Challenges in Combating COVID-19 Infodemic - Data, Tools, and Ethics

Kaize Ding
Arizona State University
kaize.ding@asu.edu

Kai Shu
Illinois Institute of Technology
kshu@iit.edu

Yichuan Li
Arizona State University
yichuan1@asu.edu

Amrita Bhattacharjee
Arizona State University
abhatt43@asu.edu

Huan Liu
Arizona State University
huan.liu@asu.edu

Abstract

While the COVID-19 pandemic continues its global devastation, numerous accompanying challenges emerge. One important challenge we face is to efficiently and effectively use recently gathered data and find computational tools to combat the COVID-19 infodemic, a typical information overloading problem. Novel coronavirus presents many questions without ready answers; its uncertainty and our eagerness in search of solutions offer a fertile environment for infodemic. It is thus necessary to combat the infodemic and make a concerted effort to confront COVID-19 and mitigate its negative impact in all walks of life when saving lives and maintaining normal orders during trying times. In this position paper of combating the COVID-19 infodemic, we illustrate its need by providing real-world examples of rampant conspiracy theories, misinformation, and various types of scams that take advantage of human kindness, fear, and ignorance. We present three key challenges in this fight against the COVID-19 infodemic where researchers and practitioners instinctively want to contribute and help. We demonstrate that these challenges can and will be effectively addressed by collective wisdom, crowd sourcing, and collaborative research.

Title of the Proceedings: "Proceedings of the CIKM 2020 Workshops October 19-20, Galway, Ireland"
Editors of the Proceedings: Stefan Conrad, Ilaria Tiddi

1 Introduction

Coronavirus disease 2019 (COVID-19) is an infectious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The World Health Organization (WHO) recently declared the COVID-19 outbreak a Public Health Emergency of International Concern (PHEIC) and a pandemic due to its high morbidity and mortality rates. As of April 15, 2020, more than 2.04 million cases have been reported across 210 countries and territories, resulting in over 133,000 deaths[1]. These numbers are continuing to rise and the health systems in many countries are overwhelmed to provide treatment. Concomitant with the pandemic are many unknowns that create a conducive environment for misinformation, fake news, political disinformation campaigns, scams, etc. Those malicious contents instigate fears or anger, capitalize on human vulnerability, and exploit human emotion, kindness, and/or wishes for miracles.

As the coronavirus spreads like fire in the world, disinformation machines also accelerate their campaigns on various fronts, rendering a new infodemic battlefield. Social media platforms such as Facebook/Instagram, Twitter, and Google/YouTube have been abused to disseminate erroneous contents. When the whole world is scrambling to fight the COVID-19 pandemic, governments and WHO also have to combat an infodemic, which is defined as "an overabundance of information — some accurate and some not—that makes it hard for people to find trustworthy sources and reliable guidance when they need it" [Don20]. The COVID-19 infodemic causes confusion, sows di-

[1] https://en.wikipedia.org/wiki/Coronavirus_disease_2019

vision, incites hatred, promotes unproven cures, and provokes social panic, which directly impacts emergency response, treatment, recovery, and financial and mental health during the difficult time of self-isolation. Therefore, combating the COVID-19 infodemic is a challenging yet imperative task to solve.

In this paper, we first present some COVID-19 related examples to illustrate the variety and range of infodemic cases in representative categories: conspiracy theories and misinformation, and scams and security attacks to reinforce the urgency and need for addressing the COVID-19 infodemic via scalable and timely solutions. We then discuss the essential challenges in designing and developing corresponding AI solutions from three perspectives: data, computational tools, and ethics. The last challenge of ethics is particularly easy to overlook when we rush to confront the immediate threats. Therefore, it is important to understand unintended consequences when developing AI solutions to ensure sustainable and healthy use and deployment. Last, we use some current efforts to demonstrate the feasibility of addressing the three challenges in combating the COVID-19 infodemic; Meanwhile, by understanding the challenges and what we have, we also appreciate the importance of collaborative research for effectively and efficiently combating the COVID-19 infodemic.

2 Examples of COVID-19 Infodemic

To illustrate what the COVID-19 infodemic looks like, how expansive, active, and devastating it is, and why it is important to thwart or mitigate its present threats, we first present various examples regarding conspiracy theories and misinformation, and scam and other security attacks.

2.1 Conspiracy theories and misinformation

With the spread of COVID-19 pandemic, the World Health Organization (WHO) recently warned of an "infodemic" of rampant conspiracy theories about the coronavirus. Those conspiracy theories have appeared in both social media and mainstream news outlets and are often intertwined with geopolitics. One example is about how the new coronavirus originated: according to a Pew Research Center survey, nearly three-in-ten Americans believe COVID-19 was a bio-weapon made in the lab. Some top 10 conspiracy theories include SARS-CoV-2 virus was created as a biologic weapon from a lab, GMOs are the culprit, COVID-19 actually doesn't exist, and coronavirus is a plot by big Pharma [Lyn20].

Coronavirus misinformation is also flooding the internet through social media, text messages, and propagated by celebrities, politicians, or other prominent public figures. According to the report in [KG20], "among outlets that repeatedly share false content, eight of the top 10 most engaged-with sites are running coronavirus stories." For instance, there are plenty of supposed "cures" on social media that will likely mislead people to risk their lives for quick fixes. Disregarding the National Institutes of Health (NIH) warning of many hearsay cures without evidence of curing being effective, there are endless claims such as herbs and teas, or something of the sort that can prevent the coronavirus. Recently, some wireless towers were damaged in the UK due to a false claim that radio waves sent by 5G technology are causing small changes to people's bodies that make them succumb to the virus.

2.2 Scam, spam, phishing, and malware

As more and more people start working or studying from home, cyber criminals recently shift focus to target remote workers. Different attacks such as scam, spam, phishing and malware, which prey on people's willingness to help, fear of supply shortage, and moments of weakness, have become increasingly active. Researchers have found that the volume of coronavirus email scams nearly tripled in one week, with almost 3% of all global spam now estimated to be COVID-19 related. During the coronavirus pandemic, as state governments and hospitals have scrambled to obtain masks and other medical supplies, scammers attempted to sell a fake stockpile of 39 million masks to a California labor union. According to The Hill [Mil20] , "Hackers are taking advantage of the increased reliance on networks to target critical organizations such as health care groups and members of the public, stealing and profiting off sensitive information and putting lives at risk."

3 Data, Tool, and Ethics Challenges

The scale, volume, and reach of the COVID-19 infodemic entails the reliance on AI and machine learning (ML) algorithms to react promptly and respond rapidly. The success of AI and ML algorithms requires large amounts of multi-modal data for their efficiency and effectiveness, which introduces *a data challenge.* Data extraction and curation from multi-source data needs different computational tools to accurately categorize and sort out various types of data, which presents *a tool challenge.* When we rush to deal with present threats, we should be aware of potential side-effects, unexpected consequences, and biases of our solutions, which suggests *an ethic challenge.* In this section, we will discuss these challenges in detail.

3.1 Data challenge

Though numerous COVID-19 data sources are available online, their datasets are available on various websites for different needs. The major data challenge of isolated data sources is the awareness of their existence. Another related issue is that they are collected from different sources or under different crawl settings. For example, Allen Institute for AI (AI2) released the scholarly articles dataset[2] collected from PMC, medRxiv and bioRxiv; LitCovid [CAL20] collected the scientific information from PubMed. Combining different data sources leads to higher quality of data and better coverage.

To address the data challenges, we need to overcome some shortcomings: *disorganization* – most of them merely list all the collected datasets on their websites without information summarizing the relationships among them; *specificity* – data collected for a specific topic, for example, Amazon provides the epidemic dataset on cloud[3] and COVID-19 GIS Hub[4] only contain the academic findings and geospatial-related datasets respectively; and *inconvenience* – most sites merely provide the reference links to the source datasets and do not provide data utility tools like covid19datahub [GA20] for easy access.

3.2 Computational tool challenge

There are existing resources that can assist users to identify malicious intent in websites. Google's Safe Browsing API, for instance, allows the user to enter a URL and check it against Google's constantly updated lists of unsafe web resources. Similar resources include isitPhishing.org, malwareurl.com, and antivirus software, among many others. Additionally, users can check malicious domain lists through different sources such as phishtank.com or the aforementioned Google's Safe Browsing lists. As many malicious sites use URL shorteners to disguise themselves, to counteract potential attacks, it would be safe to first use URL expanders to figure out what they are before clicking them. Despite the easy access of those computational tools, they are not available conveniently in a single place where different tools can be called up whenever needed.

The awareness of these existing tools and efficient use of them for quick response is vital for combating COVID-19. An associate issue is the requirement for current and frequently updated black-lists [SLH17]. As we know, it is infeasible to manually maintain a dynamically changing list of malicious URLs, with new sites being generated everyday. Therefore, it is necessary to develop AI/ML identifiers that can learn from the old malicious sites for estimating the threats of new ones.

3.3 Ethics challenge

The COVID-19 pandemic is ushering in a new era of digital surveillance since governments are employing tools that track and monitor individuals. South Korea and Israel, for instance, have demonstrated the effectiveness of harnessing different digital surveillance tools. However, such a new practice can breach data privacy in the meantime and may even remain in use after the pandemic. In this section, we discuss the potential *privacy concerns*, *trade-offs* between stringent disease monitoring and patient privacy and ethical issues behind the disruption of *civil liberties*.

Gauging the war-like severity of the coronavirus pandemic, academics, researchers, companies and non-profits alike have come forward to contribute in any possible way. However, given the rapid nature of such responses and the subsequent lack of policy checks, these otherwise novel endeavors may have ethical loopholes. In an attempt to provide a transparent view of the degree of infection and prevent community spread of the virus, many counties and states in the United States have decided to publicly release data corresponding to cases, including the number of cases per zip-code [Mal20]. Smartphone applications with geo-locating capabilities have come out for users to log their symptoms. But the use of such applications has significant *privacy concerns* [5] [Wet20]. Contact tracing has been identified as an effective way to control the spread of the virus in communities where the infection is not yet widespread or has slowed down significantly, and companies including Google and Apple are currently developing applications to make this possible. Only when a sufficient number of people use the application and voluntarily report their cases can it be used as a reliable tool of tracking. In this situation, there is an obvious *trade-off* between user health privacy and data transparency and it is challenging to identify well-defined ethical boundaries when it comes to public health during a pandemic. The success of such an app requires a majority of the population to download and use it.

4 Feasibility Discussion

In this section, we present some current efforts that address the aforementioned challenges and show that

[2] `https://allenai.org/data/cord-19`

[3] `https://aws.amazon.com/blogs/big-data/a-public-data-lake-for-analysis-of-covid-19-data/`

[4] `https://coronavirus-disasterresponse.hub.arcgis.com/`

[5] `https://privacyinternational.org/examples/apps-and-covid-19`

the three challenges are solvable with collaborative research.

For the *data challenge*, we collect the publicly available COVID-19 datasets and cluster them into several groups[6]. Under each group, researchers can reference complete datasets from different sources or settings. For example, in social media data, we gather available tweet corpus on COVID-19 [BTW+20][CLF20] with different query keywords and time spans. The hierarchy cluster structure in Figure 1 helps the researchers to quickly locate the dataset. Lastly our data repository includes areas in academics, news, social media, and epidemic reports for multi-disciplinary research. For example, if a researcher wants to analyze the influence of the news or academic findings on social media like Twitter, s/he can use the data in academic or news topics and social media.

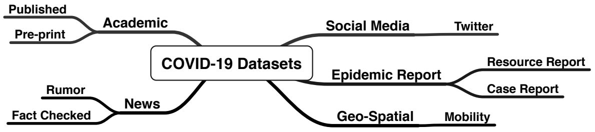

Figure 1: A taxonomy of collected datasets

To help a researcher easily access the datasets in the repository, we build a data-loader[7]. It is a Python package with a pandas Dataframe [pdt20] by calling *data = DataLoader().download(url)*. This widely used data format can help the downstream data analysis.

To tackle the tool challenge, we develop *TellMe*, a computational tool that provides an estimate if a piece of news or text is disinformation. Its input includes URLs and text, and its output is a score based on different functions of TellMe as shown in Figure 2: URL Checker, Fake News Classifier, Website Matcher, Credibility and Trusty. The Trusty [MYL09] and Credibility [AL13] scores are based on contents' social engagements that malicious users share more similarity than general users. The fake news score is returned from a state-of-the-art fake news detector [SZL+20]. The website matcher compares the input URL with websites that publish false information about the virus found by NewsGuard [BC19]. In addition, we are also in the process of developing and integrating more components (e.g., advertisement tracker, source attributor) and algorithms [DLBL19, DLL19, DLD+19] into the TellMe system.

Now, we use fake news as an example to illustrate our attempts to learn with weak social supervision to detect COVID-19 disinformation more effectively and with explainability. First, for *effective fake news detection*, we consider the relationships among publishers, news pieces, and consumers, which is motivated by existing sociological studies on journalism on the correlation between the partisan bias of publishers, the credibility of consumers, and the veracity degree of news content; and explore various auxiliary information from these relations to help detect fake news [SWL19]. Second, for *explainable fake news detection*, we aim to derive explanation of prediction results to help decision makers and practitioners; we attempt to explore user comments as a source and mine informative and relevant pieces to help explain why a piece of news is predicted as fake, and pinpoint more fictional text in news text simultaneously [SCW+19].

Figure 2: The current components of TellMe.

To tackle the ethics challenge due to the increase in government surveillance and prevalence of smartphone apps to collect and gather user/patient data, we need to take into account legitimate concerns regarding privacy and the degree to which such a regime of monitoring and enforcement will affect democracy after the pandemic ends. It requires us to understand and acknowledge the fact that there is a clear difference between standard biomedical ethics versus privacy concerns and ethics during a public health crisis. Governments and public health officials may need to take certain measures aimed at minimizing the damage caused by the virus and for the common good during this trying time, which under normal circumstances might have been inappropriate. Nevertheless, measures could be taken to avoid potential misuse of data. One possible way to have better guarantees on user privacy would be to make these contact tracing smartphone applications communicate in an encrypted peer to peer way rather than storing all the data in a central server. These technologies should also be deployed in a way that is as transparent as possible, so that the user is fully aware of what and how much personal information he/she permits the application to use. Furthermore, there is significant ongoing discussion among experts, researchers and policy-makers regarding a steady recovery into a normal functioning society. For example, the ethics research group at Harvard University makes efforts at finding solutions without compromising user privacy to keep civil liberty and democracy at the forefront.

5 Looking Ahead

The significance of combating the COVID-19 infodemic lies at protecting people from falling victims to the pandemic in this unexpected front and from disrupting otherwise already inconvenient daily routines

[6] `https://github.com/bigheiniu/awesome-coronavirus19-dataset`

[7] `https://github.com/bigheiniu/COVID-19-Dataloaders`

so as to improve our resilience in our fight to contain the pandemic. In this position paper, we show a good number of problems posed by the COVID-19 infodemic, the vast amounts of data generated in the world's effort to contain the pandemic, and the need for concerted efforts at various levels to efficiently and effectively deal with current and future challenges in medical and information fronts.

It is evident that (1) we face both immediate and future challenges in this unprecedented fight, (2) existing data will grow fast, and existing computational tools are insufficient to contain and mitigate the COVID-19 infodemic, and (3) short-term solutions can have potential long-term impact. Therefore, when we face hard choices, we need to resist the temptation to trade-off so as to minimize long-term negative impact; when we search for solutions, we should consider those employing crowdsourcing and take long views for fairness and responsibility; when we design methods, we should rely on collective wisdom and diversity to aim for robustness; and when we form teams, we should give priority to multi-disciplinary collaboration and preemptively address hidden biases. Our future will always be uncertain, but with the advancement in science and technology and with our preparedness trained and tested in our concerted efforts to contain the pandemic in all fronts, our future will surely be brighter and healthier.

Acknowledgements

This work is, in part, supported by Global Security Initiative (GSI) at ASU and by NSF grants (2029044 and 1614576). We would like to thank Denis Liu for helping develop earlier versions of TellMe and for carefully proofreading an earlier version of this paper.

References

[AL13] Mohammad-Ali Abbasi and Huan Liu. Measuring user credibility in social media. In *SBP-BRiMS*, 2013.

[BC19] S Brille and G Crovitz. Newsguard now available on microsoft edge mobile apps for ios and android, 2019.

[BTW+20] Juan M. Banda, Ramya Tekumalla, Guanyu Wang, Jingyuan Yu, Tuo Liu, Yuning Ding, and Gerardo Chowell. A large-scale covid-19 twitter chatter dataset for open scientific research – an international collaboration, 2020.

[CAL20] Q. Chen, A. Allot, and Z. Lu. Keep up with the latest coronavirus research. *Nature*, 2020.

[CLF20] Emily Chen, Kristina Lerman, and Emilio Ferrara. Covid-19: The first public coronavirus twitter dataset, 2020.

[DLBL19] Kaize Ding, Jundong Li, Rohit Bhanushali, and Huan Liu. Deep anomaly detection on attributed networks. In *SDM*, 2019.

[DLD+19] Kaize Ding, Jundong Li, Shivam Dhar, Shreyash Devan, and Huan Liu. Interspot: interactive spammer detection in social media. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 6509–6511. AAAI Press, 2019.

[DLL19] Kaize Ding, Jundong Li, and Huan Liu. Interactive anomaly detection on attributed networks. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 357–365, 2019.

[Don20] Joan Donovan. Here's how social media can combat the coronavirus 'infodemic', 2020. `https://www.technologyreview.com/s/615368/facebook-twitter-social-media-infodemic-misinformation/`.

[GA20] Emanuele Guidotti and David Ardia. Covid-19 data hub, 04 2020.

[KG20] Kornbluh and Ellen P. Goodman. Safeguarding digital democracy, 2020. `http://www.gmfus.org/publications/safeguarding-democracy-against-disinformation`.

[Lyn20] Mark Lynas. Covid: Top 10 current conspiracy theories, 2020. `https://allianceforscience.cornell.edu/blog/2020/04/covid-top-10-current-conspiracy-theories/`.

[Mal20] Laurel Mallory. Sc health officials update list of confirmed and estimated coronavirus cases by zip code, 2020. `https://www.wtoc.com/2020/04/10/dhec-releases-number-confirmed-estimated-coronavirus-cases-by-zip-code/`.

[Mil20] Maggie Miller. Virtual army rising up to protect health care groups from hackers, 2020. `https://thehill.com/policy/cybersecurity/493997-virtual-army-`

CEUR Workshop Proceedings
http://ceur-ws.org
ISSN 1613-0073

Version 1.03-ntp
2020-03-02

CEUR Workshop Proceedings
http://CEUR-WS.org
RWTH Aachen, Informatik 5, Aachen, Germany

Author Agreement to Publish a Contribution as Open-Access on CEUR-WS.org

Herewith I/we (the author(s) resp. the copyright holders) **agree** that my/our contribution:

Challenges in Combating COVID-19 Infodemic -Data, Tools and Ethics

authored by:

Kaize Ding, Kai Shu, Yichuan Li, Amrita Bhattacharjee, Huan Liu

with corresponding author

Name: Kaize Ding

Affiliation: Arizona State University

Address: 699 South Mill Ave, Tempe, AZ, 85281, US

Email: kaize.ding@asu.edu

shall be made available as an open-access publication under the **Creative Commons License Attribution 4.0 International (CC BY 4.0)**, available at https://creativecommons.org/licenses/by/4.0/legalcode, and be published as part of the proceedings volume of the event

Name and year of the event:

Mining Actionable Insights from Social Network (MAISoN 2020)

Editors of the proceedings (editors):

Ebrahim Bagheri, Huan Liu, Kai Shu, Fattane Zarrinkalam.

I/we agree that my/our contribution is made available publicly under the aforementioned license on the servers of CEUR Workshop Proceedings (CEUR-WS). I/we grant the editors, RWTH Aachen, CEUR-WS, and its archiving partners the non-exclusive and irrevocable **right to archive** my/our contribution and **to make it accessible** (online and free of charge) for **public distribution**. This granted right extends to any associated metadata of my/our contribution. Specifically, I/we license the associated metadata under a Creative Commons CC0 1.0 Universal license (public domain). I/we agree that our author names and affiliations is part of the associated metadata and may be stored on the servers of CEUR-WS and made available under the CC0 license. I/we acknowledge that the editors hold the copyright for the proceedings volume of the aforementioned event as the official collection of contributions to the event.

I/we have not included any copyrighted third-party material such as figures, code, data sets and others in the contribution to be published.

I/we warrant that my/our contribution (including any accompanying material such as data sets) does not infringe any rights of third parties, for example trademark rights, privacy rights, and intellectual property rights. I/ we understand that I/we retain the copyright to my/our contribution. I/we understand that the dedication of my/our contribution under the CC BY 4.0 license is irrevocable.
I/we understand and agree that the **full responsibility/liability** for the content of the contribution rests upon me/us as the authors of the contribution. I/we release the aforementioned editors, RWTH Aachen, persons providing the CEUR-WS service, and the archiving partners of CEUR-WS from any liability caused by the publication or archiving of my/our contribution via the servers used by CEUR-WS.
I/we have read the conditions of the Creative Commons License Attribution 4.0 International (CC BY 4.0), and agree to apply this license to my/our contribution.

Tempe, 09/15/2020, Kaize Ding

Location, Date, Signature of the corresponding author representing all authors
(Signature must be handwritten with a pen on paper)

`rising-up-to-protect-healthcare-groups-from-hackers/`.

[MYL09] Sai T Moturu, Jian Yang, and Huan Liu. Quantifying utility and trustworthiness for advice shared on online social media. In *CSE*, 2009.

[pdt20] The pandas development team. pandas-dev/pandas: Pandas, February 2020.

[SCW+19] Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. defend: Explainable fake news detection. In *KDD*, 2019.

[SLH17] Doyen Sahoo, Chenghao Liu, and Steven CH Hoi. Malicious url detection using machine learning: A survey. *arXiv preprint arXiv:1701.07179*, 2017.

[SWL19] Kai Shu, Suhang Wang, and Huan Liu. Beyond news contents: The role of social context for fake news detection. In *WSDM*, 2019.

[SZL+20] Kai Shu, Guoqing Zheng, Yichuan Li, Subhabrata Mukherjee, Ahmed Hassan Awadallah, Scott Ruston, and Huan Liu. Leveraging multi-source weak social supervision for early detection of fake news. *arXiv preprint arXiv:2004.01732*, 2020.

[Wet20] Nicole Wetsman. Personal privacy matters during a pandemic — but less than it might at other times, 2020. `https://www.theverge.com/2020/3/12/21177129/personal-privacy-pandemic-ethics-public-health-coronavirus/`.