Improving the Reliability of Health Information Credibility Assessments

Marcos Fernández-Pichel (marcosfernandez.pichel@usc.es)

Selina Meyer
Markus Bink
Alexander Frummet
David E. Losada
David Elsweiler

CiTIUS, Universidade de Santiago de Compostela and Chair for Information Science, Universität Regensburg
2nd April 2023, Dublin, ROMCIR'23

Universität Regensburg

Misinformation increasingly common
(Eysenbach, 2002)

👆 Can cause harmful consequences to people interacting with such info (Pogacar, 2017)

Shared-task competitions
(Clarke, 2020) (Clarke, 2021) (Nakov, 2021)

🏅 Impact restricted to the **quality training** data

Man Fatally Poisons Himself While Self-Medicating for Coronavirus, Doctor Says

An Arizona man and his wife ingested a fish tank cleaning additive made with the same active ingredient as chloroquine phosphate, which President Trump has referred to as a "game changer."

 (Vigdor, 2020)

Generate ground truth data is a **crucial** and **costly** process

Requires **human intervention**

TREC HM, label documents on topical relevance, **credibility** and correctness

Credibility, highly subjective and individual differences

Credibility (as defined by TREC HM): doc trustworthiness and authoritativeness, perceived by the assessors

Robust guidelines to clarify the process and produce solid benchmarks

Subjectivity *inherent to* **credibility judgments** *demands clear and specific guidelines for the development of* **test collections**

Methodology:

1) We apply **TREC HM guidelines** to a collection (Zimmerman, 2019) and evaluate agreement across raters

2) Identify **reasons** for **disagreement**

3) Create **new guidelines** that: lead to higher inter-annotator reliability and can inform about why a rater made a specific decision (traceability)

In TREC HM, assessors are provided with a list of guidelines

Information about the **number of assessors** and **agreement**, **not publicly available**

Independently judge 12 random documents from the medical domain (Zimmerman, 2019)

Pairwise Cohen's Kappa (0.25 - 0.79) with a median of κ=0.44

Krippendorff's α=0.66

Group meeting to identify the problems

Three main reasons for disagreement:

Lengthy and **unstructured** guides (i.e. *"Try to determine the amount of expertise, authoritativeness, and trustworthiness of a document"*)

Lack of **clear-cut** between credibility levels

Ambiguous concepts, not defined (i.e., *expertise, trustworthiness, ubiquity, etc.*)

A Robust and Traceable Set of Credibility Guidelines

	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.

	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

A Robust and Traceable Set of Credibility Guidelines

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.

A Robust and Traceable Set of Credibility Guidelines
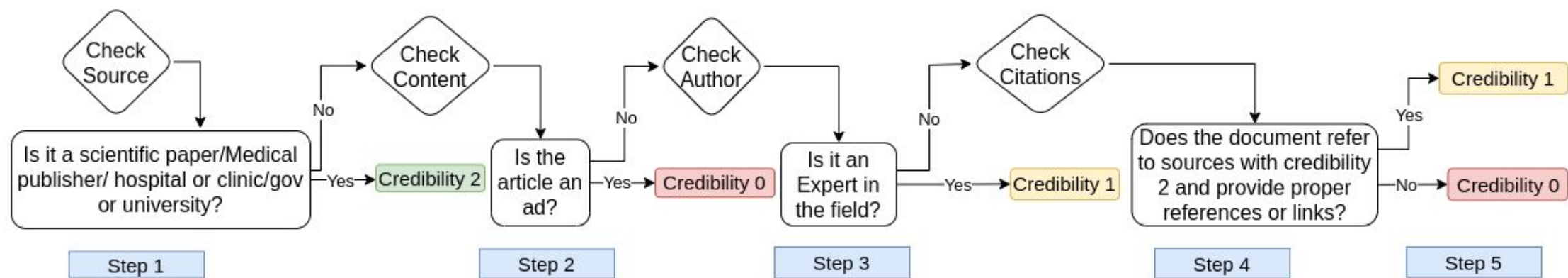
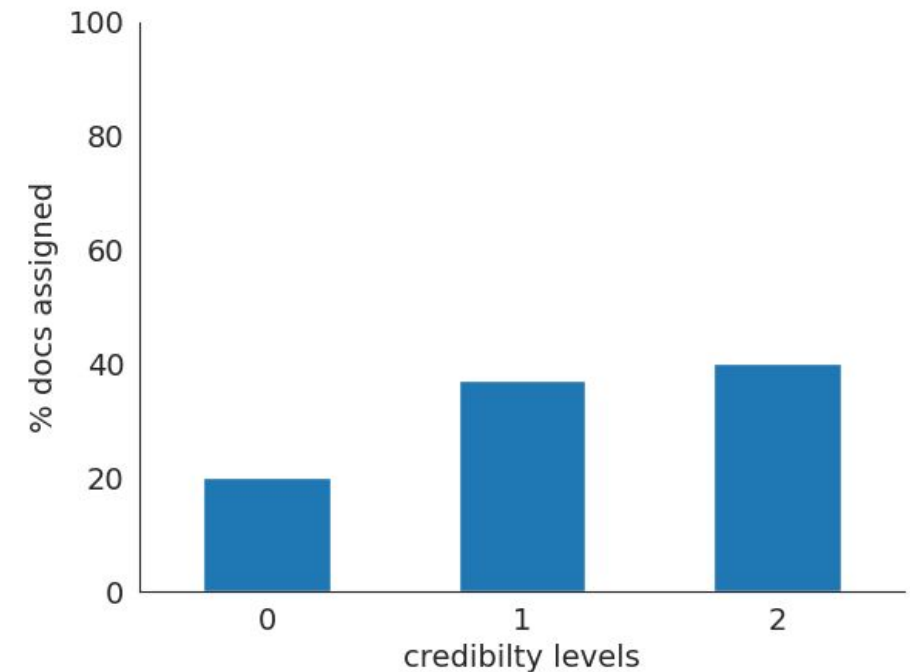
The same 12 documents annotated again by the same 4 raters

Krippendorff's **α**=0.88 (↑28%)

Pairwise Cohen's Kappa with a median of **κ**=0.89 (0.78 - 1)

Avoid bias familiarity with previous documents

Even higher Krippendorff's α=0.93

Pairwise Cohen's Kappa with a median of κ=0.88 (0.78 - 1)

Four external assessors (*e1-e4*), who were not involved in the design process

e1 **trained** in a 15-minute conversation with open questions

✏️ Taking both external and authors', Krippendorff's **α**=0.72

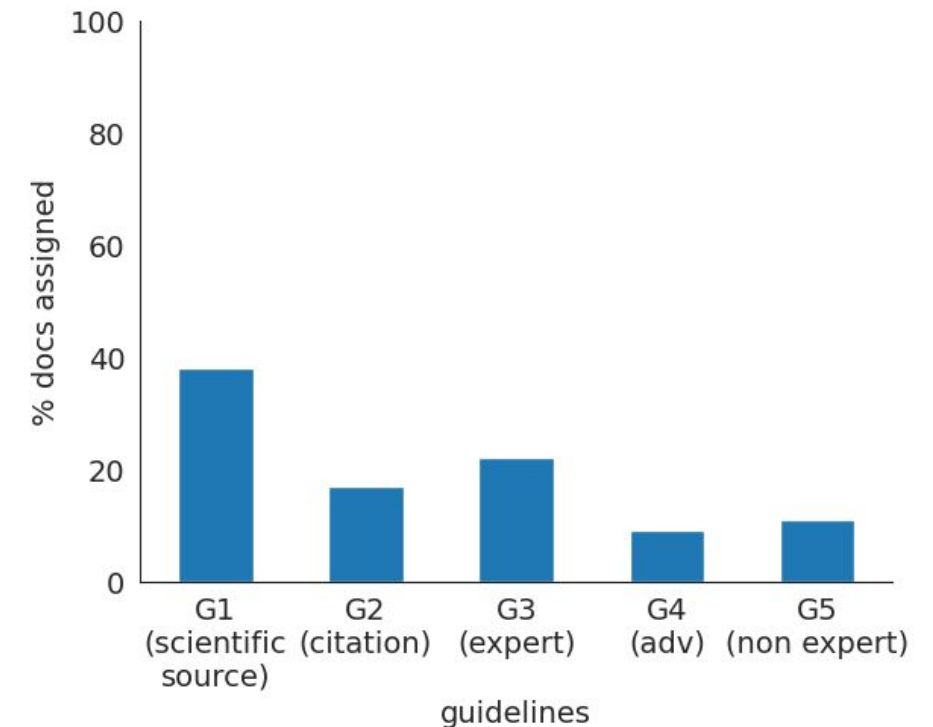⚠️ Pairwise Cohen's Kappa, important differences between raters (0.18-1):

      e1 higher agreement with the authors
      e3 low-agreement judgements

Guideline Traceability

We asked each annotator to note down the guideline

Agreement Krippendorff's α=0.77, but including the external α=0.51

Lower agreement, but explainability tool

Some annotators labelled dentistry websites as credibility 0

Amend G4 to *the website is trying to sell, and we may conclude it is a fake*

Room for **improvement**, not a definitive proposal

Significant **shift** for TREC-like initiatives

We cannot yet ascertain that these transfer to user perception of credibility

Difficulty of assessing webpages in terms of credibility

Set of **guidelines** to create **robust annotations**

Brief **training process** to the raters can be positive

Keep **polishing** these guidelines

Improve the **user perception** of what is credible