Beyond Binary Classification: Ranking for Information Access in Misinformation Contexts

David La Barbera, Gian Carlo Milanese, Georgios Peikos, Gabriella Pasi and Marco Viviani*

University of Milano-Bicocca (DISCo – IKR3 Lab), Edificio U14 (ABACUS), Viale Sarca, 336 – 20126 Milan, Italy

Abstract

The spread of misinformation has catalyzed a surge of interest in systems that assist users in identifying false claims, misleading content, and other forms of unreliable information, with Automated Fact-Checking (AFC) emerging as a prominent – yet not exclusive – approach. However, the majority of current models overwhelmingly rely on binary or categorical reliability-related labels, oversimplifying the complex, context-sensitive nature of information verification. In this paper, we argue that such reductive labeling fails to align with the interpretive, epistemic, and normative dimensions of real-world misinformation detection practices, including fact-checking. Drawing on critiques from both computational and social research, we illustrate how binary judgments not only distort reliability-related aspects, but also limits system trustworthiness and public accountability. Instead, we propose a shift from binary classification to ranking-based approaches that prioritize access to information that is contextually relevant and reliable across distinct dimensions. Grounded in Information Retrieval (IR) theory, our alternative evaluation paradigm emphasizes graded relevance, contextual utility, and ambiguity tolerance, reflecting how both laypersons and professional fact-checkers alike navigate uncertainty. By reframing misinformation detection as an issue of information access, we seek to bridge the gap between technical modeling and the practical challenges of managing fake, misleading, or harmful content.

Keywords

Misinformation, Information Retrieval, Information Access, Automated Fact-Checking, Ranking, Classification

1. Introduction

In recent years, the term "misinformation" has become pivotal in the public discourse worldwide. Misinformation, broadly defined as false, misleading, or harmful information regardless of intent, poses a serious threat to democratic processes and public trust. This distinguishes it from "disinformation", which is deliberately deceptive [1], thus rendering the latter a more narrowly defined term. The Web and social media platforms have facilitated the spread of misinformation by creating an algorithmic environment that facilitates the diffusion of emotionally charged and fabricated content, often perceived as reliable. This issue has been amplified by the rise of *Large Language Models* (LLMs) [2], which are capable of rapidly generating convincing and personalized text automatically.

Such a rapid spread of misinformation has made the verification of information in online ecosystems an urgent priority in both public discourse and academic research [3]. This encompasses a broader set of practices, including the assessment of accuracy, credibility, factuality, and contextual factors such as source reliability, content coherence, and potential manipulations [4, 5]. Regarding the fundamental concept of factuality (also referred to as factual accuracy), it denotes the degree to which information is grounded in objective facts and can be empirically verified against reality. Within this framework, fact-checking is the systematic process of assessing the factuality of claims to determine their truthfulness, which can be used within broader misinformation detection processes. However, fact-checking remains largely human-driven, making it time-consuming and resource-intensive. Consequently, organizations often struggle to keep up with the volume of claims, spurring interest in *Automated Fact-Checking* (AFC) systems to address misinformation at scale [6, 7].

Ital-IA 2025: 5th National Conference on Artificial Intelligence, organized by CINI, June 23-24, 2025, Trieste, Italy
*Corresponding author.
david.labarbera@unimib.it (D. La Barbera); giancarlo.milanese@unimib.it (G. C. Milanese); georgios.peikos@unimib.it (G. Peikos); gabriella.pasi@unimib.it (G. Pasi); marco.viviani@unimib.it (M. Viviani)
0000-0002-8215-5502 (D. La Barbera); 0000-0002-1757-0152 (G. C. Milanese); 0000-0002-2862-8209 (G. Peikos); 0000-0002-6080-8170 (G. Pasi); 0000-0002-2274-9050 (M. Viviani)

Despite technical progress in misinformation detection in general, and AFC in particular, a critical flaw persists: the field remains anchored in an oversimplified modeling paradigm that reduces complex epistemic judgments to *binary labels* (e.g., "true"/"false") [8], often by formulating the task at hand as a *binary classification* problem [9]. Although this simplification is convenient for supervised learning and benchmarking [10, 11], it does not capture the complexity of reliability as a concept in general, nor of "truth" in practice, which is often situated, partial, contested, or context-dependent [12, 13]. In the context of AFC, in particular, the mismatch between AFC's modeling assumptions and the interpretive realities of human fact-checking reveals a deeper epistemic and methodological gap. Human fact-checking is not merely a binary classification task, but a sociotechnical process involving the monitoring of diverse information channels, consultation with domain experts, management of ambiguity, and the maintenance of transparency to foster public trust [14]. The field's normative commitments – including the obligation to explain verdicts and to contextualize claims – are largely absent in AFC's current design and evaluation frameworks.

In this paper, we argue that real progress in misinformation detection requires moving beyond binary classification toward evaluation frameworks rooted in *Information Retrieval* (IR) theory. These approaches prioritize graded relevance, user-centered utility, and contextual sensitivity, allowing for partial correctness and epistemic nuance. Unlike rigid label-based systems, IR-inspired methods better reflect how both laypersons and professional fact-checkers evaluate information.

2. The Limits of Binary Classification in the Age of Misinformation

A growing body of work has criticized the use of binary labels, both in general misinformation detection and in fact-checking specifically, for flattening the epistemic and social complexity of truth and other related concepts. While convenient for automated classification, such dichotomous judgments fail to reflect the context-sensitive and negotiated nature of knowledge. As Aghajari et al. [15] show, users often perceive binary verdicts to be mismatched with their understanding of truth, particularly in domains where facts change over time or require interpretation. This discrepancy imposes a rigidity that risks undermining trust and may distort informational nuance.

The inadequacy of binary classification is especially clear at finer textual levels. As noted by Hamed et al. [16], documents often contain a mixture of truthful and misleading statements, which resist categorical summarization. Similarly, Thorne and Vlachos [17] underscore the importance of external context and background knowledge in fact-checking – factors that binary systems systematically fail to accommodate. Beyond technical concerns, labels like "true" or "fake" exert normative force. As argued by Altay et al. [18], such labels do not merely describe but shape public perception, assign blame, and justify interventions. In this light, the very categories used in AFC warrant critical reflection. Ultimately, the problem is epistemological. Scientific and social knowledge rarely lends itself to definitive judgments: even well-established facts emerge through processes of consensus and contestation [19].

To address this oversimplification, some authors have proposed finer-grained scales [20], inspired by fact-checking organizations that adopt nuanced taxonomies to capture the complexity of real-world claims. For instance, Snopes uses a diverse set of ratings, including labels such as True, Mostly True, Mixture, False, Unproven, and Outdated. This allows fact-checkers to indicate degrees of factual accuracy, the presence of misleading context, or the absence of verifiable evidence.[1] Similarly, PolitiFact relies on six distinct ratings: True, Mostly True, Half True, Mostly False, False, and the emphatic Pants on Fire.[2] The use of more than two labels is a practice also adopted in well-known datasets. Two notable datasets are Liar [21], based on PolitiFact, and FEVER [22]. Despite having more labels, both have flaws: Liar's short statements lack full context, limiting veracity assessment, while FEVER's claims, generated by modifying Wikipedia sentences, often miss the style and structure of real misinformation [23].

[1] https://www.snopes.com/fact-check-ratings/

[2] https://www.politifact.com/article/2018/feb/12/principles-truth-o-meter-politifacts-methodology-i/#How%20we%20choose%20claims

Some authors have tried to go beyond the idea of classification. Ahmadi et al. [24] applied Probabilistic Answer Set Programming (LP^{MLN}) to fact-checking by translating knowledge graph facts and mining Web data into weighted logical rules. The outcome is a probabilistic verdict with transparent derivation, rather than a true/false label. Despite its explainability, such an approach raises questions about its scalability due to the challenge of maintaining an up-to-date knowledge graph. Sufi and Alsulami [25], on the other hand, proposed decomposing complex claims into atomic semantic units and evaluating each probabilistically. Each unit's veracity is weighted by evidence quality and source credibility, yielding interpretable, granular scores per unit.

While these approaches move beyond rigid classification and enable more expressive veracity assessments, they involve trade-offs. Probabilistic and logic-based systems often depend on structured inputs, curated evidence, or expert rules, limiting scalability in open or fast-evolving domains. Though soft outputs may aid interpretability, they are often difficult to calibrate, compare, or align with user expectations – especially where ranked access to trustworthy information is key. Moreover, most models remain disconnected from IR, lacking mechanisms to retrieve or rank evidence, and thus fail to reflect how users and fact-checkers assess claims through comparative evaluation of competing information.

3. Reframing Binary Classification for Information Access

The dominant framing of misinformation detection, especially AFC, as a classification task often overlooks that the real challenge is facilitating access to reliable information. From this perspective, the goal is not merely to categorize claims, documents, or other information as, for example, true or false, but to meet the need for relevant, trustworthy, and context-appropriate information. This reframing aligns with IR principles, which view information-seeking as a dynamic, context-sensitive process to satisfy user queries.

Rather than generating fixed verdicts, an IR approach to misinformation conceptualizes the identification of reliable information as a *retrieval task*. Given a claim or an information need, the system returns a ranked list of documents – such as news articles, policy statements, or expert analyses – organized according to their relevance to the user's intent. Crucially, relevance in this context, as in many others, is a multidimensional concept [26]. In fact, in addition to *topicality* (i.e., whether the document addresses the claim), reliability-related dimensions such as *credibility*, *correctness*, *source reliability*, etc., become essential ranking signals. Ranking-based approaches informed by these dimensions can better reflect the variability and partiality of real-world information, while aligning more closely with how average individuals and professional fact-checkers interpret complex pieces of information.

In this context, some previous works have attempted to frame misinformation detection as a *ranking task* [20]. For example, between 2019 and 2022, the TREC *Health Misinformation* (HM) track served as a pivotal benchmark for investigating retrieval-based approaches to combating health-related misinformation [27]. This task challenged participating systems to move beyond simple topical relevance by incorporating additional dimensions, such as credibility – primarily related to the information source – and correctness – mainly associated with factual accuracy – into their document assessments. Within the track, many systems leveraged machine learning techniques (e.g., [28, 29]), including claim verification models, stance detection methods, and classifiers trained on credibility- or correctness-related features, to estimate the likelihood that a document is trustworthy and/or factually accurate. Alternative solutions – either as part of the track itself or by leveraging the datasets it released – involved comparing retrieved documents against expert-authored reference materials (e.g., [30, 31]). Within this line of research, the system presented in [32] computes a "genuineness" score for each candidate document by measuring its similarity to the top-k scientific articles retrieved for the same query, which is then combined with traditional topical relevance scores to produce the final ranking.

Despite sustained efforts, current IR systems considering reliability-related aspects still face notable limitations. For example, the top-performing submission in the most recent TREC HM track [33] employs a multi-stage neural re-ranking architecture that combines stance prediction – using the *Vera*

model [34] – with LLMs such as GPT-3 and GPT-4 for query reformulation and answer generation. While this system achieves improved performance on benchmark metrics, it also introduces new challenges related to the reliability and interpretability of its outputs. In particular, the use of LLMs for answer generation raises concerns about factual consistency, as these models are known to produce inaccurate or misleading content – commonly referred to as *hallucinations* – which may pose risks to users [35]. Another critical limitation is that LLMs require external support to update their underlying knowledge. This has led to the adoption of *Retrieval-Augmented Generation* (RAG) systems [36] as a promising strategy to increase the likelihood of factual accuracy in generated responses, particularly in high-stakes domains such as health, as demonstrated in [37]. However, even RAG-based approaches cannot guarantee full factual correctness [38, 39].

Recent work has explored *knowledge-driven* alternatives to LLM-based approaches by leveraging structured sources of trusted information to assess factual correctness. For instance, [40] proposes an IR system that evaluates the factual accuracy of documents by verifying health-related claims against a curated medical knowledge graph. This graph-based representation enables precise alignment between textual claims and authoritative medical knowledge. Given that Web documents often contain a mixture of accurate and potentially misleading content, the system selectively focuses its correctness assessment on claims most relevant to the user's query. This targeted evaluation helps to minimize the influence of tangential or irrelevant information. However, the effectiveness of such knowledge-driven approaches is contingent on the availability of comprehensive, high-quality knowledge graphs—resources that require ongoing expert curation and may not always reflect the most up-to-date medical evidence.

Another key challenge for current IR in misinformation contexts is the semantic ambiguity surrounding relevance dimensions related to truthfulness or, more generally, reliability. As a recent review of multidimensional ranking systems notes [26], terms like credibility, reliability, trustworthiness, objectivity, authority, and factuality are often used inconsistently. These related dimensions are frequently treated as interchangeable or arranged hierarchically—yet with varying definitions across studies. For example, reliability is sometimes seen as encompassing credibility and factuality [41], while other work frames credibility as the broader construct, including dimensions like factuality and objectivity [42]. In addition to this conceptual ambiguity, there is significant variation in the computational strategies used to estimate these dimensions. Some methods rely on source-level metadata (e.g., publisher reputation or author credentials), whereas others derive signals directly from the content of the documents (e.g., linguistic features or semantic consistency). This definitional and methodological heterogeneity hinders standardization and comparability, highlighting the need for a more unified and systematic framework to guide future research on reliability-related relevance dimensions in retrieval settings.

Moreover, another critical limitation of many IR-based systems designed to address misinformation lies in their limited support for *explainability*, an aspect that remains relatively underexplored in the field. While such systems often focus on optimizing retrieval effectiveness or, for example, factual accuracy, they typically offer minimal insight into why a particular document is considered trustworthy or untrustworthy. This lack of interpretability can hinder user trust and reduce the system's practical utility in high-stakes domains such as health or science. Recent advances in explainable health IR have begun to address this gap. For example, some systems annotate ranked results with supporting scientific evidence drawn from authoritative sources [43], while others establish semantic links between document claims and structured representations in knowledge graphs [40]. Despite their promise, these approaches also face limitations – particularly when attempting to justify negative assessments. In many cases, the absence of supporting evidence does not necessarily indicate that a claim is false, but may instead reflect the incompleteness of the underlying knowledge base. This challenge is especially pronounced in knowledge-graph-based systems, where coverage gaps are well-documented.

However, despite their current limitations, IR-based systems for addressing misinformation represent a promising direction for moving beyond the constraints of traditional frameworks. By shifting the focus from binary labeling to the retrieval and ranking of evidence along dimensions such as topical relevance, credibility, factual accuracy, etc., these systems more accurately reflect the interpretive and comparative nature of real-world information assessment.

4. Conclusions

Misinformation detection systems based on binary classification fail to capture the epistemic complexity and contextual variability of misinformation. This paper reframes misinformation detection as an information access problem, highlighting the potential of IR approaches to deliver more nuanced, user-centered responses. We advocate for systems that not only rank information according to distinct reliability-related dimensions (e.g., credibility, factual accuracy, and more), but also provide transparent explanations and assist users in navigating uncertainty. Future research should focus on integrating RAG architectures, designing explainable interfaces, and developing evaluation protocols that reflect the graded, interpretive nature of reliability-related labels.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT and Grammarly in order to: paraphrase and reword, grammar and spelling check. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] E. Broda, J. Strömbäck, Misinformation, disinformation, and fake news: lessons from an interdisciplinary, systematic literature review, Annals of the Int. Comm. Assoc. 48 (2024) 139–166.

[2] M. Garry, W. M. Chan, J. Foster, L. A. Henkel, Large language models (LLMs) and the institutionalization of misinformation, Trends in cognitive sciences (2024).

[3] G. Livraga, M. Viviani, Data confidentiality and information credibility in on-line ecosystems, in: Proceedings of the 11th international conference on management of digital ecosystems, 2019, pp. 191–198.

[4] H. Suominen, L. Goeuriot, L. Kelly, L. A. Alemany, E. Bassani, N. Brew-Sam, V. Cotik, D. Filippo, G. González-Sáez, F. Luque, et al., Overview of the clef ehealth evaluation lab 2021, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2021, pp. 308–323.

[5] R. Upadhyay, G. Pasi, M. Viviani, Health misinformation detection in web content: a structural-, content-based, and context-aware approach based on web2vec, in: Proceedings of the conference on information technology for social good, 2021, pp. 19–24.

[6] I. Vykopal, M. Pikuliak, S. Ostermann, M. Simko, Generative Large Language Models in Automated Fact-Checking: A Survey (2024). arXiv:2408.14317.

[7] L. Hu, S. Wei, Z. Zhao, B. Wu, Deep learning for fake news detection: A comprehensive survey, AI Open 3 (2022) 133–155.

[8] L. Graves, Understanding the promise and limits of automated fact-checking, 2018.

[9] G. Warren, I. Shklovski, I. Augenstein, Show me the work: Fact-checkers' requirements for explainable automated fact-checking, CHI '25, 2025.

[10] C. Thibault, J.-J. Tian, G. Peloquin-Skulski, T. L. Curtis, J. Zhou, F. Laflamme, Y. Guan, R. Rabbany, J.-F. Godbout, K. Pelrine, A guide to misinformation detection data and evaluation, 2025. arXiv:2411.05060.

[11] G. K. M, K. S. Faizz Ahmad, S. G. Pamidimukkala, A. P. Sathe, S. G.N.V.G, S. R. M, K. Ch, Hybrid optimization driven fake news detection using reinforced transformer models, Scientific Reports 15 (2025) 14782.

[12] P. B. Brandtzaeg, A. Følstad, M. Ángeles Chaparro Domínguez and, How journalists and social media users perceive online fact-checking and verification services, Journalism Practice 12 (2018) 1109–1129.

[13] S. M. Williamson, V. Prybutok, The era of artificial intelligence deception: Unraveling the complexities of false realities and emerging threats of misinformation, Information 15 (2024).

[14] H. Liu, A. Das, A. Boltz, D. Zhou, D. Pinaroc, M. Lease, M. K. Lee, Human-centered nlp fact-checking: Co-designing with fact-checkers using matchmaking for ai, Proc. of the ACM on Human-Computer Interaction 8 (2024) 1–44.
[15] Z. Aghajari, E. P. S. Baumer, D. DiFranzo, Reviewing interventions to address misinformation: The need to expand our vision beyond an individualistic focus, Proc. ACM Hum.-Comput. Interact. 7 (2023).
[16] S. K. Hamed, M. J. Ab Aziz, M. R. Yaakub, A review of fake news detection approaches: A critical analysis of relevant studies and highlighting key challenges associated with the dataset, feature representation, and data fusion, Heliyon 9 (2023) e20382.
[17] J. Thorne, A. Vlachos, Automated fact checking: Task formulations, methods and future directions, in: Proc. of the 27th International Conference on Computational Linguistics, 2018, pp. 3346–3359.
[18] S. Altay, M. Berriche, A. Acerbi, Misinformation on misinformation: Conceptual and methodological challenges, Social Media + Society 9 (2023) 20563051221150412.
[19] L. Graves, Anatomy of a Fact Check: Objective Practice and the Contested Epistemology of Fact Checking, Communication, Culture and Critique 10 (2017) 518–537.
[20] Z. Guo, M. Schlichtkrull, A. Vlachos, A survey on automated fact-checking, Transactions of the ACL 10 (2022) 178–206.
[21] W. Y. Wang, "liar, liar pants on fire": A new benchmark dataset for fake news detection, ACL, Vancouver, Canada, 2017.
[22] J. Thorne, A. Vlachos, C. Christodoulopoulos, A. Mittal, FEVER: a large-scale dataset for fact extraction and VERification, ACL, New Orleans, Louisiana, 2018.
[23] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, Y. Choi, Truth of varying shades: Analyzing language in fake news and political fact-checking, ACL, 2017, pp. 2931–2937.
[24] N. Ahmadi, J. Lee, P. Papotti, M. Saeed, Explainable fact checking with probabilistic answer set programming, CoRR (2019). arXiv:1906.09198.
[25] F. Sufi, M. Alsulami, Quantifying truthfulness: A probabilistic framework for atomic claim-based misinformation detection, Mathematics (2025).
[26] G. Peikos, G. Pasi, A systematic review of multidimensional relevance estimation in information retrieval, WIREs Data. Mining. Knowl. Discov. 14 (2024).
[27] C. L. A. Clarke, S. Rizvi, M. D. Smucker, M. Maistro, G. Zuccon, Overview of the TREC 2020 health misinformation track, volume 1266, 2020.
[28] R. Pradeep, X. Ma, X. Zhang, H. Cui, R. Xu, R. Nogueira, J. Lin, H2oloo at trec 2020: When all you got is a hammer... deep learning, health misinformation, and precision medicine, Corpus 5 (2020).
[29] L. C. Lima, D. B. Wright, I. Augenstein, M. Maistro, University of copenhagen participation in trec health misinformation track 2020, arXiv preprint arXiv:2103.02462 (2021).
[30] S. N. Gonçalves, F. Martins, Voh. colab at trec 2020 health misinformation track., in: TREC, 2020.
[31] I. B. Schlicht, A. F. M. de Paula, P. Rosso, Upv at trec health misinformation track 2021, in: The Thirtieth REtrieval Conference Proc. (TREC 2021) NIST, 2022, pp. 500–335.
[32] R. Upadhyay, G. Pasi, M. Viviani, An unsupervised approach to genuine health information retrieval based on scientific evidence, in: Web Information Systems Engineering – WISE 2022: 23rd International Conference, Biarritz, France, November 1–3, 2022, Proc., 2022.
[33] R. Pradeep, J. Lin, Towards automated end-to-end health misinformation free search with a large language model, in: European Conference on Information Retrieval, Springer, 2024, pp. 78–86.
[34] R. Pradeep, X. Ma, R. Nogueira, J. Lin, Vera: Prediction techniques for reducing harmful misinformation in consumer health search, in: Proc. of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021, pp. 2066–2070.
[35] O. Freyer, I. C. Wiest, J. N. Kather, S. Gilbert, A future role for health applications of large language models depends on regulators enforcing safety standards, The Lancet Digital Health (2024).
[36] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, H. Wang, Retrieval-augmented generation for large language models: A survey, arXiv preprint arXiv:2312.10997 (2023).
[37] R. Upadhyay, M. Viviani, Enhancing Health Information Retrieval with RAG by prioritizing topical relevance and factual accuracy, Discover Computing 27 (2025).

[38] A. Gilson, X. Ai, T. Arunachalam, et al., Enhancing large language models with domain-specific retrieval augment generation: A case study on long-form consumer health question answering in ophthalmology, CoRR (2024). arXiv:2409.13902.
[39] D. Johnson, R. Goodman, J. Patrinely, C. Stone, E. Zimmerman, R. Donald, S. Chang, S. Berkowitz, A. Finn, E. Jahangir, et al., Assessing the accuracy and reliability of ai-generated medical responses: an evaluation of the chat-gpt model, Research square (2023).
[40] G. C. Milanese, G. Peikos, G. Pasi, M. Viviani, Fact-driven health information retrieval: Integrating LLMs and knowledge graphs to combat misinformation, in: Advances in Information Retrieval, 2025.
[41] M. Fernández-Pichel, D. E. Losada, J. C. Pichel, A multistage retrieval system for health-related misinformation detection, Eng. Appl. Artif. Intell. 115 (2022) 105211.
[42] C. Lioma, B. Larsen, W. Lu, Y. Huang, A study of factuality, objectivity and relevance: three desiderata in large-scale information retrieval?, 2016.
[43] R. Upadhyay, P. Knoth, G. Pasi, M. Viviani, Explainable online health information truthfulness in consumer health search, Frontiers in Artificial Intelligence 6 (2023).