DISCLAIMER: This article contains examples of offensive and abusive language as a necessary part of illustrating its research findings.

THEME ARTICLE: CYBER SOCIAL HEALTH

Session-Based Cyberbullying Detection: Problems and Challenges

Lu Cheng, Yasin N. Silva, Deborah Hall, and Huan Liu, *Arizona State University, Tempe, AZ, 85281, USA*

Cyberbullying has become one of the most pressing online risks for young people, due in part to the rapid increase in social media use, and has raised serious concerns in society. Existing studies have examined various approaches to cyberbullying detection focusing on a single piece of text, whereas relatively little is known about cyberbullying detection within a **social media session.** ***A social media session typically consists of an initial post, images/videos, a sequence of comments that involves user interactions, user information, spatial location, and other social content. By investigating cyberbullying at the level of social media sessions, researchers can draw on data that are more complex, diverse, and crucial for understanding two defining characteristics of cyberbullying, in particular:*** **repetitive acts** ***and*** **power imbalance.** ***This article thus highlights the importance of studying session-based cyberbullying detection, identifies core challenges, and serves as a resource to help direct future research efforts.***

The number of cyberbullying instances has been rising at an alarming rate. Recent studies indicate that, overall, 36.5% of people report that they have experienced cyberbulling in their lifetime.[1] Despite some variability in working definitions of cyberbullying and different views on the degree of overlap between cyberbullying and related terms (e.g., cyberaggression), cyberbullying is generally defined as "intentional acts of aggression carried out by a group or individual using electronic communication, *repeatedly or over time* against victims who cannot easily defend themselves."[2] Crucial components of this definition are a *power imbalance* between bully and victim and the *repetition* of the aggressive acts over time. Examples of cyberbullying include repetitively sending mean or threatening messages, tricking someone into revealing personal or embarrassing information and sending it to others, and sharing explicit images/videos of others without their consent.

1089-7801 © 2020 IEEE
Digital Object Identifier 10.1109/MIC.2020.3032930
Date of publication 22 October 2020; date of current version 16 April 2021.

Existing efforts[3–6] in cyberbullying detection have been largely directed toward studying social media posts that include either a single piece of text (e.g., a single tweet) from one user or the concatenation of multiple pieces of text (e.g., multiple replies) from different users. These text-based approaches have revealed the superior power of Natural Language Processing (NLP) techniques such as *n*-grams (with or without tf-idf weighting), part-of-speech information (e.g., first and second pronouns), and sentiment information based on (polarity and profanity) lexicons for this task.

A common limitation of these methods is that they do not fully capture the unique characteristics of cyberbullying behavior—power imbalance and repetition—that might significantly improve the performance of cyberbullying detection.[7,8] Challenges also arise when applying NLP techniques and machine learning tools to cyberbullying detection, given key differences between social media data and traditional documents (e.g., newspaper articles). To illustrate, social media data are typically real-time, geospatially coded, and comprising communications between multiple individuals. The content frequently includes emotion, neologisms, and information of

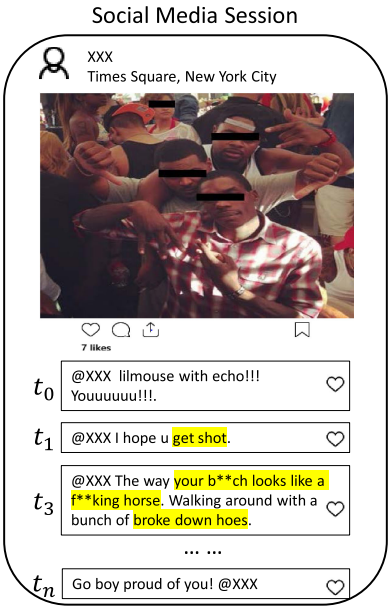

FIGURE 1. Sample social media session on Instagram.

questionable credibility (e.g., fake news, rumors). Nonstructured texts can be found in a multitude of formats, with variation in languages and styles, and posted text is often informal, short, and grammatically incorrect. Simply put, adapting conventional methods for the analysis of social media data presents unique scientific challenges.

To overcome current limitations in cyberbullying detection, a more comprehensive understanding of social media data is necessary. Taking an Instagram post as an example, a typical *social media session* consists of an initial post, images/videos, spatial location, a sequence of time-stamped comments, and other social content such as user names and the number of likes and shares.[9] Figure 1 depicts an Instagram session in which multiple bullying comments were posted and directed toward one or more victims. In contrast to text-based cyberbullying detection, this session-based analysis is grounded on inherently *hierarchical* (i.e., words form a comment, comments form a session) and *multimodal* data (i.e., text, location and image, etc.) with evolving *user interactions*. This enriched data creates an unprecedented range of possibilities for researchers to take into account as they seek to better understand cyberbullying properties. Investigations at the level of a social media session, for example, can provide invaluable insight into the imbalance of power between a bully and victim that may become increasingly evident across an entire session compared to a single text. The repetitive nature of cyberbullying can be captured by the sequence of comments (as shown in Figure 1) within a session. Examination of the hierarchy of a social media session also enables the models to differentiate the importance of media objects within a session. As a result, session-based cyberbullying detection represents both a generalization and a vital extension of text-based analysis and opens promising research directions for identifying, understanding, and ultimately preventing cyberbullying.

DEFINITION OF SESSION-BASED CYBERBULLYING DETECTION

We define session-based cyberbullying detection as *the identification of cyberbullying behavior within a social media session by leveraging multiple media objects including textual features, user interactions, spatial location, temporal information, visual cues, social network, and other social attributes, e.g., users' profile information.*

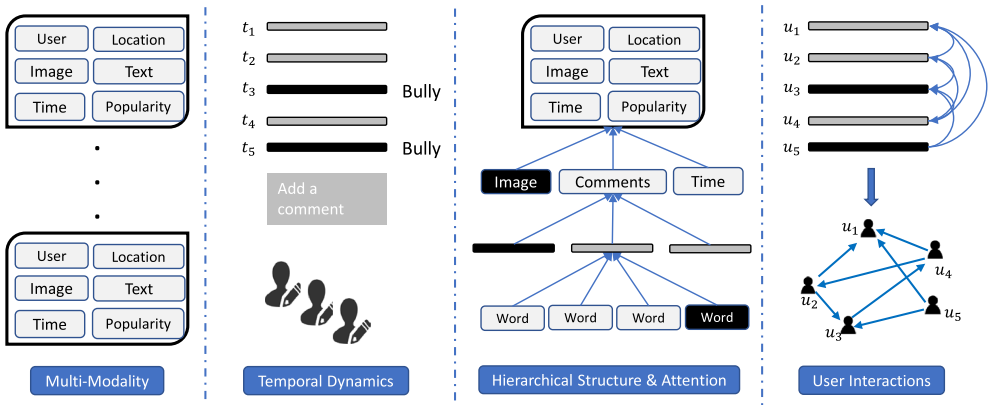

FIGURE 2. Unique features of session-based cyberbullying detection.

As illustrated in Figure 2, session-based cyberbullying detection presents four unique features. (1) *Multimodality.* Conventional models overlook critical information included in the various social media modalities that represent potential ways for bullies to interact with victims. For instance, they can post humiliating images, insulting comments or captions, or edit and then repost another user's images. (2) *Temporal Dynamics.* A social media session evolves over time as the number of comments posted by various users increases. The dynamic nature enables us to examine the repetitive aspect of cyberbullying by taking into account the full history of a social media session. (3) *Hierarchical Structure and Attention.* Social media sessions inherently exhibit a hierarchical structure: from words to comment, from comments to a session. Modeling this hierarchy improves the quality of session representation and enables us to investigate the importance of different media objects. (4) *User Interactions.* The intensive user interactions uncover *indirect* cyberbullying that occurs at the interaction-level rather than at the level of a single post or user. These evolving conversations can facilitate research that identifies the roles of different users, e.g., bully.

CHALLENGES OF SESSION-BASED CYBERBULLYING DETECTION

The identification of cyberbullying in complex sessions presents multiple challenges as well as promising opportunities that differ from those of the traditional cyberbullying detection task. In this section, we highlight several of these challenges including (1) aspects related to the characteristics of social media sessions highlighted in Figure 2, and (2) issues related to data collection such as requirements for privacy preserving and data labeling.

Multimodal Context

Social media sessions are multimodal by nature. A straightforward approach to encode multimodal context is to simply concatenate the raw feature vectors of each modality (e.g., locations, comments, images, timestamps). However, this method overlooks both *structural dependencies* among different social media sessions and *cross-modal correlations* among different modalities.[8] There are two major issues when modeling multimodal context.

- *Number of distinct feature values.* Social media data exhibits considerable variations due to its multimodal nature. There are nonattributed modalities that do not have features such as index representation of users, and attributed modalities that are associated with features such as text, locations, etc. We are often confronted with the data sparsity issue as feature types are diverse and the number of unique feature values each modality can take is overwhelmingly large. The limited size of training data for each modality can further complicate the training process.
- *Cross-modal correlation and structural dependencies.* Information in different modalities is heterogeneous and not compatible with each other. A key problem to be addressed is how to effectively encode the cross-modal correlation among different types of modalities. Since sessions are not independent and identically distributed (i.i.d.) but instead intrinsically correlated (e.g., homophily in social networks—the tendency of individuals to have ties with similar others), it is important to model the structural dependencies among sessions.

Temporal Dynamics

Emerging literature identifies cyberbullying as a continuous temporal phenomena rather than a one-off incident.[10] However, little is known regarding the temporal dynamics of cyberbullying in social media. For instance, the number, frequency, and timing of posts may vary systematically between cyberbullying and noncyberbullying sessions. The integration of temporal analysis tasks to model the evolution of and correlations among comments can be effective tools to improve the accuracy of cyberbullying detection.[7]

A few recent efforts in computer science aimed at modeling the temporal characteristics of cyberbullying. These include models that use point processes,[10] multitask learning,[7] and burst analysis.[11] Specifically, Soni and Singh[10] modeled comments as a series of Dirac delta functions and the level of activity in a social media session under the assumption that each comment boosts the activity level by an exponentially decaying amount. Their findings uncover the significantly different patterns between bullying and nonbullying sessions. In addition to using temporal features, Cheng *et al.*[7] proposed a time interval prediction task, complimentary to the cyberbullying detection task, in order to leverage the temporal characteristics. This multitask learning framework exploits the commonalities and differences across these two tasks to improve the performance of

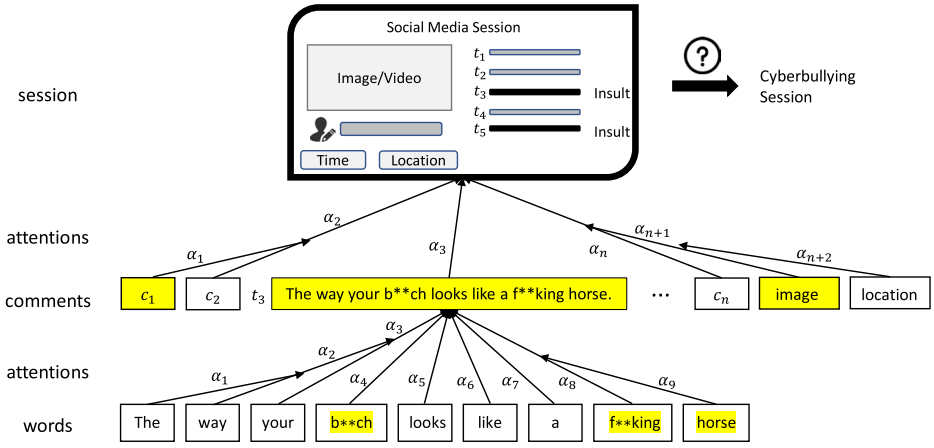

FIGURE 3. Hierarchical structure of a social media session and attention mechanisms.

cyberbullying detection. More recently, several key temporal properties of cyberbullying were presented by Gupta *et al.*[11] using descriptive and burst analyses. This study shows how important metrics such as the number of cyberbullying and noncyberbullying comments change over time in cyberbullying and noncyberbullying sessions. It also reveals the bursty nature of comments in social media sessions.

Hierarchical Structure and Attention

Each session consists of multiple media objects. To illustrate, an Instagram session includes an image, comments, timestamps, and social attributes; each comment further consists of a sequence of words and timestamps. Previous work showed that modeling session structures can improve the quality of session representations.[7] Additionally, different media objects in a session often provide the various levels of information for cyberbullying detection. In fact, the meaning of words and comments are highly dependent on their context. *"You're a f**king gay!"* and *"I'm a gay."* both include the word *gay*, however, the former sentence is more likely to be an instance of cyberbullying. Consequently, an important challenge of session-based cyberbullying detection is how to model the hierarchical structure of social media sessions while enabling mechanisms that allow paying different levels of attention to different session components.

We illustrate the hierarchical structure of a social media session and attention mechanisms in Figure 3. In this figure, a sequence of weighted words form a comment and a sequence of weighted comments, an image and a location form a session. In a supervised learning scenario, the final session vector is the input of a binary classifier which then outputs the (probability of a) label indicating whether the session is bullying.

Modeling User Interactions

Whereas traditional bullying involves face-to-face interactions intended to cause harm to others, cyberbullying often takes place throughout a series of interactions on social media platforms. The majority of the computational research related to cyberbullying, however, has been focused on detecting cyberbullying at the user or post level (e.g., detecting if an individual message is a cyberbullying interaction or not). Therefore, interventions involving reflective user interactions have yet to be studied.[12] A promising research task is to identify cyberbullying that may appear across a sequence of user interactions.

There are at least two tools to leverage when modeling user interactions. The first one is language modeling that focuses on conversational text analysis. For instance, a sequence of comments is more likely to be a candidate for cyberbullying if the underlying topic is of a sensitive nature and includes profanity, negativity, and subtlety.[12] The mathematical models for user interactions in social media sessions can be complex as the use of language in a social setting is parameterized by a rich set of characteristics, and because specificity and uniqueness play a key role in effective interaction analysis. On the other hand, machine learning for language modeling relies on abstraction, generalization, and stable patterns in the data. It is important to find a balance between these two paradigms.[12]

Another tool for studying user interactions is the construction of networks based on explicit user interactions such as replying to a comment or implicit

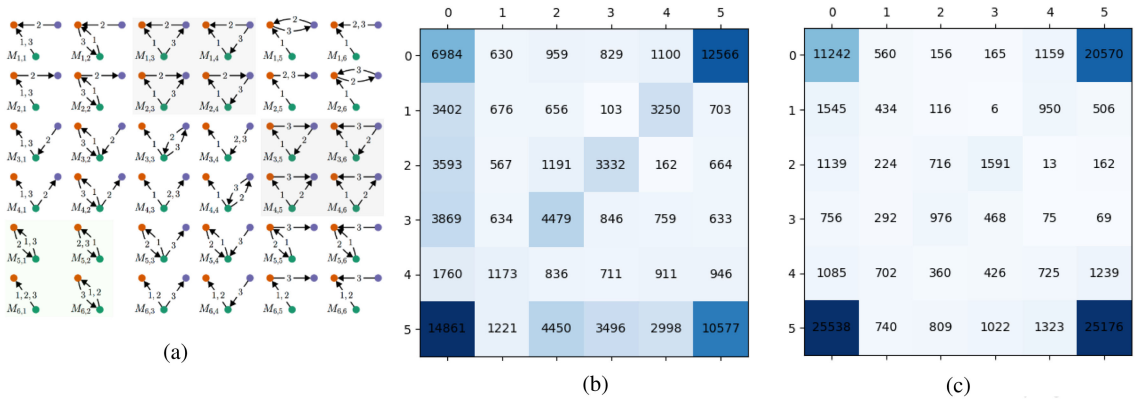

FIGURE 4. Temporal motif analysis based on the count matrix in Figure 4(a). The value in the *i*th row and *j*th column in Figure 4(b) and (c) is the number of instances of motif $M_{i,j}$ in Figure 4(a). The color of a cell indicates the fraction of the cell value w.r.t. the total number of motifs–darker tones = higher values. For example, the motif with the largest count in bullying sessions is $M_{6,1}$ in Figure 4(a) as indicated by the darkest cell (bottom left) in Figure 4(b). We can also observe that there are only four primary motifs in nonbullying sessions.

user interactions defined by specific similarity measures, e.g., content similarity. We refer to this type of network as an *interaction network*. For social networks, various analysis tools can be used to investigate social structures composed of nodes (users) and links (friendships or interactions) such as homophily and network closure. For similarity-based network, a powerful tool to leverage is the graph neural networks[13] that have gone significantly popular recently. Interaction networks can record evolving conversations among social media users. The analysis of these networks can help identify different roles (e.g., victims, bullies, defenders, assistants, and bystanders), and uncover unique as well as common user characteristics.[3]

Network motif analysis, i.e., the study of small subgraph patterns in large graphs, is a promising direction to understand the structure and function of the interaction system modeled by a graph. For instance, a dynamic user interaction graph can be represented by a series of timestamped edges.[14] The goal of temporal network motifs is to discover specific user-interaction patterns in a cyberbullying session. These patterns can take the form of a directed multigraph. Figure 4 shows some initial results of using temporal network motif analysis in the context of cyberbullying. We conducted this experiment using an Instagram dataset[9] and the SNAP tool for motif analysis (https://snap.stanford.edu/temporal-motifs/code.html). We observe the presence of more diverse types of motifs (among the 36 predefined motifs) in bullying sessions compared to nonbullying sessions, i.e., there are more types of motifs with notable count in bullying sessions [see Figure 4(b)] than in nonbullying sessions [see Figure 4(c)]. Future work can be aimed at identifying the specific interaction patterns that set apart bullying from nonbullying sessions.

Privacy Preservation

Recent cyberbullying detection models rely on access to accurate information from potential victims' social media profiles and interactions. In many real-world scenarios, however, the availability of this information is affected by the privacy-preferences of users and limitations imposed by social networking sites. When a session-based identification model is used by the potential victim, the user may want to restrict the amount and type of data available to the model. It becomes even more complex when the use of a model involves the victim's parent (who receives the results of the model). In this scenario, a teen may want to enforce additional privacy-related restrictions on the data that will be accessed by the model and the report presented to the parent. Furthermore, social media platforms such as Facebook and Instagram have recently restricted the type and amount of data an app can access. Privacy preservation is challenging as a desired solution requires (1) the understanding of users' privacy preferences (e.g., potentially conflicting preferences of parents and teens) and the connections between privacy concerns and adoption patterns of automated tools; and (2) the development of models that effectively use incomplete or anonymized data to comply with privacy requirements.

Data Labeling

Data labeling is a time-consuming and labor-intensive process. Before the process starts, it is important to select appropriate definitions of key terms that will be used during ground truth labeling. Previous literature highlighted the difficulty to differentiate among various types of misbehavior such as cyberbullying, cyberaggression, and hate speech.[9] For example, cyberaggression is broadly defined as using digital media to intentionally harm another person while cyberbullying also indicates power imbalance and repetition. Power imbalance can take on a variety of forms, including physical, social-relational, and psychological. Repetition implies bullying behavior occurs over time by sharing a negative comment or photo with multiple individuals.

Data labeling is specially challenging due to the multimodality of social media sessions. When asking human labelers to decide whether a session is cyberbullying, it is important to incorporate information from all available data components with different modalities such as images and text-based comments. Another common issue in cyberbullying data labeling is that researchers often favor collecting sessions that have higher percentages of negative content to obtain datasets with likely larger proportions of bullying instances. This is important for training the session-based cyberbullying classifiers. However, real-world datasets are markedly imbalanced, with only a small percentage of sessions containing cyberbullying content. The distribution shift between training and test data can exacerbate the performance of cyberbullying classifiers.

Finally, the definition of cyberbullying is subjective and can evolve over time.[15] Therefore, it is important to properly design effective cyberbullying labeling strategies and examine the objectivity and consistency of potential contributors in order to generate high-quality labeled datasets.

FUTURE RESEARCH DIRECTIONS

In this article, we discuss the challenges of session-based cyberbullying detection considering the leading traits of cyberbullying behavior, namely, repetition and power imbalance. To help tackle these challenges, we propose here three research directions drawing on recent advances in artificial intelligence. A repetition-centered approach is key toward effective session-based cyberbullying detection. This approach can complement earlier cyberbullying detection contributions by jointly modeling the hierarchical structure and the temporal patterns of multimodal social media sessions and by explicitly constructing user interaction networks. It is also important to enable compliance with users' privacy preferences. To this end, a key task will be the design of privacy-aware algorithms that can adapt to various levels of data availability, e.g., aggregated or anonymized data. Finally, this area of work can benefit from the development of mechanisms to address the limitations of time-consuming data-labeling and the time-evolving understanding of cyberbullying. Unsupervised models for cyberbullying detection could be useful in this domain. These research directions are not independent but rather highly interrelated. They could help realize holistic models to identify and prevent cyberbullying. In addition to the multifaceted challenges and opportunities it brings in the research domain, session-based cyberbullying detection also enables the possibility of developing practical tools, e.g., third-party antibullying apps, on social media platforms.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation under Grants 1719722, 1614576, and 2036127.

REFERENCES

1. S. Hinduja and J. W. Patchin, "Connecting adolescent suicide to the severity of bullying and cyberbullying," *J. School Violence*, vol. 18, no. 3, pp. 333–346, 2019.
2. P. K. Smith, J. Mahdavi, M. Carvalho, S. Fisher, S. Russell, and N. Tippett, "Cyberbullying: Its nature and impact in secondary school pupils," *J. Child Psychol. Psychiatry*, vol. 49, no. 4, pp. 376–385, 2008.
3. L. Cheng, J. Li, Y. Silva, D. Hall, and H. Liu, "Pi-bully: Personalized cyberbullying detection with peer influence," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, 2019, pp. 5829–5835.
4. J.-M. Xu, K.-S. Jun, X. Zhu, and A. Bellmore, "Learning from bullying traces in social media," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, 2012, pp. 656–666.
5. H. Dani, J. Li, and H. Liu, "Sentiment informed cyberbullying detection in social media," in *Proc. Eur. Conf. Mach. Learn. Princ. Pract. Knowl. Discovery Databases*, 2017, pp. 52–67.
6. L. Cheng, R. Guo, and H. Liu, "Robust cyberbullying detection with causal interpretation," in *Proc. WWW' Companion*, 2019, pp. 169–175.
7. L. Cheng, R. Guo, Y. Silva, D. Hall, and H. Liu, "Hierarchical attention networks for cyberbullying detection on the instagram social network," in *Proc. SIAM Int. Conf. Data Mining*, 2019, pp. 235–243.

8. L. Cheng, J. Li, Y. N. Silva, D. L. Hall, and H. Liu, "Xbully: Cyberbullying detection within a multi-modal context," in *Proc. Int. Conf. Web Search Data Mining*, 2019, pp. 339–347.
9. H. Hosseinmardi, S. A. Mattson, R. I. Rafiq, R. Han, Q. Lv, and S. Mishra, "Detection of cyberbullying incidents on the instagram social network," 2015, *arXiv:1503.03909*.
10. D. Soni and V. Singh, "Time reveals all wounds: Modeling temporal characteristics of cyberbullying," in *Proc. Int. AAAI Conf. Web Social Media*, 2018.
11. A. Gupta, W. Yang, D. Sivakumar, Y. N. Silva, D. L. Hall, and M. C. N. Barioni, "Temporal properties of cyberbullying on instagram," in *Proc. Companion Proc. Web Conf.*, 2020, pp. 576–583.
12. K. Dinakar, B. Jones, C. Havasi, H. Lieberman, and R. Picard, "Common sense reasoning for detection, prevention, and mitigation of cyberbullying," *TiiS*, vol. 2, no. 3, pp. 1–30, 2012.
13. J. Zhou *et al.*, "Graph neural networks: A review of methods and applications," 2018, *arXiv:1812.08434*.
14. A. Paranjape, A. R. Benson, and J. Leskovec, "Motifs in temporal networks," in *Proc. Int. Conf. Web Search Data Mining*, 2017, pp. 601–610.
15. L. Cheng, K. Shu, S. Wu, Y. N. Silva, D. L. Hall, and H. Liu, "Unsupervised cyberbullying detection via time-informed Gaussian mixture model," in *Proc. ACM Int. Conf. Inf. Knowl. Manage.*, 2020, pp. 185–194.

LU CHENG is currently working toward a fourth-year Ph.D. degree with the Computer Science and Engineering, Arizona State University, Tempe, AZ, USA. Her research interests include causal learning and data mining. She has authored or coauthored research papers in premier conferences of data mining and machine learning. She is a student member of the ACM, SIAM and AAAI. She is the corresponding author of this article. Contact her at lcheng35@asu.edu.

YASIN N. SILVA is currently an Associate Professor of computer science with the School of Mathematical and Natural Sciences, Arizona State University, Tempe, AZ, USA. His areas of interest include social media analysis, cyberbullying detection in social networks, big data, similarity-aware data analysis, and fairness and transparency in AI. Dr. Silva is an ACM and IEEE member. Contact him at ysilva@asu.edu.

DEBORAH HALL is currently an Associate Professor of psychology with the School of Social and Behavioral Sciences, Arizona State University, Tempe, AZ, USA. Her research interests include social and group identity, cyberbullying, and quantitative methods. Contact her at d.hall@asu.edu.

HUAN LIU is currently a Professor of Computer Science and Engineering with Arizona State University, Tempe, AZ, USA. He received the B.Eng. degree in computer science and electrical engineering from Shanghai Jiao Tong University, Shanghai, China, and the Ph.D. degree in computer science with the University of Southern California, Los Angeles, CA, USA. He is a Fellow of ACM, AAAI, AAAS, and IEEE. Contact him at huanliu@asu.edu.