Federal Register Notice 87 FR 15274, https://www.federalregister.gov/documents/2022/03/17/2022-05683/request-for-information-on-federal-priorities-for-information-integrity-research-and-development, May 15, 2022

Request for Information on Federal Priorities for Information Integrity Research and Development

Applied Research Laboratories, The University of Texas at Austin (ARL-UT)

DISCLAIMER: Please note that the RFI public responses received and posted do not represent the views or opinions of the U.S. Government. We bear no responsibility for the accuracy, legality, or content of the responses and external links included in this document.

APPLIED RESEARCH
LABORATORIES
The University of Texas at Austin

Response of ARL:UT to the NITRD NCO and NSF's RFI on Information Integrity R&D

Applied Research Laboratories, The University of Texas at Austin (ARL:UT)

May 12, 2022

Introduction

This document contains answers to several of the RFI questions based on our past work on information integrity. Our group, the Center for Content Understanding (CCU), is part of Applied Research Laboratories at the University of Texas at Austin (ARL:UT). ARL:UT is a Navy University Affiliated Research Center with a long history of work in acoustics, electromagnetics, and information technology. The lab works on a wide range of problems involving machine learning, including natural language processing, cybersecurity, image/video segmentation, and digital signal processing.

CCU works on various applications of machine learning and AI for the US Government, with a focus at the intersection of automated content understanding and security applications. We have completed and/or are currently working on a number of research projects related to the RFI. Our team consists of an experienced group of researchers with expertise in fabricated text detection, topic modeling, ontology-based decision support, taxonomy induction, concept extraction, transfer learning, named entity recognition, and human-guided machine learning, with many applications of these technologies in cybersecurity contexts. We have also established avenues for collaboration with external researchers in areas related to the RFI. Our team was one of four partners working with OpenAI on the detection and mitigation of malicious uses of large language models such as GPT-2. We also maintain connections with external researchers that work in areas including information integrity and the detection of disinformation campaigns.

Below are answers to the RFI questions which are most relevant to ARL:UT's previous and ongoing information integrity research. In particular, we focus on answers to questions 1, 2, and 4.

Responses to RFI Questions

Question 1: Understanding the information ecosystem

> "There are many components, interactions, incentives, social, psychological, physiological, and technological aspects, and other considerations that can be used to effectively characterize the information ecosystem. What are the key research challenges in providing a common foundation for understanding information manipulation within this complex information ecosystem?"

Ecosystems for the creation and disseminations of propaganda and disinformation have emerged in most countries.[1] A complete understanding of these information ecosystems should include the relevant actors (such as governments, social media platforms, political parties, and marketing

[1] Bradshaw, Samantha, Hannah Bailey, and P. Howard. "Industrialized disinformation: 2020 global inventory of organized social media manipulation. Computational Propaganda Research Project." (2021).

firms[2]), and their incentives, capabilities and strategies. However, we shall focus our answer on foreign state-sponsored disinformation campaigns, since these present one of the major threats to U.S. national security, economic prosperity, and individual well-being. We describe three research challenges that will be faced when characterizing–and detecting–state-sponsored disinformation campaigns.

One of the main research challenges in understanding information manipulation is that it is a moving target. Malicious actors will modify their tactics in response to policy and technological changes. Specifically, any (human and/or automated) system for the detection of disinformation operations should also be designed for and tested under adversarial conditions, i.e., under the assumption that the disinformation agents will adapt to evade detection. Techniques from cybersecurity (red teaming and adversary emulation) could be adapted to the information integrity space to help understand the effectiveness of any proposed solutions.

A related challenge for the detection of disinformation campaigns is that automated, machine-learning based approaches are generally known to degrade as they are applied in *cross-domain* settings–i.e., when conditions differ too much from their training data. For example, we found that a classifier of state-sponsored social media accounts trained on Russian tweets (IRA 2016 campaign) suffered a large drop in accuracy when tested on tweets from the Iranian Endless Mayfly operation. More work needs to be done on ensuring that automated detectors continue to perform well on new disinformation campaigns. ARL:UT has expertise in designing and evaluating machine learning systems in cross-domain settings for various natural language processing applications such as entity recognition[3] and detection of computer-generated text[4]. We believe these sorts of techniques should be studied in the context of the use of automated approaches for detecting disinformation campaigns.

Another key research challenges in understanding state-sponsored information operations is that much of the relevant data is locked behind proprietary platforms. However, many information operations also make use of proxy websites to amplify a narrative.[5] Hyperlink data is readily available, and the analysis of hyperlink networks has shown promise in characterizing malicious (misinformational) sites[6]. ARL:UT's CCU has ongoing work in collaboration with UT Austin's Global (Dis)Information Lab (GDIL) in characterizing state-sponsored disinformation campaigns

[2] Martin, D., J. Shapiro, and J. Ilhardt. "Trends in online influence efforts." Empirical Studies of Conflict Project (2020).

[3] Rodriguez, Juan Diego, Adam Caldwell, and Alex Liu. "Transfer learning for entity recognition of novel classes." Proceedings of the 27th International Conference on Computational Linguistics. 2018.

[4] Rodriguez, Juan Diego, Todd Hay, David Gros, Zain Shamsi and Ravi Srinivasan, "Cross-Domain Detection of GPT-2-Generated Technical Text", Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2022

[5] Global Engagement Center, "Pillars of Russia's disinformation and propaganda ecosystem." U.S. Department of State, Washington, D.C. (2020).

[6] Sehgal, Vibhor, et al. "Mutual hyperlinking among misinformation peddlers." arXiv preprint arXiv:2104.11694 (2021).

and identifying disinformation outlets and high-risk affinity groups from freely available information on the web. If automated disinformation detection methods are accurate enough, they could be used to help prevent the narratives from a disinformation operation from spreading too widely. In addition, using link information has some potential advantages compared to an approach that only uses the document text (e.g., more easily handling multilingual disinformation campaigns). To the best of our knowledge, most research on disinformation has focused on data on proprietary platforms (e.g., social media) or on data centered around analysis of text content. Alternate approaches would be an interesting and potentially useful area of research.

Question 2: Preserving information integrity and mitigating the effects of information manipulation

> "Strategies for protecting information integrity must integrate the best technical, social, behavioral, cultural, and equitable approaches. These strategies should accomplish a range of objectives including to detect information manipulation, discern the influence mechanisms and the targets of the influence activities, mitigate information manipulation, assess how individuals and organizations are likely to respond, and build resiliency against information manipulation. What are the key gaps in knowledge or capabilities that research should focus on, in order to advance these objectives? What are the gaps in knowledge regarding the differential impact of information manipulation and mitigations on different demographic groups?"

Given the complexity and magnitude of the problem, we believe that efforts to mitigate the effects of information manipulations will involve both technological as well as educational and policy initiatives. However, we will focus our answer here on the technological solutions to detecting information manipulation, as well as their shortcomings which should be addressed by future research.

While much disinformation is still generated by human troll farms, we believe it is likely to become increasingly automated as the AI technology for creating realistic content ("deepfakes") improves. The capabilities for creating synthetic video, audio, images and text has improved tremendously in the past few years and is becoming increasingly accessible. There are already examples of these technologies being used for malicious purposes. According to the Director of the U.S. National Counterintelligence and Security Center (NCSC), fake social media profiles with computer-generated faces have been used to recruit spies.[7] Even the mere possibility of a piece of content being a deepfake can have dangerous implications. For example, Gabon's military justified its attempted coup in part by claiming that a video of Gabon's current president was a deepfake.

Much research has been done on detecting deepfake images and video. Given that text is just as important a medium as images or video (e.g., for communicating online, for the public

[7] https://apnews.com/article/ap-top-news-artificial-intelligence-social-platforms-think-tanks-politics-bc2f19097a4c4fffaa00de6770b8a60d

record, and for the scientific enterprise), we believe more research is needed for the detection of computer-generated text. Studies have shown that people have difficulty in distinguishing real from generated text[8,9], and tools for automated detection could help defend against computer-generated disinformation. While automated detectors of computer-generated text have been built, they can degrade when faced with text which is statistically different from the text used in building the detector.[10,11,12,13]

Our group investigated whether it is possible to adapt a detector of computer generated text for scientific articles in one technical subject area to another[14]. However, much remains to be done in ensuring that detectors are reliable and robust. Current challenges include defending against perturbed versions of generated text (adversarial examples), ensuring that detectors still work with higher quality fakes (e.g., text from more powerful text generators such as GPT-3) and detecting other forms of document tampering. For example, we found that detecting paragraph replacements is possible, but further work is required to accurately detect word or phrase replacements.[15]

Finally, in another CCU project we investigated whether it was possible to watermark text by hiding a small "message" (watermark) in the text itself, rather than in the file holding the text. We devised a method to do so such that the watermark has the desirable properties of robustness (i.e., small manipulations to the text should not remove the watermark) and imperceptibility (the watermark should look similar to the original text and the watermark should not be obvious). We believe that text-based watermarking is another promising approach that could be used to verify the information integrity of text.

[8] Kreps, Sarah, R. Miles McCain, and Miles Brundage. "All the news that's fit to fabricate: AI-generated text as a tool of media misinformation." Journal of Experimental Political Science 9.1 (2022): 104-117.

[9] Clark, Elizabeth, et al. "All That's 'Human'Is Not Gold: Evaluating Human Evaluation of Generated Text." Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2021.

[10] Solaiman, Irene, et al. "Release strategies and the social impacts of language models." arXiv preprint arXiv:1908.09203 (2019).

[11] Ippolito, Daphne, et al. "Automatic Detection of Generated Text is Easiest when Humans are Fooled." Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020.

[12] Bakhtin, Anton, et al. "Real or fake? learning to discriminate machine from human generated text." arXiv preprint arXiv:1906.03351 (2019).

[13] Stiff, Harald, and Fredrik Johansson. "Detecting computer-generated disinformation." International Journal of Data Science and Analytics (2021): 1-21.

[14] Rodriguez, Juan Diego, Todd Hay, David Gros, Zain Shamsi and Ravi Srinivasan, "Cross-Domain Detection of GPT-2-Generated Technical Text", Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2022

[15] Schuster, Tal, et al. "The Limitations of Stylometry for Detecting Machine-Generated Fake News." Computational Linguistics 46.2 (2020): 499-510.

Question 4. Barriers to research

> "Information integrity is a complex and multidisciplinary problem with many technical, social, and policy challenges that requires the sharing of expertise, data, and practices across the full spectrum of stakeholders, both domestically and internationally. What are the key barriers for conducting information integrity R&D? How could those barriers be remedied?"

One significant barrier to researchers working on automated techniques to detect disinformation campaigns is the availability of suitable data. We strongly believe that making datasets available to researchers will help shape and drive research on detecting disinformation campaigns. In particular, we believe a concentrated effort on creating data should address the following concerns:

- The datasets should specifically be tied to mitigation techniques. Several recent papers have proposed various disinformation kill chains (e.g., the MITRE disinformation kill chain included in the October 2019 report "Combatting Targeted Disinformation Campaigns" funded by the ODNI and DHS). However, many widely available datasets on detecting disinformation used in machine learning studies seem to have been created without such a kill chain or specific mitigation techniques in mind. In particular, we believe many datasets are created simply because the data is relatively straightforward to gather and/or label, not because the dataset was designed with specific, deployable methods for mitigating a disinformation campaign. Creating the dataset specifically to tie into and disrupt a particular step in a disinformation kill chain would help machine learning researchers with problem formulation and algorithmic development such that the proposed approaches could more easily be deployed in practice.
- Similarly, the datasets should be purposefully designed to mirror how detection would work in practice. For example, if detecting disinformation campaigns based on the text content of the disinformation is useful to disrupt a campaign in practice, then the focus should be on creating these types of datasets. However, it is not clear if this is the case, and datasets based on other information (e.g., network effects) should be created purposefully to enable government and industry to detect and disable future disinformation campaigns.
- Crafting a good dataset can also shape future research. For example, the types of tactics, techniques, and procedures (TTPs) used in a disinformation campaign differ based on numerous factors. A benchmark dataset will only be able to capture some of these TTPs, but making the dataset widely available will naturally attract many researchers towards detection techniques for those specific TTPs. Moreover, carefully creating a separate training set and test set for automated approaches can be used to motivate techniques that can generalize across TTPs and threat actors. Similarly, it is unclear how much data would actually be available to detect emerging disinformation campaigns in actual problem settings. The amount of data alone would shape the types of machine learning approaches studied in the academic space.

Closing Comments

Solutions to mitigate the effects of disinformation should be multifaceted; in this response we have focused on the aspects of the problem for which technological approaches could play a role. R&D from academia, UARCS and FFRDCs can be useful both for understanding the disinformation ecosystem and for developing tools to discover and disrupt future information operations. However, there are a number of challenges which will prevent research efforts from being maximually beneficial when deployed:

1. Malicious actors will modify their strategies as the environment changes (in particular, to evade detection or disruption). Systems should be designed and evaluated under adversarial conditions.
2. Further effort is required to ensure that approaches continue to work effectively with novel disinformation campaigns or novel actors.
3. The availability of suitable data to develop and test mitigation techniques is essential. Effort must be taken to ensure that such data should mirror how a disinformation campaign could be disrupted in practice.
4. Deepfakes pose a major threat to information integrity. Further work is required to develop effective and robust detectors of deepfake content, or to verify the integrity of a given piece of content in other ways.

For additional information, please feel free to contact us at:

- Center for Content Understanding
- https://www.arlut.utexas.edu/ccu/index.shtml