Transparent and Explainable Information Quality Prediction

by Davide Ceolin (CWI)

Predicting the quality of the information online is a key step to contrast the spread of dis- and misinformation. Transparency and explainability of information quality prediction are key elements to increase their trustability and usefulness. We at CWI work on fostering online information quality explainability through transparent AI pipelines that combine argumentation reasoning, crowdsourcing, and logical reasoning.

The amount of information online and the impact it has on society imply the need for an automated prediction of information quality. However, different users in different contexts have very different needs and requirements. Therefore, these individuals can really benefit from the diversity of the information the Web provides. Some contexts might even require disinformation to be part of the picture, e.g., when researchers or journalists want to study, describe, analyse, or report on the nature of online disinformation itself. Therefore, the solution to the problem of disinformation, misinformation, and information excess is not to filter out low-quality information but to predict information quality to increase user awareness. As quality prediction can be perceived as subjective or biased, it is crucial that information prediction is both transparent and explainable. In other words, in order to increase the usefulness of information quality predictions, we need to win user trust in them (See Figure 1). Transparency and explainability are key ingredients to this aim. Also, explainability is a key ingredient to help identifying disinformation campaigns: by predicting diverse aspects of quality, we can unveil complex strategies that involve, for instance, the manipulation of narratives based on the combination of factual statements.

Transparency of information quality prediction guarantees that the computational steps performed to obtain quality predictions are accessible by humans so that they can follow the reasoning and understand how it has been implemented. This means that these computational steps should, ideally, implement well-known approaches from the humanities and social sciences, that are familiar to humans. For example, in our studies, we refer to argumentation theory and source criticism as useful theories on which to base our pipelines.

Computational argumentation reasoning is a vast field of AI that aims at studying how claims can be supported through diverse types of reasoning. We demonstrate that it is possible to implement argumentation reasoning through AI pipelines that combine natural language processing, machine learning, and logical reasoning, and use it to predict the quality of information items. For example, we test it on product reviews [1], and show that computational argumentation reasoning can be a useful tool to predict their quality in a similar manner as humans do. Here, we aim at transparency for the purpose of gaining actionable insights, so we are now working on understanding the implications of the choice of different implementations for the AI components involved in the pipelines (e.g., clustering algorithms, readability measures, etc.) in the information quality prediction performance.

Figure 1: Information quality prediction can be useful to users as long as labels are explainable and transparent to users (image source: flickr, “mysterious conspiracy” by ranma_tim, licensed under CC BY-ND 2.0.)

Source criticism is a well-known practice from the humanities, meant to determine the quality of information sources. Through specific checklists, scholars can determine whether a given book, journal, or article is of high enough quality to be considered as a source for their studies. We translated this practice in order to evaluate Web information items and again implemented this theory into transparent AI pipelines that use network analysis and evidential reasoning to help laypeople understand the quality of the information they consume online [2].

Information quality can be informally defined as ‘fitness for purpose’ and therefore, we can think of quality prediction as Boolean labels: information either fits a given purpose or not. However, such labels can be obtained in different manners. Transparency helps us understand the computational part of this, but then we need to understand also which aspects of quality are being considered. Information can, in fact, be more or less precise, neutral, complete, etc., and these aspects (or information quality dimensions) shed some light on

Figure 2: Transparent AI pipelines that combine natural language processing, machine learning, and logical reasoning are a tool for assessing online information quality (image source: www.pexels.com).

different and possibly independent characteristics of information. We can combine these predictions in order to understand whether a given information item fits a given purpose. However, when these aggregated predictions are presented to final users, it is important to explain them. Explaining information quality predictions means explaining which aspects of information quality were considered when predicting them. We performed experiments involving both experts and laypeople to this aim, using crowdsourcing platforms. These contributors were asked to evaluate statements and documents regarding the vaccination debate in one case, and regarding political statements in another one. Results show that we can obtain consistent results from human contributors [3].

These lines of research will be further investigated in the recently started Eye of the Beholder project [L1], which is led by Davide Ceolin and is a collaboration between CWI, the Netherlands eScience Center, and the University of Amsterdam. The project aims at extending an existing platform for transparent AI pipelines in order to provide media studies scholars with a tool to predict the quality of online information items in a transparent and explainable manner.

Link:
[L1] https://kwz.me/hjN

References:

[1] D. Ceolin, et al: “Assessing the Quality of Online Reviews Using Formal Argumentation Theory”, Int. Conf. on Web Engineering (ICWE) 2021: 71-87.

[2] D. Ceolin, F. Doneda, G. Primiero: “Computable Trustworthiness Ranking of Medical Experts in Italy during the SARS-CoV-19 Pandemic”, ACM 1st Int. Conf. on Information Technology for Social Good (GoodIT 2021): 271-276.

[3] M. Soprano, et al.: “The many dimensions of truthfulness: Crowdsourcing misinformation assessments on a multidimensional scale”, Information Processing and Management 58(6): 102710 (2021), Elsevier.

Please contact:

Davide Ceolin, CWI, The Netherlands
Davide.Ceolin@cwi.nl

SPOTTED: Systematic Mapping of Detection Approaches on Data Sources for Enhanced Cyber Defence

by Manuel Kern and Florian Skopik

In the last decade there was a clear paradigm shift from focusing only on prevention and protection to also including detection and response. While prevention and protection are indispensable to enable a baseline security, it is presumed that attackers have already compromised systems to some extent (“presumption of compromise”). The fact that professional attackers often operate in the network over a long period of time has long been known in cyber security research. A key pillar of a holistic security approach is therefore the early detection of attackers in the network. But still, the average time to detect attackers remains high. In the course of a study commissioned by IBM Security [L1], the average time it takes to detect a data breach is quantified with a time period of 212 days, five days longer than the year before.

It was in late 2020 that the disastrous case of SolarWinds became public [L2]. State attackers abused the update mechanism of a security solution to infiltrate thousands of organisations up to the highest state level and to move unnoticed in the networks of the attacked organisations for many months. Kaspersky [L3] reports that compared to costs of a cyberattack with immediate remediation, recovery costs are four times higher if remediation is per-