Improving the Reliability of Health Information Credibility Assessments

Marcos Fernández-Pichel[1,*], Selina Meyer[2], Markus Bink[2], Alexander Frummet[2], David E. Losada[1] and David Elsweiler[2]

[1]Centro de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, Santiago de Compostela, Spain
[2]Chair for Information Science, Regensburg University, Regensburg, Bavaria, Germany

Abstract

The applicability of retrieval algorithms to real data relies heavily on the quality of the training data. Currently, the creation process of training and test collections for retrieval systems is often based on annotations produced by human assessors following a set of guidelines. Some concepts, however, are prone to subjectivity, which could restrict the utility of any algorithm developed with the resulting data in real world applications. One such concept is credibility, which is an important factor in user's judgements on whether retrieved information helps to answer an information need. In this paper, we evaluate an existing set of assessment guidelines with respect to their ability to generate reliable credibility judgements across multiple raters. We identify reasons for disagreement and adapt the guidelines to create an actionable and traceable annotation scheme that i) leads to higher inter-annotator reliability, and ii) can inform about why a rater made a specific credibility judgement. We provide promising evidence about the robustness of the new guidelines and conclude that they could be a valuable resource for building future test collections for misinformation detection.

Keywords

reliability, credibility assessments, health-related content

1. Introduction

Misinformation on the Internet is becoming increasingly common [1, 2]. When interacted with, such information can cause people to make decisions with potentially harmful consequences [3], especially when the information is related to health [4]. Multiple prestigious venues regularly organise shared-task competitions with the goal of developing retrieval or classification algorithms that counteract misinformation on the web [5, 6, 7]. While these initiatives are highly valuable to fostering research in misinformation detection, the potential impact of the algorithms developed by participants may be restricted by the quality of the provided

ROMCIR 2023: The 3rd Workshop on Reducing Online Misinformation through Credible Information Retrieval, held as part of ECIR 2023: the 45th European Conference on Information Retrieval, April 2-6, 2023, Dublin, Ireland
*Corresponding author.
marcosfernandez.pichel@usc.es (M. Fernández-Pichel); selina.meyer@ur.de (S. Meyer); markus.bink@ur.de (M. Bink); alexander.frummet@ur.de (A. Frummet); david.losada@usc.es (D. E. Losada); david.elsweiler@ur.de (D. Elsweiler)
0000-0002-6560-9832 (M. Fernández-Pichel); 0000-0002-4736-2565 (S. Meyer); 0000-0002-5982-7104 (A. Frummet); 0000-0001-8823-7501 (D. E. Losada); 0000-0002-5791-0641 (D. Elsweiler)

training and test data. Generating ground truth data is a crucial and costly process, as it often requires the intervention of human assessors for creating complex assessments. For instance, in the TREC Health Misinformation Track judge(s)[1] are asked to label documents on topical relevance, credibility, and correctness[2]. Some of these dimensions, however, are hard to assess and annotation practices vary across competitions. One such hard to assess concept is credibility, which has been shown to be highly subjective and susceptible to individual differences [8]. Here we adopt the term "credibility", as defined by the TREC track: the document's trustworthiness and authoritativeness, as perceived by the assessors [6]. This is intrinsically a subjective concept, but robust guidelines can help in clarifying the label creation process and, thus, produce more solid benchmarks.

In this study, we try to shed light on the difficulty to create credibility assessments and propose new guidelines to produce more robust judgements. We do this by applying the current TREC Health Misinformation credibility guidelines to a series of health-related web documents and evaluating agreement across multiple raters. We then identify reasons for rater disagreement and adapt the existing guidelines to create an actionable and traceable annotation scheme that i) leads to higher inter-annotator reliability, and ii) can inform about why a rater made a specific credibility judgement. We provide promising evidence about the robustness of the new guidelines and conclude that they could be a valuable resource for building test collections for misinformation detection and thus mitigating the impact of health misinformation on the web.

2. Related work

Research on credibility examines how characteristics, such as expertise or trustworthiness, affect the "believability" of an information source [9, 10]. Credibility judgements for web pages have been studied extensively and are known to be influenced by aesthetics and impressions of professionalism [11, 12]. Credibility is a largely subjective concept and can be strongly influenced by an individual rater's personal disposition, e.g. how suspicious they are by nature, their propensity to risk [12], as well as their reading abilities [13]. There is mixed evidence with respect to the influence of topical knowledge and expertise [14, 15].

The subjectivity inherent to credibility judgements demands clear and specific guidelines for the development of test collections, so that robust annotations can be obtained that accurately reflect population judgements overall. One attempt at creating such guidelines was presented by Nabożny et al. [16] in the context of medical misinformation. In contrast to our aims, these authors directed their annotation protocol towards medical experts and focused on sentence credibility, marked mainly by the presentation of factual and reliable information, rather than web page credibility. Zhang and colleagues [17] proposed a set of article credibility indicators, which included hard to spot indicators such as logical fallacies and tone. Their approach, which requires trained annotators, assigned content and context based annotation to individuals with different levels of training. The results showed low inter-rater agreement on some of the items, again highlighting the difficulty of credibility annotation.

[1] it is unclear from the TREC overview if multiple assessors were employed

[2] https://trec-health-misinfo.github.io/docs/TREC-2021-Health-Misinformation-Track-Assessing-Guidelines_Version-2.pdf

We examine here a set of guidelines for judging web page credibility which, to our knowledge, have not yet been publicly evaluated, namely the TREC Health Misinformation track guidelines[2]. We then adapt this framework to create new guidelines which show first evidence of leading to higher inter-rater reliability with minimal to no training in preliminary tests, thus enabling non-domain experts to judge the credibility of health-related web pages.

3. Evaluating Guidelines

According to the TREC Health Misinformation track's guidelines, a highly credible document in the context of health should be "unquestionably trustworthy and authoritative", whereas for low-credibility documents "there is little evidence to believe or trust the information source", with medium-credibility documents located in-between. To help determine the credibility level of a document, human assessors are provided with an extensive list of guidelines to follow[2]. Information about the number of assessors recruited to judge the credibility of documents in this TREC collection is not publicly available. There is also no public information about inter-rater agreement in the track. To evaluate the reliability of these judgements, four of the authors of this paper followed these guidelines to independently judge 12 randomly chosen web documents from an existing collection of documents from the medical domain [18, 19]. Next, we calculated pairwise, linear-weighted Cohen's Kappa to evaluate agreement between single raters, and Krippendorff's Alpha for ordinal scales to evaluate agreement between all raters. Kappa-values ranged between 0.25 and 0.79, with a median of $\kappa = 0.44$. Krippendorff's α was 0.6. These Kappa values indicate only moderate agreement on average [20], and α falls below the lowest conceivable limit of $\alpha \geq 0.667$ for reliable annotations (as defined by Krippendorff [21]).

Next, the four annotators discussed their judgements in a group meeting to identify the problems with the current guidelines. Three main reasons were found behind the low agreement between raters. First, the lengthy and unstructured nature of the guidelines. In some cases, it was difficult to pinpoint which part of the guidelines had led the annotators to decide on a certain judgement. The guidelines are unnumbered, despite consisting of 13 bullet points, and some items could have been broken down into multiple related aspects (for example, "Try to determine the amount of expertise, authoritativeness, and trustworthiness of the document"). This makes the process less traceable and causes raters to focus on different aspects, leading to divergent credibility judgements. Second, the lack of a clear-cut differentiation between levels of credibility. While three levels of credibility are defined, the guidelines give no indication on how these definitions relate or where cut-offs lie. It is left for the annotator to decide, how important each guideline item is for the credibility of the document. Third, the use of ambiguous concepts as a way to judge credibility. The guidelines introduce new, difficult to define and comparably subjective concepts to act as credibility indicators. These include e.g. trustworthiness, authoritativeness, expertise, and ubiquity. These issues informed the development of the new guidelines proposed below.

	Label	Guideline	Step
G1	2	Source is a scientific paper, or a Medical publisher or hospital/clinic or government website or university.	1
G2	1	Document is citing the information they provide in their articles. They provide links or specific references to their sources. They cite sources with credibility 2 (i.e. medical publications and/or lab studies).	4
G3	1	Document is written by an expert in the field/someone qualified to write this document (irrespective of publishing venue).	3
G4	0	The document is actually for advertising or marketing purposes. If so, the website might be biased or a scam designed to trick people into fake treatments or into buying medical products that do not live up to their claim.	2
G5	0	The information posted by a non-expert person providing a medical product review or providing medical advice without proper citations (links/list of references).	5
G6	0	The website provides or states claims that go against well-known medical consensus (e.g. smoking cigarettes does not cause cancer).	5

NOTE: It is generally allowed to look up authors to check whether they have the required knowledge to be regarded as an expert and look up websites to find out if they are legitimate.

Table 1
New proposed guidelines for web content credibility labelling.

4. A Robust and Traceable Set of Credibility Guidelines

The TREC guidelines were taken as a starting point and the individuals involved in the initial annotation iteratively revised the guidelines in multiple discussion sessions. At the end of this process, the original guidelines had been condensed and adapted to six guidelines for webpage credibility labelling (see Table 1). The new guidelines provide clear credibility score recommendations based on the fulfilment of a number of criteria, summarise the TREC guidelines in a way that reflects the most important aspects and, at the same time, decrease ambiguity. The full guidelines are presented in this paper both in written and in flowchart form to facilitate and speed up the evaluation with a visual tool, see Figure 1 and Table 1. Each guideline can be mapped to a specific step in the flowchart. The new guidelines forgo the introduction of complicated concepts as much as possible and rely mostly on measurable indicators.

The same 12 documents from the initial evaluation were then annotated again by the same four raters using the new guidelines. This led to an increase in agreement by 28%, resulting in a Krippendorff's α of 0.88 and a median Cohen's κ of 0.89 (max: $\kappa = 1$, min: $\kappa = 0.78$), indicating almost perfect agreement.

4.1. Evaluation with a New Sample of Webpages

Since the assessors were already familiar with the initial 12 documents and had used them to inform the guideline development, we needed to extend the evaluation of the new guidelines to other documents. To that end, a new, previously unseen sample of 12 webpages was randomly drawn from the same document collection and again annotated by the four raters. An even higher agreement of $\alpha = 0.93$ and median $\kappa = 0.88$ (min: $\kappa = 0.78$, max: $\kappa = 1$) was achieved on

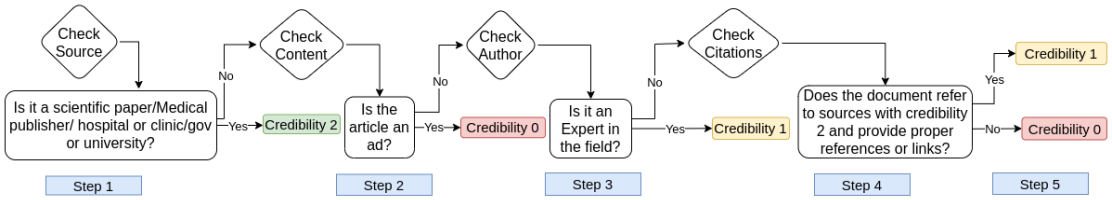

Figure 1: New proposed guidelines in flowchart form.

these new documents. Of this selection of documents, 40% were labelled with credibility 1, 37% with credibility 2 (the highest credibility), and 20% as non-credible at all. This evaluation with the new sample demonstrates the effectiveness of our guidelines and, more importantly, the general need for more specific instructions that produce more robust labels for concepts prone to subjective differences.

4.2. Evaluation with External Assessors

Four external assessors ($e1$ - $e4$), who were not involved in the design process, were recruited to further evaluate the guidelines. These assessors were recruited from our research group but were not involved in this project. They are familiar with search technologies, but they are non-experts in the medical domain. $e1$ was trained in a 15-minute conversation, in which open questions were answered. The three remaining assessors were not trained (in order to test how well the annotation process works with no prior knowledge). Krippendorff's α was then recalculated, taking both the authors' and external assessors' judgements into account. While the score decreased compared to the agreement between only the original assessors, at $\alpha = 0.72$ it is still substantial and 12% higher than on the original guidelines. However, the pairwise Cohen's Kappa scores revealed significant differences in agreement between different raters (with scores ranging between $\kappa = 0.18$ and $\kappa = 1$). As expected, $e1$ obtained a higher agreement with the authors. This suggests that credibility judgements, while subjective at first glance, can become more objective with a short training. Assessor $e3$ produced comparably low-agreement judgements, while the agreement for the remaining assessors yielded a Krippendorff's α of 0.82 and $\alpha = 0.80$ (considering all external assessors but $e3$). In practice, we could even consider to remove $e3$'s judgments, as excluding low-agreement workers is a common procedure in the literature [22]. We include a discussion of these individual differences and an error analysis in Section 5 along with some general conclusions.

4.3. Guideline Traceability

One of the goals of our new guidelines was to make raters' decisions traceable and explainable. We thus asked each annotator to note down not only a credibility label, but also the guideline they based their decision on. This allows the systematic evaluation of the quality of credibility judgements and can reveal potential misunderstandings and reasons for disagreements, enabling the incremental improvement of the guidelines. Agreement on the guidelines was calculated using Krippendorff's α for nominal scales and was high at $\alpha = 0.77$. Including the external

assessors led to a decreased agreement score of $\alpha = 0.51$. However, not considering $e3$'s judgements would lead to an increase in agreement ($\alpha = 0.59$). While agreement on the guidelines is lower than on the credibility labels, they mainly serve as an explainability tool and were used for the error analysis included in Section 5. Observing the concrete decisions over the documents, 38% were labelled using G1 (source has a scientific basis), 17% under G2 (proper citations), 22% under G3 (written by an expert), 9% under G4 (advertising purposes), 11% under G5 (written by a non-expert without citations), and surprisingly none felt under G6 category (claims against well-known medical consensus).

5. Discussion

From analysing the assessors' input, we found some differences in interpretation of the guidelines between $e3$ and the other assessors. Sources of disagreement were, whether dentistry websites should be judged as clinic sources and websites such as "*MedicineNet*" as medical publishers (G1). Thus, a possible improvement could be to remove this terminology from G1 as the term "*government website*" already encompasses health publishers like the CDC or the WHO. On the other hand, some dentistry websites displaying their number and the invitation to book an appointment next to blog articles were labelled as advertisements (G4) by some annotators whereas others interpreted them as written by medical experts (G3). Therefore, amending G4's wording to "the *website/article* is trying to sell a product, and we may conclude from its content that it is a fake" would be an improvement.

The main finding of this study is that well-defined guidelines lead to higher quality labels and more robust agreement among the judges. While there is still room for improvement in the proposed guidelines, we have observed that in our experiment even a brief teaching process can lead to more coherent label. Nevertheless, due to the limited number of reviewers and annotated documents (which are sort of unavoidable, in our framework) we have promising signals. This is a significant shift for TREC-like initiatives, raising the quality of the labels generated and increasing the realism of evaluation. In addition, we consider the process' increased cost and difficulty are both manageable.

One of the limitations of this study is that we cannot yet ascertain that the proposed guidelines sufficiently transfer to the actual credibility of a document. While they certainly reflect quality and may cover specific aspects of credibility, past research has shown that users seldom judge credibility based on source or quality [11, 23]. We plan to address this by comparing real users' subjective credibility judgements with annotations based on our guidelines. Additionally, we want to see how applying these guidelines to judge the credibility of health related websites affects experimental results from prior years and to what extent introducing these guidelines to users can improve their ability to judge the credibility and quality of websites.

6. Conclusions

In this paper, we have demonstrated the difficulty of assessing webpages in terms of credibility. Our main contribution is a set of guidelines to create robust annotations that can be further improved by providing brief training to the raters. In future work, we intend to keep polishing

these guidelines and run a user study to understand the relationship between credibility and the assigned labels in more detail. We are also interested in providing laypeople with the guidelines to see whether this improves their ability to judge the credibility/quality of web contents. We hope that the proposed tool can not only improve annotation processes for producing high-quality training data, but also have a positive impact on users' perceptions.

Acknowledgements

The authors thank the support obtained from: i) project PLEC2021-007662 (MCIN/AEI/10.13039/501100011033, Ministerio de Ciencia e Innovación, Agencia Estatal de Investigación, Plan de Recuperación, Transformación y Resiliencia, Unión Europea-Next GenerationEU), and ii) the Xunta de Galicia - Consellería de Cultura, Educación, Formación Profesional e Universidades (Centro de investigación de Galicia accreditation 2019-2022 ED431G-2019/04 and Reference Competitive Group accreditation 2022-2025, ED431C 2022/19) and the European Union (European Regional Development Fund - ERDF).

References

[1] G. Eysenbach, Infodemiology: The epidemiology of (mis) information, The American Journal of Medicine 113 (2002) 763–765.

[2] S. Y. Rieh, Judgment of information quality and cognitive authority in the web, Journal of the American society for Information Science and Technology 53 (2002) 145–161.

[3] F. A. Pogacar, A. Ghenai, M. D. Smucker, C. L. Clarke, The positive and negative influence of search results on people's decisions about the efficacy of medical treatments, in: Proceedings of the ACM SIGIR Int. Conf. on Theory of Information Retrieval, 2017, pp. 209–216.

[4] N. Vigdor, Man fatally poisons himself while self-medicating for coronavirus, doctor says, 2020. URL: https://www.nytimes.com/2020/03/24/us/chloroquine-poisoning-coronavirus.html, [accessed June 9, 2022].

[5] C. Clarke, M. Maistro, M. Smucker, Overview of the trec 2021 health misinformation track, in: Proceedings of the Thirtieth Text REtrieval Conference, TREC, 2021.

[6] C. Clarke, M. Maistro, M. Smucker, G. Zuccon, Overview of the trec 2020 health misinformation track, in: Proceedings of the Twenty-Nine Text REtrieval Conference, TREC, 2020, pp. 16–19.

[7] P. Nakov, G. Da San Martino, T. Elsayed, A. Barrón-Cedeno, R. Míguez, S. Shaar, F. Alam, F. Haouari, M. Hasanain, N. Babulkov, et al., The clef-2021 checkthat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news, in: European Conference on Information Retrieval, Springer, 2021, pp. 639–649.

[8] M. Kąkol, M. Jankowski-Lorek, K. Abramczuk, A. Wierzbicki, M. Catasta, On the subjectivity and bias of web content credibility evaluations, in: Proceedings of the 22nd international conference on world wide web, 2013, pp. 1131–1136.

[9] B. J. Fogg, Persuasive technology: using computers to change what we think and do, Ubiquity 2002 (2002) 2.

[10] A. L. Ginsca, A. Popescu, M. Lupu, et al., Credibility in information retrieval, Foundations and Trends in Information Retrieval 9 (2015) 355–475.
[11] B. J. Fogg, Prominence-interpretation theory: Explaining how people assess credibility online, in: CHI'03 extended abstracts on human factors in computing systems, 2003, pp. 722–723.
[12] D. H. McKnight, C. J. Kacmar, Factors and effects of information credibility, in: Proceedings of the ninth international conference on Electronic commerce, 2007, pp. 423–432.
[13] C. Hahnel, F. Goldhammer, U. Kröhne, J. Naumann, The role of reading skills in the evaluation of online information gathered from search engine environments, Computers in Human Behavior 78 (2018) 223–234.
[14] M. S. Eastin, Credibility assessments of online health information: The effects of source expertise and knowledge of content, Journal of Computer-Mediated Communication 6 (2001) JCMC643.
[15] J. Unkel, A. Haas, The effects of credibility cues on the selection of search engine results, Journal of the Association for Information Science and Technology 68 (2017) 1850–1862.
[16] A. Nabożny, B. Balcerzak, A. Wierzbicki, M. Morzy, M. Chlabicz, et al., Active annotation in evaluating the credibility of web-based medical information: Guidelines for creating training data sets for machine learning, JMIR medical informatics 9 (2021) e26065.
[17] A. X. Zhang, A. Ranganathan, S. E. Metz, S. Appling, C. M. Sehat, N. Gilmore, N. B. Adams, E. Vincent, J. Lee, M. Robbins, et al., A structured response to misinformation: Defining and annotating credibility indicators in news articles, in: Companion Proceedings of the The Web Conference 2018, 2018, pp. 603–612.
[18] M. Bink, S. Zimmerman, D. Elsweiler, Featured snippets and their influence on users' credibility judgements, in: ACM SIGIR Conference on Human Information Interaction and Retrieval, 2022, pp. 113–122.
[19] S. Zimmerman, A. Thorpe, C. Fox, U. Kruschwitz, Privacy nudging in search: Investigating potential impacts, in: Proceedings of the 2019 Conference on Human Information Interaction and Retrieval, 2019, pp. 283–287.
[20] M. L. McHugh, Interrater reliability: the kappa statistic, Biochemia medica 22 (2012) 276–282.
[21] K. Krippendorff, Content analysis: An introduction to its methodology, Sage publications, 2018.
[22] D. Feng, S. Besana, R. Zajac, Acquiring high quality non-expert knowledge from on-demand workforce, in: Proceedings of the 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources (People's Web), 2009, pp. 51–56.
[23] J. Caverlee, L. Liu, Countering web spam with credibility-based link analysis, in: Proceedings of the twenty-sixth annual ACM symposium on Principles of distributed computing, 2007, pp. 157–166.