Misinformation Needs a Data Community: the NewsQA project

Authors: Connie Moon Sehat, Research Community Lead at Credibility Coalition (connie@hackshackers.com); Ellen Zegura, Fleming Professor in the School of Computer Science at Georgia Tech (ewz@cc.gatech.edu)

Collaborator: Jeff Jarvis, Director of the Tow-Knight Center for Entrepreneurial Journalism and The Leonard Tow Professor of Journalism Innovation at City University New York, jeff.jarvis@journalism.cuny.edu

What is the societal problem you seek to address?

How do we increase the flow of factual news and information worldwide in ways that are sensitive to the freedom of expression and opinion? And, implicit to this, can the totality of content generation both contain reliable information and be economically viable?

Many are familiar with the current problems surrounding misinformation and disinformation proliferation on the internet, and improving the ability of algorithms to recommend accurate or quality news content online is considered an important part of this puzzle.[1] To this end, there are a number of projects and funding opportunities to seed solutions; earlier catalogs and upcoming mapping projects affiliated with Credibility Coalition, the RAND Corporation, and the Social Science Research Council will demonstrate this diversity.[2]

[1] For viability of "signal" or indicator-based solutions to this problem, ref. Amy Zhang, Aditya Ranganathan, Sarah Emlen Metz, Scott Appling, Connie Moon Sehat, Norman Gilmore, Nick B. Adams, Emmanuel Vincent, Jennifer 8. Lee, Martin Robbins, Ed Bice, Sandro Hawke, and David Karger. *A Structured Response to Misinformation: Defining and Annotating Credibility Indicators in News Articles*. The Web Conference, April 2018. The paper also contains a literature review related to the definition of misinformation, readers' perception of misinformation, computational models on information credibility, as well as the viability of human annotation.

[2] Example of catalogs: https://misinfocon.com/catalogue-of-all-projects-working-to-solve-misinformation-and-disinformation-f85324c6076c, https://foundation.mozilla.org/en/campaigns/eu-misinformation/resources/; Credibility Coalition early post on project https://misinfocon.com/our-toolkit-for-people-and-teams-tackling-misinformation-online-9e6d240f3136, SSRC announcement https://www.ssrc.org/programs/component/media-democracy/disinformation-research-mapping-initiative.

But the overall attention to current carriers of misinformation and disinformation is not the same as attention to the methods of delivery for factual information. Our current ecosystem places incredible responsibility on news and information consumers, to know what sources can be trusted, to be suspicious, to understand that there are fact checking services to use when seems suspect, and then to understand which of these fact checking sites can be trusted. Platforms spew largely non-curated content, and the human has to serve as a highly skilled librarian.

Take for example the 2016 US Elections, in which according to the Oxford Internet Institute there existed a one-to-one ratio of "junk news" to professional news shared by voters over Twitter; evaluations of the recent 2019 EU elections indicated a range of experiences with Poland having a 21:13 ratio of junk to professional in Twitter space but Spain only 1:28.[3] This places an incredible demand upon humans to wade through poor information, given that we also know that consistent exposure to narratives (no matter how poorly supported) can be influential to the ways that humans think.[4]

How is it done today, and what are the limits of current practice?

One of the challenges in solving this problem is that there are in fact a number of connected but open questions whose answers are unknown or ill-defined. These questions need to be addressed before significant advances can be made:

- **Where are we now?** (***The question of description***): What precisely is the character or quality of news and information propagated on the internet as a whole, and in particular local contexts? Although there are increasing studies in certain local/isolated examples, they do not yet capture the larger picture, partly because there is a lack of data related to being able to understand the problem.

[3] US: P N. Howard et al., "Social Media, News and Political Information during the US Election: Was Polarizing Content Concentrated in Swing States?," Data Memo 2017.8 (Oxford, United Kingdom: Project on Computational Propaganda, Oxford Internet Institute, Oxford University, 2018), EU Elections: Nahema Marchal et al., "Junk News During the EU Parliamentary Elections: Lessons from a Seven-Language Study of Twitter and Facebook," Data Memo 2019.3 (Oxford, United Kingdom: Project on Computational Propaganda, Oxford Internet Institute, Oxford University, 2019).

[4] In this case, framing, agenda setting, and priming are key aspects of communication that take advantage of repetition. See Maxwell McCombs, "A Look at Agenda-Setting: Past, Present and Future," Journalism Studies 6, no. 4 (November 1, 2005): 543–57, doi:10.1080/14616700500250438.; Dietram A. Scheufele and David Tewksbury, "Framing, Agenda Setting, and Priming: The Evolution of Three Media Effects Models," Journal of Communication 57, no. 1 (March 1, 2007): 9–20, doi:10.1111/j.0021-9916.2007.00326.x; Robert M. Entman, "Framing: Toward Clarification of a Fractured Paradigm," Journal of Communication 43, no. 4 (December 1, 1993): 51–58, doi:10.1111/j.1460-2466.1993.tb01304.x.

- **What should be different?** (***The question of prescription***): What exactly news quality mean? Different communities — journalists, knowledge authorities, public readers — define quality information differently, though there are overlaps.
- **How are we supposed to get there?** (***The question of transformation***): What would it take for social media and search engines to change from *Where we are now* to *What should be different?* This solution is not only a technological one, but a social and economic one; a sub-question would also address the challenge of internet economics vis-a-vis quality journalism which is currently quite reliant on certain advertising models.

But in order to advance understanding around these questions, to try and address the larger question, we need central locations of data sharing and collaborative exchange. These are in the process of being developed in small ways but no resources are being advanced at the global scale of platform reach. Our efforts to understand the issues and create solutions should be as systemic and large as the magnitude of the problem.

What is new in your approach and why do you think it will be successful?

The Craig Newmark Graduate School of Journalism at City University of New York has advocated, along with Hacks/Hackers and the Credibility Coalition, that creating a data resource that is shared with the platforms and ad ecosystem, along with researchers, and that foments progress is key:

- In order to *describe* the current situation adequately, we need to have more data collected and shared related to the possible "signals" of misinformation as well as quality news — from journalistic standards and prizes to fact-checked articles to inflammatory language.[5]
- In order to *prescribe* solutions, we need to have collaborative discussions on the exact nature and character of quality news — alongside the freedom of expression and opinion in different contexts — from different stakeholder communities.
- In order to *establish* transformation, key stakeholders, from readers to journalists to technological platforms, need to be part of the conversation in building holistic solutions to a very complicated problem.

[5] The definition of a *signal* in this context can be found in W3C Credible Web Community Group, "Technological Solutions to Improving Credibility Assessment on the Web: Draft Community Group Report 10 October 2011," https://credweb.org/report/20181011

To this end, we have built and are completing our first year of the “News QA” (News Quality Aggregator) project, a database that also aims to be the foundation of an ecosystem around the question of news quality. We have reached a Minimum Viable Product database of over 90 “signals” or units of information that may bear on news and information quality against 13,000 US-based news and information sites. These signals range among several types, including observed internet traffic data, to self-reported journalist corrections policies, or endorsement of media outlets by external authorities.[6] Currently available via an API to a small set of advisors, NewsQA aims to provide the raw streams of information with which to understand what combination of factors may be associated with more reliable information or quality journalism in different contexts.

However, beyond this technical achievement, equally important to the project concept is the community and social aspect: how do we create an ecosystem of exchange that continually builds the database and furthers insights and results? How does the database support the community in answering the key questions above, and further factual information flows that also respect expression and opinion?

We have therefore tried to lay the groundwork for the future growth of the project, which should scale across countries, languages, topics of misinformation emergencies, alongside a robust community of advisors, users, and researchers to address all three open questions, and ultimately the larger challenge.

So far, the positive feedback has been overwhelming -- while it has not been easy, we have established loose cooperation and enthusiasm around this first stage of the project. Many conversations with different platform, academic/research, private AI enterprise, non-profit organizations, and advertising players have taken and continue to take place. Some of those we currently can confirm include advising relationships or partnerships with members of the following organizations:

- Facebook Journalism Project, Facebook News
- Google Research
- Alphabet Jigsaw
- Wikipedia (as funders of the upcoming Wikipedia North American Conference, to be held on Reliability of Information in collaboration with Credibility Coalition)

We are as well developing partnerships with the following organizations:

- Reporters Without Borders, with their 19 country project on media ownership
- Duke Reporter’s Lab and ClaimReview, for fact-checked information globally

[6]More types of signals are described in the project’s inaugural post, https://medium.com/@TKCUNY/aggregating-signals-of-quality-in-news-3fc1d009dc19

- Center for Community and Ethnic Media at CUNY, for diversity in news representation as a signal of quality information
- TED's CIVIC initiative around Health Misinformation (https://www.wired.com/story/claire-wardle-ted-2019-crowdsource-against-misinformation/)
- First Draft News
- The Social Science Research Council, for best practices in data sharing and research requirements

In addition, because of connections through related projects, we know of collaborations that can feed into this effort, including for example with The Carter Center and the Mozilla Foundation.

The biggest reason that this project has been challenging is because at its core, NewsQA needs to motivate a number of different commercial and non-commercial interests to give data to this project without any promise of financial gain. In the current market, data equals revenue. By creating cooperation around the possibility of shared insights into the data through invited researcher access and publication, we have been able to generate interest and openness to the project.

Our project, however, is really only at the beginning stages. Critical next steps include seeding the data further around specific topics, cementing the kind of open licensing infrastructure and protocol exchange that we have provisionally developed, and nurturing productive research that will make the data useful for civic ends. While we have some of this funding in sight, NewsQA would benefit from greater exposure and more support.

Who cares? If you are successful, what difference will it make?

With success over time, we think that

- *reporters* and *journalists* who are dedicated to truthful, accurate information will care because their stories will take greater precedent in social media and online search;
- *platforms and ad-agencies* will care because they can have access to centralized, informative data to help their algorithmic development alongside a community of participants that understand the challenges that they face;

- *researchers* will care because of their access to data on news and information quality and support towards understanding the misinformation/disinformation problem better;
- *expert communities* such as health authorities will care because of opportunities to develop stronger links to factual information in news;
- *journalistic outlets* may benefit in the long run to the extent that the question of economic sustainability is able to be addressed;
- *citizens* who care about being adequately informed about events and the world will benefit because their news feeds and experience change for the better.

What might be intermediate and long-term metrics or goals for success?

Preserving freedom of expression and opinion in balance with factual information (on the internet) is a tall order. It has been and will be an ongoing struggle, and cannot be solved with any single database. However, we believe the resource can aid this struggle. Intermediate metrics include:

- Confirmed participation by platforms, ad agencies and networks with access to the database;
- Access to researchers with white paper/publications that follow;
- Confirmed partner organizations contributing data - yearly growth of data;
- Establishment and growth of expert community communication related to news and information quality;
- The defining of news quality flow metrics and baselines per market (e.g. country/language), that are mindful of local requirements related to freedom of expression and opinion.

Longer metrics and goals for success include:

- Sustainable nonprofit model established not through paying for data (our experience and ongoing conversations show that this is not a good long-term way forward) but through a cooperative effort among foundations and organizations able to sustain it;
- Productive exchanges on viable models for economic sustainability for transparent, fact-based journalism;
- Improved news quality flows in defined markets or series of produced research in which data flows are not improving;
- Ethical data sharing models and long-term data preservation in conversation with archiving/library communities;
- Papers and news articles that reference vocabularies and metrics the News QA project, showing its influence.