Review of the dissertation

# “Credibility Evaluation of Online Health Information using Human in the Loop Machine Learning”

by Aleksandra Nabożny

Reviewer: Michal Ptaszynski,
Kitami Institute of Technology, Japan

Kitami, Japan, 2023/05/29

## General overview

The doctoral dissertation titled "Credibility Evaluation of Online Health Information using Human in the Loop Machine Learning" addresses the pressing need for evaluating the credibility of medical content on the Internet. The dissertation effectively highlights the challenges posed by the lack of relevant data and emphasizes the importance of human-expert involvement in decision-making to mitigate the risks associated with false medical recommendations. The author also takes up the challenge to create an expert-supported, semi-automated system for capturing and tagging unreliable medical texts.

## Experimental Design and Data Collection

The dissertation showcases a commendable effort in conducting three comprehensive experiments to collect the necessary data for evaluating online health information. The inclusion of multiple analyses, each described in a separate article attached to the dissertation, demonstrates a systematic approach to research. The initial experiment focusing on single sentences and their assessment with and without context provides valuable insights into the challenges faced by domain experts. The subsequent experiment exploring different methods for enriching sentence context exhibits an adaptive approach, highlighting the author's commitment to refining the evaluation process. However, further details regarding the experimental design, such as sample size per expert and selection criteria, would enhance research transparency.

## Identification of an Efficient Unit of Text

One of the dissertation's notable contributions is the identification of an efficient unit of text for evaluating online health information. Defining a unit consisting of three consecutive sentences with keywords not only enables more accurate assessments but also streamlines the evaluation process. This finding is particularly valuable as it bridges the gap between the need for human involvement and the necessity for semi-automation. However, additional explanation regarding the

selection of this specific unit and its validation for different medical domains would strengthen the argument and bolster the credibility of the dissertation.

# Critical Analysis and Future Directions

While the dissertation presents compelling findings, a more extensive critical analysis would further enrich the research. Much of the analysis is already included in the attached papers, but, since these are mostly conference papers written within a specific page limit, an encompassing analysis within the main text of the dissertation would help grasp the contributions of the study. Moreover, although the author addresses a number of limitations, a deeper and more direct analysis of present and potential limitations and challenges encountered during the experiments, as well as discussing alternative approaches or methodologies, would contribute to a more well-rounded study. Additionally, suggesting more realistic and straightforward future directions and potential advancements in the field would provide a roadmap for subsequent research endeavors. For specific comments regarding the improvement of the dissertation, please refer to the “Specific Comments and Questions” section below.

# Clarity of Presentation

The dissertation exhibits clarity in conveying the study's overall objectives and outcomes. The writing style is mostly precise and concise, allowing readers to grasp the main points effortlessly. However, certain aspects, such as the specific techniques employed for semi-automation and the role of machine learning, could be further elaborated to enhance understanding. Comments regarding the clarification of several paragraphs were attached in the “Specific Comments and Questions” section below.

# Conclusion

In conclusion, the doctoral dissertation on "Credibility Evaluation of Online Health Information using Human in the Loop Machine Learning" offers a substantial contribution to the field of evaluating health-related content on the Internet. The systematic approach to data collection, the identification of an efficient unit of text, and the emphasis on human involvement in the evaluation process are commendable. Addressing certain areas for improvement, such as clarifying the meaning of the most important terms used in the dissertation (credibility, online health information, human in the loop machine learning, etc.), or providing supportive evidence for several decision choices (e.g., the three-sentence long context considered as optimal), would enhance the dissertation's overall impact.

# Specific Comments and Questions

## Overall

1. Credibility is taken in this dissertation for granted as a concept equally understood by both - the authors, the readers of the dissertation, and the experts involved in the annotation. However, it is not so obvious, and as the main concept of the whole dissertation should be explained with caution and care. For example, typically, credibility refers to the status of an entity (e.g., a bank, an organization, a website - things that can have some background to

provide that credibility). Thus typically a set of three out-of-context sentences cannot be credible *per se*. Perhaps the author of those sentences could be a credible person, or the publication where those three sentences appeared could be a credible publication venue, but not the sentences alone. For such a social context independent span of text, like one or three sentences, usually one would talk about “soundness”, “being convincing”, or simply “true or untrue”. Thus, since the author uses the word “credible” with a specific meaning, this meaning needs to be clearly defined in the dissertation.

2. It is not certain what is meant by “online health information”. As this is the second most important concept in the dissertation, it should be properly defined and used consistently. First of all, the term is immediately confusing, because it can be parsed in two ways: [Online [health information]] or [[Online health] information]. The interpretation used in this dissertation is the first one but it is just as reasonable to interpret it as “information about online health”, which could refer to the literal physical health of a human but could also refer to a state of online community (e.g., good online health of an online community means that it is in stable growth, has little harmful information, and few toxic users).
   Moreover, „medical content” is considered here the same as „online health information”. Not every health-related information on the Internet is medical content. For example, a depression patient writing how they feel recently on a private blog is not medical information but is health-related information. This needs to be defined, corrected, or - at best - unified to either one - if it is to be considered an important technical term for this dissertation.

3. The notion of “human in the loop ML” is used in the dissertation as a somehow unique concept, not widely applied, which is misleading, and requires proper clarification. First of all, the notion of “human in the loop ML” typically refers to a situation where the human is actively involved in the process of improving the ML model (for a good summary of HITL-ML field see: Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J., & Fernández-Leal, Á. (2023). Human-in-the-loop machine learning: A state of the art. Artificial Intelligence Review, 56(4), 3005-3054.). However, in this dissertation, the “human in the loop” simply means that humans-experts provided annotations for the collected data. This is the overwhelmingly common situation in ML, especially in NLP, and represents the most typical example of passive learning - which is the antithesis of HITL-ML. If the author wishes to place her work within the actual HITL-ML, she must clearly explain what in the process will mean that the human is “in the loop”. For example, how will the experts be continuously involved in improving the ML system? Will they look separately at high probability yet erroneous classifications? Will they look at low probability classifications and discuss extending the context for classification for some cases? How actively will the human-experts be involved in the final system?

4. Regarding the fixed three-sentence frame as the optimal context length for health-related information.
   It is highly dubious whether such a fixed sentence span would be optimal for all health-related information. It is rarely the case that “one size fits all”, especially in health-related fields. Instead of proposing a supposedly universally optimal sentence text/span, it would be more useful to specify different text spans depending on the field (e.g., pediatry vs. psychology). What if a statement only becomes fully understandable after 4 sentences? Also, how can we be sure that those three sentences are always the most relevant?
   Additionally, in: *“Chapter 1. Introduction”*, *“1.2.1 Defining an optimal unit for labeling online health information in terms of credibility.”*: *“These shorter text units could be selected with a recommendation algorithm for expert annotation.”*

What if the relevant context sentences are scattered around the article? The present approach either assumes that the relevant sentences will always appear in a row, which is wrong, or that there will be a system responsible for the extraction of such relevant sentences for the annotator to annotate. Since this is a potentially separate research topic, it should be mentioned in Future Works.
Moreover, the use of a single sentence as a unit here is also dubious. A sentence could be very short. Word-long. If the length of a sentence is not taken into consideration, this could cause a lack of context. Why not a paragraph? Paragraphs have been used widely in research applying medical/clinical data. See for example the following research:
- Lee, M., Cimino, J., Zhu, H. R., Sable, C., Shanker, V., Ely, J., & Yu, H. (2006). Beyond information retrieval—medical question answering. In AMIA annual symposium proceedings (Vol. 2006, p. 469). American Medical Informatics Association.
- Zhai, H., Lingren, T., Deleger, L., Li, Q., Kaiser, M., Stoutenborough, L., & Solti, I. (2013). Web 2.0-based crowdsourcing for high-quality
- August, T., Wang, L. L., Bragg, J., Hearst, M. A., Head, A., & Lo, K. (2022). Paper plain: Making medical research papers approachable to healthcare consumers with natural language processing. arXiv preprint arXiv:2203.00130.

# In: Abstract

5. *“Evaluating the credibility of medical content on the Internet is becoming increasingly urgent in the 21st century.”*
   Why? What makes it so urgent? There needs to be a logical consequence in reasoning. Therefore, this part needs an additional sentence explaining why this is an urgent case. For example, maybe there has been an increase in health-related information on the Internet or an increase in health-related conspiracy theories. This either needs to be explained in the dissertation, or a reference should be given to in which paper exactly it is explained.

6. *“However, countless amounts of data published daily online do not allow for manual evaluation of their content by domain experts.”*
   Why would domain experts even need to evaluate such information? Is there a bureau that does that as part of its work? If not - then no expert is obliged to evaluate anything on the Internet unless they are specifically hired to do so. If you wish to propose creating such an enterprise - to have experts constantly sit somewhere and evaluate the reliability of online health-related articles, regardless if it even is realistic or not - this needs to be put as one of the main premises in the abstract.

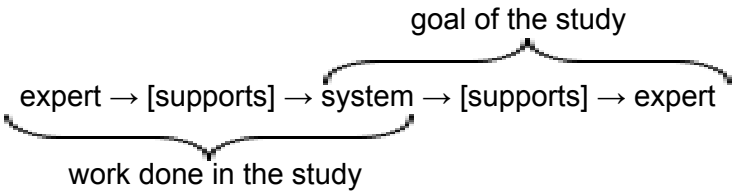
7. *“decisions based on false medical recommendations can be so severe that the final classification of credibility ought to be made by a human.”*
   How severe exactly? In the abstract give at least a simple example, and expand on that in the dissertation. Also, eventually, it is always the human who makes the decision to either use the piece of medical advice or not. So, „human” in this context should be replaced with „expert” because it is not just „any human”.

8. *“This work takes the essential steps toward creating an expert-supported, semi-automated system for capturing and tagging unreliable medical texts appearing on the Web.”*
   If eventually the final decision is done by an expert, then the work only asks the question „to what extent the work of an expert can be made more efficient/automated with technology?”. However, if it is still the expert that has to make the final decision, then there either already has to be an environment in action where such experts already exist and actively work to evaluate

the medical content / online health information, or such an environment has to be proposed, e.g., as one of the future goals of this study.

9. *“The second article also describes an analysis that detects rhetorical patterns that mislead experts, distorting their credibility assessment.”*
   This is one of the most interesting gems of this study and should be highlighted more. Typically experts are considered infallible, however, it is well known, that even two doctors can have different opinions about one case (thus the importance of the “second opinion”). It should be highlighted more that the experts can also be fooled, and therefore need assistance to make better decisions. This idea should also be mentioned in Future Works as one of the important future paths. The fact that experts can be fooled means they could be fooled by someone on purpose, especially if that person used the rhetorical patterns analyzed in the study. Especially, this idea could be used in the education of both laypeople and experts. Moreover, this could be extended to the analysis of such phrases or patterns by generative language models applied in writing health-related content automatically.
   Also regarding: *“The qualitative analysis of the obtained credibility labels indicates that cognitive biases, to some extent, distort the medical expert assessment.”*
   How about then providing the experts with a system that would not just suggest whether a text is credible or not, but additionally point out those sentences which might negatively influence the decision? That would have much wider applicability than a simple classifier and would immediately have important educational value.

10. *“Therefore, the efforts focused on maximizing the throughput of the expert-supported assessment system.”*
    From the whole dissertation, it is clear that the main focus is put on creating the assessment system to help experts, not the other way around, so it should rather be „assessment system-supported expert”, or „expert-supporting assessment system”. See the image below for better understanding.




11. *“the results of the experiments allowed for the isolation of fragments of medical texts - three sentences.”*
    This phrasing suggests some specific three sentences. Rather: „a span of three sentences”.

## In: Abstract (PL)

12. *“Drugi artykuł opisuje również analizę polegającą na wykryciu schematów retorycznych, które […]”*
    "Schematy retoryczne" to specjalistyczny termin z analizy dyskursu, który nie odpowiada temu, który jest w angielskiej wersji (rhetorical patterns) ani temu faktycznie użytemu w dysertacji. Prosze sie upewnić, czy na pewno chodzi o "schematy retoryczne”, czy też o ogólnie pojęte wzroce zdaniowe.

13. *“które pojawiają się w niewiarygodnych treściach medycznych”*
    Raczej: ”w treściach medycznych o niskiej niewiarygodnosci”. Slowo „niewiarygodny” ma tutaj wydzwiek kolokwialny.

# In: Chapter 1 Introduction

14. *“https://www.zippia.com/advice/how-many-people-use-the-internet/”*
    Use a more reliable source of information, than a blog article. For example: https://ourworldindata.org/internet

15. *“It should also be noted that experts and non-specialists are burdened”*
    Either „experts and non-experts” or „experts as well as laypeople”.

16. In *“Figure 1.1: The workflow.”, “medical disinformation”* (also throughout the dissertation)
    In the Introduction you mention that „false medical information” is usually posted without harmful intent, thus, according to your own definition, it usually is „medical misinformation”. Are you specifically focusing here on false information with harmful intent, or should this be changed to misinformation? And if you specifically decide to continue to use the term “medical disinformation”, how can you specify if a text was written with malicious intent?

17. *“The dataset of around 10,000 annotations of statements related to selected medical subjects such as psychiatry or cardiology had to be constructed.”*
    Why exactly ten thousand? Also, if the ten thousand contains multiple medical areas, there will be only around two thousand samples per area, which is very scarce for medical data.

18. *“subject matter experts should be involved and provided with a clear and precise annotation protocol”*
    Experts should not only be provided with a precise annotation protocol but should also be involved in creating the protocol. An annotation protocol created by laypeople for experts will inherently limit the capabilities of experts.

# In: Chapter 2 Literature Review

19. *“One of the first projects in the field of computer science to deal with the problem of credibility was the Reconcile project carried out between 2014 and 2017”*
    There has been plenty of projects studying Web information credibility before 2014, going back even to the early 2000s. Do a more diligent field survey. For example, check at least the following papers:
    - Wathen, C. N., & Burkell, J. (2002). Believe it or not: Factors influencing credibility on the Web. Journal of the American society for information science and technology, 53(2), 134-144.
    - Akamine, S., Kawahara, D., Kato, Y., Nakagawa, T., Inui, K., Kurohashi, S., & Kidawara, Y. (2009, August). Wisdom: A web information credibility analysis system. In Proceedings of the ACL-IJCNLP 2009 Software Demonstrations (pp. 1-4).
    - Castillo, C., Mendoza, M., & Poblete, B. (2011, March). Information credibility on twitter. In Proceedings of the 20th international conference on World wide web (pp. 675-684).

20. *“Merriam-Webster dictionary 1. It states that credibility is ”the quality or power of inspiring belief”.”*
    Citing a general English language dictionary as more reliable than specialized literature is unprofessional. Instead, find a scientific source that has proposed a coherent and sufficiently broad definition that answers your needs. This could be a work from linguistics, sociology, or other. Using a general, or layperson’s dictionary causes serious problems. For example, in the definition in question:
    - How should „power” be understood/defined? A force? Quality has a static value (=it „is”), but

power has an active value (it „causes and effect”). Mixing those two together makes it very confusing how one should imagine what credibility is.
- Which meaning of „belief” is used in this definition?
- What does „inspiring” mean here? Causing (from no belief to belief)? Reinforcing (from weak belief to strong belief)?
The above problems occur because this definition is written for laypeople. Thus it is grounded in „common sense” and not very precise.

21. *“Researchers at the University of California confirmed that […]”*
    This is not a news article, but a scientific dissertation. Cite previous work properly. For example: „Metzger, Flanagin, and Medders (2010)”, or „Metzger et al. (2010)”, etc.

22. *“Thus, the review of works that automate the assessment process is also part of chapter2.2.2.”*
    1. Chapter or section?
    2. add space.

23. *“In the field of research on Web credibility, some works distinguish truthfulness as a separate characteristic. Sometimes it is treated only as an element of credibility.”*
    Add sources.

24. *“Below, the implicit features and exmplicit fact-checking approach will be discussed in more detail.”*
    exmplicit → explicit

25. *“"Liar, liar pants on fire" - a decade-long, 12.8K manually labeled short statements in various contexts which provides detailed analysis report and links to source documents for each case [42].”*
    1. A dataset cannot be „decade-long”. It is not a unit to measure the length of datasets. Perhaps you mean „a decade in the making”?
    2. “manually labeled short” → “manually labeled a dataset of short”

26. *“Thirdly, maintaining structured Knowledge Bases of verified claims is costly, and even in a perfect scenario of a Base that contains every possible claim stated in the World Wide Web quick detection of misinformation in the new emergent topic would be hard to accomplish.”*
    Why? If the claims are indexed in a database, searching for relevant claims for comparison with the new one would be trivially fast.

27. *“Excellent example of message credibility assessment is studies related to”*
    → “An excellent example of message credibility assessment are studies related to”

28. *“Surprisingly, the tools used by medical journalists and practitioners are not very common within the computer science society.”*
    What do you mean by „computer science society”? If this does not refer to some specific association, such as Computer Society of IEEE, you cannot speak for the whole of the field, unless you have unquestionable evidence. Perhaps you want to state that such tools are not often used in CS studies. If so, show a significant number of related papers that you reviewed and show in how many of those such tools were used on not used.

29. *“In [86] The Authors undertake the task of classifying health-related press releases. Working on a collection of articles from reliable and unreliable sources, The Authors distinguished”*

Why the uppercase? (this inconsistency appears also in several other places in the dissertation)

30. *“The Authors take into account: URLs, titles, keywords, text, images, tags, authors, date news reviews rating, the ground truth of rating criteria, explanations of the ground truth, category, summary, descriptions, source, social engagements, tweets about the news source, as well as the tweet’s replies, retweets, user network, profiles, timelines, followings, and followers.”*
You list up what features they used, but do not discuss how their reported method compared to other methods.

31. *“or the (infamous) GPT-3 [100]”*
This is not a blog or a popular article. Do not use colloquial language in your dissertation.

32. *“The context is usually coded in such a way that is not possible to interpret by a human but only by a neural network that processes this context.”*
This is not very scientific. Check how the context is processed in transformers. If it was impossible to interpret by a human the whole method would be inherently unexplainable. The problem with transformers is not that they are inexplainable, but rather that the operations on context are so overwhelmingly numerous that it would be an impossible task to track the processing of each word separately.

33. *“The Authors take advantage of the popular deep-learning architectures with recurrent neural networks and pay close attention to solving this problem.”*
What do you mean by „pay close attention to solving this problem”? Do they solve the problem or not? Do other authors comparatively not pay close attention to solving their problems?

34. *“For a potential crowd-sourced expert-in- the-loop annotation system, using shorter message text units instead of full-length articles would substantially increase the capacity of the system.”*
That might be true, but it introduces an additional problem - how to extract those sets of sentences accurately?

35. *“more effortless than assessing the whole document.”*
Something can be effortless or not. There is no such thing as „more effortless” or „less effortless”. Change to „more efficient” or „less time consuming”, etc.

36. *“Most of the related work focuses on full automation, which, in my opinion, is too dangerous”*
This is not an opinion piece, but a scientific dissertation. Delete “in my opinion”.

37. “To sum up, I conclude from the literature review that…” →
To sum up, I conclude from the literature review that the following are the issues the research community has not yet tackled.

38. “curating a credibility large dataset” → curating a large dataset annotated with credibility

39. “are the issues the research community has not yet tackled.”
Delete. Do not split a sentence between a long list.

## In: 3 Contributions

40. *“I conducted several experiments to investigate the credibility evaluation of different text units and to create datasets.”*
    What datasets exactly - remind the reader here. Medical information credibility datasets?

41. *“2. EUvsDisinfo is the flagship project of the European External Action Service’s East StratCom Task Force(opens in a new tab).”*
    What does “opens in a new tab” mean here? This looks like directly copied from a website.

42. *“Contribution 4. Topical classifiers of credibility of medical sentence triplets with 90% precision incredibleclass”*
    The question is whether the classifier should pinpoint the credible texts or not credible ones? If the majority of research is focused on finding out sentence patterns that are often used by non-credible sources, the primary goal of the classifier should be to detect non-credible texts. So, the 90% goal should be put on precision in the non-credible class. Ultimately it is not possible to „detect credibility”, unless deeper fact-checking is done, or a human expert is employed. If only simple topical classifiers are applied, the only thing that can be automatically detected is „a way of writing typical for non-credible sources”.

43. *“the evaluation of the models should go beyond their accuracy and include subjective judgments.”*
    Instead of „accuracy vs. subjective judgments” one could argue for quantitative vs qualitative evaluation.

## In: Chapter 4 Discussion

44. *“credibility classifiers for sentence triplets can pre-filter the data to remove triplets evaluated as credible with high certainty.”*
    Are high certainty/probability/confidence classifications always correct?

45. *“4.2 Limitations”*
    One major limitation not mentioned here is - not using any methods of fact-checking, but only relying on classifiers learning bare words and word patterns. This limits the detection of non-credible texts to those which are „written in a non-credible way”. So, any article with credible information, but using word patterns similar to those in non-credible articles will inherently be classified as non-credible and vice versa. The practical scope of this limitation needs to be studied in the future. Although false alarms (false positives) can be easily verified by experts, misses (false negatives) will eventually cause the slipping through of the non-credible information to the public.

46. *“classifiers built upon such datasets and, most importantly, help to disambiguate”*
    Delete „to”.

47. *“An example of the effect appears intuitive:”*
    Change to „is, for example” or a similar phrase. Never assume in a scientific publication that something is obvious, or intuitive. Explanation or exemplification is always necessary.

48. “I propose a relationship exists”
    I propose that a…