj MN, Mashyakhy M, Ariffin Z et al. Quality and readability of web-based Arabic health information on denture hygiene: an infodemiology study. *J Contemp Dent Pract* 2020; 21: 956–960.
10. Alnaim L. Evaluation breast cancer information on the Internet in Arabic. *J Cancer Educ* 2019; 34: 810–818.
11. Halboub E, Al-Ak'hali MS, Al-Mekhlafi HM et al. Quality and readability of web-based Arabic health information on Covid-19: an infodemiological study. *BMC Public Health* 2021; 21: 1–7.
12. Bánki F, Thomas SJ, Main B et al. Communication of information about oral and oropharyngeal cancer: the quality of online resources. *Oral Surg* 2017; 10: 4–10.
13. Silberg WM, Lundberg GD and Musacchio RA. Assessing, controlling, and assuring the quality of medical information on the internet: Caveant lector et viewor—let the reader and viewer beware. *Jama* 1997; 277: 1244–1245.
14. Boyer C, Selby M, Scherrer JR et al. The health on the net code of conduct for medical and health websites. *Comput Biol Med* 1998; 28: 603–610.
15. Charnock D, Shepperd S, Needham G et al. Discern: an instrument for judging the quality of written consumer health information on treatment choices. *J Epidemiol Commun Health* 1999; 53: 105–111.
16. Stvilia B, Mon L and Yi YJ. A model for online consumer health information quality. *J Am Soc Inf Sci Technol* 2009; 60: 1781–1791.
17. Eysenbach G and Köhler C. How do consumers search for and appraise health information on the world wide web? Qualitative study using focus groups, usability tests, and in-depth interviews. *BMJ* 2002; 324: 573–577.
18. Chou WYS, Oh A and Klein WM. Addressing health-related misinformation on social media. *Jama* 2018; 320: 2417–2418.
19. Wand Y and Wang RY. Anchoring data quality dimensions in ontological foundations. *Commun ACM* 1996; 39: 86–95.
20. Al-Jefri M, Evans R, Uchyigit G et al. What is health information quality? Ethical dimension and perception by users. *Front Med (Lausanne)* 2018; 5: 260.

21. Tao D, LeRouge C, Smith KJ et al. Defining information quality into health websites: a conceptual framework of health website information quality for educated young adults. *JMIR Human Factors* 2017; 4: e6455.
22. Sun Y, Zhang Y, Gwizdka J et al. Consumer evaluation of the quality of online health information: systematic literature review of relevant criteria and indicators. *J Med Internet Res* 2019; 21: e12522.
23. Zhang Y, Sun Y and Xie B. Quality of health information for consumers on the web: a systematic review of indicators, criteria, tools, and evaluation results. *J Assoc Inf Sci Technol* 2015; 66: 2071–2084.
24. Kinkead L, Allam A and Krauthammer M. Autodiscern: rating the quality of online health information with hierarchical encoder attention-based neural networks. *BMC Med Inform Decis Mak* 2020; 20: 1–13.
25. Risk A, Dzenowagis J et al. Review of internet health information quality initiatives. *J Med Internet Res* 2001; 3: e848.
26. Al-Jefri MM, Evans R, Ghezzi P et al. Using machine learning for automatic identification of evidence-based health information on the web. In *Proceedings of the 2017 International Conference on Digital Health*. pp.167–174.
27. Meppelink CS, Hendriks H, Trilling D et al. Reliable or not? An automated classification of webpages about early childhood vaccination using supervised machine learning. *Patient Educ Couns* 2021; 104: 1460–1466.
28. Samuel H and Zaïane O. Medfact: towards improving veracity of medical information in social media using applied machine learning. In *Advances in Artificial Intelligence: 31st Canadian Conference on Artificial Intelligence, Canadian AI 2018, Toronto, ON, Canada, May 8–11, 2018, Proceedings 31*. Springer, pp.108–120.
29. Bal R, Sinha S, Dutta S et al. Analysing the extent of misinformation in cancer related tweets. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14, pp.924–928.
30. Ruder S, Peters ME, Swayamdipta S et al. Transfer learning in natural language processing. In *Proceedings of the 2019 conference of the North American chapter of the Association for computational linguistics: Tutorials*. pp.15–18.
31. Vaswani A, Shazeer N, Parmar N et al. Attention is all you need. *Adv Neural Inf Process Syst* 2017; 30: 15–18.
32. Luong MT, Pham H and Manning CD. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:150804025* 2015.
33. Wolf T, Debut L, Sanh V et al. Huggingface's transformers: state-of-the-art natural language processing, 2020. 1910.03771.
34. Robillard JM, Alhothali A, Varma S et al. Intelligent and affectively aligned evaluation of online health information for older adults. In *AAAI Workshops*. pp.15–18.
35. Oroszlányová M, Teixeira Lopes C, Nunes S et al. Predicting the quality of health web documents using their characteristics. *Online Inform Rev* 2018; 42: 1024–1047.
36. Alasmari A, Alhothali A and Allinjawi A. Hybrid machine learning approach for Arabic medical web page credibility assessment. *Health Informatics J* 2022; 28: 14604582211070998.
37. Di Sotto S and Viviani M. Health misinformation detection in the social web: an overview and a data science approach. *Int J Environ Res Public Health* 2022; 19: 2173.
38. Upadhyay R, Pasi G and Viviani M. Health misinformation detection in web content: a structural-, content-based, and context-aware approach based on web2vec. In *Proceedings of the conference on information technology for social good*. pp.19–24.
39. Upadhyay R, Pasi G and Viviani M. Vec4cred: a model for health misinformation detection in web pages. *Multimed Tools Appl* 2022; 82: 5271–5290.
40. Page MJ, McKenzie JE, Bossuyt PM et al. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *Int J Surg* 2021; 88: 105906.
41. Kitchenham B, Charters S et al. Guidelines for performing systematic literature reviews in software engineering, 2007.
42. Ouzzani M, Hammady H, Fedorowicz Z et al. Rayyan—a web and mobile app for systematic reviews. *Syst Rev* 2016; 5: 1–10.
43. Brennan P and Silman A. Statistical methods for assessing observer variability in clinical measures. *BMJ: Brit Med J* 1992; 304: 1491.
44. Nidhra S, Yanamadala M, Afzal W et al. Knowledge transfer challenges and mitigation strategies in global software development—a systematic literature review and industrial validation. *Int J Inf Manage* 2013; 33: 333–355.
45. Al-Rawashdeh M, Keikhosrokiani P, Belaton B et al. IoT adoption and application for smart healthcare: a systematic review. *Sensors* 2022; 22: 5377.
46. Goeuriot L, Suominen H, Kelly L et al. Overview of the clef ehealth evaluation lab 2020. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 11th International Conference of the CLEF Association, CLEF 2020, Thessaloniki, Greece, September 22–25, 2020, Proceedings 11*. Springer, pp.255–271.
47. Robillard JM and Feng TL. Health advice in a digital world: quality and content of online information about the prevention of Alzheimer's disease. *J Alzheimers Dis* 2017; 55: 219–229.
48. Kilgarriff A, Baisa V, Bušta J et al. The sketch engine: ten years on. *Lexicography* 2014; 1: 7–36.
49. Choudhary A and Arora A. Linguistic feature based learning model for fake news detection and classification. *Expert Syst Appl* 2021; 169: 114171.
50. Horne B and Adali S. This just in: fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. In *Proceedings of the international AAAI conference on web and social media*. 1, pp.759–766.
51. Markowitz DM and Hancock JT. Linguistic traces of a scientific fraud: the case of diederik stapel. *PLoS ONE* 2014; 9: e105937.
52. Pérez-Rosas V, Kleinberg B, Lefevre A et al. Automatic detection of fake news. *arXiv preprint arXiv:170807104* 2017.
53. Gupta A, Kumaraguru P, Castillo C et al. Tweetcred: real-time credibility assessment of content on twitter. In *Social Informatics: 6th International Conference, SocInfo 2014, Barcelona, Spain, November 11–13, 2014. Proceedings 6*. Springer, pp.228–243.

54. Choi W and Stvilia B. Web credibility assessment: conceptualization, operationalization, variability, and models. *J Assoc Inf Sci Technol* 2015; 66: 2399–2414.
55. Hong T. The influence of structural and message features on web site credibility. *J Am Soc Inf Sci Technol* 2006; 57: 114–127.
56. Rieh SY and Belkin N. Interaction on the web: scholars' judgement of information quality and cognitive authority. In *Proceedings of the annual meeting-American society for information science*, Vol. 37. Citeseer, pp. 25–38.
57. Campos R, Mangaravite V, Pasquali A et al. Yake! Keyword extraction from single documents using multiple local features. *Inf Sci (Ny)* 2020; 509: 257–289.
58. Mihalcea R and Tarau P. Textrank: bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*. pp.404–411.
59. Zerrouki T. pyarabic, an Arabic language library for Python. https://pypi.python.org/pypi/pyarabic,year=2010.
60. Richardson L. Beautiful soup documentation, 2007.
61. Minaee S, Kalchbrenner N, Cambria E et al. Deep learning-based text classification: a comprehensive review. *ACM Comput Surv (CSUR)* 2021; 54: 1–40.
62. Lee J, Yoon W, Kim S et al. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 2020; 36: 1234–1240.
63. Yunianto I, Permanasari AE and Widyawan W. Domain-specific contextualized embedding: a systematic literature review. In *2020 12th International Conference on Information Technology and Electrical Engineering (ICITEE)*. IEEE, pp.162–167.
64. Morency LP, Liang PP and Zadeh A. Tutorial on multimodal machine learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorial Abstracts*. pp.33–38.
65. Baltrušaitis T, Ahuja C and Morency LP. Multimodal machine learning: a survey and taxonomy. *IEEE Trans Pattern Anal Mach Intell* 2018; 41: 423–443.
66. Feng J, Zou L, Ye O et al. Web2vec: phishing webpage detection method based on multidimensional features driven by deep learning. *IEEE Access* 2020; 8: 221214.
67. Soliman AB, Eissa K and El-Beltagy SR. Aravec: a set of Arabic word embedding models for use in Arabic nlp. *Procedia Comput Sci* 2017; 117: 256–265.
68. Mikolov T, Chen K, Corrado G et al. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:13013781* 2013.
69. Devlin J, Chang MW, Lee K et al. Bert: pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:181004805* 2018.
70. Pennington J, Socher R and Manning CD. Glove: global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. pp.1532–1543.
71. Sa Knight and J Burn. Developing a framework for assessing information quality on the world wide web. *Inform Sci* 2005; 8: 162–164.
72. Ng A. *Machine learning yearning: technical strategy for AI engineers in the era of deep learning*. deeplearning.ai, 2019.
73. Pan X, Lin Y and He C. A review of cognitive models in human reliability analysis. *Qual Reliab Eng Int* 2017; 33: 1299–1316.
74. Sørensen K, Pelikan JM, Röthlin F et al. Health literacy in Europe: comparative results of the European health literacy survey (HLS-EU). *Eur J Public Health* 2015; 25: 1053–1058.
75. Renggli C, Rimanic L, Hollenstein N et al. Evaluating Bayes error estimators on real-world datasets with feebee. *arXiv preprint arXiv:210813034* 2021.
76. Bal R, Sinha S, Dutta S et al. Analysing the extent of misinformation in cancer-related tweets. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. pp. 924–928.
77. Eysenbach G, Powell J, Kuss O et al. Empirical studies assessing the quality of health information for consumers on the world wide web: a systematic review. *Jama* 2002; 287: 2691–2700.
78. Maki A, Evans R and Ghezzi P. Bad news: analysis of the quality of information on influenza prevention returned by google in English and Italian. *Front Immunol* 2015; 6: 616.
79. Yaqub M and Ghezzi P. Adding dimensions to the analysis of the quality of health information of websites returned by Google: cluster analysis identifies patterns of websites according to their classification and the type of intervention described. *Front Public Health* 2015; 3: 204.
80. Schwarz J and Morris M. Augmenting web pages and search results to support credibility assessment. In *Proceedings of the SIGCHI conference on human factors in computing systems*. pp.1245–1254.
81. Sondhi P, Vydiswaran VV and Zhai C. Reliability prediction of webpages in the medical domain. In *ECIR*, Vol. 12. Springer, pp.219–231.
82. Allam A, Schulz PJ and Krauthammer M. Toward automated assessment of health web page quality using the discern instrument. *J Am Med Inform Assoc* 2017; 24: 481–487.
83. Cui L and Lee D. Coaid: COVID-19 healthcare misinformation dataset. *arXiv preprint arXiv:200600885* 2020.
84. Zhou X, Mulay A, Ferrara E et al. Recovery: a multimodal repository for COVID-19 news credibility research. In *Proceedings of the 29th ACM international conference on information & knowledge management*. pp.3205–3212.
85. Dai E, Sun Y and Wang S. Ginger cannot cure cancer: battling fake health news with a comprehensive data repository. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. pp.853–862.
86. Lopes CT and Ribeiro C. Measuring the value of health query translation: an analysis by user language proficiency. *J Am Soc Inf Sci Technol* 2013; 64: 951–963.
87. Stratton SJ. Population research: convenience sampling strategies. *Prehosp Disaster Med* 2021; 36: 373–374.
88. Etikan I, Musa SA, Alkassim RS et al. Comparison of convenience sampling and purposive sampling. *Am J Theor Appl Stat* 2016; 5: 1–4.
89. Sezerer E and Tekir S. A survey on neural word embeddings. *arXiv preprint arXiv:211001804* 2021.

90. Almeida F and Xexéo G. Word embeddings: a survey. *arXiv preprint arXiv:190109069* 2019.
91. Khattak FK, Jeblee S, Pou-Prom C et al. A survey of word embeddings for clinical text. *J Biomed Inform* 2019; 100: 100057.
92. Pastore R and Scheirer C. Signal detection theory: considerations for general application. *Psychol Bull* 1974; 81: 945.
93. Maguire H. Book review: data quality: concepts, methodologies and techniques by C. Batini and M. Scannapieco. *Int J Inform Qual* 2007; 1: 444–450.
94. Jasem Z, AlMeraj Z and Alhuwail D. Evaluating breast cancer websites targeting arabic speakers: empirical investigation of popularity, availability, accessibility, readability, and quality. *BMC Med Inform Decis Mak* 2022; 22: 126.
95. Leung JY, Ni Riordain R and Porter S. Readability and quality of online information regarding dental treatment for patients with ischaemic heart disease. *Br Dent J* 2020; 228: 609–614.
96. Ölçer D and Taşkaya Temizel T. Quality assessment of web-based information on type 2 diabetes. *Online Inform Rev* 2022; 46: 715–732.
97. Hornik K. Approximation capabilities of multilayer feedforward networks. *Neural Netw* 1991; 4: 251–257.
98. LeCun Y and Misra I. Self-supervised learning: the dark matter of intelligence. *Meta AI* 2021; 23: 3–4.
99. Dai AM and Le QV. Semi-supervised sequence learning. *Adv Neural Inf Process Syst* 2015; 28: 2–8.
100. Peters ME, Neumann M, Iyyer M et al. Deep contextualized word representations. corr abs/1802.05365 (2018). *arXiv preprint arXiv:180205365* 2018.
101. Radford A, Narasimhan K, Salimans T et al. Improving language understanding with unsupervised learning. Technical report, OpenAI, 2018.
102. Howard J and Ruder S. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:180106146* 2018.
103. Antoun W, Baly F and Hajj H. Arabert: transformer-based model for Arabic language understanding. *arXiv preprint arXiv:200300104* 2020.
104. Kenton JDMWC and Toutanova LK. Bert: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, Vol. 1. p. 2.
105. da Costa LS, Oliveira IL and Fileto R. Text classification using embeddings: a survey. *Knowl Inf Syst* 2023; 65: 2761–2803.
106. Alammary AS. Bert models for Arabic text classification: a systematic review. *Appl Sci* 2022; 12: 5720.