Original Paper

Visualizing the Interpretation of a Criteria-Driven System That Automatically Evaluates the Quality of Health News: Exploratory Study of 2 Approaches

Xiaoyu Liu[1,2], MBA, PhD; Hiba Alsghaier[1], MSc; Ling Tong[3], BSc; Amna Ataullah[1]; Susan McRoy[1], PhD

[1]Department of Computer Science, University of Wisconsin Milwaukee, Milwaukee, WI, United States

[2]School of Health Sciences, Southern Illinois University Carbondale, Carbondale, IL, United States

[3]Department of Health Informatics and Administration, University of Wisconsin Milwaukee, Milwaukee, WI, United States

Corresponding Author:
Susan McRoy, PhD
Department of Computer Science
University of Wisconsin Milwaukee
Engineering and Mathematical Sciences Bldg 1275
3200 N Cramer St
Milwaukee, WI, 53211
United States
Phone: 1 414 229 6695
Email: mcroy@uwm.edu

Abstract

Background: Machine learning techniques have been shown to be efficient in identifying health misinformation, but the results may not be trusted unless they can be justified in a way that is understandable.

Objective: This study aimed to provide a new criteria-based system to assess and justify health news quality. Using a subset of an existing set of criteria, this study compared the feasibility of 2 alternative methods for adding interpretability. Both methods used classification and highlighting to visualize sentence-level evidence.

Methods: A total of 3 out of 10 well-established criteria were chosen for experimentation, namely whether the health news discussed the costs of the intervention (the cost criterion), explained or quantified the harms of the intervention (the harm criterion), and identified the conflicts of interest (the conflict criterion). The first step of the experiment was to automate the evaluation of the 3 criteria by developing a sentence-level classifier. We tested Logistic Regression, Naive Bayes, Support Vector Machine, and Random Forest algorithms. Next, we compared the 2 visualization approaches. For the first approach, we calculated word feature weights, which explained how classification models distill keywords that contribute to the prediction; then, using the local interpretable model-agnostic explanation framework, we selected keywords associated with the classified criterion at the document level; and finally, the system selected and highlighted sentences with keywords. For the second approach, we extracted sentences that provided evidence to support the evaluation result from 100 health news articles; based on these results, we trained a typology classification model at the sentence level; and then, the system highlighted a positive sentence instance for the result justification. The number of sentences to highlight was determined by a preset threshold empirically determined using the average accuracy.

Results: The automatic evaluation of health news on the cost, harm, and conflict criteria achieved average area under the curve scores of 0.88, 0.76, and 0.73, respectively, after 50 repetitions of 10-fold cross-validation. We found that both approaches could successfully visualize the interpretation of the system but that the performance of the 2 approaches varied by criterion and highlighting the accuracy decreased as the number of highlighted sentences increased. When the threshold accuracy was ≥75%, this resulted in a visualization with a variable length ranging from 1 to 6 sentences.

Conclusions: We provided 2 approaches to interpret criteria-based health news evaluation models tested on 3 criteria. This method incorporated rule-based and statistical machine learning approaches. The results suggested that one might visually interpret an automatic criterion-based health news quality evaluation successfully using either approach; however, larger differences may arise when multiple quality-related criteria are considered. This study can increase public trust in computerized health information evaluation.

(JMIR AI 2022;1(1):e37751) doi: 10.2196/37751

KEYWORDS

health misinformation; machine learning; local interpretable model-agnostic explanation; LIME; interpretable artificial intelligence; AI

Introduction

Background

The internet has grown in popularity as a source for the public to learn about their health and investigate potential treatments for their health conditions. It is estimated that 80% of internet users consult web-based health information before making decisions [1]. Web-based media outlets such as social media feeds, forum threads, blogs, and newspapers have made information access and sharing easier. However, this has also accelerated the propagation of misleading information. Misinformation about health has been detected on different social media sites, such as Twitter [2-5], Facebook [6-9], YouTube [10-13], Pinterest [14,15], and Weibo [16,17]. Waszak et al [18] found that 40% of the most frequently shared links on social media contained medical information related to the most common diseases and causes of death were classified as fake news. In addition, the spread of health-related misinformation is not confined by geography. A series of studies have reported and studied health misinformation in different geographic settings, such as in the United States [19-21], China [16,17,22,23], India [24], and Italy [25,26]. With the rise of seeking health information on the internet, the concerns and health-related harm cases regarding misinformation have increased [27-29].

Unlike other types of misinformation, health-related misleading information, especially claims of efficacy about health interventions, such as medical treatments, tests, products, or procedures, can cause immediate actual harm to real people. The public and patients may be misled into making bad decisions that could result in severe consequences regarding people's quality of life and even the risk of mortality. This negative influence has been observed in many countries worldwide, despite cultural, regulatory, and geographic variances [30]. When the COVID-19 pandemic started in 2019, health misinformation was further exacerbated globally as more people increasingly turned to social media to confirm possible symptoms and share treatment plans [31]. Misleading and erroneous information, information of low quality such as conspiracy theories, poorly sourced medical advice, and information trivializing the virus has not only contributed to widespread misconceptions about the novel coronavirus but also caused public panic, catastrophic consequences of public health, and even people's distrust in public health institutions at the global level [32,33].

To address this public health crisis, continuing efforts to counteract health misinformation are being made across a wide range of disciplines and organizations. Detection and fact-checking work that relies on human effort is limited in scope, considering the high volume of fake news generated on the internet. Many attempts have been made to leverage artificial intelligence (AI) to analyze enormous amounts of information generated daily on a scale that would be impossible for humans to handle [34]. AI-powered automated detection methods, in comparison with people, are faster, more efficient, and may be deployed on targeted platforms at a low cost and on larger scale, by replicating human intelligence using data-driven analysis by computers [35]. When combating misinformation, AI technology may distinguish between accurate and misleading information using terms or word patterns associated with misinformation as cues from a relatively small set of articles that have been previously annotated by experts. Therefore, AI techniques can automate the process of detection of misleading information, which is conventionally performed manually.

Related Work

In recent years, there has been an increasing trend in AI-based studies attempting to address health misinformation. The choice of health topic is a critical factor to consider, as it requires domain understanding and knowledge to assess the quality of health information and confirm the presence of misinformation. Health topics incorporated in past misinformation detection studies either focused on a specific topic, such as vaccination [36-38], Zika [39], autism [40], COVID-19 [41-44], or a collection of miscellaneous health conditions and lifestyle choices [45-50]. Health misinformation resides in various information outlets. Existing studies have proposed the detection of false, misleading health news on platforms such as Twitter [37,39,51,52], websites [36,45,46], and web-based forums [48,49].

Setting an appropriate benchmark for evaluating and annotating health information is unavoidable when developing detection systems. On the basis of the benchmark and objectives of this study, previous work on misinformation classification can be briefly categorized into a veracity-based approach or a criteria-based approach. Studies that follow a veracity-based approach involved training classifiers to assess the truth of each health-related claim using data that have been annotated to indicate whether the claim can be validated or refuted by finding a similar statement using a trusted source. These supporting sources might be experts from a third-party fact-checking organization (eg, Snopes [53]), medical and health-related professional organizations (eg, World Health Organization [54]), academic or research institutions (eg, John Hopkins Medicine [55]), and the federal government (eg, CDC [56]) which are typically considered as the officially sanctioned sources of bona fide accurate information and play an active role in myth debunking. For example, Ghenai and Mejova [39] proposed a novel pipeline that combines health experts, crowdsourcing, and machine learning (ML) to capture rumors on Twitter. The model was created using 13 million tweets concerning Zika infection between February 2016 and the Summer Olympics and rumors outlined by the World Health Organization and Snopes. The study found that rumor-related topics have a particularly burst behavior. The results demonstrated the feasibility of using automated techniques to remove rumor-bearing tweets when a questionable topic was detected.

In contrast, studies that followed the criteria-based approach looked at misinformation based on various quality-indicating criteria predefined by research. An example of such criteria might be the reliability or unreliability of the source; the rationale is that *intuitively, a news article published on an unreliable website and forwarded by unreliable users is more likely to be fake news than news posted by authoritative and credible users* [57]. For example, Liu et al [50] predefined a list of reliable and unreliable websites from which health-related articles from various sources on the Chinese Internet society were extracted for data set construction. Experiments were performed based on various ML classifiers using manually extracted features and text-classification modeling. The best performance among all models reached a precision of 0.8374. Other approaches were based on the idea that news that does not satisfy certain items on an assessment checklist for health information quality can be considered untrustworthy. For instance, Shah et al [37] used a 7-point checklist adapted from 2 validated tools, the DISCERN and Quality Index for health-related Media Reports checklists, to manually appraise the credibility of 474 web pages after sampling from 143,003 unique vaccine-related web pages shared on Twitter between January 2017 and March 2018. According to previous studies, the best-performing classifiers could distinguish between low, medium, and high credibility with an accuracy of 78% and labeled low-credibility web pages with a precision of >96%. Al-Jefri et al [58] and Afsana et al [59] both developed 10 classifiers to automatically evaluate the quality of health news based on the criteria developed by HealthNewsReview.org. However, the latter’s models demonstrated better classification performance owing to the inclusion of more features. In summary, veracity-based studies examined the authenticity of the news. The criteria-based approach focused on the characteristics of the news content, but the results did not make claims about the veracity of information.

In addition to the wide range of themes and strategies in detecting misinformation identified in the literature, methodologically, current studies also show the effectiveness of AI-based algorithms in classifying misinformation and quality information. Traditional ML algorithms, including Logistic Regression [40,47,52,60], support vector machine [37,40,47,50], decision tree [52,61], and random forest (RF) [37,39,41,48,60] have been widely applied in these studies, yielding effective and accurate performance. More recent studies have shown improved performance on large data sets by incorporating deep learning techniques, including convolutional neural networks [49,61], bidirectional encoder representations from transformers [43], and long short term memory [42,44,61]. As part of the modeling process, feature engineering has also been a critical step in improving the performance of classifiers. Zhao et al [57] reviewed and summarized 12 features used in health misinformation detection models. These features were grouped into 4 subsets: linguistic, topic, sentiment, and behavioral features.

Compared with traditional human fact-checking, an AI-based model consists of an algorithm that can automatically learn latent patterns and relationships from the data. However, one of the major challenges is the lack of a human-understandable rationale to support the results of classification tasks. Approaches that attempt to address this concern are often called “interpretable ML,” “explainable ML,” or “explainable AI” [62]. Open-source software with implementations of various interpretable ML methods are also available, such as local interpretable model-agnostic explanation (LIME) [63], Shapley Additive Explanations [64], Eli5 [65], and InterpretML [66], etc. These tools have been applied to various tasks, including image classification and text classification. With interpretations or visualized cues, users can verify the model and determine whether it meets their expectations. In addition, users can discover knowledge, justify predictions, and improve the performance of models using interpretable ML methods. Therefore, interpretable AI improves the trust and usability of the classifiers.

However, to date, only a small body of research has incorporated explainable AI models to combat health misinformation [43,67]. All of these studies on health information classification were veracity-based. A knowledge gap remains regarding the effectiveness of constructing an interpretable, criteria-driven classification system to help users evaluate the quality of health information.

Objective of This Study

We aimed to address the aforementioned concerns and needs by creating an interpretable, criteria-driven system to assist the public in evaluating the quality of health news to mitigate the adverse consequences of health misinformation. Previous work using the HealthNewsReview.org data set and ML classifiers at the document level found that 3 criteria (cost, harm, and conflict) are more accurately classifiable among the 10 criteria, using linguistic features [58,68]; therefore, we selected these 3 criteria for this exploratory study. Our study, because it addressed interpretability, also focused on the use of features that are directly visualizable (linguistic features), excluding less visualizable features (such as average sentence length), which sometimes improved classification accuracy.

As an exploratory study, we opted to test 2 possible interpretation approaches, using 3 criteria. The evaluation results for the criteria will be visually explained with highlighted sentences as cues to enhance interpretability and reliability. As the number of highlighted sentences may affect the overall visual representation and effectiveness of the interpretation, we also attempted to determine the ideal range for the number of highlighted sentences.

Methods

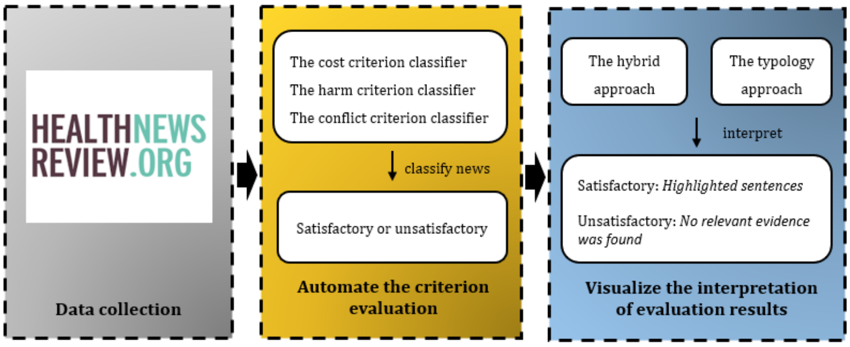
Overview

The experiment consisted of 3 components, as illustrated in Figure 1. In the first component, we collected reviewed health news from HealthNewsReview.org [69] to build the data set for modeling. Each criterion review result provided by HealthReview.org was treated as a classification target. The second component was a supervised document classification task that automated the criteria evaluation process. Each health news article was categorized automatically at the document level using established criteria and the output was binary (satisfactory or unsatisfactory). “Satisfactory” meant the entire health news met the given criterion and “unsatisfactory” meant the opposite.

The last component visualized and interpreted the evaluation results provided by the health news quality-evaluation system. For example, for the criterion “Does the news adequately explain or quantify the harms of the intervention?” the method highlighted sentences that described the harms of intervention to help users quickly understand how well the criterion was met.

We examined 2 approaches to achieve this goal. The first was a hybrid approach (the hybrid approach). It was inspired by principles from rule-based systems, where patterns are cospecified by LIME and experts. The second approach (the typology approach) was a supervised sentence typology classification method, where hand-labeled training data are analyzed algorithmically to build models that can detect similar patterns when applied to unseen data.

Figure 1. Overview of the exploratory experiment.


HEALTHNEWS REVIEW.ORG
Data collection
The cost criterion classifier
The harm criterion classifier
The conflict criterion classifier
classify news
Satisfactory or unsatisfactory
Automate the criterion evaluation
The hybrid approach
The typology approach
interpret
Satisfactory: Highlighted sentences
Unsatisfactory: No relevant evidence was found
Visualize the interpretation of evaluation results

Data Description and Collection

The data set that we used was adapted from an existing resource created by HealthNewsReview.org [69]. HealthNewsReview.org is a web-based project that reviewed articles from 2005 to 2018. Their team of experts rated the claims about health care interventions to improve the quality of health care information. Their rating instrument included 10 criteria used by the Australian and Canadian Media Doctor sites, and its interreviewer reliability was tested using a random sample of 30 stories [70]. HealthNewsReview.org included reviews of news stories from leading US media and news releases from institutes. The contents included efficacy claims about specific treatments, tests, products, or procedures. The news pieces were assessed using a standard rating system. At least 2 reviewers reviewed each news story. The reviewers were selected based on their years of experience in the health domain, spanning the fields of journalism, medicine, health services research, public health, or as patients, and each of them signed an industry-independent disclosure agreement. For each news story or news release reviewed, the criteria were scored as “satisfactory,” “unsatisfactory,” or “not applicable.” Total scores were posted for articles with ≤2 “not applicable” ratings and were expressed as proportions. It was acknowledged that increasing the diversity and independence of the reviewers could have reduced the potential for bias in the assessments. By the time the project ended, the website had accumulated 2616 health story reviews and 606 news release reviews.

For this study, we crawled health story news reviews and news release reviews, as archived by HealthNewsReviews.org, complying with the robots.txt. We scraped news contents that corresponded to the acquired reviews. Then, we visualized the results for the three selected criteria: (1) “Does the news adequately discuss the costs of the intervention?” (the cost criterion), (2) “Does the news adequately explain or quantify the harms of the intervention?” (the harm criterion), and (3) **“**Does the news identify conflicts of interest?” (the conflict criterion).

Automating the Criterion Evaluation

All 3 criteria applied to both news types, so we merged the 2 types of news content and treated them uniformly. We also combined health news that was scored as “unsatisfactory” or “not applicable” and named them as “unsatisfactory.” We preprocessed all news content via multiple text processing techniques, including removal of nonword elements (numbers, assented characters, and punctuation) and stop words, tokenization, stemming, and lemmatization. Then, we converted the textual representation into a vector space model using term frequency–inverse document frequency (TF-IDF).

We chose 4 representative algorithms: logistic regression, naive Bayes, support vector machine, and RF, from which we selected the best base algorithm that was suitable for automating the criterion evaluation. The 4 algorithms are commonly used in health misinformation classification tasks, as evident in previous studies [36,38,39,46,51,59], and were found to be effective. We applied RandomSearch to determine the optimal model hyperparameters for building the classifier. For our study, we defined the best classifier output from RandomSearch as the feature count, hyperparameter, and algorithm combination that produced the highest mean 5–cross-validated area under the curve (AUC) score. The performance of the classifier was further evaluated through 50-repeated 10–fold-validation.

Visualizing the Interpretation of Evaluation Result

We experimented with 2 approaches to visualize the interpretation of the evaluation results. The desired outcome was that all highlighted sentences were relevant to the examined criterion and provided evidence to assist end users in comprehending and validating the evaluation results. To determine what qualified a sentence as evidence, we strictly adhered to the criteria definitions and review guidelines provided by HealthNewsReview.org [71-73]. For example, as per the explanation of the harm criterion provided by HealthNewsReview.org, satisfactory health news on the harm criterion should “include a discussion of harms and side effects, as well any measured ‘adverse events’ in a study” [71]. The measured “adverse events” can be addressed by a discussion of “both frequency of side effects and severity of side effects” and a discussion of “both major and minor side effects” [71].

The Hybrid Approach

The hybrid approach combined the interpretable AI technique, LIME, rule-based systems, and supervised document classification. LIME, proposed in 2016 by Ribeiro et al [74], belonged to a family of local model-agnostic methods, a type of interpretable AI method. It is used to explain the individual predictions of black-box ML based on a surrogate model, which is trained to approximate the predictions of the underlying black-box model [74,75]. The intuition of LIME is based on the idea that the behavior of a black-box model can be learned by perturbing the input. Specifically, a modified data set is generated by LIME through permutation by removing word features, corresponding to which predictions are obtained from the black-box model. Words with feature weights >0 indicate that the removal of such words affects the prediction result. For a negative case, no nonzero weight was estimated because regardless of which word was removed, the predicted evaluation result remained the same. Thus, an explanation can be generated by approximating the underlying model with a more interpretable model (such as a linear model or decision tree), learned locally on perturbations of the original instance [75]. Owing to the local fidelity nature of LIME, it does not guarantee a good global approximation [76]. A critique LIME often receives is that it lacks “stability” [77]. There are cases in which the surrogate model built by LIME can predict the instance correctly but provide incorrect reasons [75]. To address the instability of LIME, adding manually selected keywords can reduce the risk of obtaining incorrect keywords for highlighting. In this approach, we adopted the LIME method to facilitate the interpretable result of the predicted criterion evaluation. The Python packages used for implementing LIME algorithms were ELI5 [65] and LIME [63] application programming interface packages.

The explanation of the classification model for each criterion using the hybrid approach consisted of 3 steps. First, an ML classifier classifies health news as satisfactory or unsatisfactory based on the chosen criterion. Then, the classification model learned the difference of word distribution in satisfactory or unsatisfactory instances from the collection of health news document sets. LIME highlighted keywords in texts that contributed to the prediction. The keywords were also ranked using a weighted score, indicating their contribution to the prediction. Finally, we combined the keywords that contributed to a satisfactory prediction with a list of manually selected keywords, as shown in Table 1. The manual selection of the keywords was based on a consensus among the annotators who had taken part in the processes of evidence extraction for the typology approach.

We then extended the highlighting from the keyword to the sentence level to enhance the final visual representation. Sentences containing keywords with more weight were prioritized for highlighting. By default, manually selected keywords outweighed any keywords automatically picked by LIME.

Table 1. Lists of manually selected keywords for the cost, harm, and conflict criteria.

Criterion	Manually selected keywords
The cost criterion	Price, cost, charge, insurance, and pay
The harm criterion	Side effect, adverse reaction, adverse event, complication, and risk
The conflict criterion	Fund, sponsor, grant, spokesman, professor, and director

The Typology Approach

The typology approach was a sentence-level text-classification task. This approach was inspired by the study of persuasive communication and rhetoric. Reynolds and Reynolds [78] distinguished between statistical, testimonial, anecdotal, and analogical evidence. Hoeken and Hustinx [79] put forward 4 types of evidence in argumentation: individual examples, statistics, causal explanations, and expert opinions. Subsequent studies showed that machines can detect various types of evidence. For example, Fiok et al [80] built a classification model to automatically identify the evidence of respect in Twitter communication. There were 2 types of sentences in each health news item in our study. In the harm criterion, the first type of sentences was the evidence that supported the predicted evaluation result. Sentences of this type contained a description of side effects, including the symptoms, severity, and frequency of the symptoms. The second type of sentence referred to those that could not justify why a piece of certain health news satisfied a given criterion. Therefore, they were not characterized as evidence.

To implement the typology approach, for each criterion task, we designed and experimented with the typology approach in 2 stages. The first stage was to build an annotated data set of the sentence evidence. We extracted sentence evidence from health news that was evaluated as satisfactory by HealthNewsReview.org. A total of 3 people performed the sentence extraction tasks. The project investigator provided training and clarification to the other 2 extractors. The sentence extraction guideline fully adopted the criteria explained by HealthNewsReivew.org [71-73]. Two people performed most of the extraction work. Another individual worked as an independent reviewer to resolve disagreement. When combining the extracted sentences, sentences picked by the 2 extractors were characterized as evidence. If a sentence was extracted by an extractor but not picked by the other, an independent third person was invited to resolve the disagreement. All approved sentences were considered positive. To build the negative class, we randomly selected the same number of sentences irrelevant to the evaluation of the pertinent criterion. An interannotator agreement was assessed using both simple counts and the percentage of the final quantity of the evidence in the total extracted items to address the relatively small sample size. The interrater agreement was also in line with the expectations of other studies [81]. The second stage involved building a supervised ML classifier. We followed the same steps as for automating the criterion evaluation.

For the final visual representation, the sentence classifier was applied to health news content to identify sentence evidence. Sentences with a higher probability of being categorized as evidence by the classifier were prioritized for highlighting purposes.

Evaluating and Optimizing the 2 Approaches

For each criterion's interpretation, we evaluated 2 visualization approaches to determine how accurately each scheme highlighted the sentences that supported the prediction result. The evaluation was conducted using 20 test cases. The selection of 20 test cases was based on the observation that the true positive health news counts in the test set (30% of the data set) ranged from 20 to 70, depending on the task criterion type. We measured the accuracy of 2 highlighting schemes by calculating the percentage of correctly highlighted evidence for all highlighted sentences. A total of 3 people evaluated the correctness of the highlighted sentence in accordance with each criterion's guideline. An independent reviewer was invited to handle any disputes.

As the number of highlighted sentences may affect the highlighting accuracy and thus the final visual representation, we calculated a spectrum of accuracies of both the highlighting approaches when the number of highlighted sentences increased from 1. A threshold was then selected with the lowest accuracy to determine the optimal range of sentence counts for highlighting.

Results

Classification Model Performance

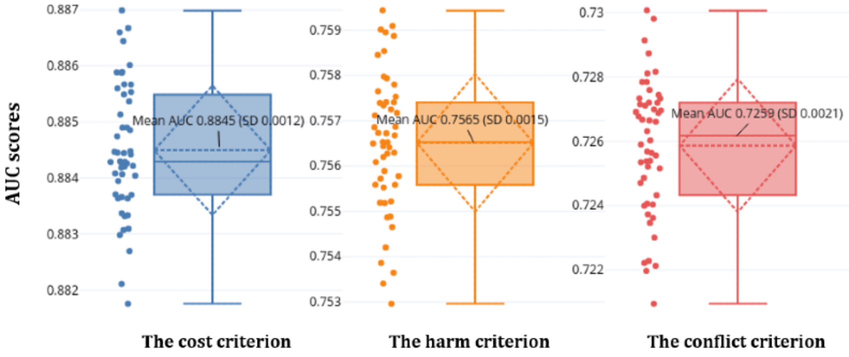
After removing dead links (to inaccessible news content), the acquired data set yielded 1453 stories and 579 news releases. Among the 2032 health news instances, the satisfactory or unsatisfactory instance ratios for the cost, harm, and conflict criteria were 25.03% (405/1618), 44.71% (625/1398), and 98.14% (1002/1021), respectively. Of the 4 experimental algorithms, RF was found to be the most effective in automating the evaluation of all the 3 criteria, as shown in Multimedia Appendix 1, despite the fact that the feature count varied according to the criterion. Table 2 shows the set of optimal hypermeters that RandomSearch selected for each criterion classifier.

For the cost, harm, and conflict criteria, Figure 2 shows that the average AUCs were 0.8845, 0.7565, and 0.7259, respectively, after 50 repeated 10-fold validations.

Table 2. Hyperparameters selected by RandomSearch for each criterion evaluation classifier.

Criteria	Base classifier	Word feature count, n	Hyperparameters
The cost criterion	Random forest	1000	("n_estimators": 600, "min_samples_split": 2, "min_samples_leaf": 4, "max_features": "sqrt," "max_depth": 10, and "bootstrap": false)
The harm criterion	Random forest	2000	("n_estimators": 1400, "min_samples_split": 10, "min_samples_leaf": 4, "max_features": "auto," "max_depth": 90, "bootstrap": false)
The conflict criterion	Random forest	1000	("n_estimators": 1200, "min_samples_split": 10, "min_samples_leaf": 1, "max_features": "auto," "max_depth": 20, and "bootstrap": true)

Figure 2. The performance of the cost, harm, and conflict criterion classifiers was measured with 10-fold cross-validated area under the curve (AUC) scores with a total of 50 repetitions.



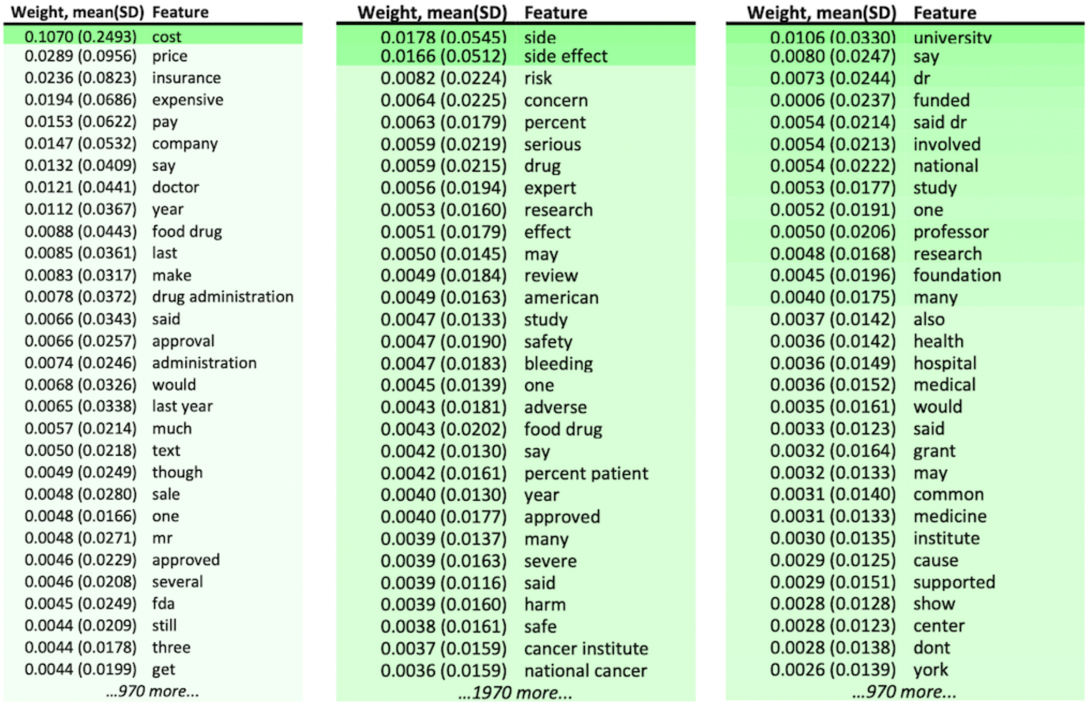
Interpretable Model Performance

The Visual Interpretation by the Hybrid Approach

The LIME Text Explainer visualized how different word features contributed to the evaluation results for each classifier. Figure 3 illustrates the top 30 bigram or unigram word features that contributed to the classification learned from the entire data set related to a given criterion. For example, the binary word feature with the highest weight in the harm criterion classification was “side effect.” Words that directly indicate the harm of intervention, such as “risk,” “concern,” “bleeding,” and “harm,” also ranked among the top features. Similarly, words that are commonly used to describe the intervention costs and insurance coverage such as “cost,” “insurance,” “expensive,” and “pay” were also observed high in contribution to the evaluation for the cost criterion. For the conflicts criterion, the words were descriptive of one’s affiliations such as “university,” “dr,” and “professor” stand out. The keyword “funded,” which directly discloses funding information, also ranked high.

Figure 4 shows how LIME performed first-level visualization on a sample health news that was rated as satisfactory on the harm criterion. The classifier predicted the sample health news with a positive result of 65% probability. The words marked in orange were picked by LIME and explained as they contributed to the positive classification results of the model. Certain words were also highlighted in blue despite being scarce in number, indicating the likelihood of an unsatisfactory prediction. On the basis of the prediction result, the words “adverse,” “reaction,” “risk,” “adverse,” “serious,” and “administration,” were ranked among the most predictive words in the satisfactory classification result. A snapshot of the final visualized representation is shown in Figure 5, after highlighting sentences containing the keywords selected by LIME and the human expert. The 2-level visual interpretation cases for the cost and conflict criteria can be found in the Multimedia Appendix 2.

Figure 3. Top 30 word features with their feature weights in 3 criteria (the cost, harm, and conflict criteria) classifiers. The word feature weights signify how much discriminatory information each word contributes to the classification task by random forest algorithm.

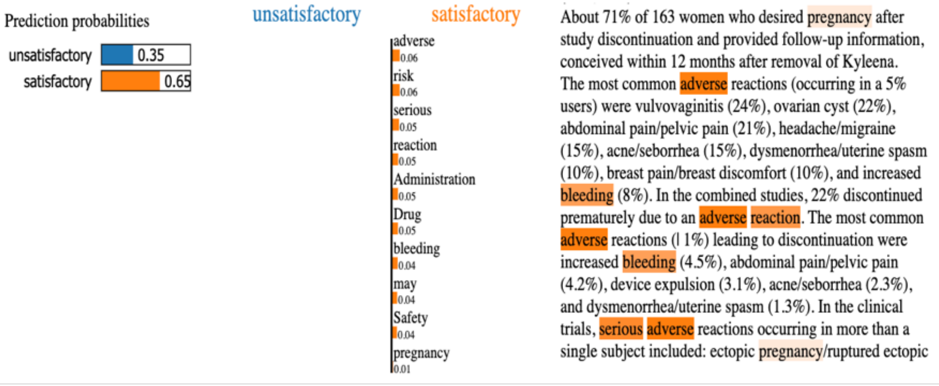


Weight, mean(SD)	Feature
0.1070 (0.2493)	cost
0.0289 (0.0956)	price
0.0236 (0.0823)	insurance
0.0194 (0.0686)	expensive
0.0153 (0.0622)	pay
0.0147 (0.0532)	company
0.0132 (0.0409)	say
0.0121 (0.0441)	doctor
0.0112 (0.0367)	year
0.0088 (0.0443)	food drug
0.0085 (0.0361)	last
0.0083 (0.0317)	make
0.0078 (0.0372)	drug administration
0.0066 (0.0343)	said
0.0066 (0.0257)	approval
0.0074 (0.0246)	administration
0.0068 (0.0326)	would
0.0065 (0.0338)	last year
0.0057 (0.0214)	much
0.0050 (0.0218)	text
0.0049 (0.0249)	though
0.0048 (0.0280)	sale
0.0048 (0.0166)	one
0.0048 (0.0271)	mr
0.0046 (0.0229)	approved
0.0046 (0.0208)	several
0.0045 (0.0249)	fda
0.0044 (0.0209)	still
0.0044 (0.0178)	three
0.0044 (0.0199)	get
...970 more...	

Weight, mean(SD)	Feature
0.0178 (0.0545)	side
0.0166 (0.0512)	side effect
0.0082 (0.0224)	risk
0.0064 (0.0225)	concern
0.0063 (0.0179)	percent
0.0059 (0.0219)	serious
0.0059 (0.0215)	drug
0.0056 (0.0194)	expert
0.0053 (0.0160)	research
0.0051 (0.0179)	effect
0.0050 (0.0145)	may
0.0049 (0.0184)	review
0.0049 (0.0163)	american
0.0047 (0.0133)	study
0.0047 (0.0190)	safety
0.0047 (0.0183)	bleeding
0.0045 (0.0139)	one
0.0043 (0.0181)	adverse
0.0043 (0.0202)	food drug
0.0042 (0.0130)	say
0.0042 (0.0161)	percent patient
0.0040 (0.0130)	year
0.0040 (0.0177)	approved
0.0039 (0.0137)	many
0.0039 (0.0163)	severe
0.0039 (0.0116)	said
0.0039 (0.0160)	harm
0.0038 (0.0161)	safe
0.0037 (0.0159)	cancer institute
0.0036 (0.0159)	national cancer
...1970 more...	

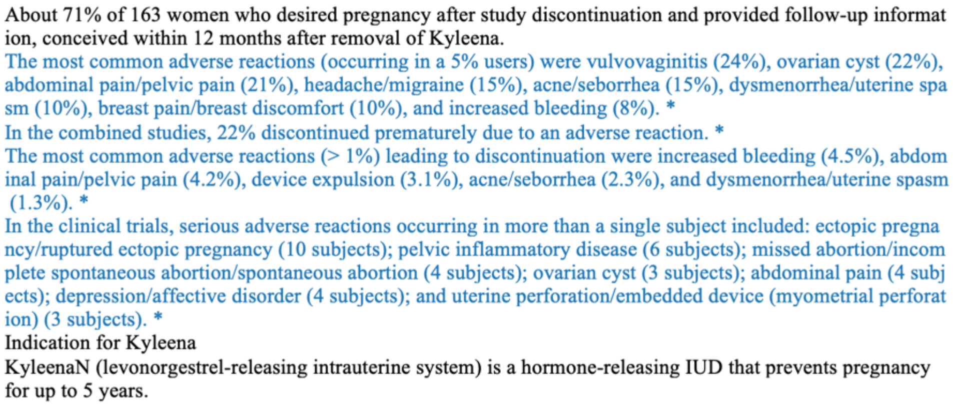
Weight, mean(SD)	Feature
0.0106 (0.0330)	university
0.0080 (0.0247)	say
0.0073 (0.0244)	dr
0.0006 (0.0237)	funded
0.0054 (0.0214)	said dr
0.0054 (0.0213)	involved
0.0054 (0.0222)	national
0.0053 (0.0177)	study
0.0052 (0.0191)	one
0.0050 (0.0206)	professor
0.0048 (0.0168)	research
0.0045 (0.0196)	foundation
0.0040 (0.0175)	many
0.0037 (0.0142)	also
0.0036 (0.0142)	health
0.0036 (0.0149)	hospital
0.0036 (0.0152)	medical
0.0035 (0.0161)	would
0.0033 (0.0123)	said
0.0032 (0.0164)	grant
0.0032 (0.0133)	may
0.0031 (0.0140)	common
0.0031 (0.0133)	medicine
0.0030 (0.0135)	institute
0.0029 (0.0125)	cause
0.0029 (0.0151)	supported
0.0028 (0.0128)	show
0.0028 (0.0123)	center
0.0028 (0.0138)	dont
0.0026 (0.0139)	york
...970 more...	

Figure 4. Lime text explainer visualizes word’s contribution to a satisfactory prediction on the harm criterion using random forest algorithm.



About 71% of 163 women who desired pregnancy after study discontinuation and provided follow-up information, conceived within 12 months after removal of Kyleena. The most common adverse reactions (occurring in a 5% users) were vulvovaginitis (24%), ovarian cyst (22%), abdominal pain/pelvic pain (21%), headache/migraine (15%), acne/seborrhea (15%), dysmenorrhea/uterine spasm (10%), breast pain/breast discomfort (10%), and increased bleeding (8%). In the combined studies, 22% discontinued prematurely due to an adverse reaction. The most common adverse reactions (| 1%) leading to discontinuation were increased bleeding (4.5%), abdominal pain/pelvic pain (4.2%), device expulsion (3.1%), acne/seborrhea (2.3%), and dysmenorrhea/uterine spasm (1.3%). In the clinical trials, serious adverse reactions occurring in more than a single subject included: ectopic pregnancy/ruptured ectopic

Figure 5. Example of a highlighting scheme for the harm criterion by the hybrid approach.

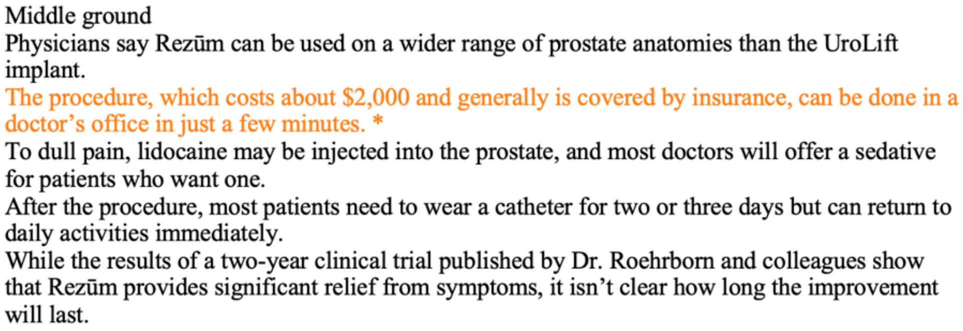


About 71% of 163 women who desired pregnancy after study discontinuation and provided follow-up informat
ion, conceived within 12 months after removal of Kyleena.
The most common adverse reactions (occurring in a 5% users) were vulvovaginitis (24%), ovarian cyst (22%),
abdominal pain/pelvic pain (21%), headache/migraine (15%), acne/seborrhea (15%), dysmenorrhea/uterine spa
sm (10%), breast pain/breast discomfort (10%), and increased bleeding (8%). *
In the combined studies, 22% discontinued prematurely due to an adverse reaction. *
The most common adverse reactions (> 1%) leading to discontinuation were increased bleeding (4.5%), abdom
inal pain/pelvic pain (4.2%), device expulsion (3.1%), acne/seborrhea (2.3%), and dysmenorrhea/uterine spasm
(1.3%). *
In the clinical trials, serious adverse reactions occurring in more than a single subject included: ectopic pregna
ncy/ruptured ectopic pregnancy (10 subjects); pelvic inflammatory disease (6 subjects); missed abortion/incom
plete spontaneous abortion/spontaneous abortion (4 subjects); ovarian cyst (3 subjects); abdominal pain (4 subj
ects); depression/affective disorder (4 subjects); and uterine perforation/embedded device (myometrial perforat
ion) (3 subjects). *
Indication for Kyleena
KyleenaN (levonorgestrel-releasing intrauterine system) is a hormone-releasing IUD that prevents pregnancy
for up to 5 years.

The Visual Interpretation by the Topology Approach

The interannotator agreement rates on evidence extraction for the cost, harm, and conflicts criteria were 72.04%, 72.24%, and 77.91%, respectively. The extraction task for each criterion yielded 201 (cost criterion), 318 (harm criterion), and 694 (conflict criterion) sentences in the positive class. We randomly selected the same number of sentences as the negative class to build the classification data sets. Following the same approach applied to the automation of criterion evaluation, which included base classifiers, word feature count selection, and hyperparameters tuning using RandomSearch, the classifiers of the 3 criteria attained an average AUC of 0.8791 (cost criterion), 0.7232 (harm criterion), and 0.8951 (conflict criterion) with 50 repetitions of 10–cross-fold validations. Figure 6 shows the result of applying the classifier to each sentence in the document and highlighting positive sentence instances that supported a cost criterion evaluation.

Figure 6. Example of a highlighting scheme for the cost criterion by the typology approach.



Middle ground
Physicians say Rezūm can be used on a wider range of prostate anatomies than the UroLift
implant.
The procedure, which costs about $2,000 and generally is covered by insurance, can be done in a
doctor's office in just a few minutes. *
To dull pain, lidocaine may be injected into the prostate, and most doctors will offer a sedative
for patients who want one.
After the procedure, most patients need to wear a catheter for two or three days but can return to
daily activities immediately.
While the results of a two-year clinical trial published by Dr. Roehrborn and colleagues show
that Rezūm provides significant relief from symptoms, it isn't clear how long the improvement
will last.

The Overall Performance and Optimization of the 2 Approaches

As the total number of highlighted sentences increased from 1, we calculated the varying rates of accurately highlighted sentences, as shown in Table 3. The numbers with footnotes suggest that the relevant approach could obtain a better result (accuracy >75%) within a certain number of sentences for highlighting.

According to Table 3, the accuracy of both approaches declined as the number of highlighted sentences increased. When both approaches highlighted the same number of sentences, the hybrid approach outperformed the typology approach in most scenarios. Typology, however, performed more accurately when the target was to pick <3 sentences to justify the harm criterion evaluation. When the threshold for highlighting accuracy was set at 75%, the optimal window size for the typology approach to achieve relatively better interpretation results was 2, 4, and 1 for the cost, harm, and conflict criteria, respectively. Comparatively, the hybrid approach still produced comparable outcomes when the window size for each criterion was extended by 2.

Table 3. The accuracy of both approaches for interpreting each criterion evaluation; maximum highlighting sentence count.

Number	Criterion and approach					
	Cost		Harm		Conflict	
	Typology (%)	Hybrid (%)	Typology (%)	Hybrid[a] (%)	Typology (%)	Hybrid (%)
1	80.00[a]	100.00[a]	100.00[a]	90.00	75.00[a]	90.00[a]
2	75.00[a]	92.50[a]	97.50[a]	90.00	72.50	87.50[a]
3	60.78	86.67[a]	86.67[a]	90.00	66.67	81.67[a]
4	66.67	76.25[a]	76.39[a]	85.00	61.11	68.75
5	60.00	67.00	72.94	81.00	55.29	59.00
6	54.55	59.65	66.67	75.83	56.41	50.93

[a] The relevant approach could obtain a better result (accuracy >75%) within a certain number of sentences for highlighting.

Discussion

Principal Findings

This study experimented with 2 AI-based approaches to visualize the interpretation of a criteria-based system designed to assist users in systematically evaluating the quality of health news.

The findings of our experiments were 3-fold. First, we found that both the hybrid and typology approaches could achieve the desired visualization result to justify the predicted evaluation result, despite the nature of the 2 approaches being differentiated. With 20 tests for each criterion, the performance of the hybrid approach was slightly better than that of the typology approach. Second, we were able to locate a window size to predetermine the sentences to be highlighted for a better visualization result for each criterion. The hybrid approach showed a higher capacity to reliably choose more sentences when the accuracy criterion was set at 75%. Third, the feasibility of the rule-based strategy to enhance LIME's interpretation work was supported by our observation during evidence extraction for the typology approach that specific words or phrases such as "adverse effect," "danger," "death," and "side effect" appeared repeatedly in the evaluation of the harm criterion; keywords such as "cost," "price," and "insurance" frequently appeared for the cost criterion evaluation; and "spokesman," "funding," and "sponsor" were typically used to disclose the conflicts of interests.

A Comparison of the 2 Approaches

The hybrid approach demonstrated both good accuracy and efficiency in visualizing the automatic model's interpretation for evaluating the 3 criteria. Compared with the typology approach, it was advantageous in saving manual effort because it did not require sentence extraction. We also observed that the hybrid approach tended to pick fewer sentences but with higher accuracy when not limiting the maximum number of sentences to be highlighted. By contrast, the typology approach selected more sentences, but only a few were relevant to the criterion.

However, the hybrid approach also had inherent weaknesses. The highlight scheme in the hybrid approach was to locate the sentence in which keywords were present. The drawback of this scheme was that it sometimes failed to discern the semantic differences between a sentence about the risk of the intervention and a sentence that described the benefits of the intervention by relieving or preventing adverse conditions. For example, in one of the test cases, the sentence, "Moreover, the study verified that long-term use of bisphosphonate drugs reduces the risk of typical osteoporosis fractures by 24 percent." was incorrectly highlighted. The sentence contained keywords, including "risk" and "fractures," which are relevant to adverse symptoms. However, it introduced how bisphosphonates are expected to benefit patients by decreasing the risk of negative outcomes. The other weakness associated with the hybrid approach was that it failed to distinguish between the intervention and stock prices. Both types of sentences typically shared many keywords that described the values associated with the intervention.

By contrast, the typology approach performed somewhat better at handling expressions with more lexical variations. For example, sentences, "Last fall the Food and Drug Administration issued a 'safety update' urging doctors and patients to be on the lookout for the problem." and "These medications are now linked to a growing number of complications, ranging in seriousness from nutrient deficiencies, joint pain and infections to bone fractures, heart attacks and dementia." were successfully picked by the typology approach; whereas they were missed by the hybrid approach, as keywords in those sentences were less commonly used to describe side effects. The typology approach distilled relevant information from text documents through sentence extraction by human experts. This information was key to building a knowledge base for the identification of sentences about side effects. We anticipated that the typology approach will be more robust and stable than the hybrid approach when visualizing the interpretation of criteria that are less keyword-reliant. For example, 1 of the 10 criteria, "Does the news compare the new approach with existing alternatives?" examined whether health news included a discussion on alternatives. Sentences that supported a satisfactory evaluation result may have been less likely to be observed with repetitive keywords than with the experimental criteria.

Limitations

This exploratory study had some limitations. The first limitation was that we only considered the TF-IDF values of words as features for building both the document- and sentence-level classifiers. We acknowledged that the performance of our document-level classification model was lower compared with similar studies that adopted the same data set from HealthNewsReview.org. The performance of our doc-level classification models for the harm, cost, and conflict criteria were 0.71, 0.82, and 0.67, respectively, when measured by F_1 and 0.76, 0.88, and 0.72 when measured by AUC. The performance was better compared with a study by Al-Jefri et al [58] that focused on building health news quality classification models. The precision performance for classifying the harm, cost, and conflict criteria was reported to be 74.61, 77.61, and 70.89, respectively. The study incorporated more features, such as TF-IDF, comparative forms, and named-entity recognition tags and strategically changed the feature selections for different criterion classification tasks. In another study by Afsana et al [59], which also aimed to achieve the same research goal, the performance of their models for the harm, cost, and conflict measures by weighted F_1-score was reported 0.84, 0.899, and 0.835, respectively. However, superior performance was achieved through extensive work on feature engineering with 53,012 features applied. Considering that the key focus of this study was to experiment with 2 interpretation approaches, which both mentioned studies lack, we believed that the current performance of models was effective in serving the purpose of the study. In the future, we will incorporate some work on feature engineering for both document-level classification and especially the typology approach, which is embodied as a sentence-level classifier.

The second limitation is the simple rules of the hybrid approach. The hybrid approach takes advantage of both human knowledge and an autogenerated keyword list generated by the LIME. However, existing rules provided by human experts were keyword-based and did not contain complex rules for handling various expression variants. As part of the future plan, we will implement more complex rules for the hybrid to address the weak spots of the hybrid to enable it to distinguish different types of sentences when they share similar lexicons but different semantics.

A further limitation of the study was the absence of a user study to investigate how the final visual interpretation generated by the 2 interpretation approaches would increase user trust in a black-box model, particularly in the context of evaluating the quality of health news to mitigate misinformation. However, we have an ongoing user study to investigate whether a criteria-based system with visualized interpretation for evaluating health news quality will increase the trust of users compared with the system without interpretation. As of the completion of this study, the user study is still in the recruitment phase.

Comparison With Prior Work

Our study addressed the public’s need to help evaluate the quality of health news and the typical opaqueness of an AI approach. The significance of this study is illustrated in 2 ways.

First, compared with previous interpretability work in suggested health-related misinformation detection systems, our work on adding the interpretability of a health misinformation system is innovative. To our knowledge, the current state of the art in explainable misinformation detection systems mostly looks to provide explanations for veracity predictions concerning inputs to the system. Our study fills a gap in the literature by explaining a criteria-based system for health misinformation. Moreover, developing an interpretable module on a criteria-based model is advantageous. The criteria-based approach inherently looks for the linguistic characteristics of health news, such as the presence or absence of crucial information, whereas a veracity-based system may face a challenge to be interpreted based on the linguistic features of text alone. In addition, we believe that our study exhibited a greater level of readability of the interpretation than the existing interpretation work on health misinformation, such as Alharbi et al [80] for fake news. The interpretation level achieved in the study by Alharbi et al remained at the word level, with both positive and negative words highlighted and dispersed throughout the articles; whereas our study presented 2 approaches to achieve sentence-level visualized interpretation, which demonstrated higher levels of readability to end users.

Second, this exploratory study demonstrated great potential for the development of a criteria-based system for evaluating the quality of health news as a way to counteract health misinformation. Compared with a veracity-based health misinformation detection system, a criteria-based system demonstrated high generalizability in handling health information on various topics. Most existing veracity-based fake news detectors are built on linguistic cues, leading to a lack of generalizability across topics, languages, and domains [82]. This weakness was also proven in a study by Gerts et al [41], as the team found a huge variation in the classifier performance (F_1-scores between 0.347 and 0.857) on 4 conspiracy topics and more narrowly defined topics could increase performance. In comparison, the idea of a criteria-driven system was to evaluate the quality of health news based on evidence for specified criteria. The evaluation procedure did not require a significant amount of domain knowledge. Thus, this type of system can be adapted to handle a variety of health news stories on various themes, as it did not rely on a data set with a strictly defined topic. In addition, an interpretable, criteria-based system may address the complexity and multidimensional attributes of the health information disorder [83-85]. Automatic tools for evaluating health misinformation have proven promising owing to their high accuracy and fast processing speed. However, existing studies are still predominantly binary classification tasks. This places a great challenge in identifying health misinformation, as the binary label is insufficient to represent the complicated evaluation process of health news in actual practice. This is especially the case with the veracity-based classification. Human-based fact-checking involves extensive knowledge understanding, inference, and source tracking, which remains a challenge, even in deep learning methods. This is because fabricated news is intended to mirror the truth to deceive readers; as a result, without cross-referencing and high-level inference, it might be impossible to determine the authenticity of news stories by text analysis alone [82]. Although it does not provide veracity-level health news validation for users, it has the potential to provide another way of combating health misinformation by improving users' critical thinking about health news, as the slogan on HealthNewsReview.org indicates.

Conclusions

In this study, we described an interpretable, criteria-based strategy for evaluating the quality of health news. We explored 2 methods for visualizing the interpretation of the system. To aid in the exploration, an experiment was developed by comparing rule-based and statistical ML approaches. Our results suggested that either approach can successfully automate criterion-based health news quality ratings, with visual evidence supporting model explanation. This study has the potential to increase public trust in computer-assisted reviews of health information. We intend to expand on this study by applying 2 visualization approaches to more criteria and focusing on improving the performance of the classification model.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Performance of different base classifiers for automating 3 criteria evaluation.
[DOCX File , 27 KB-Multimedia Appendix 1]

Multimedia Appendix 2

The 2-level visual interpretation cases for the cost and the conflict criteria.
[DOCX File , 11543 KB-Multimedia Appendix 2]

References

1. Fox S. Health Topics. Pew Research Center. 2011 Feb 1. URL: https://www.pewresearch.org/internet/2011/02/01/health-topics-2/ [accessed 2021-09-21]
2. Becker BF, Larson HJ, Bonhoeffer J, van Mulligen EM, Kors JA, Sturkenboom MC. Evaluation of a multinational, multilingual vaccine debate on Twitter. Vaccine 2016 Dec 07;34(50):6166-6171. [doi: 10.1016/j.vaccine.2016.11.007] [Medline: 27840012]
3. Bonnevie E, Goldbarg J, Gallegos-Jeffrey AK, Rosenberg SD, Wartella E, Smyser J. Content themes and influential voices within vaccine opposition on Twitter, 2019. Am J Public Health 2020 Oct;110(S3):S326-S330. [doi: 10.2105/AJPH.2020.305901] [Medline: 33001733]
4. Hanson CL, Cannon B, Burton S, Giraud-Carrier C. An exploration of social circles and prescription drug abuse through Twitter. J Med Internet Res 2013 Sep 06;15(9):e189 [FREE Full text] [doi: 10.2196/jmir.2741] [Medline: 24014109]
5. Jamison A, Broniatowski DA, Smith MC, Parikh KS, Malik A, Dredze M, et al. Adapting and extending a typology to identify vaccine misinformation on Twitter. Am J Public Health 2020 Oct;110(S3):S331-S339. [doi: 10.2105/AJPH.2020.305940] [Medline: 33001737]
6. Buchanan R, Beckett RD. Assessment of vaccination-related information for consumers available on Facebook. Health Info Libr J 2014 Sep;31(3):227-234 [FREE Full text] [doi: 10.1111/hir.12073] [Medline: 25041499]
7. Faasse K, Chatman CJ, Martin LR. A comparison of language use in pro- and anti-vaccination comments in response to a high profile Facebook post. Vaccine 2016 Nov 11;34(47):5808-5814. [doi: 10.1016/j.vaccine.2016.09.029] [Medline: 27707558]
8. Johnson SB, Parsons M, Dorff T, Moran MS, Ward JH, Cohen SA, et al. Cancer misinformation and harmful information on Facebook and other social media: a brief report. J Natl Cancer Inst 2022 Jul 11;114(7):1036-1039 [FREE Full text] [doi: 10.1093/jnci/djab141] [Medline: 34291289]
9. Seymour B, Getman R, Saraf A, Zhang LH, Kalenderian E. When advocacy obscures accuracy online: digital pandemics of public health misinformation through an antifluoride case study. Am J Public Health 2015 Mar;105(3):517-523. [doi: 10.2105/AJPH.2014.302437] [Medline: 25602893]
10. Abukaraky A, Hamdan AA, Ameera MN, Nasief M, Hassona Y. Quality of YouTube TM videos on dental implants. Med Oral Patol Oral Cir Bucal 2018 Jul 01;23(4):e463-e468 [FREE Full text] [doi: 10.4317/medoral.22447] [Medline: 29924766]
11. Basch CH, Zybert P, Reeves R, Basch CE. What do popular YouTube videos say about vaccines? Child Care Health Dev 2017 Jul;43(4):499-503. [doi: 10.1111/cch.12442] [Medline: 28105642]
12. Biggs TC, Bird JH, Harries PG, Salib RJ. YouTube as a source of information on rhinosinusitis: the good, the bad and the ugly. J Laryngol Otol 2013 Aug;127(8):749-754. [doi: 10.1017/S0022215113001473] [Medline: 23866821]
13. Röchert D, Neubaum G, Stieglitz S. Identifying political sentiments on YouTube: a systematic comparison regarding the accuracy of recurrent neural network and machine learning models. In: Proceedings of the 2nd Multidisciplinary International Symposium on Disinformation in Open Online Media. 2020 Presented at: MISDOOM '20; October 26–27, 2020; Leiden, The Netherlands p. 107-121 URL: https://link.springer.com/chapter/10.1007/978-3-030-61841-4_8 [doi: 10.1007/978-3-030-61841-4_8]

14. Guidry J, Jin Y, Haddad L, Zhang Y, Smith J. How health risks are pinpointed (or not) on social media: the portrayal of waterpipe smoking on Pinterest. Health Commun 2016;31(6):659-667. [doi: 10.1080/10410236.2014.987468] [Medline: 26512916]
15. Guidry JP, Carlyle K, Messner M, Jin Y. On pins and needles: how vaccines are portrayed on Pinterest. Vaccine 2015 Sep 22;33(39):5051-5056. [doi: 10.1016/j.vaccine.2015.08.064] [Medline: 26319742]
16. Li A, Huang X, Jiao D, O'Dea B, Zhu T, Christensen H. An analysis of stigma and suicide literacy in responses to suicides broadcast on social media. Asia Pac Psychiatry 2018 Mar;10(1):e12314. [doi: 10.1111/appy.12314] [Medline: 29383880]
17. Xiao L, Chen S. Misinformation in the Chinese Weibo. In: Proceedings of the 12th International Conference on Social Computing and Social Media. Design, Ethics, User Behavior, and Social Network Analysis. 2020 Presented at: HCII '20; July 19–24, 2020; Copenhagen, Denmark p. 407-418 URL: https://link.springer.com/chapter/10.1007/978-3-030-49570-1_28
18. Waszak PM, Kasprzycka-Waszak W, Kubanek A. The spread of medical fake news in social media – the pilot quantitative study. Health Policy Technol 2018 Jun;7(2):115-118. [doi: 10.1016/j.hlpt.2018.03.002]
19. Chua AY, Banerjee S. Intentions to trust and share online health rumors: an experiment with medical professionals. Comput Human Behav 2018 Oct;87:1-9. [doi: 10.1016/j.chb.2018.05.021]
20. Radzikowski J, Stefanidis A, Jacobsen KH, Croitoru A, Crooks A, Delamater PL. The measles vaccination narrative in Twitter: a quantitative analysis. JMIR Public Health Surveill 2016 Jan 4;2(1):e1 [FREE Full text] [doi: 10.2196/publichealth.5059] [Medline: 27227144]
21. Vraga EK, Bode L. Using expert sources to correct health misinformation in social media. Sci Commun 2017 Sep 14;39(5):621-645. [doi: 10.1177/1075547017731776]
22. Chen B, Shao J, Liu K, Cai G, Jiang Z, Huang Y, et al. Does eating chicken feet with pickled peppers cause avian influenza? Observational case study on Chinese social media during the avian influenza a (H7N9) outbreak. JMIR Public Health Surveill 2018 Mar 29;4(1):e32 [FREE Full text] [doi: 10.2196/publichealth.8198] [Medline: 29599109]
23. Li Y, Zhang X, Wang S. Fake vs. real health information in social media in China. Proc Assoc Info Sci Tech 2017 Oct 24;54(1):742-743. [doi: 10.1002/pra2.2017.14505401139]
24. Leong AY, Sanghera R, Jhajj J, Desai N, Jammu BS, Makowsky MJ. Is YouTube useful as a source of health information for adults with type 2 diabetes? A South Asian perspective. Can J Diabetes 2018 Aug;42(4):395-403.e4. [doi: 10.1016/j.jcjd.2017.10.056] [Medline: 29282200]
25. Aquino F, Donzelli G, De Franco E, Privitera G, Lopalco PL, Carducci A. The web and public confidence in MMR vaccination in Italy. Vaccine 2017 Aug 16;35(35 Pt B):4494-4498. [doi: 10.1016/j.vaccine.2017.07.029] [Medline: 28736200]
26. Bessi A, Zollo F, Del Vicario M, Scala A, Caldarelli G, Quattrociocchi W. Trend of narratives in the age of misinformation. PLoS One 2015 Aug 14;10(8):e0134641 [FREE Full text] [doi: 10.1371/journal.pone.0134641] [Medline: 26275043]
27. Vraga EK, Bode L. Defining misinformation and understanding its bounded nature: using expertise and evidence for describing misinformation. Polit Commun 2020 Feb 06;37(1):136-144. [doi: 10.1080/10584609.2020.1716500]
28. Chou WS, Oh A, Klein WM. Addressing health-related misinformation on social media. JAMA 2018 Dec 18;320(23):2417-2418. [doi: 10.1001/jama.2018.16865] [Medline: 30428002]
29. Suarez-Lledo V, Alvarez-Galvez J. Prevalence of health misinformation on social media: systematic review. J Med Internet Res 2021 Jan 20;23(1):e17187 [FREE Full text] [doi: 10.2196/17187] [Medline: 33470931]
30. Bianco V. Countering Online Misinformation Resource Pack. UNICEF Regional Office for Europe and Central Asia. 2020 Aug. URL: https://www.unicef.org/eca/media/13636/file [accessed 2021-10-26]
31. Fighting misinformation in the time of COVID-19, one click at a time. World Health Organization. 2021 Apr 27. URL: https://www.who.int/news-room/feature-stories/detail/fighting-misinformation-in-the-time-of-covid-19-one-click-at-a-time [accessed 2022-09-21]
32. Bridgman A, Merkley E, Zhilin O, Loewen PJ, Owen T, Ruths D. Infodemic pathways: evaluating the role that traditional and social media play in cross-national information transfer. Front Polit Sci 2021 Mar 29;3:20. [doi: 10.3389/fpos.2021.648646]
33. Cui L, Lee D. CoAID: COVID-19 healthcare misinformation dataset. arXiv 2020 May 22 [FREE Full text]
34. Marr B. Fake News Is Rampant, Here Is How Artificial Intelligence Can Help. Forbes. 2021 Jan 25. URL: https://www.forbes.com/sites/bernardmarr/2021/01/25/fake-news-is-rampant-here-is-how-artificial-intelligence-can-help/ [accessed 2021-12-15]
35. Burns E, Laskowski N, Tucci L. What is Artificial Intelligence (AI)? - AI Definition and How it Works. SearchEnterpriseAI. URL: https://searchenterpriseai.techtarget.com/definition/AI-Artificial-Intelligence [accessed 2021-10-26]
36. Meppelink CS, Hendriks H, Trilling D, van Weert JC, Shao A, Smit ES. Reliable or not? An automated classification of webpages about early childhood vaccination using supervised machine learning. Patient Educ Couns 2021 Jun;104(6):1460-1466 [FREE Full text] [doi: 10.1016/j.pec.2020.11.013] [Medline: 33243581]
37. Shah Z, Surian D, Dyda A, Coiera E, Mandl KD, Dunn AG. Automatically appraising the credibility of vaccine-related web pages shared on social media: a Twitter surveillance study. J Med Internet Res 2019 Nov 04;21(11):e14007 [FREE Full text] [doi: 10.2196/14007] [Medline: 31682571]
38. Wang Z, Yin Z, Argyris YA. Detecting medical misinformation on social media using multimodal deep learning. IEEE J Biomed Health Inform 2021 Jun;25(6):2193-2203. [doi: 10.1109/JBHI.2020.3037027] [Medline: 33170786]

39. Ghenai A, Mejova Y. Catching Zika fever: application of crowdsourcing and machine learning for tracking health misinformation on Twitter. In: Proceedings of the 2017 IEEE International Conference on Healthcare Informatics. 2017 Jul 12 Presented at: ICHI '17; August 23-26, 2017; Park City, UT, USA p. 518. [doi: 10.1109/ichi.2017.58]
40. Zhao Y, Da J, Yan J. Detecting health misinformation in online health communities: incorporating behavioral features into machine learning based approaches. Inf Process Manag 2021 Jan;58(1):102390. [doi: 10.1016/j.ipm.2020.102390]
41. Gerts D, Shelley CD, Parikh N, Pitts T, Watson Ross C, Fairchild G, et al. "Thought I'd share first" and other conspiracy theory tweets from the COVID-19 infodemic: exploratory study. JMIR Public Health Surveill 2021 Apr 14;7(4):e26527 [FREE Full text] [doi: 10.2196/26527] [Medline: 33764882]
42. Abdelminaam DS, Ismail FH, Taha M, Taha A, Houssein EH, Nabil A. CoAID-DEEP: an optimized intelligent framework for automated detecting COVID-19 misleading information on Twitter. IEEE Access 2021 Feb 9;9:27840-27867 [FREE Full text] [doi: 10.1109/ACCESS.2021.3058066] [Medline: 34786308]
43. Ayoub J, Yang XJ, Zhou F. Combat COVID-19 infodemic using explainable natural language processing models. Inf Process Manag 2021 Jul;58(4):102569 [FREE Full text] [doi: 10.1016/j.ipm.2021.102569] [Medline: 33776192]
44. Kolluri NL, Murthy D. CoVerifi: a COVID-19 news verification system. Online Soc Netw Media 2021 Mar;22:100123 [FREE Full text] [doi: 10.1016/j.osnem.2021.100123] [Medline: 33521412]
45. Dhoju S, Main Uddin Rony M, Ashad Kabir M, Hassan N. Differences in health news from reliable and unreliable media. In: Companion Proceedings of The 2019 World Wide Web Conference. 2019 May Presented at: WWW '19; May 13-17, 2019; San Francisco, CA, USA p. 981-987. [doi: 10.1145/3308560.3316741]
46. Saengkhunthod C, Kerdnoonwong P, Atchariyachanvanich K. Detection of unreliable medical articles on Thai websites. In: Proceedings of the 13th International Conference on Knowledge and Smart Technology. 2021 Presented at: KST '21; January 21-24, 2021; Bangsaen, Thailand p. 102-107. [doi: 10.1109/kst51265.2021.9415756]
47. Dito FM, Alqadhi HA, Alasaadi A. Detecting medical rumors on Twitter using machine learning. In: Proceedings of the 2020 International Conference on Innovation and Intelligence for Informatics, Computing and Technologies. 2020 Presented at: 3ICT '20; December 20-21, 2020; Sakheer, Bahrain p. 1-7. [doi: 10.1109/3ict51146.2020.9311957]
48. Kinsora A, Barron K, Mei Q, Vydiswaran VG. Creating a labeled dataset for medical misinformation in health forums. In: Proceedings of the 2017 IEEE International Conference on Healthcare Informatics. 2017 Presented at: ICHI '17; August 23-26, 2017; Park City, UT, USA p. 456-461. [doi: 10.1109/ichi.2017.93]
49. Parfenenko Y, Verbytska A, Bychko D, Shendryk V. Application for medical misinformation detection in online forums. In: Proceedings of the 2020 International Conference on e-Health and Bioengineering. 2020 Presented at: EHB '20; October 29-30, 2020; Iasi, Romania p. 1-4. [doi: 10.1109/ehb50910.2020.9280120]
50. Liu Y, Yu K, Wu X, Qing L, Peng Y. Analysis and detection of health-related misinformation on Chinese social media. IEEE Access 2019 Oct 14;7:154480-154489. [doi: 10.1109/ACCESS.2019.2946624]
51. Elhadad MK, Li KF, Gebali F. Detecting misleading information on COVID-19. IEEE Access 2020 Sep 9;8:165201-165215 [FREE Full text] [doi: 10.1109/ACCESS.2020.3022867] [Medline: 34786288]
52. Khanday A, Khan QR, Rabani ST. Identifying propaganda from online social networks during COVID-19 using machine learning techniques. Int J Inf Technol 2021;13(1):115-122 [FREE Full text] [doi: 10.1007/s41870-020-00550-5] [Medline: 33145473]
53. Snopes-Medical Archives. Snopes.com. URL: https://www.snopes.com/fact-check/category/medical/ [accessed 2022-02-15]
54. World Health Organization. URL: https://www.who.int [accessed 2022-02-15]
55. Johns Hopkins Medicine, based in Baltimore, Maryland. URL: https://www.hopkinsmedicine.org/ [accessed 2022-02-15]
56. CDC Works 24/7. Centers for Disease Control and Prevention. 2022. URL: https://www.cdc.gov/index.htm [accessed 2022-02-15]
57. Zhou X, Zafarani R. A survey of fake news: fundamental theories, detection methods, and opportunities. ACM Comput Surv 2021 Sep;53(5):1-40. [doi: 10.1145/3395046]
58. Al-Jefri M, Evans R, Lee J, Ghezzi P. Automatic identification of information quality metrics in health news stories. Front Public Health 2020 Dec 18;8:515347 [FREE Full text] [doi: 10.3389/fpubh.2020.515347] [Medline: 33392124]
59. Afsana F, Kabir MA, Hassan N, Paul M. Automatically assessing quality of online health articles. IEEE J Biomed Health Inform 2021 Feb;25(2):591-601. [doi: 10.1109/JBHI.2020.3032479] [Medline: 33079686]
60. Hawa S, Lobo L, Dogra U, Kamble V. Combating misinformation dissemination through verification and content driven recommendation. In: Proceedings of the 3rd International Conference on Intelligent Communication Technologies and Virtual Mobile Networks. 2021 Presented at: ICICV '21; February 4-6, 2021; Tirunelveli, India p. 917-924. [doi: 10.1109/icicv50876.2021.9388406]
61. Choudrie J, Banerjee S, Kotecha K, Walambe R, Karende H, Ameta J. Machine learning techniques and older adults processing of online information and misinformation: a covid 19 study. Comput Human Behav 2021 Jun;119:106716 [FREE Full text] [doi: 10.1016/j.chb.2021.106716] [Medline: 34866770]
62. Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable AI: a review of machine learning interpretability methods. Entropy (Basel) 2020 Dec 25;23(1):18 [FREE Full text] [doi: 10.3390/e23010018] [Medline: 33375658]
63. Ribeiro MT. lime: Local Interpretable Model-Agnostic Explanations for machine learning classifiers. GitHub. 2021 Jul 30. URL: http://github.com/marcotcr/lime [accessed 2022-02-15]

64. Lundberg S. shap: A unified approach to explain the output of any machine learning model. GitHub. URL: https://github.com/slundberg/shap [accessed 2022-02-15]
65. Korobov M, Lopuhin K. eli5: Debug machine learning classifiers and explain their predictions. GitHub. URL: https://github.com/eli5-org/eli5 [accessed 2022-02-15]
66. InterpretML Team. interpret: Fit interpretable machine learning models - Explain blackbox machine learning. GitHub. URL: https://github.com/interpretml/interpret [accessed 2022-02-15]
67. Kotonya N, Toni F. Explainable automated fact-checking for public health claims. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. 2020 Oct 19 Presented at: EMNLP '20; November 16-18, 2020; Virtual p. 7740-7754. [doi: 10.18653/v1/2020.emnlp-main.623]
68. Zuo C, Zhang Q, Banerjee R. An empirical assessment of the qualitative aspects of misinformation in health news. In: Proceedings of the 4th Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda. 2021 Presented at: NLP4IF '21; June 6, 2021; Virtual p. 76-81. [doi: 10.18653/v1/2021.nlp4if-1.11]
69. HealthNewsReview. URL: https://www.healthnewsreview.org/ [accessed 2022-02-15]
70. Schwitzer G. How do US journalists cover treatments, tests, products, and procedures? An evaluation of 500 stories. PLoS Med 2008 May 27;5(5):e95 [FREE Full text] [doi: 10.1371/journal.pmed.0050095] [Medline: 18507496]
71. Criterion #3 Does the story adequately explain/quantify the harms of the intervention? HealthNewsReview. URL: https://www.healthnewsreview.org/about-us/review-criteria/criterion-3/ [accessed 2022-02-15]
72. Criterion #1 Does the story adequately discuss the costs of the intervention? HealthNewsReview. URL: https://web.archive.org/web/20220629205927/http://www.healthnewsreview.org/about-us/review-criteria/criterion-1/ [accessed 2022-09-05]
73. Criterion #6 Does the story use independent sources and identify conflicts of interest? HealthNewsReview. URL: https://web.archive.org/web/20220629202851/https://www.healthnewsreview.org/about-us/review-criteria/criterion-6/ [accessed 2022-09-05]
74. Ribeiro MT, Singh S, Guestrin C. "Why should i trust you?": explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016 Feb 16 Presented at: KDD '16; August 13-17, 2016; San Francisco, CA, USA p. 1135-1144.
75. Ribeiro MT, Singh S, Guestrin C. Local interpretable model-agnostic explanations (LIME): an introduction. O'Reilly Media. 2016 Aug 12. URL: https://www.oreilly.com/content/introduction-to-local-interpretable-model-agnostic-explanations-lime/ [accessed 2022-02-15]
76. Molnar C. Interpretable Machine Learning. URL: https://christophm.github.io/interpretable-ml-book/ [accessed 2022-02-15]
77. Zafar MR, Khan N. Deterministic local interpretable model-agnostic explanations for stable explainability. Mach Learn Knowl Extr 2021 Jun 30;3(3):525-541. [doi: 10.3390/make3030027]
78. Reynolds RA, Reynolds JL. Evidence. In: Dillard JP, Pfau M, editors. The Persuasion Handbook: Developments in Theory and Practice. Thousand Oaks, CA, USA: Sage Publications; 2002:427-445.
79. Hoeken H, Hustinx L. The relative persuasiveness of anecdotal, statistical, causal, and expert evidence. In: Proceedings of the 5th Conference of the International Society for the Study of Argumentation. 2002 Presented at: ISSA '02; June 26-28, 2002; Amsterdam, The Netherlands p. 497-502 URL: https://repository.ubn.ru.nl/handle/2066/82921
80. Fiok K, Karwowski W, Gutierrez E, Liciaga T, Belmonte A, Capobianco R. Automated classification of evidence of respect in the communication through Twitter. Appl Sci 2021 Feb 01;11(3):1294. [doi: 10.3390/app11031294]
81. Freund AJ, Giabbanelli PJ. Are we modeling the evidence or our own biases? A comparison of conceptual models created from reports. In: Proceedings of the 2021 Annual Modeling and Simulation Conference. 2021 Presented at: ANNSIM '21; July 19-22, 2021; Fairfax, VA, USA p. 1-12. [doi: 10.23919/annsim52504.2021.9552054]
82. Sharma K, Qian F, Jiang H, Ruchansky N, Zhang M, Liu Y. Combating fake news: a survey on identification and mitigation techniques. ACM Trans Intell Syst Technol 2019 Apr;10(3):1-42. [doi: 10.1145/3305260]
83. Wardle C, Derakhshan H. Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making. Council of Europe. 2017. URL: https://edoc.coe.int/en/media/7495-information-disorder-toward-an-interdisciplinary-framework-for-research-and-policy-making.html [accessed 2021-10-26]
84. Habgood-Coote J. Stop talking about fake news!. Inquiry 2018 Aug 11;62(9-10):1033-1065 [FREE Full text] [doi: 10.1080/0020174x.2018.1508363]
85. Wang Y, McKee M, Torbica A, Stuckler D. Systematic literature review on the spread of health-related misinformation on social media. Soc Sci Med 2019 Nov;240:112552 [FREE Full text] [doi: 10.1016/j.socscimed.2019.112552] [Medline: 31561111]

Abbreviations

AI: artificial intelligence
AUC: area under the curve
LIME: local interpretable model-agnostic explanation
ML: machine learning
RF: random forest

TF-IDF: term frequency–inverse document frequency

Edited by K El Emam, B Malin; submitted 04.03.22; peer-reviewed by P Giabbanelli, L Bošnjak, B Puladi; comments to author 06.06.22; revised version received 22.09.22; accepted 11.11.22; published 20.12.22

Please cite as:
Liu X, Alsghaier H, Tong L, Ataullah A, McRoy S
Visualizing the Interpretation of a Criteria-Driven System That Automatically Evaluates the Quality of Health News: Exploratory Study of 2 Approaches
JMIR AI 2022;1(1):e37751
URL: https://ai.jmir.org/2022/1/e37751
doi: 10.2196/37751
PMID: