When the Bad is Good and the Good is Bad: Understanding Cyber Social Health Through Online Behavioral Change

Ugur Kursuncu, University of South Carolina, Columbia 29208, SC, USA
Hemant Purohit, George Mason University, Fairfax 22030, VA, USA
Nitin Agarwal, University of Arkansas Little Rock, Little Rock 72204, AR, USA
Amit Sheth, University of South Carolina, Columbia 29208, SC, USA

Online platforms have facilitated the exchange of harmful content at an unprecedented scale in the form of disinformation, hate speech, cyberbullying, and extremism. Its insidious impact has been observed in events related to health, disaster recovery, elections, finance, climate communication, and terrorism. These trends have led to a rising prominence of social media analytics in academia, public health, politics, and homeland security, using computational techniques. In the pursuit of understanding online malevolent behavior, research in social media analytics has seen significant development of advanced techniques.[1] Yet, it has been challenging to detect, monitor, counter, and overcome the malevolent behavior by ill-intentioned actors due to the complex nature of social media, and other large-scale sociotechnical infrastructures.[2] As harmful content is rich in subjectivity and emotion, it is challenging to understand individual messages in terms of its features and the human decision-making processes that foster its diffusion. Hence, the meaning of language varies depending on the source's intent and the state of the target's belief system, allowing the bad to be perceived as good due to positive social construction and vice versa.[3] Additional complexity arises when bad actors use coordinated actions with intentions for harming other individuals for a variety of goals from manipulation to harassment. This becomes even more dangerous if malicious groups or state actors orchestrate their actions,[4] involving bots[5] and human actors, to disseminate misinformation and persuade individuals on its truthfulness by malicious groups, threatening an individual, our society, or democratic institutions at large.[6] These efforts have led to negative consequences. Recent examples demonstrating the urgency to address this scourge include:

1) the successful campaigns spreading mis/disinformation around human health (e.g., COVID-19-see https://cosmos.ualr.edu/covid-19, Zika);
2) cyberbullying, harassment, and hate speech by individuals and groups;[7]
3) extremist groups (e.g., ISIS, white supremacy) spreading their propaganda.

> *In spite of significant progress in technologies to fight negative uses of social media, it has been challenging to detect, monitor, counter, and overcome the malevolent behaviors and use by ill-intentioned actors.*

Such information distorts the existing belief in the memory of each individual, challenging the human tendency to avoid conflict with the existing belief. Moreover, the repetition of exposure to such content helps consolidate these beliefs in memory. A carefully constructed sequence of harmful content persuades the target causing a change in behavior. Researchers can characterize individuals based on different facets and dimensions, such as their intentions, biases, socio-cultural affinity, and motivations through learning representations from thick data.[8] For instance, the individual who propagates misinformation will be represented differently from the recipient individual targeted to this misinformation. On the other hand, it is crucial to have a theoretical grounding of the design of computational models in well-established social theories concerning neural, psycholinguistic, and

1089-7801 © 2021 IEEE
Digital Object Identifier 10.1109/MIC.2020.3045232
Date of current version 3 February 2021.

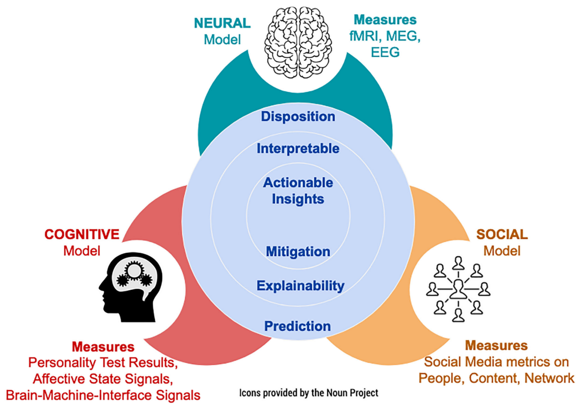

FIGURE 1. Conceptual design that demonstrates modeling at cognitive, neural, and social levels for cumulative measurements in prediction, explainability, and mitigation of misinformation.

cognitive processes in the decision-making of humans. The insights derived from this thick data modeling approach will contextualize the big data analytics at larger scale and provide deeper insights (see thick data modeling section for more.)

Harmful content flows through online communities, where information diffusion of such content occurs through complex dynamic interactions. What constitutes harmful, cyberbullying, or misinformation varies based on the existing belief system of an individual as well as the diffusion of the content. Hence, modeling the cognitive processes of individuals driving decision-making and actions requires incorporating a multidimensional understanding of messages including politics, religion, hate among others. In this case, experimental evaluation of theories, techniques, models/algorithms, and provenance of information, and trust perception of its source are fundamental in each stage of diffusion (see the diffusion of new harmful content for more). Human behavior changes upon exposure to harmful content online, due to the process involving information leading to persuading a human to take an action. This behavioral change might occur gradually at cognitive, neural, and social levels. Figure 1 illustrates a conceptual design to demonstrate modeling this behavioral change through cumulative measurements at these levels. In the figure, the innermost layer tasks (e.g., mitigation, deriving actionable insights) leverage the power of the outer layers (e.g., explainability, disposition) to give insights into mitigation of misinformation.[9] These (online/offline) measurements performed at cognitive, neural, and social levels (e.g., fMRI, state signals, and social media metrics) will provide context for a richer and sensible analysis. This approach can be applied to model specific use cases: i) cyberbullying and/or harassment leading to mental health issues, ii) extremist propaganda leading to radicalization, iii) misinformation and disinformation leading to damaging one's own or others' health and well-being.

In the following sections, we describe thick data modeling and its utility to understand the content, its flow in a network, the trust and provenance factors, and the diffusion of harmful content. Then, we discuss how the insights from these analyses would provide richer context to the big data analysis, especially for combating malicious attacks online. Before we briefly introduce the accepted articles, we also provide a brief discussion on the fairness of these approaches, ethical considerations, and their implications in the society.

THICK DATA MODELING FOR UNDERSTANDING MESSAGE CONTENT

Deriving insights from big data alone may lead to overlooking important details, given the complexity of human behavior online. Hence, a thick data modeling approach to analyze cognitive, neural, and social dimensions will provide a contextual understanding of big data analysis.[10,11] To understand the message content, multidimensionality in models requires operationalizing abstract models of behavior from contextual dimensions, such as culture, politics, and psychology, rendering them computationally accessible. Further, modeling new as well as existing information environments for humans online mandates a holistic approach that cultivates representation of misinformation to understand its effects on the existing memory and cognitive processes. An individual or a group attempts to change the belief system of its target, harass, or bully, and in some cases incite the target to carry out an action. This can be a benign action, such as buying a product or a destructive action such as radicalizing a disoriented or lonely individual into a violent extremist. Information online is perceived by the human brain based on the existing belief, accordingly, changing patterns in behavior, language, and cognitive processes. For instance, even if the individual does not adopt the misinformation due to the existing belief, the influence of such misinformation continues (Liar's Dividend).[12]

As harmful content is usually subjective depending on the context, assessment of the actual meaning for a concept is crucial for the reliable analysis. Different semantics of concepts affect an individual's decision-making. This mandates learning such representations from structured and unstructured factual knowledge resources as prior. For instance, in the context of Islamist extremism, the true meaning of "jihad" can be propagated as harming others in the name of religion; on the other hand, in the context of the religion of Islam, it can be self-struggle to become a better person or fight for self-defense.[13] Hence, distinguishing these semantic differences in the assessment of such narratives and incorporating prior knowledge in contemporary models is essential. It is unlikely that a language model built from a large corpus will provide a clear context as it would be possible by using a knowledge graph or ontology.[14]

FLOW OF HARMFUL INFORMATION, TRUST, AND PROVENANCE

Gradually minimizing the spread of incorrect beliefs via introducing corrective information has been found useful.[15] Sources of corrective information are mainstream news media, certain government sources, sociocultural transcripts, treatise (e.g., Quran, Bible phrases), trusted collective intelligence (e.g., Wikipedia), and trusted data sources (e.g., USAfacts.org). Other approaches to debunking,[16] such as rebuttal, factual elaboration, identifying intentions, as well as trust and provenance-based information credibility, have been effective as well. Propagation of misinformation across different platforms poses challenges to measure trust and capture provenance for information. Modeling trust and provenance require explainability and assessing the veracity of information and their impact on decision making. However, since messages between source and target could employ different features over time, each stage of diffusion offers unique content characteristics for different contextual dimensions. Hence, we need to better understand how the message with the misinformation is adopted and/or propagated by an individual. For instance, recent studies[13] of online radicalization by ISIS via persuasive tactics and strategies demonstrated the need for such an approach. These studies have observed that neurolinguistic, cognitive, and behavior changes were largely contextualized by religious, hate, and ideological dimensions. As these problems require incorporating theories from social science concerning human behavior, thick data modeling can be an essential approach to decipher these complex patterns.

THE DIFFUSION OF NEW HARMFUL CONTENT

As harmful content flows through the online social networks, we need to have a framework to conceptually model the dynamics of diffusion of this information. Prior research (Rogers, 2010)[17] describes the diffusion of new information in five stages: i) Acquiring new knowledge: exposition to new information, ii) Persuasion: a favorable attitude is formed, iii) Adoption: the result of persuasion, iv) Taking action: on the adopted information (e.g., propagation through the network), v) Confirmation: reinforcing the information via the outcome of the action. This process is gradual and primarily influenced by the contextual information in online communities. As the information environment changes upon exposure to new information, contextual predictive factors vary in each stage. Hence, developing theories concerning how the diffusion of misinformation takes place online will require a specific focus on the causal chain of events that triggers the transition to the next stage. For understanding this diffusion process, sometimes orchestrated by groups, a pressing need exists to develop robust mechanisms to capture such discourse online and derive insights utilizing novel approaches that go beyond current statistical and network science approaches. These models are mostly dependent on existing datasets that contain inherent biases and lack the most current information. Hence, this leads to another significant challenge in developing methods for automated dynamic assessment of harmful narratives as per the dynamic nature of events occurring in a fast-paced world.

CONTEXTUALIZED BIG DATA ANALYSIS TO COMBAT MALICIOUS BEHAVIORS

To this end, the measurements mostly concern smaller scales of data points, and the outcome might also be translating to a smaller effect. On the other hand, the information derived from these outcomes will provide contextual understanding of the bigger grand scheme of the problem at hand. Specifically, to understand online behavioral change and its reflection at the society level, we need to design complex system studies to test the feasibility and efficacy of possible combatting approaches. This design will require formally characterizing information environments to understand malicious attacks or campaigns and modeling the gradual diffusion processes to understand persuasive harmful dissemination campaigns. As described earlier, we can test hypotheses with smaller data, concerning the driving factors of these attacks and the diffusion process. Such complex system study will provide insights on how and more importantly why a particular information adoption process does or does not work. This will further inform counter-tactics and strategies.

FAIRNESS AND ETHICAL CONSIDERATIONS

The implications of these analyses on sensitive issues impact individuals as well as society; hence, it is imperative to develop fair algorithms and models taking ethical considerations into account. Bias is usually inherent in the data or introduced in the processing, algorithmic design, or evaluation phases.[18] As researchers, we have very limited understanding of the implications of computational models that we design for harmful online behaviors. Thus, the very models we design might inadvertently reinforce and amplify the biases. The impact of such pitfalls in fair algorithmic design might be elevated when deployed in an application used by millions of people. For instance, a recent study[13] showcased how bias in verified data for extremism might lead to potentially unfair social discrimination against innocent individuals. While removing biases from data in entirety may not be possible, we need to explore solutions that will mitigate bias and promote fairness in the model.

IN THIS ISSUE

We received 23 submissions for this special issue, which were reviewed by at least three reviewers. Nine articles were accepted based on the quality of the analysis, results, and presentation. In this issue, the following four articles will appear, while the remaining five articles will appear in the upcoming issue.

The article "Effect of Conformity on Perceived Trustworthiness of News in Social Media" examines how a mix of supportive and critical comments on social media platforms influence the readers' perception of the trustworthiness of news articles and how they decide to trust and act. The authors conducted a user survey and confirmed that Facebook readers display conformity to the majority view by adjusting their personal opinions. They investigate if other comments influence the comments in a thread or they are independent. They found that conformity is an important factor in posting a comment. Specifically, their findings include: i) users tend to conform to the majority opinion, ii) less confident users are more likely to conform to the majority, iii) supportive majority leads to more echo and fact-checking from conforming supportive users, iv) critical majority leads to more reporting from conforming critical users, and finally, v) following majority leads to a better decision on trustworthiness.

In "Cyberbullying Detection with Fairness Constraints," the authors address the problem of cyberbullying detection with fairness in focus, using a theoretically grounded approach of constraint optimization. Specifically, they aim to improve the performance of machine learning algorithms without reinforcing unintended social biases by guiding the model training with fairness constraints. They experimented using multiple datasets for cyberbullying to validate their proposed method, demonstrating the value in fairness constraints without sacrificing performance.

In "Misinformation Sharing on Twitter During Zika: An Investigation of the Effect of Threat and Distance," the authors recognize the evolving aspect of misinformation over time and use the Zika virus as a use case for modeling the problem of misinformation detection. They employed an extensive feature engineering capturing multiple dimensions concerning the dynamic nature of misinformation spread formulated through the nonhomogeneous poisson process. The predictive features include information on the source of information (user), content, the medium of dispersion, network transmission, and temporal information.

In "HateClassify: A Service Framework for Hate Speech Identification on Social Media," the authors reframed hate detection as a multilabel (e.g., hate, offensive, nonoffensive) classification problem utilizing a sequential CNN architecture. They crowdsourced the ground truth data for hate speech rather than relying on the interpretation of selected individuals. Their framework employs the active learning paradigm as the model is continuously updated with new crowd-sourced data.

IN THE UPCOMING ISSUE

In "Towards Hate Speech Detection at Large via Deep Generative Modeling," the authors developed an approach to improve supervised hate speech detection on social media by creating a large dataset of hate speech from a small seed set. They introduced a big ground truth dataset and assessed the generalizability of models to the variability in communications with hate speech. This article attempts to overcome the lack of diversity and improve coverage in the input dataset, and the data imbalance. The authors employ

GPT-2 fine tuned on the existing labeled datasets, to generate a larger diverse hate speech dataset. They also perform a comparative analysis on the inductive biases of DL methods during training on individual hate-speech datasets.

The article "Emotional Communication during Crisis Events: Mining Structural OSN Patterns" highlights the importance of online emotional contagion in order to understand and support the emotional state of online users during disasters. The authors analyzed the structural network patterns on Twitter that arise as people exchange emotional messages online, using the emotion-exchange motifs which provide insights on the variability of emotions in different contexts across different types of disasters.

The article "Session-based Cyberbullying Detection: Problems and Challenges" provides an overview and road-map of research for cyberbullying. The authors define session-based cyberbullying and describe challenges characterizing cyberbullying through features that account for multimodality, temporality, hierarchical structure, and user interactions.

In "Approaches for Fake Content Detection: Strengths and Weaknesses to Adversarial Attacks" authors provide a folksonomy of models for fake content detection with their characteristics that are susceptible to different adversarial attacks, and the ways to mitigate the impact of these attacks on the model. They categorize these models based on key characteristics that would potentially affect the adversarial attack performance, and briefly explain the current state of research in these categories.

The article "Analysing Public Opinion and Misinformation in a COVID-19 Telegram Group Chat" analyzes the content posted in a Telegram channel from Singapore, with a particular focus on the spread and interaction with misinformation related to COVID-19. The authors specifically investigate how opinions of users in the group change over time and how users react to fact-checked information. They analyze misinformation through sentiment, topics, and psychological features, and notably show how negative sentiment increased when the alert level was raised by the government. Although anxiety seemed to decrease as the disease progressed, anger and sadness started to increase. The authors found that misinformation was either largely denied or challenged by users, and skepticism toward misinformation was a driving factor in this behavior. The insights and observations concerning how the sentiment and emotions toward misinformation evolve during pandemic, provide avenues for further research.

ACKNOWLEDGMENTS

Figure 1 had emerged during interactions that also involved Huan Liu, Amanuel Alambo, Amit Almor, Douglas Wedell, Marco Valtorta, and Valerie Shalin. This work was supported in part by the NSF under Grant CNS 1513721, Grant OIA-1946391, Grant OIA-1920920, Grant IIS-1815459, Grant CMMI-2029719, U.S. Office of Naval Research (N00014-19-1-2336), Air Force Research Lab, Army Research Office (W911NF-17-S-0002, W911NF-16-1-0189), and Arkansas Research Alliance. The opinions expressed are those of the authors and do not reflect those of the sponsors.

REFERENCES

1. U. Kursuncu, G. Manas, U. Lokala, K. Thirunarayan, A. Sheth, and I. Budak Arpinar, "Predictive analysis on Twitter: Techniques and applications," in *Emerging Research Challenges and Opportunities in Computational Social Network Analysis and Mining*, pp. 67–104, New York, NY, USA: Springer-Verlag, 2019.
2. H. Purohit and R. Pandey, "In *to Appear*. *Intent Mining For the Good, Bad & Ugly Use of Social Web: Concepts, Methods, and Challenges*, in *Emerging Research Challenges and Opportunities in Computational Social Network Analysis and Mining*, N. Agarwal, N. Dokoohaki, and S. Tokdemir eds., New York, NY, USA: Springer, 2019.
3. R. Pandey, H. Purohit, B. Stabile, and A. Grant, "Distributional semantics approach to detect intent in twitter conversations on sexual assaults," in *Proc IEEE/WIC/ACM Int. Conf. Web Intell.*, pp. 270–277, Dec. 2018.
4. N. Agarwal and K. K. Bandeli, "Examining strategic integration of social media platforms in disinformation campaign coordination," *J. Nato Defence Strategic Commun.*, vol. 4, pp. 173–206, 2018.
5. N. Agarwal, A. Samer, R. Galeano, and R. Goolsby, "Examining the use of botnets and their evolution in propaganda dissemination," *J. NATO Defence Strategic Commun.*, vol. 2, pp. 87–112, 2017.
6. O. Katherine, D. Lazer, R. E. Robertson, and C. Wilson, "Misinformation in action: Fake news exposure is linked to lower trust in media, higher trust in government when your side is in power," Harvard Kennedy School, *Misinf. Rev.*, 2020.
7. T. Wijesiriwardene *et al.*, "ALONE: A dataset for toxic behavior among adolescents on Twitter," in *Proc Int. Conf. Social Inf.*, 2020, pp. 427–439.
8. F. Jinan, "Envisioning insight-driven learning based on thick data analytics with focus on healthcare," *IEEE Access*, vol. 8, pp. 114998–115004, 2020.

9. G. Pennycook and D. G. Rand, “Fighting misinformation on social media using crowdsourced judgments of news source quality,” *Proc. Nat. Acad. Sci.*, vol. 116, no. 7, 2019, pp. 2521–2526.
10. G. Latzko-Toth, C. Bonneau, and M. Millette, “Small data, thick data: Thickening strategies for trace-based social media research,” in *SAGE Handbook Social Media Research Methods*. Newbury Park, CA, USA: Sage, 2017, pp. 199–214.
11. J. Fiaidhi, “Envisioning insight-driven learning based on thick data analytics with focus on healthcare,” *IEEE Access*, vol. 8, pp. 114998–115004, 2020.
12. L, Stephan, U. K. H. Ecker, C. M. Seifert, N. Schwarz, and J. Cook, “Misinformation and its correction: Continued influence and successful debiasing,” *Psychol. Sci. Public Int.*, vol. 13, no. 3, 2012, pp. 106–131.
13. U. Kursuncu *et al.*, “Modeling Islamist extremist communications on social media using contextual dimensions: Religion, ideology, and hate,” in *Proc. ACM Hum.-Comput. Interact.*, vol. 3, 2019, pp. 1–22.
14. U. Kursuncu, M. Gaur, and A. Sheth, “Knowledge infused learning (K-IL): Towards deep incorporation of knowledge in deep learning,” in *Proc. AAAI Spring Symp. Combining Mach. Learn. Knowl. Eng. Pract.*, vol. 2600, 2020, pp. 1–10.
15. U. K. H. Ecker, S. Lewandowsky, B. Swire, and D. Chang, “Correcting false information in memory: Manipulating the strength of misinformation encoding and its retraction,” *Psychcol. Bull. Rev.*, vol. 18, no. 3, pp. 570–578, 2011.
16. K. Sharma, F. Qian, H. Jiang, N. Ruchansky, M. Zhang, and Y. Liu, “Combating fake news: A survey on identification and mitigation techniques,” *ACM Trans. Intell. Syst. Technol.*, vol. 10, no. 3,pp. 1–42, 2019.
17. E. M. Rogers, *Diffusion of Innovations*. New York, NY, USA: Simon and Schuster, 2010.
18. A. Olteanu, C. Castillo, F. Diaz, and E. Kiciman, “Social data: Biases, methodological pitfalls, and ethical boundaries,” *Front. Big Data*, vol. 2, p. 13, 2019.

UGUR KURSUNCU is a Postdoctoral Fellow with the Artificial Intelligence (AI) Institute, the University of South Carolina, Columbia, SC, USA. His primary research interests include knowledge-infused learning in AI, social computing, and health informatics, focusing on building socially responsible human-centered intelligent systems. Contact him at kursuncu@mailbox.sc.edu.

HEMANT PUROHIT is an Assistant Professor of Information Sciences and Technology with George Mason University, Fairfax, VA, USA. His research interests include social computing and natural language understanding with a focus on designing systems for community resilience against natural crises (e.g., hurricanes), societal crises (e.g., migration, violence), and human crises (e.g., terrorism, cyber attacks). Contact him at hpurohit@gmu.edu

NITIN AGARWAL is the Jerry L. Maulden-Entergy Chair and Distinguished Professor of Information Science with University of Arkansas Little Rock, Little Rock, AR, USA. He is the Founding Director of the Collaboratorium for Social Media and Online Behavioral Studies. His research interests include social computing, (deviant) behavior modeling, social-cyber forensics, health informatics, data mining, and privacy. Contact him at nxagarwal@ualr.edu

AMIT SHETH is the Founding Director with Artificial Intelligence Institute, University of South Carolina, Columbia, SC, USA. His current core AI research is in knowledge-infused learning and explanation, and translational research includes personalized and public health, social good, education, and future manufacturing. He is a fellow of IEEE, AAAI, and AAAS. Contact him at amit@sc.edu.