NORDIS – NORdic observatory for digital media and information DISorders

A method for auditing fact-checking tools and databases

Date: 28-02-2022

Final version

Action No:	2020-EU-IA-0189
Project Acronym:	NORDIS
Project title:	NORdic observatory for digital media and information DISorders
Start date of the project:	01/09/2021
Duration of the project:	24
Project website address:	https://datalab.au.dk/nordis
Authors of the deliverable	Ida Anthonj Nissen Jessica Gabriele Walter Marina Charquero-Ballester Anja Bechmann DATALAB, Aarhus University
Activity number	Activity 2
Task	Task 2.1

Reviewers: Matteo Magnani (Uppsala University, Sweden); Carl-Gustav T. Linden, Duc Tien Dang Nguyen (University of Bergen, Norway)

Funding:
The research was funded by European Horizon 2020 grant number 825469 and 825652, and EU CEF grant number 2394203.

1.0 Overview of this report

This report outlines the results published in the scientific paper:

> Nissen, I. A., Walter, J. G., Charquero-Ballester, M., & Bechmann, A. (2022). Digital Infrastructures of COVID-19 Misinformation: A New Conceptual and Analytical Perspective on Fact-Checking. *Digital Journalism*. https://www.tandfonline.com/doi/full/10.1080/21670811.2022.2026795

The deliverable is twofold and consists of the published academic paper (open access) and the executive summary, which is an extract of the paper. This report consists of the executive summary with reference to the full paper online for further details.

2.0 Executive Summary

Purpose and aim

The influence of digital media has increased during the COVID-19 pandemic and with it the focus on online misinformation. Misinformation related to the pandemic can have consequences for the health of the population and for adherence to government measures. Therefore, fact-checking organizations have investigated stories that potentially spread misinformation and published their investigations online, with the goal to curb the negative impact of misinformation on society. Those fact-checked stories are also collectively published by fact-checking databases, which form new overarching infrastructures of fact-checking. This study sets out to conceptualize fact-checking as overarching digital infrastructures by comparing two such overarching infrastructures, which differ in technical and economic terms: Poynter and Google Fact-check Explorer (referred to as 'Google' onwards). The Poynter infrastructure is from the #CoronaVirusFacts Alliance from the IFCN (international fact-checking network) at the Poynter institute, a non-profit organization, whereas the Google infrastructure is provided by a for-profit company. Our aim was to examine overlaps and differences and thereby to detect biases in the two infrastructures. We first looked at the number of overlapping stories and whether they agreed in the applied rating. To further compare the two infrastructures, we looked in detail at the following parameters. Who debunks misinformation? Where was misinformation published? Who published misinformation? Which content was published? We established a methodology for each of those questions, which can be used by other researchers analyzing similar data. Our comparison will advance transparency of fact-checking infrastructures and thereby enhance trust in fact-checking.

Definitions of terms

We here outline how we understand and use specific terms in the scientific paper.

- **Database**: Database refers to a collection of fact-checked stories, which can be extracted from fact-checking infrastructure. A database can contain stories from single or several fact-checker organizations.
- **Infrastructure**: Infrastructure refers to the system behind a database of fact-checked stories. The infrastructure of a fact-checking organization provides the database, determines which stories it contains, and arranges the presentation of the database (e.g., website).
- **Overarching infrastructure**: overarching infrastructures are the infrastructures resulting from the joining effort of several fact-checkers. Thus databases of overarching infrastructures contain contributions of several fact-check organizations. This requires larger considerations, of, for instance standards, and the impact of the infrastructure potentially has a larger scope. Infrastructures can be regional, national or international. In the scientific paper we argue for a conceptual distinction between overarching infrastructure and single fact-checking infrastructure in order to emphasize infrastructure considerations that cut across single fact-checker organizations.
- **Digital media:** In the context of the report, 'digital media' refers to all digital media, communication and interaction services encompassing online platforms such as social media and digital versions of partisan (blogs) and legacy media. The report's focus on digital fact-checking infrastructure thus encompasses integration with such sources.

Methodology

Our methodology is divided into two components; we compared the two infrastructures by looking at the overlap (component 1) and differences (component 2), see figure 1 and table 1. For the first component, we analyzed the overlapping stories and associated ratings. The method used for identifying the overlapping stories was a computational technique (sentence embedding using BERT, which performed better than a simple string-matching algorithm), followed by a manual check of the top 500 matches with the highest similarity scores. This combination of computational and manual approaches worked well by first reducing the number of possible matching stories to 500 matches, which was feasible for a manual check to sort out the false positive matches. To assess the ratings of the overlapping stories, we categorized the many different ratings by establishing six rating categories: false, partly false, mixed, partly true, true, and undefined.

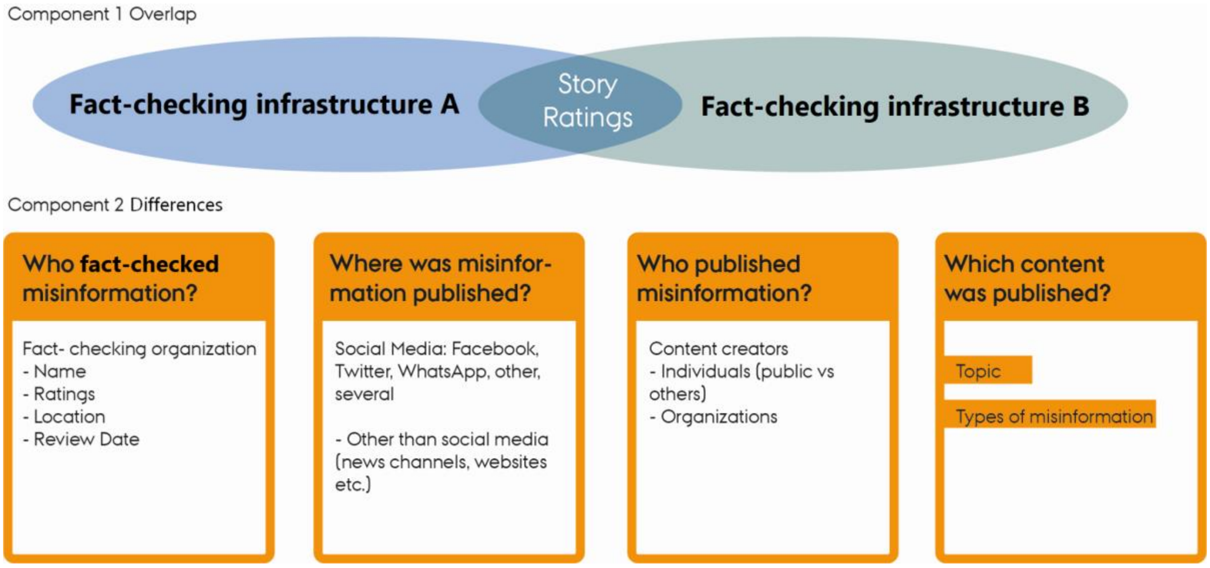

Figure 1. A holistic methodological approach for the comparison of fact-checking infrastructures. Source: (Nissen et al., 2022).

For the second component, we assessed the differences between the two infrastructures by analyzing four parameters. (1) Fact-checking organizations: we compared the infrastructures in terms of the fact-checking organizations, the applied ratings (organized into the previously established rating categories), and the geographical location of the of the fact-checking organizations. (2) Platforms: we organized the platforms into several categories: the three biggest social media platforms (Facebook, Twitter, WhatsApp), other specific social media platforms, several different social media platforms, social media and other, no social media platform mentioned, and not available. (3) Content creators: we distinguish between individuals, public persons (e.g. politicians and celebrities) and organizations (e.g. public authorities and news channels). (4) Types and topics of misinformation: we coded the types of misinformation manually into six categories: cure, prevention & treatment; conspiracy; political measures; vaccine & test kits; virus characteristics & numbers; and other. Furthermore, we used a computational approach with BERT topic modeling to identify the topics.

Table 1. Methods used for analyses by component and parameter. Source: (Nissen et al., 2022).

Component	Method	Description
Overlap	Natural Language Processing	Document embedding and Jaccard similarity for automated detection of overlap

	Manual rating	Rating automatically detected overlap into 'truly overlapping' and 'not overlapping'; categorization of used ratings into five categories → one rater
	Descriptive analyses	Comparing rating categories for overlapping stories
Differences		
Fact-checking organization	Descriptive analyses	Comparing numbers of different fact-checking organizations, distribution of ratings and rating categories, location and review date, mean number of ratings per organization
	Manual rating	Categorization of location into six continents; categorization of used ratings into five categories → one rater
Publishing Platforms	Manual rating & descriptive analyses	Categorization into seven categories → one rater; distribution by database
Content creator	Manual rating & descriptive analyses	Categorization into four categories → one rater; distribution by database
Types of misinformation	Manual rating & descriptive analyses	Categorization into six categories → two raters; computation of inter-rater reliability; distribution by database
Topics	Natural language processing	Topic modeling with BERT for sentences

Limited overlap of infrastructures regarding published stories

We found that the infrastructures showed limited overlap in the stories they contained. We looked at all published English stories from February 1 to September 30 in 2020 related to the term “COVID-19”. Only 291 stories were the same in the Poynter infrastructure (containing 8719 stories) and the Google infrastructure (containing 1217 stories). The stories were not necessarily checked by the same fact-checker or with the same rating. This result means that the two databases provide different infrastructures of fact-checking. They are supplementary in the stories they contain, as there is yet no unifying global infrastructure of fact-checked stories.

Agreement on ratings

We found that numerous different ratings were applied in both infrastructures (e.g., false, needs context, misleading). The Poynter and Google dataset contained 25 and 71 different ratings, respectively. The diverse ratings used in both infrastructures arise from the variety of contributing fact-checkers, which differ in presentation of fact-checked stories and used rating system. To be able to compare the ratings of the two infrastructures, we therefore needed to establish six rating categories: false, partly false, mixed, partly true, true, and undefined. Subsequently, we compared the rating categories of the 291 overlapping stories and found that the two infrastructures largely agreed upon the rating category. Namely, 90% fell into the same rating category, 7% had a missing rating category and only 3% (10 stories) differed in the rating category. The many different ratings indicate a need for improved standardization of rating categories across fact-checkers, potentially pre-defined in the user interface of databases.

Poynter contributors are more numerous and more geographically diverse

We found that some fact-checkers contributed to both infrastructures, in that three of the five top contributors were the same. However, the other two of the five top contributors differ, as did the amount of fact-checked stories that fact-checkers contributed with. Overall, we found that Poynter has more contributing fact-checkers (97 compared to only 50 for Google). The contributing fact-checkers to Poynter also contribute on average with more stories than the fact-checkers contributing to Google. This indicates that Poynter has a more exhaustive infrastructure for COVID-19 misinformation, whereas Google covers a broader range of misinformation stories not restricted to COVID-19 misinformation.

When looking at the continent of the contributing fact-checkers' location, we found that Google had many contributors located in North America and Asia, but none from South America. Poynter had a more diverse contribution from all continents. This confirms the international orientation of Poynter, whereas Google is oriented towards North America and Asia. Overall, the differences in location of the various fact-checkers illustrate differences in the adoption of those two infrastructures.

Funding influences the selection of fact-checked stories regarding publishing platform

When examining the platform on which the claims were originally published, Poynter was dominated by Facebook stories. Nearly half of the fact-checked stories were based on claims published on Facebook. This can be explained in that many IFCN fact-checkers are part of the Meta (formerly known as Facebook) Third-Party Fact-checking Program and partly funded by them. This finding indicates that these fact-checkers are the backbone of the Poynter infrastructure. As such, restrictions for fact-checking within the system provided by Meta influence the Poynter infrastructure. Furthermore, we found that the information about the publishing platform was often not available (22%). An increase in documentation would increase research possibilities to improve our understanding of where misinformation is propagated. Also, the amount of stories published on Facebook in the Poynter infrastructures challenges to some extent the independence of fact-checkers, as Facebook contributes with funding to the IFCN.

Information about misinformation creator was often not obtainable

We further analyzed the creators behind the misinformation. Overwhelmingly, we could not extract information about the content creator. This was the case for the vast majority of the Poynter stories (89%) and for the majority of the Google stories (64%). Hence, both fact-checking infrastructures need to improve documentation about who claimed the misinformation. Of the stories where we could extract the content creator, we found that more stories in Google are based on claims from public persons than in Poynter, which probably affects how the detected misinformation spreads.

Similar misinformation content is present in both infrastructures

To compare the content of the stories, we first manually established categories based on the published stories. These categories were the following: cures, prevention and treatment; virus characteristics and numbers; vaccine and test kits; political measures; conspiracies; and other. We found that the two infrastructures contained a similar distribution over those categories, except for Google having significantly more stories about virus characteristics and numbers, whereas Poynter showed a trend towards more stories about political measures. As these results were based on the stories from March 2020 only, we also analyzed the entire dataset using machine learning (BERT topic modeling) without predefined categories. Here, we found that the detected stories were also similar in both infrastructures. One exception was that Google had two topics about vaccines and Poynter none, but a possible explanation is that those topics did not cluster into topics with our method in the Poynter infrastructure. To conclude, both infrastructures are comparable regarding the coverage of topics and types of misinformation.

Implications for EDMO and NORDIS

Both EDMO and NORDIS aim at improving digital fact-checking infrastructure, and can hence build on our results of lacking documentation of biases. By implementing the presented method and making associated results visible, we argue this will strengthen transparency on an infrastructure level in connection with the Data Science Institute (DSI) at the London School of Economics and Political Science (e.g. Truly Media and future similar DSI fact-checking infrastructure) and thus hopefully have a positive effect on trust. Our paper provides a blueprint for how to increase such transparency on the infrastructure level and our case study uncovered some biases in the overarching infrastructures. Implementations towards more transparency at the level of the individual fact-checking organization or of the overarching infrastructure are within the ability of EDMO and associated fact-checkers and hubs and would hence improve the trust in the work of fact-checkers.

NORDIS

We have developed various methodological approaches for our comparison. Those methods can be used by other researchers studying fact-checking to show overlaps and biases in different infrastructures and for different topics.

EDMO provides the platform Truly Media for collaboration for fact-checking organizations, which thereby also serves as an overarching infrastructure as those analyzed in this article. Therefore, methods proposed in our article to increase transparency of the overarching infrastructures could also be applied by Truly Media.

Conclusion

We have provided a blueprint for how to increase transparency in fact-checking digital infrastructure by focusing on accounting for overlaps and differences in infrastructures and have thus visualized biases. We have shown through a case study of covid-19 at scale how fact-checking, as digital infrastructures that have downstream effects in society when used by a diverse set of stakeholders (e.g., the general public, news media, researchers), are biased by favoring some content types, platforms, and continents over others (bear in mind that we looked at English language tweets from all countries worldwide). These inherent infrastructural biases, along with the lack of overlap and the use of a variety of different rating labels, do not contradict each other but rather result in the Poynter and Google Fact-Check Explorer infrastructures supplementing each other. By systematically comparing infrastructures of COVID-19 misinformation, we have analyzed how such infrastructures 'color' stakeholders' and society's beliefs of what is false in different ways, because they disclose very different stories. The differences between (and biases within) infrastructures can be explained by their characteristics (organizational structure, eligibility rules, funding). Hence, making biases of the overarching infrastructure visible will increase transparency for stakeholders.

3.0 References

Nissen, I. A., Walter, J. G., Charquero-Ballester, M., & Bechmann, A. (2022). Digital Infrastructures of COVID-19 Misinformation: A New Conceptual and Analytical Perspective on Fact-Checking. *Digital Journalism*. https://www.tandfonline.com/doi/full/10.1080/21670811.2022.2026795