Can User Characteristics Predict Norm Adherence on Social Media? Exploring User-Centric Misinformation Interventions

Shangde Gao
Department of Urban and Regional Planning
Florida Institute for Built Environment Resilience
University of Florida
gao.shangde@ufl.edu

Yan Wang
Department of Urban and Regional Planning
Florida Institute for Built Environment Resilience
University of Florida
yanw@ufl.edu

ABSTRACT

The work-in-progress explores user-centric misinformation interventions on social media as such tools are limited. Knowledge of relationships between online user characteristics and their expressed adherence to a desired norm (i.e., rejecting misinformation or supporting factual information) is understudied with limited integrated multi-modal machine-learning models to infer demographic and sociopsychological characteristics. Thus, we piloted 9,331 Twitter users tweeting COVID-19 vaccines between May 1, 2020, and April 30, 2021. Employing a CNN-LSTM framework, our model analyzes user biographies, profile images, and pre-COVID historical tweets to infer user traits over 90% accuracy for individual characteristics and an overall accuracy of 85.61%, which outperforms existing tools and other designs. Further, using multi-logistic regression, we identified significant predictors of users' adherence to desired norms, such as gender and pre-pandemic prosocial content engagement, while finding no significant age correlation. Our findings illuminate pathways for targeted misinformation mitigation strategies during critical public health crises.

Keywords

Misinformation, deep learning, social media, social norm, user-centric

INTRODUCTION

Human life has become increasingly digitized, and a massive amount of information is disseminated in an online environment. At the same time, with inaccurate and misleading content, misinformation spreads widely in the physical and digital realm and becomes a threat to society (Del Vicario et al., 2016). Misinformation can mislead individuals' perceptions and behaviors widely (Swire-Thompson & Lazer, 2020). To tackle the threat of misinformation, prebunking strategies have been proposed to achieve pre-exposure implementation and long-lasting effects on preventing the adoption of misinformation (Lewandowsky & van der Linden, 2021). Grounded in "inoculation", i.e., pre-warnings can make individuals immune from misinformation, prebunking strategies can be promising for protecting individuals from misinformation with broad and diverse topics (Cook et al., 2017; Ecker et al., 2022). Prebunking can be conducted by social norm interventions, which are particularly for population-level inoculation (Maertens et al., 2020). Social norm interventions are based on the social psychological phenomenon that people tend to conform to surrounding norms to avoid sanction or confer benefit (Constantino et al., 2022; Gao et al., 2022).

Existing practices of social norm interventions are usually designed for the general public, neglecting the heterogeneity of online users regarding their sociodemographic characteristics, psychological traits, perceptions of online information, and the capacity for detecting misinformation (Chadwick et al., 2023). For instance, questionaries among university students in Singapore have identified a significant gender difference in the

prevalence of misinformation sharing (Chen et al., 2015). A survey with 113 randomly selected social media users indicates that users between 18 and 34 years old are more likely to share misinformation (Adaji, 2023). Individuals with different influences in the online environment may also share misinformation with different tendencies (Legros & Cislaghi, 2020). Also, tendencies to generate prosocial content have been identified as influencing individuals' motivation to circulate only truthful, beneficial information (Zhu et al., 2023). Without being tailored for online users, social norm interventions may only be effective for users with certain characteristics instead of the whole population of online users.

Compared to interventions that are generally designed for the public, user-centric or tailored interventions are designed to align with the heterogeneous characteristics of online users and may potentially amplify the effectiveness of interventions (Kaufman et al., 2022). User-centric or tailored designs of interventions have been successfully applied in other domains of information dissemination and communications, such as tailored and targeted communications of hazard information and health-related information (Gao & Wang, 2021; Kreuter & Wray, 2003). However, existing models that estimate social media user characteristics have not reached high-level accuracy (Wang et al., 2019). They also have not been designed to estimate users' social psychological characteristics, such as tendencies to generate prosocial content. Also, existing studies have not comprehensively investigated the relationship between user characteristics and behaviors of sharing or rejecting misinformation, providing insufficient knowledge about selecting targeted users who may need to be motivated to adhere to the *desired norm*, expressed in tweets of rejecting online misinformation or supporting the factual information.

We aim to answer the following two research questions:

RQ 1: How can we accurately infer online users' demographic and social psychological characteristics based on their features in profile images, biography text, and online posts?

RQ 2: What are the relationships between online user characteristics (including age, gender, tendency to generate prosocial content, and influence on social media) and users' adherence to the desired norms?

Focusing on the topic of COVID-19 vaccines on Twitter, we start by capturing 9,331 sample users' expressed adherence to the desired norm from May 1, 2020, to April 30, 2021. We infer online users' demographic and social psychological characteristics by developing a multi-modal deep learning model. The input data includes users' profile information and pre-COVID tweets posted between January 1, 2019, and December 31, 2019. Then, we use multi-logistic regression to examine the relationship between user characteristics and adherence to the desired norm. The preliminary study paves the way for the design of user-centric social norm interventions and contributes to the existing knowledge body in crisis informatics in two ways. First, it develops a comprehensive model designed to accurately infer a diverse set of user characteristics by analyzing their online profiles and social media posts. By integrating different data features, the model processes textual, visual, and behavioral data to predict user traits with high accuracy. Second, our analysis reveals significant correlations between user characteristics and their propensity to adhere to the desirable norm. The findings offer actionable insights for developing user-centric misinformation mitigation strategies, particularly critical in preemptive responses to future pandemics and large-scale crises.

LITERATURE REVIEW

Existing Prediction Models of Online User Characteristics

Due to the privacy policies of online platforms, the demographic characteristics of online users are usually not recorded and are not publicly accessible (Wang et al., 2019). Existing studies have proposed various methods for predicting online user characteristics, mainly focusing on demographic characteristics, especially genders and ages (or age groups). The predictions are mainly based on biographies, online posts, or profile images of online users, and the basic method of prediction is machine learning (especially deep learning) models. For example, support vector machine (SVM) has been widely used to predict the gender and age groups of users on Twitter and Facebook (Chen et al., 2021; Fink et al., 2021). Deep learning models have also been proposed to infer the gender and age of online users based on users' profile information, such as the M3 inference, which is a multi-modal deep learning model utilizing users' profile information (Wang et al., 2019).

However, existing tools may not be effective at predicting the characteristics of English-speaking users due to the biases in input features and datasets. For example, some tools rely on profile images as the primary feature for inferring users' demographics (Golder et al., 2022; Priadana et al., 2020), and they may not perform well with users with default profile images or no profile images. Also, some tools are developed based on certain platforms, such as Instagram and Facebook (Çoban et al., 2021; Priadana et al., 2020), which may affect their accuracy in inferring the characteristics of other platforms, such as Twitter. In another case, as one of the models with the

highest inference accuracy, M3 inference can still be improved to predict English Twitter users' age and gender, as its current accuracy for age prediction is around 40% and for gender is around 80%.

Meanwhile, online users' social psychological behaviors, such as tendencies to generate prosocial content, have not been predicted by previous studies. In particular, previous studies have proposed questionnaires to collect individuals' responses to some sociopsychological questions and estimated their tendencies to engage in prosocial behaviors accordingly (Lavertu et al., 2020; Perach et al., 2023). This strategy is not suitable for large populations due to the cost and may not apply to online users who are anonymous and may not participate in online surveys.

Online User Characteristics and Information Dissemination

Online users' demographic characteristics have been found to be related to their behaviors of information dissemination and the spreading of misinformation in the online environment. In some experiments with randomly sampled social media users, it was observed that female users and users between the ages of 18 and 34 have shown relatively higher tendencies to share misinformation without detecting it (Balakrishnan, 2022; Xiang et al., 2023). In addition to demographic characteristics, individuals' perceptions and dissemination of misinformation may be influenced by their social psychological traits, such as the fear of being attacked, desire to protect their self-image, lack of self-efficacy, lack of accountability, personal traits, extraversion, empathic concern, and tendencies to being prosocial (Gurgun et al., 2022; Perach et al., 2023). Among these characteristics, the tendency to be prosocial may be related to detecting misinformation, according to the experiments in China and Hungary (Orosz et al., 2023; Sun & Ma, 2023; Zhu et al., 2023). Behaviors of being prosocial include but are not limited to generating content for supporting or helping others in pursuit of developing and maintaining harmonious relationships (Van Rijsewijk et al., 2016). Online users with varying levels of influence in online environments may also have different tendencies to share or reject misinformation. Some online users, such as online news media, celebrities, journalists, and other populations with large followings, may have a greater impact on online discussions than the ones with low-level influence (Andrighetto et al., 2013; Legros & Cislaghi, 2020). Low-influence users' attitudes toward online misinformation are likely to align with the norms that are proposed and shaped by high-influence users (Constantino et al., 2022; Wu et al., 2011).

However, studies on the relationships between user characteristics and the adoption of misinformation-related beliefs may still be limited due to their reliance on self-reported data from survey participants, while individuals' responses in the surveys may deviate from their actual behaviors and beliefs (Takeuchi et al., 2023). Online users' historical posts can provide observed information about users' tendency to reject or share misinformation, but such data has not been sufficiently investigated by existing studies. Also, existing findings about the relationships between being prosocial and detecting misinformation may not be fully applied to English-speaking Twitter users, who are not covered by their experiment samples. Additionally, existing studies have mainly examined users' tendency or behaviors to detect misinformation, which does not always equal to rejecting misinformation content or expressing adherence to the norm of rejecting misinformation or supporting factual information (Moravec et al., 2018).

METHODOLOGY

Data Collection and Preprocessing

We collected Twitter data related to COVID-19 using keywords such as "covid", "coronavirus", and "sars-cov-2" through Twitter's publicly available Streaming API, as detailed in our previous studies (Gao et al., 2022; Gao et al., 2023; Wang et al., 2017; Wang et al., 2021). To ensure privacy, results were reported only at the population level, addressing potential data privacy concerns. From the large dataset, we randomly sampled 10,000 users who tweeted about COVID-19 vaccines between May 1, 2020, and April 30, 2021. After removing suspended or protected user accounts, 9,331 users remained in our study sample. The data set includes user profiles, especially profile images and biographies. Additionally, we retrieved the sample users' historic tweets from the pre-COVID period between January 1, 2019, and December 31, 2019, using a web-scraping tool named Nitter Scraper (DGNSREKT, 2023). Both the user profile and historical tweets were employed to infer demographic characteristics and tendencies of generating prosocial content using a CNN-LSTM-based multi-modal model. Notably, we leveraged a different dataset – Wikidata – to train the model because the dataset includes self-reported year-of-birth and genders of 6,739 users (Liu & Singh, 2021). Details can be found in the Model training section. Though some users may have deleted their posts, this retrospective data collection approach relies on the assumption that tendencies of being prosocial are consistent over time (Eisenberg, 2023), minimizing the impact of deleted posts on the overall analysis.

Developing an Integrated Model to Uncover Users' Demographic and Social-Psychological Characteristics

We developed an integrated multi-modal model (Figure. 1) to infer online users' gender (male or female), age groups (under 18, 18 to 44, 45 to 64, 65 or older), and tendency to generate prosocial content ("yes" or "no").

Model Input

The input of our integrated model includes the profile images and users' biographies and pre-COVID tweets, i.e., the text, topics, and sentiment, which can be distinguished among users with different ages and genders (Holmberg & Hellsten, 2015; Nguyen et al., 2021). The user images were directly input to the image-related module. The text of users' biographies and pre-COVID tweets were preprocessed before inputting into the text-related module by tokenizing words, lowering each word, and removing English stopwords, punctuations, URLs, and @usernames (Yao & Wang, 2020). After preprocessing, the text of biographies and pre-COVID tweets were embedded into two vectors based on the corpus of biographies and pre-COVID tweets of all users, then input into the text-related module. Specifically, the embedded process converted each piece of input text into a dense vector, in which each element represented the index of a word in the corpus with a 1×60 fixed-size embedding. The topics of users' biographies and the topics of their pre-COVID tweets were determined by Latent Dirichlet Allocation (LDA) topic modeling for each user (Yao & Wang, 2020). We set the possible count of topics as eight to obtain the lowest perplexity score of topic modeling. The LDA algorithm started by calculating the probability of each word in the corpus belonging to a certain topic and assigning the words to the topics with the highest probability. Each user's biography and generated tweets were then assigned the topics with the highest frequency in their respective texts. The sentiments of users' biographies and pre-COVID tweets were detected with VADER, a keyword-based and rule-based model (Hutto & Gilbert, 2014). The topics and sentiments of users' biographies and pre-COVID tweets were also input into the text-related module of our model.

Model Structure

The multi-modal model comprises two pipelines, which process the image input and the text-based inputs separately. The first pipeline includes a Lenet-based convolutional neural network (CNN) module, which is widely used for processing imagery inputs (Lecun et al., 1998). The CNN module consists of three sets of "convolution + max pooling" layers for processing the raw profile images, of which the size is 400×400. These layers are followed by a flatting layer that turns the 2D data into a 1D vector, which is the output of the first pipeline.

The second pipeline contains two parallel processes for biography-related inputs and pre-COVID tweet-related inputs. For the biography-related inputs, the word vector is processed by a Long-Short Term Memory (LSTM) layer, which is typically utilized for capturing information from text data (Sundermeyer et al., 2012). The topic and sentiment of user biographies are transformed into two dummy vectors. The output vector of the LSTM layer and the dummy vectors of topics and sentiments are then aggregated with a dense layer, which reduces the total dimension of these three vectors and avoids co-linearity. The processing of pre-COVID tweet-related inputs is the same as the processing of biography-related inputs. The outputs of the second pipeline include one vector representing users' biographies and one vector representing users' pre-COVID tweets.

The output vectors of the two pipelines are concatenated into one vector and input into a three-layer feedforward neural network, which generates a dummy variable indicating the user's characteristics. By considering users' gender (two types), age groups (four types), and tendency to generate prosocial content (two types), an online user can belong to one of 16 classes ($2 \times 4 \times 2$). For example, one class of users can be male, under 18, and tend to generate prosocial content.

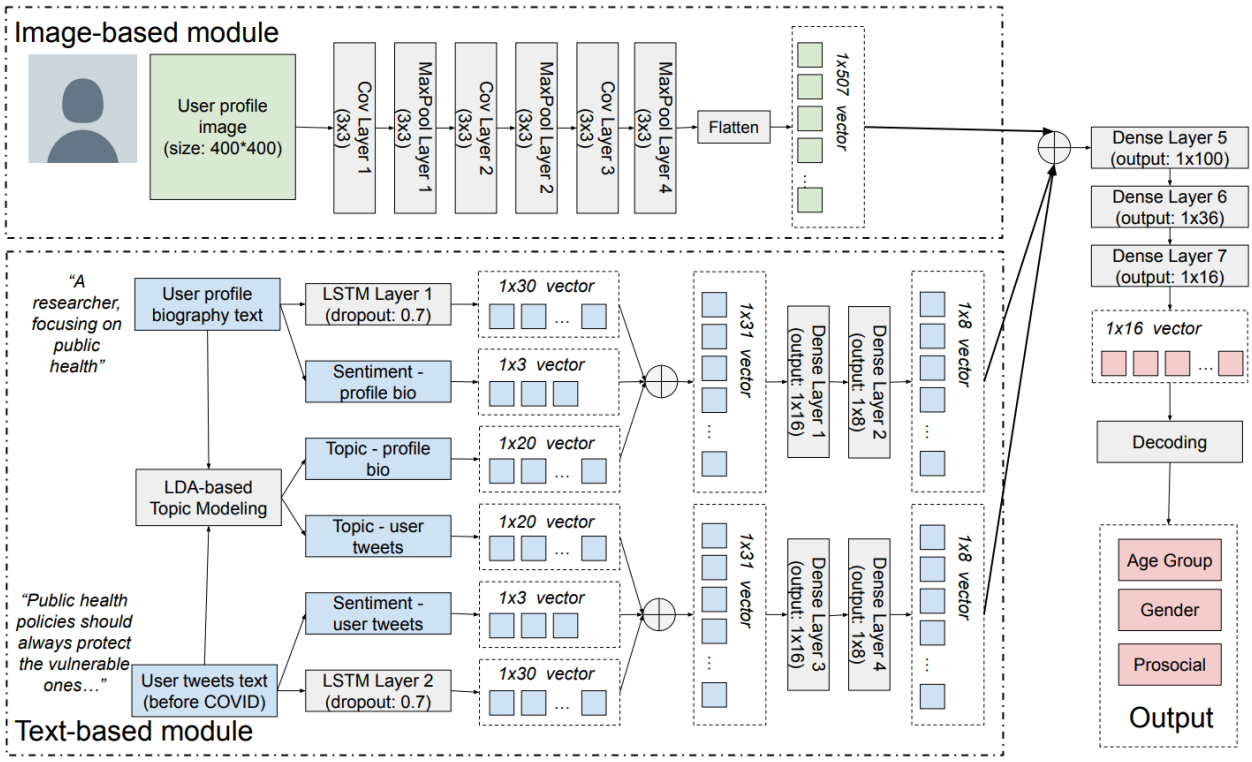

Figure 1. Architecture of the multi-modal model

Model Training

To train the model shown in Figure. 1, we used Wikidata as the training dataset containing users with self-reported year-of-birth and gender (Liu & Singh, 2021). We utilized the Nitter scraper to retrieve profile images, biographies, and pre-COVID tweets generated by the users in Wikidata between January 1, 2019, and December 1, 2019. We can retrieve the complete information from 6,739 users from Wikidata. To determine whether the combined corpora of the tweets and biographies of users in training data were prosocial or not, we labeled each user's tweet corpora based on manual reviewing and the judgment of GPT-4 (OpenAI, 2023), following existing studies (Li et al., 2024; Mujahid et al., 2023). A text was labeled as prosocial content if it contained positive, friendly, supporting, and gratitude expressions (Lysenstøen et al., 2021). An example is "How great would it be if we took the time to sit down with our veterans, visit and thank them for their service." The distribution of users in each class in the training data is shown in Figure. 2.

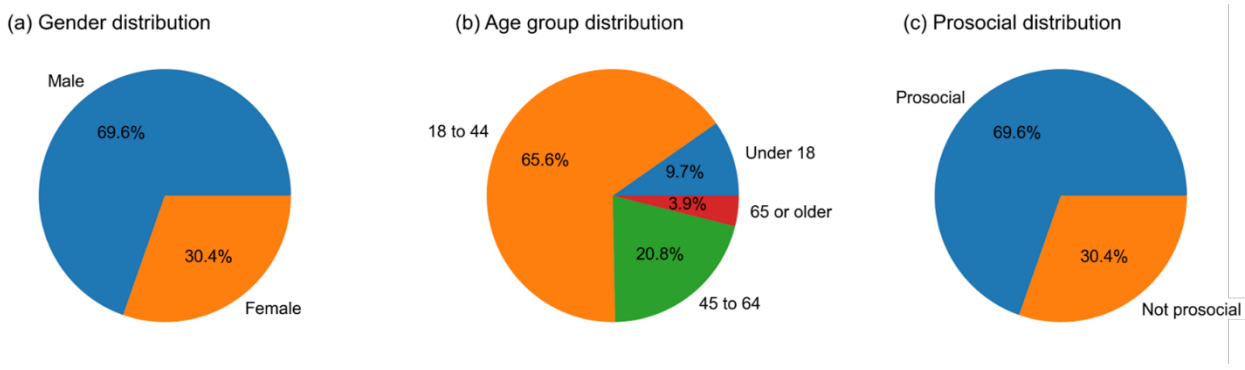

Figure 2. Distribution of genders, age groups, and users generating prosocial content in the training data

To train the multi-modal model, we split the training data into two sets, which separately took 70% (for modeling training) and 30% (for model validation) of the training dataset. The training process consisted of 20 epochs to avoid overfitting. The batch size was set to 100 to reach the balance between the times of updating model parameters and model accuracy (Smith et al., 2017). We selected Adam as the optimizer and calculated the model loss with categorical cross-entropy.

Tracking Online Users' Expressed Adherence to the Desired Norm

Leveraging the trained multi-modal model in the prior session, we can predict the user characteristics of 9,331 users. Then, we determined each user's expressed adherence to the desired norm. We started by determining if each tweet of each user expressed adherence to the desired norm. A tweet was considered as not adhering to the desired norm if its content delivered manipulated information about the side effects of COVID-19 vaccines, such as "do not take COVID-19 vaccines, they made children killed." In contrast, we considered a tweet as adhering to the desired norm if it supported COVID-19 vaccination or rejected related misinformation, such as "take COVID-19 vaccines to protect you and your family." We trained an LSTM classifier for the text classification based on 2,000 manually labeled tweets (Table 1) (Gao et al., 2023). The tweets were randomly sampled from all the real-time tweets we previously collected, which were not generated by our studied 9,331 users. LSTM was used because of its high capacity to capture phrase-level and sentence-level feature patterns (Sundermeyer et al., 2012). The validated accuracy and loss of the LSTM classifier during training are shown in Figure. 3, reaching 0.8892 and 0.2292 separately after training. Also, the RMSE of the classification outcomes is 0.3719, and the F1 score is 0.8706. These metrics indicate that our LSTM classifier has a satisfactory performance.

Table 1. Proportions of tweets expressing or not expressing adherence to the desired norm in training and testing data of the LSTM classifier

Dataset	Size	Count/Proportion of Tweets Expressing Adherence to the Desired Norm	Count/Proportion of Tweets Not Expressing Adherence to the Desired Norm
Training data	1,600 tweets	209 tweets (13.06%)	1,391 tweets (86.94%)
Testing data	400 tweets	57 tweets (14.25%)	343 tweets (85.75%)

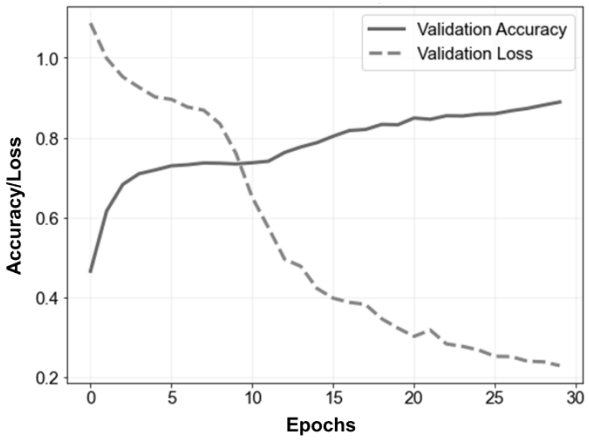

Figure 3. The accuracy and loss of the LSTM classifier during training

After determining users' expressed beliefs delivered in each tweet, we specified each user's expressed adherence to the desired norm in the twelve-month studied period (represented by $Adherence$). We first specified their vaccination-relevant tweets each month. We labeled users' tweets corpora in one month as expressing adherence to the desired norm if all of their tweets about COVID-19 vaccines in that month expressed adherence to the desired norm. Based on the monthly adherence to the desired norm, we categorized users into two groups: users consistently expressing adherence to the desired norm ($Adherence = 1$), and users not consistently expressing adherence to the desired norm ($Adherence = 0$).

Multi-Logistic Regression for Examining the Relationship between Online User Characteristics and Expressed Adherence to the Desired Norm

We examined the relationship between the user characteristics (age, gender, tendency to generate prosocial content, and count of followers) and $Adherence$ with multi-logistic regression based on 9,331 Twitter users (**Eq. 1**). One characteristic of online users is considered as significantly correlated with $Adherence$ if the $p - value$ of its coefficient is lower than 0.01 or 0.001, representing low to high levels of significance separately.

$$Adherence \sim Age + Gender + Prosocial + Influence \quad (1)$$

RESULTS

Prediction Outcomes of User Characteristics and Monthly Patterns of Expressed Adherence to the Desired Norm

With the multi-modal model, we inferred users' genders, age groups, and tendencies to generate prosocial content. The statistical distribution of user characteristics is shown in Figure. 4, which is similar to the distribution among the users in the training data. 65.0% of the analyzed users are male. Most of the analyzed users belong to the age group of 18-44 (77.0%) or 45-64 (17.2%). 77.5% of the analyzed users tend to generate prosocial content.

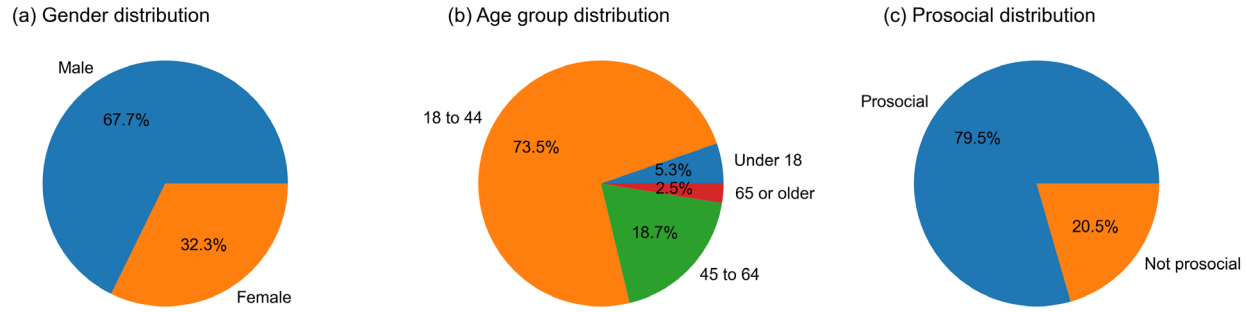

Figure 4. Distribution of genders, age groups, and users generating prosocial content among the analyzed users

After estimating the analyzed users' expressed adherence to the desired norm, we categorized users into two groups (Figure. 5a). While most of our analyzed users posted less than 200 tweets in the studied period (Figure. 5b), 44.3% of the analyzed users consistently expressed adherence to the desired norm.

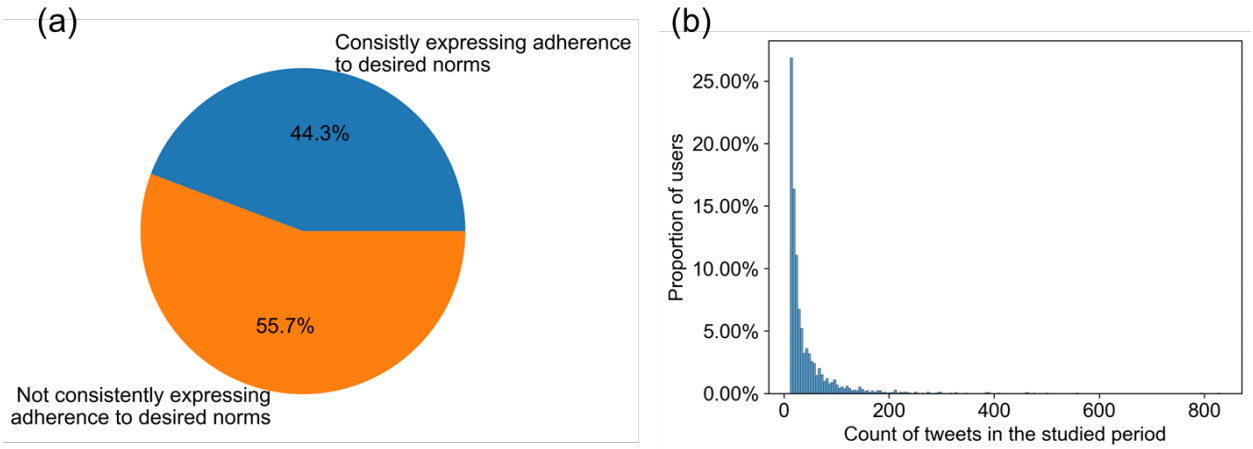

Figure 5. Proportions of users who consistently expressed adherence to the desired norm (a) and proportion of users with different counts of tweets

Evaluation of Multi-Modal CNN-LSTM Model Performance

We evaluated the multi-modal model that infers online user characteristics based on its accuracy and loss and ablation studies. In terms of accuracy and loss, our model achieved an accuracy of 85.61% when predicting online users' gender, age group, and tendency to generate prosocial content simultaneously, and its accuracy exceeds 90% for predicting each characteristic separately (Figure. 6 and Table 2). We also conducted ablation studies (Table 2), i.e., examining the prediction performance of each component of the multi-modal model and their combinations, to investigate their contributions to the prediction process. The multi-modal model has a much higher accuracy than each component classifier and each of their combinations. Among the model components, the image-based module (CNN classifier) has the major contribution and can solely achieve an overall accuracy of 60.83%, which is much higher than the text-based module.

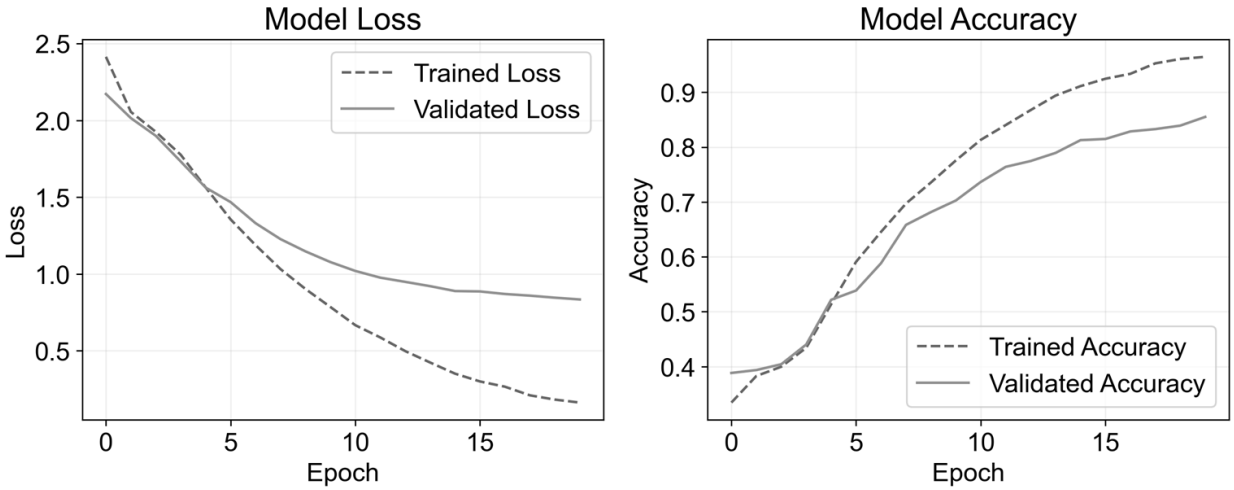

Figure 6. Loss and accuracy of the multi-modal model

Table 2. Ablation studies of the multi-modal model

Model	Inferring gender	Inferring age groups	Inferring prosocial	Overall
CNN (profile image)	0.7597	0.7538	0.8426	0.6083
LSTM (profile bio)	0.7284	0.6683	0.8029	0.5194
LSTM (tweet text)	0.6889	0.6577	0.8012	0.4486
CNN (profile image) + LSTM (profile bio)	0.7717	0.7200	0.8139	0.5710
CNN (profile image) + LSTM (tweet text)	0.7581	0.7084	0.7993	0.5480
LSTM (profile bio) + LSTM (tweet text)	0.7421	0.6889	0.8059	0.4753
Proposed model	**0.9069**	**0.9162**	**0.9467**	**0.8561**

Correlation between Online User Characteristics and Expressed Adherence to the Desired Norm

Before conducting correlation analysis, we examined the inter-correlation between online user characteristics and indicators of influence in online environments (Figure. 7). Notably, the user characteristics and influence in online environments are not significantly correlated and can be input into the regression analysis.

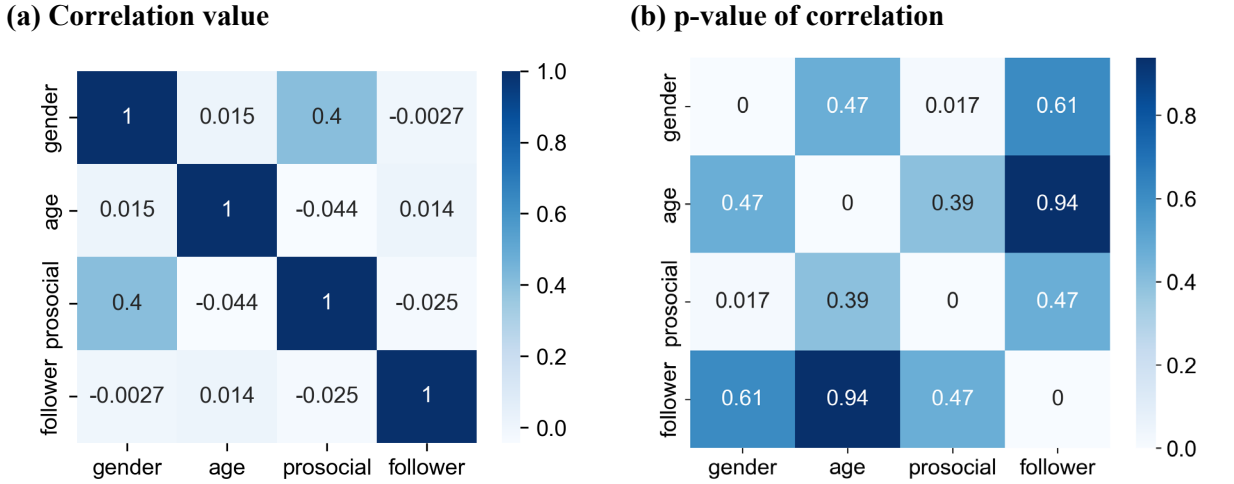

Figure 7. Correlation values and $p-value$ between user characteristics

We then conducted multi-logistic regression to examine the relationship between user characteristics and users' adherence to the desired norm ($Adherence$) (Table 3). The regression outcomes indicate that gender ($p-$

$value = 0.008$) and the tendency to generate prosocial content ($p - value = 0.003$) are significantly correlated with consistently expressed adherence to the desired norm. A user is more likely to consistently express adherence to the desired norm if the user is male or tends to generate prosocial content before the pandemic. In contrast, female users and users who do not generate prosocial content are less likely to express adherence to the desired norm.

Table 3. Multi-logistic regression outcomes

	Coefficient	**Confident Interval (lower 2.5% boundary)**	**Confident Interval (upper 97.5% boundary)**	$\boldsymbol{p - value}$
Intercept	-0.6695	-1.2020	-0.1370	0.0140
Gender	**-0.3097**	**-0.5380**	**-0.0820**	**0.0080****
Age	-0.1283	-0.3230	0.0660	0.1960
Prosocial	**0.3792**	**0.1320**	**0.6260**	**0.0030****
Count of Followers	0.0000	0.0000	0.0000	0.3250
Significance Levels: 0 '***' 0.001 '**' 0.01				

DISCUSSION AND FUTURE WORK

The research is among the first to explore the idea of user-centric normative misinformation intervention. The preliminary study contributes to the knowledge body of crisis informatics in online environments through two major outcomes. First, we address the need for an integrated online user characteristics inference model and propose a new multi-modal deep-learning model for predicting demographic and social psychological characteristics. With an overall accuracy of 85.61% and an accuracy exceeding 90% for each user characteristic, our model outperforms existing tools for inferring online users' demographic and social psychological characteristics based on publicly accessible data. Second, our statistical analysis has identified a significant and positive relationship between users' gender and their tendency to generate prosocial content and expressed adherence to the desired norm. The correlation highlights the necessity of paying attention to the users who are female and do not tend to generate pro-social content for social norm interventions. We expect that more customized and user-centric prebunking social norm interventions will alleviate the adverse social consequences of misinformation more effectively in the digital environment.

Our further research will continue to explore and test user-centric misinformation mitigations to advance the knowledge body of crisis informatics following the work-in-progress paper. First, we will expand the scope of online user characteristics for predicting online users' expressed adherence to the desired norm and misinformation content, such as self-esteem, optimism, and expressed satisfaction with daily life. Second, we will integrate the impact of individuals' exposure to their information environments to understand the motivation for changing expressed belief in vaccination. Particularly, online users' adoption of the desired norm is potentially related to the expressed beliefs of their social network members. Third, we will investigate other topics of online misinformation, such as healthy eating, threats and protective measures against communicable diseases, and risky behaviors (such as the use of drugs and smoking). Lastly, our future work will focus on designing user-centric interventions to mitigate online misinformation before the next pandemic or crisis. For example, we will explore the strategies to break the "echo chamber" among users who do not tend to generate prosocial content. We will also examine the effectiveness of spreading messages that deliver the desired norm or prosocial content to encourage users to reject misinformation.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. 2323794. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

Adaji, I. (2023). Age Differences in the Spread of Misinformation Online. *European Conference on Social Media*, 10(1), 12–19. https://doi.org/10.34190/ecsm.10.1.1156

Andrighetto, G., Castelfranchi, C., Mayor, E., McBreen, J., Lopez-Sanchez, M., & Parsons, S. (2013). (Social) Norm Dynamics. *Normative Multi-Agent Systems*, 135–170. https://doi.org/10.4230/DFU.VOL4.12111.135

Balakrishnan, V. (2022). Socio-demographic Predictors for Misinformation Sharing and Authenticating amidst the COVID-19 Pandemic among Malaysian Young Adults. *Information Development*, 026666692211189. https://doi.org/10.1177/02666669221118922

Chadwick, A., Vaccari, C., & Hall, N.-A. (2023). What Explains the Spread of Misinformation in Online Personal Messaging Networks? Exploring the Role of Conflict Avoidance. *Digital Journalism*, 1–20. https://doi.org/10.1080/21670811.2023.2206038

Chen, X., Sin, S.-C. J., Theng, Y.-L., & Lee, C. S. (2015). Why Do Social Media Users Share Misinformation? *Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries*, 111–114. https://doi.org/10.1145/2756406.2756941

Chen, X., Wang, Y., Agichtein, E., & Wang, F. (2021). A Comparative Study of Demographic Attribute Inference in Twitter. *Proceedings of the International AAAI Conference on Web and Social Media, 9*(1), 590–593. https://doi.org/10.1609/icwsm.v9i1.14656

Çoban, Ö., İnan, A., & Özel, S. A. (2021). Facebook Tells Me Your Gender: An Exploratory Study of Gender Prediction for Turkish Facebook Users. *ACM Transactions on Asian and Low-Resource Language Information Processing, 20*(4), 1–38. https://doi.org/10.1145/3448253

Constantino, S. M., Sparkman, G., Kraft-Todd, G. T., Bicchieri, C., Centola, D., Shell-Duncan, B., Vogt, S., & Weber, E. U. (2022). Scaling Up Change: A Critical Review and Practical Guide to Harnessing Social Norms for Climate Action. *Psychological Science in the Public Interest, 23*(2), 50–97. https://doi.org/10.1177/15291006221105279

Cook, J., Lewandowsky, S., & Ecker, U. K. H. (2017). Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence. *PLOS ONE, 12*(5), e0175799. https://doi.org/10.1371/journal.pone.0175799

Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H. E., & Quattrociocchi, W. (2016). The spreading of misinformation online. *Proceedings of the National Academy of Sciences, 113*(3), 554–559. https://doi.org/10.1073/pnas.1517441113

DGNSREKT. (2023). Nitter scraper. https://nitter-scraper.readthedocs.io/en/latest/#docker-engine

Ecker, U. K., Sanderson, J. A., McIlhiney, P., Rowsell, J. J., Quekett, H. L., Brown, G. D., & Lewandowsky, S. (2022). Combining refutations and social norms increases belief change. *Quarterly Journal of Experimental Psychology*, 174702182211117. https://doi.org/10.1177/17470218221111750

Eisenberg, N. (2003). Prosocial behavior, empathy, and sympathy. In *Well-Being Positive Development Across the Life Course*. Taylor and Francis.

Fink, C., Kopecky, J., & Morawski, M. (2021). Inferring Gender from the Content of Tweets: A Region Specific Example. *Proceedings of the International AAAI Conference on Web and Social Media, 6*(1), 459–462. https://doi.org/10.1609/icwsm.v6i1.14320

Gao, S., & Wang, Y. (2021). Assessing the impact of geo-targeted warning messages on residents' evacuation decisions before a hurricane using agent-based modeling. *Natural Hazards, 107*(1), 123–146. https://doi.org/10.1007/s11069-021-04576-1

Gao, S., Wang, Y., & Thai, M. T. (2023). Investigating the Dynamics of Social Norm Emergence within Online Communities. https://doi.org/10.48550/ARXIV.2301.00453

Gao, S., Wang, Y., & Webster, G. D. (2022). Causal Modeling of Descriptive Social Norms from Twitter and the Physical World on Expressed Attitudes Change: A Case Study of COVID-19 Vaccination. *Cyberpsychology, Behavior, and Social Networking, 25*(12), 769–775. https://doi.org/10.1089/cyber.2022.0153

Golder, S., Stevens, R., O'Connor, K., James, R., & Gonzalez-Hernandez, G. (2022). Methods to Establish Race or Ethnicity of Twitter Users: Scoping Review. *Journal of Medical Internet Research, 24*(4), e35788. https://doi.org/10.2196/35788

Gurgun, S., Arden-Close, E., Phalp, K., & Ali, R. (2022). Online silence: Why do people not challenge others when posting misinformation? *Internet Research*. https://doi.org/10.1108/INTR-06-2022-0407

Holmberg, K., & Hellsten, I. (2015). Gender differences in the climate change communication on Twitter. *Internet Research, 25*(5), 811–828. https://doi.org/10.1108/IntR-07-2014-0179

Hutto, C., & Gilbert, E. (2014). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *Proceedings of the International AAAI Conference on Web and Social Media, 8*(1), 216–225. https://doi.org/10.1609/icwsm.v8i1.14550

Jiang, L. C., Chu, T. H., & Sun, M. (2021). Characterization of Vaccine Tweets During the Early Stage of the COVID-19 Outbreak in the United States: Topic Modeling Analysis. *JMIR Infodemiology, 1*(1), e25636. https://doi.org/10.2196/25636

Kaufman, J., Bagot, K. L., Tuckerman, J., Biezen, R., Oliver, J., Jos, C., Ong, D. S., Manski-Nankervis, J., Seale, H., Sanci, L., Munro, J., Bell, J. S., Leask, J., & Danchin, M. (2022). Qualitative exploration of intentions, concerns and information needs of vaccine-hesitant adults initially prioritised to receive COVID-19 vaccines in Australia. *Australian and New Zealand Journal of Public Health, 46*(1), 16–24. https://doi.org/10.1111/1753-6405.13184

Kreuter, M., & Wray, R. (2003). Tailored and targeted health communication: Strategies for enhancing information relevance. *American Journal of Health Behavior*, 27, S227–S232. https://doi.org/10.5993/ajhb.27.1.s3.6

Lavertu, L., Marder, B., Erz, A., & Angell, R. (2020). The extended warming effect of social media: Examining whether the cognition of online audiences offline drives prosocial behavior in 'real life.' *Computers in Human Behavior*, 110, 106389. https://doi.org/10.1016/j.chb.2020.106389

Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE, 86*(11), 2278–2324. https://doi.org/10.1109/5.726791

Legros, S., & Cislaghi, B. (2020). Mapping the Social-Norms Literature: An Overview of Reviews. *Perspectives on Psychological Science, 15*(1), 62–80. https://doi.org/10.1177/1745691619866455

Lewandowsky, S., & van der Linden, S. (2021). Countering Misinformation and Fake News Through Inoculation and Prebunking. *European Review of Social Psychology, 32*(2), 348–384. https://doi.org/10.1080/10463283.2021.1876983

Li, L., Fan, L., Atreja, S., & Hemphill, L. (2024). "HOT" ChatGPT: The Promise of ChatGPT in Detecting and Discriminating Hateful, Offensive, and Toxic Comments on Social Media. *ACM Transactions on the Web*, 3643829. https://doi.org/10.1145/3643829

Liu, Y., & Singh, L. (2021). Age Inference Using A Hierarchical Attention Neural Network. *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 3273–3277. https://doi.org/10.1145/3459637.3482055

Lysenstøen, C., Bøe, T., Hjetland, G. J., & Skogen, J. C. (2021). A Review of the Relationship Between Social Media Use and Online Prosocial Behavior Among Adolescents. *Frontiers in Psychology, 12*, 579347. https://doi.org/10.3389/fpsyg.2021.579347

Maertens, R., Anseel, F., & van der Linden, S. (2020). Combatting climate change misinformation: Evidence for longevity of inoculation and consensus messaging effects. *Journal of Environmental Psychology, 70,* 101455. https://doi.org/10.1016/j.jenvp.2020.101455

Minaei, M., Mouli, S. C., Mondal, M., Ribeiro, B., & Kate, A. (2021). Deceptive Deletions for Protecting Withdrawn Posts on Social Media Platforms. *Proceedings 2021 Network and Distributed System Security Symposium. Network and Distributed System Security Symposium, Virtual.* https://doi.org/10.14722/ndss.2021.23139

Moravec, P., Minas, R., & Dennis, A. R. (2018). Fake News on Social Media: People Believe What They Want to Believe When it Makes No Sense at All. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3269541

Mujahid, M., Kanwal, K., Rustam, F., Aljedaani, W., & Ashraf, I. (2023). Arabic ChatGPT Tweets Classification Using RoBERTa and BERT Ensemble Model. *ACM Transactions on Asian and Low-Resource Language Information Processing, 22*(8), 1–23. https://doi.org/10.1145/3605889

Nguyen, D., Gravel, R., Trieschnigg, D., & Meder, T. (2021). "How Old Do You Think I Am?" A Study of Language and Age in Twitter. *Proceedings of the International AAAI Conference on Web and Social Media, 7*(1), 439–448. https://doi.org/10.1609/icwsm.v7i1.14381

OpenAI. (2023). *GPT-4: Generative Pre-trained Transformer 4.* Retrieved from https://openai.com/gpt-4.

Orosz, G., Paskuj, B., Faragó, L., & Krekó, P. (2023). A prosocial fake news intervention with durable effects. *Scientific Reports, 13*(1), 3958. https://doi.org/10.1038/s41598-023-30867-7

Perach, R., Joyner, L., Husbands, D., & Buchanan, T. (2023). Why Do People Share Political Information and Misinformation Online? Developing a Bottom-Up Descriptive Framework. *Social Media + Society, 9*(3), 20563051231192032. https://doi.org/10.1177/20563051231192032

Priadana, A., Maarif, M. R., & Habibi, M. (2020). Gender Prediction for Instagram User Profiling using Deep Learning. *2020 International Conference on Decision Aid Sciences and Application (DASA)*, 432–436. https://doi.org/10.1109/DASA51403.2020.9317143

Smith, S. L., Kindermans, P.-J., Ying, C., & Le, Q. V. (2017). Don't Decay the Learning Rate, Increase the Batch Size. https://doi.org/10.48550/ARXIV.1711.00489

Sun, M., & Ma, X. (2023). Combating health misinformation on social media through fact-checking: The effect of threat appraisal, coping appraisal, and empathy. *Telematics and Informatics, 84,* 102031. https://doi.org/10.1016/j.tele.2023.102031

Sundermeyer, M., Schlüter, R., & Ney, H. (2012). LSTM neural networks for language modeling. *Interspeech 2012*, 194–197. https://doi.org/10.21437/Interspeech.2012-65

Swire-Thompson, B., & Lazer, D. (2020). Public Health and Online Misinformation: Challenges and Recommendations. *Annual Review of Public Health, 41*(1), 433–451. https://doi.org/10.1146/annurev-publhealth-040119-094127

Takeuchi, M., Niimi, J., & Hoshino, T. (2023). Handling the Inconsistency between Self-Report and the Actual Behavior: Validity of Excluding Survey Participants with Insufficient Effort Responding. *International Journal of Market Research*, 14707853231209933. https://doi.org/10.1177/14707853231209933

Van Rijsewijk, L., Dijkstra, J. K., Pattiselanno, K., Steglich, C., & Veenstra, R. (2016). Who helps whom? Investigating the development of adolescent prosocial relationships. *Developmental Psychology, 52*(6), 894–908. https://doi.org/10.1037/dev0000106

Wang, Y., Wang, Q., & Taylor, J. E. (2017). Aggregated responses of human mobility to severe winter storms: An empirical study. *PLOS ONE, 12*(12), e0188734. https://doi.org/10.1371/journal.pone.0188734

Wang, Z., Hale, S., Adelani, D. I., Grabowicz, P., Hartman, T., Flöck, F., & Jurgens, D. (2019). Demographic Inference and Representative Population Estimates from Multilingual Social Media Data. *The World Wide Web Conference*, 2056–2067. https://doi.org/10.1145/3308558.3313684

Wang, Y., Hao, H., & Platt, L. S. (2021). Examining risk and crisis communications of government agencies and stakeholders during early-stages of COVID-19 on Twitter. *Computers in Human Behavior, 114*, 106568. https://doi.org/10.1016/j.chb.2020.106568

Wu, S., Hofman, J. M., Mason, W. A., & Watts, D. J. (2011). Who says what to whom on twitter. *Proceedings of the 20th International Conference on World Wide Web*, 705. https://doi.org/10.1145/1963405.1963504

Xiang, H., Zhou, J., & Wang, Z. (2023). Reducing Younger and Older Adults' Engagement with COVID-19 Misinformation: The Effects of Accuracy Nudge and Exogenous Cues. *International Journal of Human–Computer Interaction*, 1–16. https://doi.org/10.1080/10447318.2022.2158263

Yao, F., & Wang, Y. (2020). Tracking urban geo-topics based on dynamic topic model. *Computers, Environment and Urban Systems, 79*, 101419. https://doi.org/10.1016/j.compenvurbsys.2019.101419

Zhu, Z., Zhang, N., Ding, M., & Chen, L. (2023). Accountability mobilization, guanxi and social media-induced polarization: Understanding the bystander's prosocial punishment to misinformation spreader. *Information Systems Journal*, isj.12486. https://doi.org/10.1111/isj.12486