HUSO 2023

The Ninth International Conference on Human and Social Analytics

ISBN: 978-1-68558-066-7

March 13th - 17th, 2023

Barcelona, Spain

HUSO 2023 Editors

Lasse Berntzen, University of South-Eastern Norway, Norway

HUSO 2023

Forward

The Ninth International Conference on Human and Social Analytics (HUSO 2023), held between March 13th and March 17th, 2023, continued a series of events bridging the concepts and the communities dealing with emotion-driven systems, sentiment analysis, personalized analytics, social human analytics, and social computing.

The recent development of social networks, numerous ad hoc interest-based virtual communities, and citizen-driven institutional initiatives raise a series of new challenges in considering human behavior, both in personal and collective contexts.

There is a great possibility to capture particular and public opinions, allowing individual or collective behavioral predictions. This also raises many challenges, on capturing, interpreting, and representing such behavioral aspects. While scientific communities face now new paradigms, such as designing emotion-driven systems, dynamicity of social networks, and integrating personalized data with public knowledge bases, the business world looks for marketing and financial prediction.

We take here the opportunity to warmly thank all the members of the HUSO 2023 technical program committee, as well as all the reviewers. The creation of such a high-quality conference program would not have been possible without their involvement. We also kindly thank all the authors who dedicated much of their time and effort to contribute to HUSO 2023. We truly believe that, thanks to all these efforts, the final conference program consisted of top-quality contributions. We also thank the members of the HUSO 2023 organizing committee for their help in handling the logistics of this event.

We hope that HUSO 2023 was a successful international forum for the exchange of ideas and results between academia and industry and for the promotion of progress in the field of human and social analytics.

HUSO 2023 Chairs

HUSO 2023 Steering Committee

Christian Bourret, University of Paris East - Marne la Vallée (UPEM), France
Els Lefever, Ghent University, Belgium
Dennis J. Folds, Lowell Scientific Enterprises (LSE), USA
Nitin Agarwal, University of Arkansas at Little Rock, USA

HUSO 2023 Publicity Chairs

José Miguel Jiménez, Universitat Politecnica de Valencia, Spain
Sandra Viciano Tudela, Universitat Politecnica de Valencia, Spain

HUSO 2023
Committee

HUSO 2023 Steering Committee

Christian Bourret, University of Paris East - Marne la Vallée (UPEM), France
Els Lefever, Ghent University, Belgium
Dennis J. Folds, Lowell Scientific Enterprises (LSE), USA
Nitin Agarwal, University of Arkansas at Little Rock, USA

HUSO 2023 Publicity Chairs

José Miguel Jiménez, Universitat Politecnica de Valencia, Spain
Sandra Viciano Tudela, Universitat Politecnica de Valencia, Spain

HUSO 2023 Technical Program Committee

Paul Abbiati, Founding Fellow of the EUROPEAN LAW INSTITUTE, Austria
Nitin Agarwal, University of Arkansas at Little Rock, USA
Harry Agius, Brunel University London, UK
Balbir Barn, Middlesex University, London, UK
Christian Bourret, Université Gustave Eiffel, France
Anita Chandra, IIT Patna, India
Dickson Chiu, The University of Hong Kong, Hong Kong
Claudia Da Re, Gustave Eiffel University - Marne-la-Vallée Campus, France
Chen Ding, Ryerson University, Canada
Birgitta Dresp-Langley, Centre National de la Recherche Scientifique, France
Thierry Edoh, RFW-Universität Bonn, Germany
Fernanda M. Eliott, Noyce Science Center - Grinnell College, USA
Diana Florea, Lucian Blaga University of Sibiu, Romania
Silvia Florea, Lucian Blaga University of Sibiu, Romania
Dennis J. Folds, Lowell Scientific Enterprises (LSE), USA
Matteo Francia, University of Bologna, Italy
Enrico Gallinucci, University of Bologna, Italy
Luca Giraldi, EMOJ srl, Ancona, Italy
Damian Gordon, Technological University Dublin, Ireland
Denis Gracanin, Virginia Tech, USA
Shatha Ali A. Hakami, University of Birmingham, UK
Tzung-Pei Hong, National University of Kaohsiung, Taiwan
Rodrigo Huerta-Quintanilla, Cinvestav, Mexico
Emanuele Iannone, University of Salerno, Italy
Emilio Insfran, Universitat Politecnica de Valencia, Spain
Yasushi Kambayashi, Nippon Institute of Technology, Japan
Dimitris K. Kardaras, Athens University of Economics and Business, Greece
Hassan A. Karimi, University of Pittsburgh, USA
Konstantin Kuzmin, Rensselaer Polytechnic Institute (RPI), USA
Georgios Lappas, University of Western Macedonia, Greece

Copyright Information

Table of Contents

Leveraging Topic Modeling and Toxicity Analysis to Understand China-Uyghur Conflicts 1
Connice Trimmingham, Samuel Dayo Banjo, and Nitin Agarwal

Digital Practices and Uses in Secondary Schools: Achieving Digital Literacy 7
Christian Bourret and Hafida Hammadi

An Investigation of One Sided Follow Relations between Twitter Users Concerned with Tweets Disclosing Submitters' Personal Information 9
Yasuhiko Watanabe, Toshiki Nakano, Hiromu Nishimura, and Yoshihiro Okada

Examining Content and Emotion Bias in YouTube's Recommendation Algorithm 15
Obianuju Okeke, Mert Can Cakmak, Billy Spann, and Nitin Agarwal

Comparing Toxicity Across Social Media Platforms for COVID-19 Discourse 21
Nahiyan Bin Noor, Niloofar Yousefi, Billy Spann, and Nitin Agarwal

Complex Behavior Vs. Design - Interpreting AI: Reminders from Synthetic Psychology 27
Elliot Swaim and Fernanda Eliott

Leveraging Topic Modeling and Toxicity Analysis to Understand China-Uyghur Conflicts

Connice Trimmingham, Samuel Dayo Banjo, Nitin Agarwal
COSMOS Research Center
University of Arkansas at Little Rock
Little Rock, AR 72204, USA
e-mail: ctrimmingham@ualr.edu, sbanjo@ualr.edu, nxagarwal@ualr.edu

Abstract— **The growth of social networking sites, coupled with the widespread use of mobile technology, has led to the spread of various forms of toxicity. Although social media platforms provide valuable tools for meaningful interactions, political arguments, often fraught with complex mix of emotions, can quickly devolve into flame wars or partisan bickering. This article shifts attention eastward to examine how the media/information environment is being manipulated for advancing political agendas in the Indo-Pacific region. We analyzed 3,239,249 tweets discussing issues related to China and Uyghur. We explained the user relation phenomena by assessing their emerging social structures. We extracted influential topics using the Latent Dirichlet Allocation (LDA) topic modeling approach. Toxicity analysis and bot assessment were performed to examine the nature of discourse about the China and Uyghur issues. Our findings indicate a strong correlation between tweets with high toxicity and bot activity, particularly in relation to emerging events such as the existence of internment camps and news about forced Uyghur laborers in China and the Chinese Communist Party network.**

Keywords-Information operations; Indo-Pacific; Twitter; Social Media; Uyghur; Toxicity Analysis.

I. Introduction

China has been in the global spotlight for its economic strategies, investments acquisitions, and policy reinforcement. However, recently, China's reputation has been globally questioned for its targeted, inhumane, and oppressive policies towards the Uyghur population in Xinjiang [1]. From a geographical standpoint, Xinjiang is an autonomous region that measures one-sixth of China's western border and home to a Chinese Ethnic-Muslim minority. From a political perspective, Xinjiang houses an extensive potential for mineral exploitation in natural resources, such as oil, gas, and agricultural production [2].

China's Uyghur conflict has existed for decades; its universal debate however, has recently surfaced with the unprecedented evolution of online social networks. Although religious beliefs, customs, and practices have been tolerated in China to some extent, the degree of tolerance has varied considerably from time to time with the change in the political climate [3]. The use of coercion is not uncommon in Chinese history as far as religious groups are concerned [4]. Extant literature has shown that policies towards Xinjiang are similar to the policies that were directed towards Tibet [4]. To cope with these policies, the Uyghur group attempted to separate themselves from the Chinese government and develop their own identity [4]. This independent movement threatens the viability of the unified communist system established by the People's Republic of China [1].

It is pertinent to study the China Uyghur conflict as it highlights the connection between a strong authoritarian state, a terrorist threat, and a minority group [5]. However, the implications of these dynamics are potentially far-reaching, as they promise to complicate China's rise in central Asia. Many western literatures describe Chinese politics as authoritarian; while this view is not inaccurate, it is incorrect to assume that Chinese citizens have been content to be despotically ruled [6]. As a multi-ethnic state with a vast majority of Han Chinese and various minorities, the Chinese government considers any nationalist or independence movements as an attack towards China's unified communist system and economic growth [7].

There are significant scholarly works on leveraging the Internet to gain more, and better information. Despite these possibilities, extant literature has shown that algorithmic and filtering features of social media platforms have driven users to an "echo chamber" whereby they are exposed to more of what they want and like, as opposed to what they need or should see [8][9]. This can shift their narratives on world issues as users disregard any narrative about topics that are ideologically unpleasant. The pervasiveness of partisan animosity on social media also exacerbates this issue [5]. Researchers have attempted to assess the definition and representation of identities and the leveraging power of minorities versus a superior state in the negotiation process [5], [10]. The research in [11] concluded that the approaches utilized by both parties in the 'David and Goliath' duel for a contested region mostly affect the minority group due to low availability of resources and strategies.

This article will firstly offer theoretical background about this topic, and then engage in turn with how we leveraged topic modeling, toxicity analysis and bot assessment to understand the China-Uyghur issue. The remainder of the article is structured as follows. In section II, a few extant literature and analytical frameworks relating to China and Uyghur issues are reviewed. Next, the empirical study and the findings are discussed in Section III. Lastly, we discuss conclusions, limitations, and directions for future work in Section IV.

II. Literature Review

This section describes the extant literature on this topic and the theoretical framework we used for this study.

A. China and the Uyghurs

It is important for the Uyghur diaspora to establish links with the international community and create awareness in the West, especially amongst non-governmental organizations and human rights activists, so that it can exert some pressure on the Chinese state to correct the plight of the Uyghurs. Researchers have argued that Beijing's strategies in Xinjiang with respect to the Uyghur issues at the domestic, regional, and international levels are characterized with multiple contradictions [5], [12]. They further reasoned that China's approach to Xinjiang domestically contributed to the internationalization of the issue [12]. However, others have argued that China faced the prospect of Xinjiang becoming its own West Bank if it fails to re-strategize to a softer approach to integrate the region [12]. They argue that China has explicitly framed episodes in world events such as the 9/11 crisis to shift the narratives towards Uyghur rebellion as "terrorism" and boost their international and regional sympathy [12].

Researchers have also explored how the increasing complexity of the conflicts between Uyghur and China indicates the potential for Uyghur violence to escalate [12], [13]. This is specifically in light of the reported inception of a state-initiated mass 'reeducation' campaign for Uyghur and other Muslim minorities across the province [13]. They argued that, by reportedly sending Xinjiang's Muslim population to 'vocational education centers,' China's attempts to 'prevent extremism' may lead to a resurgence of ethnic unrest in Xinjiang [13].

B. Toxicity analysis on Social Media

Toxicity analysis has been used to understand the pulse of society on hot-button issues [14]. In a study conducted in [14], the researchers evaluated five categories of toxicity on comments posted on pro-and anti-NATO channels on YouTube. They demonstrated that anti-NATO channels comments were more toxic when compared to pro-NATO channels comments. Researchers have also aimed to characterize and predict the behavior of toxic users in online discussions [15]. They found topical predictions of toxic response with semantic shifts from parent comments in their study. Another study analyzes online toxicity with a case modeling approach [16]. The authors developed an epidemiological model to study and evaluate the spread of toxicity on YouTube. They applied the Susceptible, Toxic, Recovered, Susceptible (STRS) model to detect similarities between toxicity propagation on YouTube and the spread of a disease within a population. In another study, the authors evaluated the role of toxicity on tweets about societal issues such as the wearing of face masks during the COVID-19 pandemic [17]. Their results showed that tweets with pro-mask hashtags that supported wearing masks were less toxic compared to tweets who spread news about COVID-19 on YouTube.

C. Network Analysis

Tighter government regulations on online activities can make users seek a more democratic channel/outlet. However, Song et al. [18] found an increased success of China's Internet repression where the Chinese Twitter proved to be small, lacking an accessible and diverse network due to China's sophisticated Internet content control regime. This coincided with the debate on the Chinese government approach to public diplomacy. Huang et al. [19] demonstrated how the Chinese government utilizes communication channels, specifically a small number of Twitter accounts, to amplify its public diplomacy network and promote China's international influence. Huang et al. [19] further explained that China's robust Twitter network function on "timid polyphony" centered around its closest friends with expansion outward to include other alliances. Researchers have also shown how public leaders such as politicians utilize micro-blogging platforms like Twitter to gain rapid attention compared to other traditional ways of communication. Khan et al. [20] demonstrated that understanding the supporters' network of opinion leaders helps in predicting the type of relationship between supporters of the leaders.

D. Bot Analysis

Bot and botnet activities have the ability to shift narratives, opinions, and behavior of humans, especially within the political landscape where hot-button issues are debated. Ferrara et al. [21] explained that there are economic and political incentives for injecting social bots into online ecosystems. Some bots act with the objective of forming and growing an audience to exert influence. Further, research in technographic approach argues that the agency of bots should be seen not only as computing units but as interlocutors and informants [22]. Their study of chatbots development in China proved that elevated disruptive technologies such as artificial intelligence and big data are critical factors in state security and narrative control in China [22]. Another study on computational propaganda, domestic automation and opinion manipulation utilizing 1.1 million hashtags on Twitter associated with China and Chinese politics showed a large amount of automation [23]. This automation, however, was more aligned with anti-Chinese state perspectives [23].

III. Methodology

This section focuses on our study design, which consists of the data collection and approaches applied for this research.

E. Data Collection and Processing

To understand the online universal conversation specific to China and Uyghur, we collected data tailored towards narratives containing a set of preliminary key phrases such as "China" and "Uyghur." This allowed us to query and truncate our data to tweets that focus on key issues relating to both China and the Uyghur group. This approach functioned as a filter for refining our data and eliminating any term or outliers irrelevant to our research. We extracted metadata from users and posts on Twitter utilizing our in-house Twitter API crawler. All tweets collected were posted between 2020–2021. Table 1 shows the breakdown of the total tweets extracted for China and Uyghurs, respectively.

TABLE I. FREQUENCY OF TWEETS FOR CHINA AND UYGHUR

Narrative	Tweets	Users
China	1,508,016	768,855
Uyghur	1,731,233	762,364

We applied this date range based on peak periods of tweets cross referencing to specific events and news relating to China and Uyghur.

F. *Topic Modeling*

To understand the influential topics in our dataset, we applied Latent Dirichlet Allocation (LDA) topic modeling on the extracted tweets. We first tokenized each tweet into sentences, and sentences into words with the removal of punctuation and stopwords. Words were lemmatized and stemmed to their root form. The model was initially trained on a random number of topics and later decreased and ranked to the top 4 topics based on the coherence score of the topic distribution. Topic modeling revealed topic 1 and topic 2 as top topics with distinct overlaps in China narrative. Both topics contained trending words relating to communism, policing, and the Chinese Communist Party.

TABLE II. TOP TOPICS WITHIN CHINA AND UYGHUR NARRATIVE.

	China			Uyghur		
Topic	Word 1	Word 2	%	Word 1	Word 2	%
1	Communist	CCP	0.68	Home	Force	0.53
2	Positive	Chin	0.15	Education	Jalan	0.35
3	AMP	Papua	0.15	Genocide	Stop	0.07
4	Youth	Muslims	0.09	Uyghur	China	0.01

Topic 1 has top words such as "home" and "force" with highest distributions within Uyghur narrative and relate to the reinforcement of forced Labor on Uyghur Muslims. Table 2 shows top words relating to China and Uyghur along with their respective distributions.

G. *Toxicity Analysis*

Since tweets contain a wealth of information about the thoughts and feelings of people, it is imperative to analyze the toxicity of tweets discussing China-Uyghurs conflicts. By definition, online toxicity can be seen as any online harassment that silences important voices in a discourse or forces marginalized people offline [24]. Toxic tweets were evaluated using natural language processing techniques specifically, Google perspective API which utilizes machine learning to detect toxic comments and Detoxify, a pre-trained model trained to minimize bias while detecting toxic sentences [24], [25]. Detoxify was trained on 3 Jigsaw challenges: *Toxic comment classification, Unintended Bias in Toxic comments, and Multilingual toxic comment classification* aimed to detect harmful content online [25]. Both techniques are multilingual and offer a probability score between 0 and 1 with a higher score indicating a higher toxicity.

Final toxicity scores were averaged and aggregated monthly within the period of January 2020 to December 2022. The results were then multiplied by topic distribution scores to get the toxicity per topic. Figure 1 shows the volatility of toxic tweets across the top 4 topics relating to China. The most influential topic, Topic 1, had the highest toxicity relative to other topics. This pattern is explainable through the semantics of trending words in Topic 1, which revealed top conversations relating to communist, Chinese Communist Party (CCP), and Chinese government. This signals that events in this period relating to these top words triggered negative interests of Twitter users which correlates to the high toxicity of tweets. Distinct events within this period that coincided with various spikes include: "The 50 independent United Nations Human Rights experts highlighting their concern on the situation in China relating but not limited to forced labor" [26], [27].

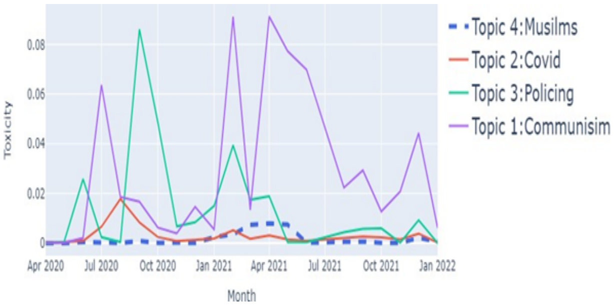


Figure 1. China Toxicity Trend within the period of 2020 - 2022.

Similarly, a high and volatile toxicity with noticeable spikes across the period was found within the Uyghur narrative, demonstrating an ongoing discussion of issues and events on these topics throughout the trend's lifecycle. Noticeable events during this period that coincide with these topics include "Officials denied the existence of internment camps, or alternatively justify them as poverty alleviation and stability maintenance efforts" and "uncovered evidence by the New York Times that reveal that Uyghur laborers, many who are interned forcibly, are involved in making personal protective equipment that are shipped all around the world [28], [29]."

H. Network Analysis and Bot Analysis

Understanding the connective relationships within both narratives helps to discover information flows and any concerted tactics about our topics. We leveraged network analysis tools such as NetworkX and Gephi to analyze and visualize social networks of both narratives [30], [31]. We utilized peak points found in our monthly tweets frequency reports to study each narrative social structure. Extreme overlaps were found in tweets posted within various peak points to news events on top topics. We discovered that the behavioral trend of tweets frequency in both narratives increased and/or decreased at the same rate. Due to computational expenses of running network graphs on our full data, we applied a random sampling technique to approximate the period each narrative tweets trend began rising.

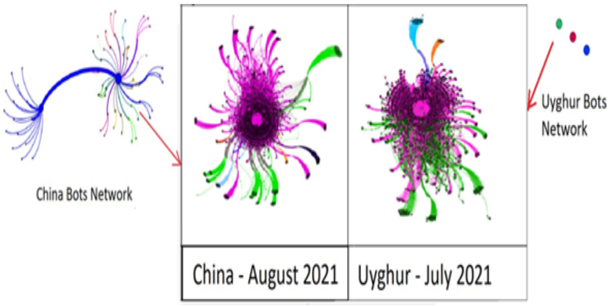


Figure 2. Network of users with bot CAP scores above 0.90 within period of 2020 - 2021

The China network focused on tweets posted in the period of August 2021 to September 2021 and references various events in August 2021 such as “Children of Detained Uyghurs parents held in Welfare schools in China’s Xinjiang” [32]. Additionally, the Uyghur network looked at July 2021 to August 2021, referencing events such as “president Xi praises Xinjiang armed police for counter terrorism effort in Uyghur territory” [33]. The biggest rise for both narratives was seen in July 2021 to November 2021. To measure the quality of division of both networks, a modality community detection algorithm was applied where higher modularity value indicated strong, distinct communities with relatively dense connections. The top 3 communities were color coded purple, green, and blue according to ranking (see Figure 2). A total of 16 dense communities were detected within a corpus of 12,292 users in China network from a modularity class of 0.673. The Uyghur network was less dense than the China network with a total of 17 communities and a modularity weight of 0.536 within 5,059 users. The majority of users within both networks had less than 500 connections with a relatively low average degree. However, about 10% of these connections had a following count of 1000 or greater. This was seen through China’s network top contributor @PaulS- mall4eva with 3 39,9 followers and connections such as @PinkRangerLB who had 100 followers. The Uyghur network had top contributor @RAbdiAnalyst maintaining identifiers such as Chief Analyst, geopolitics, and strategy with a following count of 255,000.

The nature and range of bot behaviors makes it universally difficult to define a bot [34]. To balance false positives and negatives, we applied the Complete Automation Probability (CAP) of 0.90 or higher to raw bot-scores to detect bots. CAP is a probability calculation developed by Observatory on Social Media project API ***Botometer*** that utilized Bayes’ theorem to estimate the overall prevalence of bots on a score of 0 to 1 [34]. Higher scores equate to higher probability of bot-like activity. Figure 2 highlights 18 bot communities mirroring the China network and 3 bot communities with no relations in the Uyghur network, while Figure 3 and Figure 4 show bot activities trend co-relating to toxicity on topics within the China network.

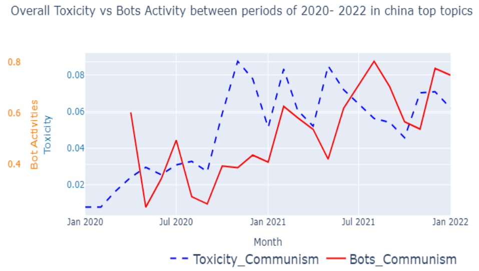


Figure 3. Overall Toxicity vs Bot Toxicity within topic 1 about Communism.

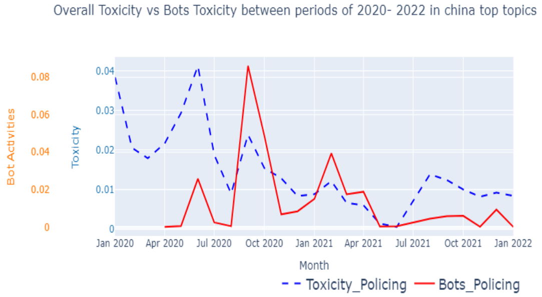


Figure 4. Overall toxicity vs Bot Toxicity within topic 3 about communism.

Overall, toxicity is directly proportional and highly comparable to bot activity in communism topics while it is relatively low but still comparable to topics on policing. These findings imply that bot activities jolted narratives toxicity and shifted opinions against communism issues in China. Future research can explore the intention of these accounts’ generation.

IV. Conclusions and Future Work

In this study, we examine the prevalence of toxicity in the China-Uyghur dilemma on Twitter. To gain insight into the viewpoints of actors in the network, we focus on top topics related to the two focal narratives and utilize network analysis tools such as Network-X and Gephi. Our network analysis is based on theoretical frameworks used in previous studies and employs modularity to detect communities. This paper contributes to the ongoing research on the online dialogue of diplomacy, identity, and policies within an authoritative state and their impact on the rights of minority groups. It provides an interoperable methodology to understand relevant topics, identify toxicity, and detect top contributors within the network. We found that actors in the network include those who push the Uyghur identity conflict beyond China's borders, neutral actors such as news agencies, and initiators who raise awareness of Uyghur issues. The topics within the network range from forced labor, genocide, education, communism, politics, and policing, and their differences provide an overlapping representation of the China-Uyghur network. Our findings suggest the presence of probable anti-China communities with top contributors and smaller connections discussing relevant topics. Further study is necessary to understand the evolution of these networks over time. The data for this study was collected after these events relating to China-Uyghur narratives had unfolded. Future research can investigate the use of social network analysis techniques to make real-time inferences about emerging socio-political issues.

Acknowledgment

This research is funded in part by the U.S. National Science Foundation (OIA-1946391, OIA-1920920, IIS-1636933, ACI-1429160, and IIS-1110868), U.S. Office of the Under Secretary of Defense for Research and Engineering (FA9550-22-1-0332), U.S. Office of Naval Research (N00014-10-1-0091, N00014-14-1-0489, N00014-15-P-1187, N00014-16-1-2016, N00014-16-1-2412, N00014-17-1-2675, N00014-17-1-2605, N68335-19-C-0359, N00014-19-1-2336, N68335-20-C-0540, N00014-21-1-2121, N00014-21-1-2765, N00014-22-1-2318), U.S. Air Force Research Laboratory, U.S. Army Research Office (W911NF-20-1-0262, W911NF-16-1-0189, W911NF-23-1-0011), U.S. Defense Advanced Research Projects Agency (W31P4Q-17-C-0059), Arkansas Research Alliance, the Jerry L. Maulden/Entergy Endowment at the University of Arkansas at Little Rock, and the Australian Department of Defense Strategic Policy Grants Program (SPGP) (award number: 2020-106-094). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding organizations. The researchers gratefully acknowledge the support.

References

[1] G. Bovingdon, "Autonomy in Xinjiang: Han nationalist imperatives and Uyghur discontent,", East-West-Center Washington, 2004.

[2] S. Glen, "Sources of Conflict in the Xinjiang Uyghur Autonomous Region: History, Power, and Uyghur Identity Flux?,", Griffith Asia Quarterly, vol. 3, no. 1, pp. 630-2051, 2015.

[3] M. Julienne, M. Rudolf, and J. Buckow, "Beyond Doubt: The Changing Face of Terrorism in China". The Diplomat. [Online]. Available from: https://thediplomat.com/2015/05/beyond-doubt-the-changing-face-of-terrorism-in-china/ May 28, 2015, [retrieved: March 8, 2023].

[4] Z. Raza, "China's 'political re-education' camps of Xinjiang's Uyghur Muslims," Asian Affairs, vol. 50, no. 4, pp. 488-501, 2019.

[5] A. Lecours and N. Geneviève, eds., "Dominant Nationalism, Dominant Ethnicity: Identity," Federalism, and Democracy, vol. 15, Peter Lang, 2009

[6] K. Mukherjee, "The Uyghur question in contemporary China," Strategic Analysis, vol. 34, no. 3, pp. 420-435, 2010.

[7] E. Hyer, "China's policy towards Uighur nationalism," Journal of Muslim Minority Affairs, vol. 26, no. 1, pp. 75-86, 2006

[8] M. Kent, "Managerial rhetoric as the metaphor for the World Wide Web," Critical Studies in Media Communication, vol. 18, no. 3, pp. 359-375, 2001.

[9] M. Kent, "Using social media dialogically: Public relations role in reviving democracy," Public Relations Review, vol. 39, no. 4, pp. 337-345, 2013.

[10] E. Davis, "Uyghur Muslim ethnic separatism in Xinjiang, China," Asian Affairs: An American Review, vol. 35, no. 1, pp. 15-30, 2008.

[11] S. Glen, "Sources of Conflict in the Xinjiang Uyghur Autonomous Region: History, Power, and Uyghur Identity Flux?," Griffith Asia Quarterly, vol. 3, no. 1, pp. 630-2051, 2015.

[12] M. Clarke, "China and the Uyghurs: the 'Palestinization' of Xinjiang," Middle East Policy, vol. 22, no. 3, pp. 127-146, 2015.

[13] N. Soliev, "Uyghur violence and Jihadism in China and beyond," Counter Terrorist Trends and Analyses, vol. 11, no. 1, pp. 71-75, 2019.

[14] A. Obadimu, T. Khaund, E. Mead, T. Marcoux, and N. Agarwal, "Developing a socio-computational approach to examine toxicity propagation and regulation in COVID-19 discourse on YouTube," Information Processing & Management, vol. 58, no. 3, pp. 102660, 2021.

[15] H. Almerekhi, H. Kwak, B. J. Jansen, and J. Salminen, "Detecting toxicity triggers in online discussions." In Proceedings of the 30th ACM conference on hypertext and social media, Sep 2019, pp. 291-292, doi: 10.1145/3342220.3344933.

[16] A. Obadimu, E. Mead, M. Maleki, and N. Agarwal, "Developing an epidemiological model to study spread of toxicity on YouTube." In Social, Cultural, and Behavioral Modeling: 13th International Conference, SBP-BRiMS 2020, Washington, DC, USA, October 18–21, 2020, Proceedings 13, 2020, pp. 266-276.

[17] P. Pascual-Ferrá, N. Alperstein, D. J. Barnett, and R. N. Rimal, "Toxicity and verbal aggression on social media: Polarized discourse on wearing face masks during the COVID-19 pandemic," Big Data & Society, vol. 8, no. 1, pp. 20539517211023533, 2021.

[18] S. Y. Song, R. Faris, J. Kelly, "Beyond the Wall: Mapping Twitter in China," Berkman Center Research Publication, no. 2015-14, 2015.

[19] Z. Huang and W. Rui, "Building a network to 'tell China stories well': Chinese diplomatic communication strategies on Twitter," International Journal of Communication, vol. 13, pp. 2984-3007, 2019.

[20] A. Khang et al. ,"Predicting politician's supporters' network on Twitter using social network analysis and semantic analysis," Scientific Programming, vol. 2020, pp. 1-17, 2020.

[21] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini, "The rise of social bots," Communications of the ACM, vol. 59, no. 7, pp. 96-104, 2016.

[22] Y. Xu, "Programmatic dreams: Technographic inquiry into censorship of Chinese chatbots," Social Media + Society, vol. 4, no. 4, 2018, doi: 2056305118808780.

[23] G. Bolsover and P. Howard, "Chinese computational propaganda: Automation, algorithms and the manipulation of information about Chinese politics on Twitter and Weibo," Information, Communication & Society, vol. 22, no. 14, pp. 2063-2080, 2019.

[24] Perspective API. Google. [Online]. Available from: https://perspectiveapi.com [retrieved: 03, 2023]

[25] Detoxify. PyPI - the Python Package Index. [Online]. Available from: https://pypi.org/project/detoxify/ [retrieved: 03, 2023].

[26] M. Young, The Technical Writer's Handbook, University Science, Mill Valley, CA, 1989.

[27] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini, "The rise of social bots," Communications of the ACM, vol. 59, no. 7, pp. 96-104, 2016.

[28] United Nations News. "1067312,". [Online]. Available from https://news.un.org/en/story/2020/06/1067312. 06,2020. [Retrieved: 03, 2023].

[29] New York Times. China Investigates Mask Factory Using Forced Labor of Uighur Muslims. [Online]. Available from: https://www.nytimes.com/2020/07/19/world/asia/china-mask-forced-labor.html. 07 2020. [Retrieved: 03, 2023].

[30] NetworkX. Network analysis in Python. [Online]. Available from: https://networkx.org/ Retrieved: 03, 2023].

[31] Gephi. The open Graph Viz-Platform. [Online]. Available from: https://gephi.org/. [Retrieved: 03, 2023].

[32] Uyghur human rights project. Timelines: Forced Labour. [online]. Available from: https://xinjiang.sppga.ubc.ca/timelines/forced-labour/ [Retrieved: 03, 2023].

[33] S. Issabayeva, "President XI praises Xinjiang Armed Police for 'Counter Terrorism' efforts in Uyghur territory," World Uyghur Congress. [Online]. Available from: https://www.uyghurcongress.org/en/president-xi-praises-xinjiang-armed-police-for-counter-terrorism-efforts-in-uyghur territory/?utm_source=rss&utm_medium=rss&utm_campaign =president-xi-praises-xinjiang-armed-police-for-counter-terrorism-efforts-in-uyghur-territory 07. 07 2023. [Retrieved: 03, 2023]

[34] Botometer. Frequently Asked Questions. [Online]. Available from: https://botometer.osome.iu.edu/faq [Retrieved: 03, 2022].

Digital Practices and Uses in Secondary Schools

Achieving Digital Literacy

Christian Bourret
DICEN IDF – Gustave Eiffel University
Marne-la-Vallée - France
E-mail : christian.bourret@univ-eiffel.fr

Hafida Hammadi
DICEN IDF – Gustave Eiffel University
Marne-la-Vallée -France
Email : hafi_12@yahoo.fr

***Abstract*— In the Georges Seurat Middle School in Courbevoie (Ile-de-France region), our study focuses on the description and analysis of students' and teachers' digital activities in the classroom and online, as well as the digital content offered in the school environment (OZE92) in face-to-face, distance learning, and hybrid courses.**

Keywords- digital literacy, digital workspace, digital practices, digital education, media literacy.

I. INTRODUCTION

This work presents qualitative research conducted with both students and teachers. We intend to complete this component and begin the quantitative study's reflection phase. The results show that group work motivation, digital knowledge, and abilities are actively being acquired. Hence, is it possible to enhance and deepen digital literacy in secondary schools.

The objective of our study is to assess how secondary school students and teachers interact with and use technology, teach digital literacy to students (digital citizens), how to instruct students in digital literacy, improve and upgrade professional teaching methods and contribute to the development of the professor's digital teaching strategies for secondary schools.

II. SCIENTIFIC POSITIONING

Our project is based on the interdisciplinary field of information, communication, and education sciences, as well as the position illustrated by F. Bernard regarding the articulation of the four dimensions of the link, meaning, knowledge, and action [1]. We shall employ the concepts of the situation (particularly learning and observation) and socio-technical devices or mediating artifacts. We shall begin with A. Mucchielli's " Dynamic situational contextualization method" and the various contexts that he proposes for a situation [2]. Then, we move on to A. Mucchielli's "Situation and Communication" (2010), where he discusses the "genesis of meanings" using a semiotic method based on "contextualization." [3]. Based on the socio-constructive approach, we shall examine social interactions in training [4].

We also analyze group interactions in the frame of the Adaptive Structuration Theory (AST) as Scott Poole argues that group members intentionally adopt rules and resources to achieve goals. Poole holds that group members have an impact on outcomes. In the seven necessary pieces of knowledge known for future education, we can refer to Edgar Morin [5].

III. METHODOLOGY

Our research applies a mixed methodology:

A. *Qualitative approach* (observations, interviews, and content analysis) and

B. *Quantitative approach* (based on questionnaires).

Alex Mucchielli's "Constructivist approach to communications" includes a qualitative study. It is a random sample of 397 students from various classes. This study is preliminary and based on 219 hours of observations in the classroom. The goal of qualitative research is to understand, analyze, and quantify the barriers to the growth of digital uses, practices, contents, devices, and so on in the context of schools.

This study relied on several digital projects that were carried out in different classes including media class, WebTv, Green delegate project, “O Lab Citizen”, Mediatiks competitions (online newspaper, photo-reportage), and others.

IV. RESULTS

3.1. Students

The observations of our first study in the Digital Collaborative Space (DCS) are as follows: learners are more confident in DCS after observing digital practices. The enjoyable aspects of DCS encourage frequent participation. Online collaboration makes group work easier. Students are provided with the necessary homework tools, allowing them to complete their assignments in an optimal and efficient manner. Students can use the chat to ask questions and receive personalized responses. DCS allows teachers to engage students in personalized and differentiated instruction. All of the DCS's tools promote communication between students and teachers as well as between peers. The teacher's role has changed; he or she now supervises and facilitates the students' learning. Students learn to work independently as well as in groups.

Students having trouble are assisted by the pair in the construction of knowledge within the framework of the DCS, which produces motivation and the desire to make an effort to complete the task.

3.1.1. DCS's difficulties

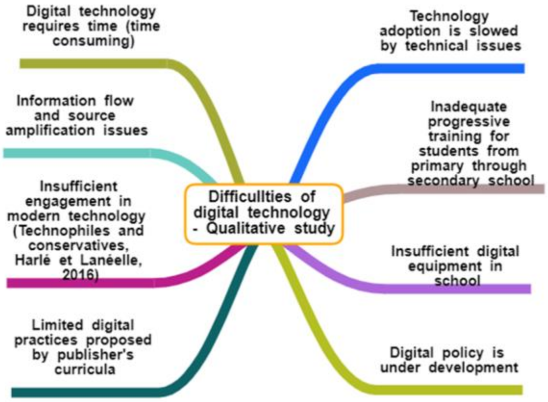
Progressive digital training constitutes one of the hardest subjects in the curriculum for students. There are differences in how students are trained because some academic subjects do not employ common digital technology. Also, the learner's level of digital literacy is influenced by the environment in which they live.

3.2. *Teachers*

In a setting where knowledge is created collaboratively within a group under the guidance of the teacher, the latter's role is modified, and he/she positions themselves as a facilitator and guide. The teacher is no longer the exclusive source of knowledge. Conversation and exchanges are encouraged through various communication technologies such as synchronous documents, chat, commenting, email, and so on. Within the parameters of the synchronous document, the teacher may adopt the immediate correction. This enables the student to receive individualized remediation. Students are encouraged to work both independently and in groups in a flipped classroom. By adapting objectives and content to the abilities and challenges of each student, the teacher in DCS can implement differentiated instruction. The teacher has access to the student's work at the same time. As a result, the student is encouraged to participate in the project and to demonstrate his or her involvement by name. All the tools required to complete the project are provided in the collaborative space where it is set up online (resources, etc.). In DCS, teachers can implement competency-based instruction by focusing on specific skills such as online information validation. Students are taught how to use the "OZE92" digital workspace's resources and services (ENT). They are also given information-documentary training (online research methodology, legal aspects of information, validation of online information, etc.). Students work exclusively on digital material during this experiment; paper copies are not necessary.

3.2.1 Teacher interviews – Results

Six teachers of mathematics, French, arts, sciences, history, and languages were questioned on the difficulties associated with the use of digital technology in classrooms.


Difficullties of digital technology - Qualitative study
Digital technology requires time (time consuming)
Information flow and source amplification issues
Insufficient engagement in modern technology (Technophiles and conservatives, Harlé et Lanéelle, 2016)
Limited digital practices proposed by publisher's curricula
Technology adoption is slowed by technical issues
Inadequate progressive training for students from primary through secondary school
Insufficient digital equipment in school
Digital policy is under development

V. Conclusion

As a preliminary conclusion, we have found out that both individual and group work on the DCS is highly effective. The projects are progressing rapidly, and the objectives are being met. Also, we trust that the adoption of progressive digital citizen training should be applied in compulsory education. The equipment of schools and the training of teachers would be essential elements to guarantee the success of this process. It is worth mentioning that the school's director confirmed that digital technology represents one of the school's pillars of policy and its priority.

References

[1] F. Bernard, "Information and communication sciences is a decentralized and open field of study", in BOUZON (Arlette) under the dir. de, *Debatable organizational communication. Domains, concepts, and perspectives*, Paris, L'Harmattan, 2006, pp. 33 – 46.

[2] A. Mucchielli and C. Noy, "Communication studies: Constructivist approaches", Collection U, Ed. Armand Colin, 2005, p. 240, chap. 6, p. 113-124.

[3] A. Mucchielli, "Situation and communication", Ed. Ovadia, Nice, 2010, pp. 170.

[4] L. Begue, O. Desrichard, "Social psychology treatise. The study of interpersonal interactions", Brussels, De Boeck, 2013, pp. 849.

[5] E Morin, The seven pieces of knowledge known for future education, Ed. Seuill- UNESCO 1999, p. 71.

An Investigation of One Sided Follow Relations between Twitter Users Concerned with Tweets Disclosing Submitters' Personal Information

Yasuhiko Watanabe, Toshiki Nakano, Hiromu Nishimura, and Yoshihiro Okada
Ryukoku University
Seta, Otsu, Shiga, Japan
Email: watanabe@rins.ryukoku.ac.jp, t180450@mail.ryukoku.ac.jp,
t160405@mail.ryukoku.ac.jp, okada@rins.ryukoku.ac.jp

***Abstract*—Nowadays, many people use a Social Networking Service (SNS). Most SNS users are careful in protecting the privacy of personal information: name, age, gender, address, telephone number, birthday, etc. However, some SNS users disclose their personal information that can threaten their privacy and security even if they use unreal name accounts. In this study, we investigated Twitter users who gave likes to tweets disclosing submitters' personal information that potentially threatened submitters' privacy and security. We collected 318 tweets promising to disclose submitters' personal information. Then, we investigated the one sided follow relations between the submitters of these 318 tweets and users who gave likes to them. The results of our survey showed that giving likes to tweets promising to disclose submitter's personal information is not a sufficient trigger to get to follow users. Submitters were careful to follow unfamiliar users even if the users followed them and gave likes to their tweets. Also, users were careful to follow unfamiliar users even if the users followed them and gave likes to the same tweets.**

Keywords–personal information; Twitter; SNS; privacy risk; one sided follows; unreal name account user.

I. Introduction

Nowadays, many people use a Social Networking Service (SNS) to communicate with each other and try to enlarge their circle of friends. SNS users are generally concerned about potential privacy risks. To be specific, they are afraid that unwanted audiences will obtain information about them or their families, such as where they live, work, and play. As a result, SNS users are generally careful in disclosing their personal information. They disclose their personal information only when they think the benefits of doing so are greater than the potential privacy risks. However, some SNS users, especially young users, disclose their personal information on their profiles, for example, real full name, gender, hometown and full date of birth, which can potentially be used to identify details of their real life, such as their social security numbers. In order to discuss the reasons why some SNS users disclose their personal information willingly, it is important to investigate who their intended readers are. However, it is difficult to ask them who their intended readers are. To solve this problem, it is important to investigate who gave responses to their SNS messages disclosing their personal information. This is because, if submitters felt unwanted audiences read and gave responses to their SNS messages disclosing their personal information, they would delete them. In order to investigate who gave responses to SNS messages disclosing submitters' personal information, we investigate Twitter users who gave likes to tweets disclosing submitters' personal information. Furthermore, we investigate follow relations between users concerned with a tweet disclosing submitter's personal information. In other words, we investigate

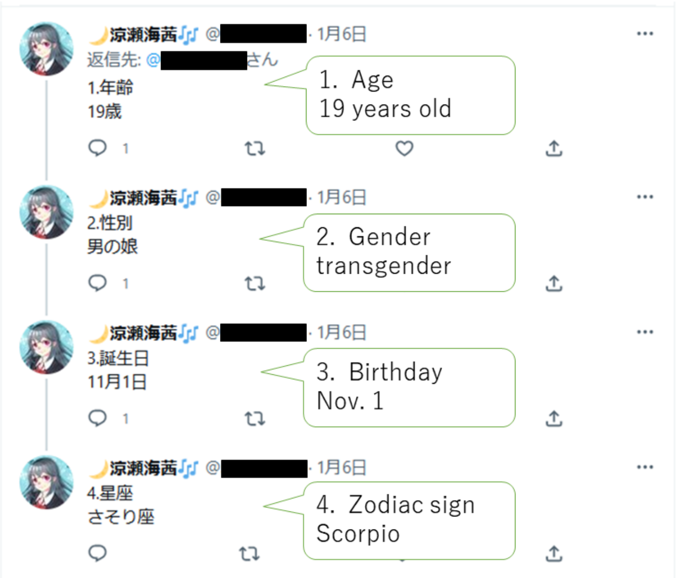
- whether a submitter followed users who gave likes to his/her tweets disclosing his/her personal information,
- whether users who gave likes to submitter's tweet disclosing his/her personal information followed the submitter, and
- whether each user who gave a like to a tweet disclosing submitter's personal information followed every other user who gave a like to the same tweet.

In our previous work, we reported mutual follow relations and no follow relations between users concerned with a tweet disclosing submitter's personal information [1]. In this study, we investigate one sided follow relations between them. It is important to investigate one sided follow relations between users because they are bound to happen in the process of acquaintance between users who do not follow each other. By using the results of the investigation, we discuss the groups of submitters and users who gave likes to tweets disclosing submitters' personal information.

The rest of this paper is organized as follows: in Section II, we survey the related works. In Section III, we show how to collect tweets where submitters seemingly disclosed their personal information honestly and detect users who gave likes to them. In Section IV, we investigate one sided follow relations between users concerned with a tweet disclosing submitter's personal information and discuss the groups of submitters and users who gave likes to tweets disclosing submitters' personal information. Finally, in Section V, we present our conclusions.

II. Related work

Personally identifiable information is defined as information which can be used to distinguish or trace an individual's identity such as social security number, biometric records, etc. alone, or when combined with other information that is linkable to a specific individual, such as date and place of birth, mother's maiden name, etc. [2] [3]. Internet users are generally concerned about unwanted audiences obtaining personal information. Fox et al. reported that 86% of Internet users are concerned that unwanted audiences will obtain information about them or their families [4]. Also, Acquisti and Gross reported that students expressed high levels of concern for general privacy issues on Facebook, such as a stranger finding out where they live and the location and schedule of their classes, and a stranger learning their sexual orientation,



🌙涼瀬海茜🎶 @[redacted] · 1月6日
返信先: @[redacted]さん
1.年齢
19歳

1. Age
19 years old

🌙涼瀬海茜🎶 @[redacted] · 1月6日
2.性別
男の娘

2. Gender
transgender

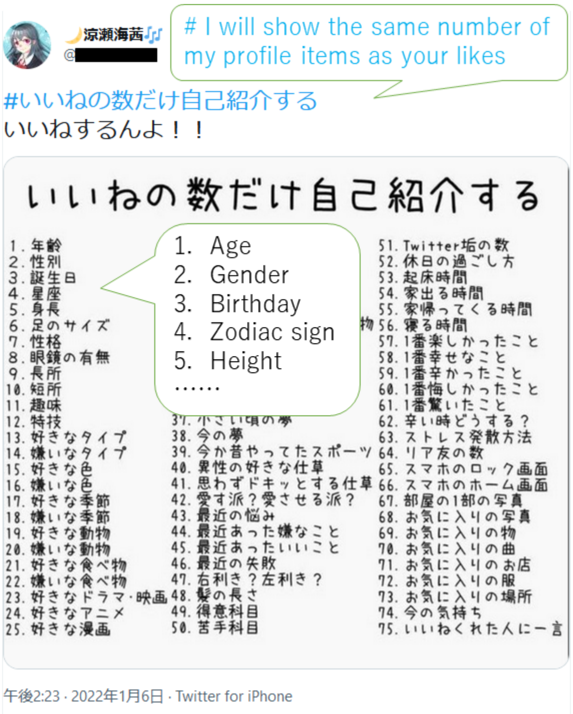
🌙涼瀬海茜🎶 @[redacted] · 1月6日
3.誕生日
11月1日

3. Birthday
Nov. 1

🌙涼瀬海茜🎶 @[redacted] · 1月6日
4.星座
さそり座

4. Zodiac sign
Scorpio

Figure 1. An unreal name account user, *Suzuse*, disclosed her personal profile items in her tweets.



🌙涼瀬海茜🎶
@[redacted]

I will show the same number of my profile items as your likes

#いいねの数だけ自己紹介する
いいねするんよ！！

いいねの数だけ自己紹介する

1. 年齢
2. 性別
3. 誕生日
4. 星座
5. 身長
6. 足のサイズ
7. 性格
8. 眼鏡の有無
9. 長所
10. 短所
11. 趣味
12. 特技
13. 好きなタイプ
14. 嫌いなタイプ
15. 好きな色
16. 嫌いな色
17. 好きな季節
18. 嫌いな季節
19. 好きな動物
20. 嫌いな動物
21. 好きな食べ物
22. 嫌いな食べ物
23. 好きなドラマ·映画
24. 好きなアニメ
25. 好きな漫画

37. 小さい頃の夢
38. 今の夢
39. 今か昔やってたスポーツ
40. 異性の好きな仕草
41. 思わずドキッとする仕草
42. 愛す派？愛させる派？
43. 最近の悩み
44. 最近あった嫌なこと
45. 最近あったいいこと
46. 最近の失敗
47. 右利き？左利き？
48. 髪の長さ
49. 得意科目
50. 苦手科目

51. Twitter垢の数
52. 休日の過ごし方
53. 起床時間
54. 家出る時間
55. 家帰ってくる時間
56. 寝る時間
57. 1番楽しかったこと
58. 1番幸せなこと
59. 1番辛かったこと
60. 1番悔しかったこと
61. 1番驚いたこと
62. 辛い時どうする？
63. ストレス発散方法
64. リア友の数
65. スマホのロック画面
66. スマホのホーム画面
67. 部屋の1部の写真
68. お気に入りの写真
69. お気に入りの物
70. お気に入りの曲
71. お気に入りのお店
72. お気に入りの服
73. お気に入りの場所
74. 今の気持ち
75. いいねくれた人に一言

1. Age
2. Gender
3. Birthday
4. Zodiac sign
5. Height
……

午後2:23 · 2022年1月6日 · Twitter for iPhone

Figure 2. A tweet promising to disclose the same number of submitter's personal profile items as likes to it.

name of their current partner, and their political affiliations [5]. However, Internet users, especially young users, tend to disclose personal information on their profiles, for example, real full name, gender, hometown and full date of birth, which can potentially be used to identify details of their real life, such as their social security numbers. As a result, many researchers discussed the reasons why young users willingly disclose personal information on their SNS profiles. Barnes argues that Internet users, especially teenagers, are not aware of the nature of the Internet and SNSs [6]. Barth et al. highly questioned whether privacy as a concept is already implanted in SNS users' perception and social representation [7]. Obar and Oeldorf-Hirsch reported that individuals often ignore privacy and terms of service policies for SNSs [8]. Viseu et al. reported that many online users believe the benefits of disclosing personal information in order to use an Internet site are greater than the potential privacy risks [9]. On the other hand, Acquisti and Gross explain this phenomenon as a disconnection between the users' desire to protect their privacy and their actual behavior [5]. Also, Livingstone points out that teenagers' conception of privacy does not match the privacy settings of most SNSs [10]. Alshaikh et al. reported that SNS users were worried about their individual information security especially when SNS organizations changed their privacy terms [11]. Joinson et al. reported that trust and perceived privacy had a strong affect on individuals' willingness to disclose personal information to a website [12]. Also, Tufekci found that concern about unwanted audiences had an impact on whether or not students revealed their real names and religious affiliation on MySpace and Facebook [13]. The authors also think that most students are seriously concerned about their privacy and security. However, they often underestimate the risk of their online messages and submit them. For example, Watanabe et al. focused on unreal name Twitter users who promised to disclose their personal profile items, analyzed the details of their personal profile items disclosed by themselves, especially their ages, genders, and heights, and showed that most of the submitters disclosed their ages, genders, and heights honestly [14].

III. A COLLECTION OF TWEETS DISCLOSING SUBMITTERS' PERSONAL INFORMATION

It is difficult to collect tweets disclosing submitters' personal information, such as tweets in Figure 1, directly. To solve this problem, we focused on tweets where submitters promised their audiences to disclose the same number of their own personal profile items as likes to their tweets. Figure 2 shows a tweet submitted by *Suzuse* on January 6, 2022. Both in Figure 1 and Figure 2, her screen name is redacted for privacy. Figure 2 shows that *Suzuse* promised her audiences to disclose the same number of her personal profile items as likes to her tweet. Actually, as shown in Figure 1, *Suzuse* submitted four replies disclosing her four personal profile items to her tweet shown in Figure 2 on January 6, 2022. Watanabe et al. reported that Twitter users seemingly disclosed their personal information honestly when they promised to do it, such as *Suzuse*'s tweet in Figure 2 [14]. As a result, it is easy to collect tweets disclosing submitters' personal profile items when we collect tweets promising to disclose submitters' personal profile items. Furthermore, they often used the same sentence in their tweets, like a game password, as shown in Figure 2, *# I will show the same number of my profile items as your likes*. In order to collect tweets promising to disclose submitters' personal profile items, we used the shared sentence as key to collect them. To be specific, we collected these tweets by using Twitter API v2 [15]. Twitter API v2 helps us to collect tweets where the given sentence is used. Also, Twitter API v2 helps us to collect user accounts who submitted a specific tweet and who gave likes to it. Furthermore, it helps us to collect

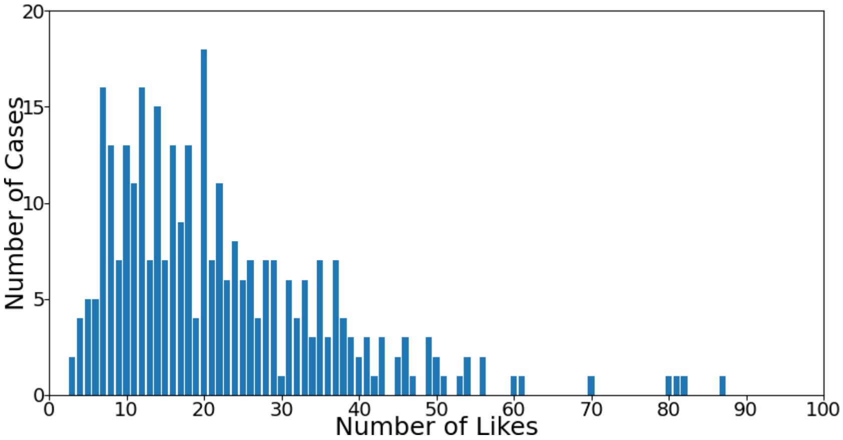


Figure 3. The histogram of the number of likes given to the 318 tweets promising to disclose submitters' personal information.

user accounts who are followed by a specific user. Every 10 PM, we tried to collect user accounts and their tweets

- that contained *# I will show the same number of my profile items as your likes*
- that were submitted in the past 24 hours, and
- that were given one or more likes.

After we obtained the tweets promising to disclose submitters' personal profile items, we tried to collect

- user accounts who gave likes to the obtained tweets and
- user accounts followed by the submitters of the obtained tweets and the users who gave likes to them

once daily for a week. Finally, we collected 318 Japanese tweets promising to disclose submitters' personal information. These 318 tweets were submitted from December 30, 2021 to January 31, 2022 by 317 users. One out of the 317 users submitted two tweets promising to disclose his personal information on January 12 and 17, 2022. These 318 tweets were given 7060 likes by 6325 users within a week after they were submitted. Figure 3 shows the histogram of the number of likes given to the obtained 318 tweets promising to disclose submitters' personal information. Figure 4 shows the daily number of likes given to the obtained 318 tweets in the investigation period. Day N in Figure 4 means that N days have passed since the obtained tweet was submitted and our investigation started. Day 6 was the last day of the investigation period. Figure 4 shows that 77 % of likes were given on Day 0. 30 tweets out of the 318 tweets were deleted within a week after they were submitted.

IV. One sided follow relations between users concerned with tweets disclosing submitters' personal information

In this section, we investigate one sided follow relations between users who communicated through tweets disclosing submitters' personal information. To be specific, we survey

- Twitter users who submitted tweets promising to disclose the same number of their own personal profile items as likes and
- Twitter users who gave likes to these tweets

and investigate

- whether an user who submitted tweets promising to disclose his/her personal information followed users who gave likes to his/her tweets,
- whether users who gave likes to tweets promising to disclose submitter's personal information followed the submitter, and
- whether users who gave likes to a tweet promising to disclose submitter's personal information followed each other.

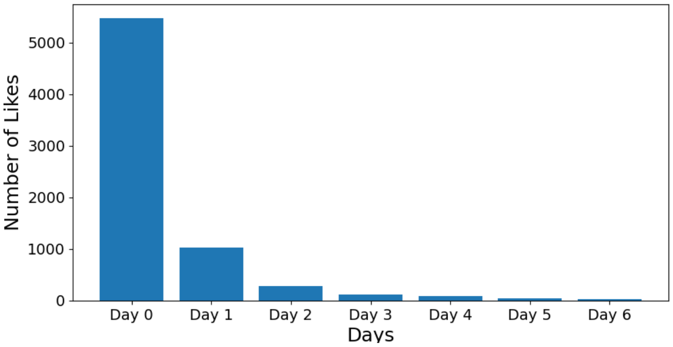


Figure 4. The daily number of likes given to the obtained 318 tweets since the tweets were submitted.

After collecting user accounts of submitters and users who gave likes to submitters' tweets, we analyze the relations between them. The relations between a submitter and an user who gave a like to submitter's tweet can be classified into four types:

- mutual follow relation: the submitter and the user mutually followed each other.
- one sided follow relation (from the submitter): the submitter followed the user, however, the user did not follow the submitter.
- one sided follow relation (to the submitter): the user followed the submitter, however, the submitter did not follow the user.

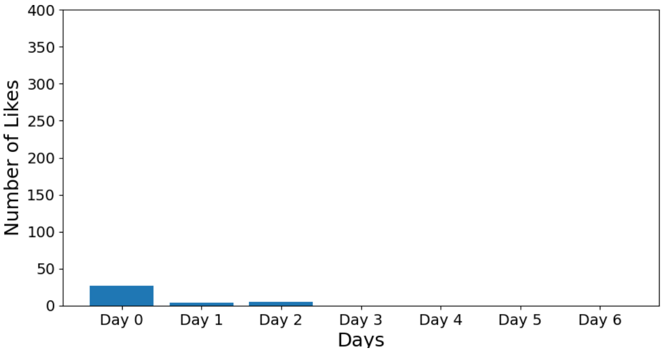


Figure 5. The daily number of likes given by users who did not follow submitters but were followed by the submitters in the investigation period.

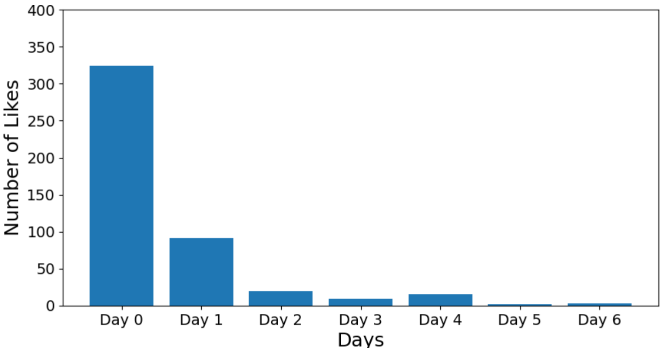


Figure 6. The daily number of likes given by users who followed submitters but were not followed by the submitters in the investigation period.

- no follow relation: the submitter and the user did not follow each other.

Figure 5 shows the daily number of likes given by users who did not follow submitters but were followed by the submitters in the investigation period. On the other hand, Figure 6 shows the daily number of likes given by users who followed submitters but were not followed by the submitters in the investigation period. Figure 5 and Figure 6 show that users who did not follow submitters but were followed by the submitters gave less likes than those who followed submitters but were not followed by the submitters. Furthermore, we analyze the relations among users who gave likes to submitter's tweet. They can also be classified into three types: mutual follow relation, on sided follow relation, or no follow relation.

Let us consider one example. As shown in Figure 2, a Twitter user, *Suzuse*, submitted a tweet promising her audiences to disclose the same number of her own personal profile items as likes on January 6, 2022 at 2:23 PM. We detected her tweet on the same day at 10:00 PM, and then, recorded that she received ten likes and submitted ten replies disclosing her ten personal profile items on January 6, 2022. After that, every 10 PM, we tried to check whether someone gave likes to her tweet and analyzed the relations between *Suzuse* and users who gave likes to her tweet by the day. For example, on January 7, 2022, we detected one more user gave a like to her tweet and recorded that *Suzuse* received eleven likes from eleven users by the day. Then, we analyzed the relations between *Suzuse* and each of the eleven users and confirmed that each of the eleven users followed *Suzuse* and she followed seven of them. As a result, the relations between *Suzuse* and the seven users were mutual follow relations. On the other hand, the relations between *Suzuse* and the other four users were one sided follow relations: these four users followed *Suzuse* but she did not follow them. Furthermore, we analyzed the relations among the eleven users who gave likes to her tweet by January 7, 2022. There were 55 cases to choose two out of the eleven users. In three cases out of the 55 cases, two users followed each other. On the other hand, in the other 52 cases, two users did not follow each other. As a result, the relation of three cases were mutual follow relations and the relations of the other 52 cases were no follow relations. On January 12, 2022, we confirmed that eleven users gave eleven likes to her tweet on January 6, 2022, as shown in Figure 2, and finished the investigation on her tweet.

A. *One Sided Follow relations between submitters and users who gave likes to submitters' tweets*

At first, we discuss the cases where submitters followed users who gave likes to their tweets, but the users did not follow the submitters. We call the follow relations between these submitters and users *one sided follow relations (from submitters)*. In order to discuss this type of follow relation, we introduce the ratio of one sided follow relations (from submitters) between a submitter and users who gave likes to his/her tweet. Suppose that the number of users who gave likes to tweet t is n and m of them do not follow the submitter of tweet t but are followed by him/her. Then, the ratio of one sided follow relations (from submitters) between the submitter of tweet t and the users who gave likes to it, $P_{OSFfromS}(t)$, is defined as follows:

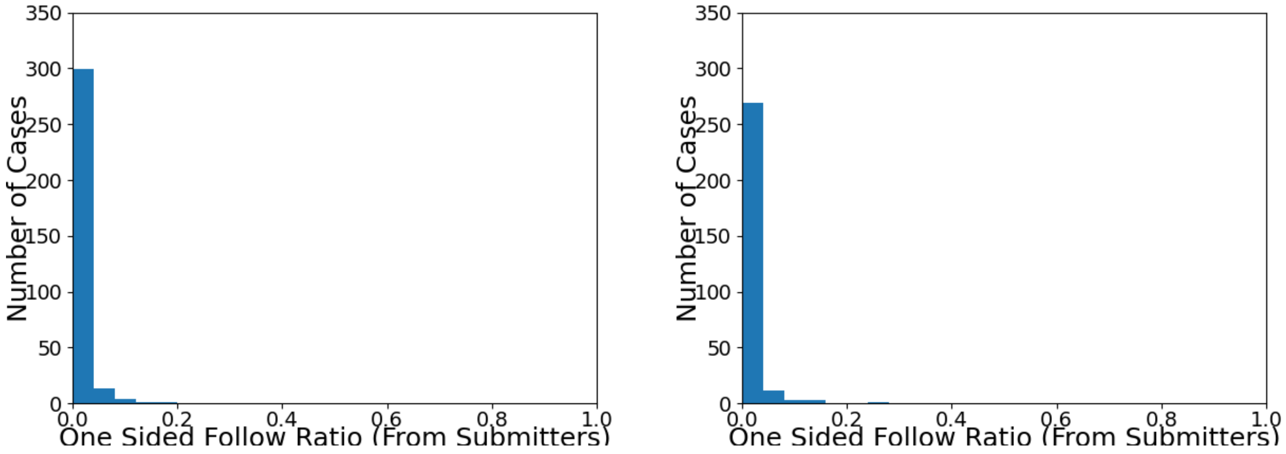
$$P_{OSFfromS}(t) = \frac{m}{n}$$

Figure 7 shows the distribution of the ratio of one sided follow relations (from submitters) between the submitters of the obtained 318 tweets and the users who gave likes to them. Furthermore, Figures 7 (a) and (b) shows the distribution of them investigated on the Day 0 and Day 6, respectively. As shown in Figure 7, the ratio in each case was less than 0.2. There were few cases where two or more users who had one sided follow relations (from submitters) with a submitter gave likes to his/her tweet promising to disclose his/her personal information.

Next, we discuss the cases where submitters did not follow users who gave likes to submitters' tweets, but the users followed the submitters. We call the follow relations between these submitters and users *one sided follow relations (to submitters)*. In order to discuss this type of follow relation, we introduce the ratio of one sided follow relations (to submitters) between a submitter and users who gave likes to his/her tweet. Suppose that the number of users who gave likes to tweet t is n and m of them follow the submitter of tweet t but are not followed by him/her. Then, the ratio of one sided follow relations (to submitters) between the submitter of tweet t and the users who gave likes to it, $P_{OSFtoS}(t)$, is defined as follows:

$$P_{OSFtoS}(t) = \frac{m}{n}$$

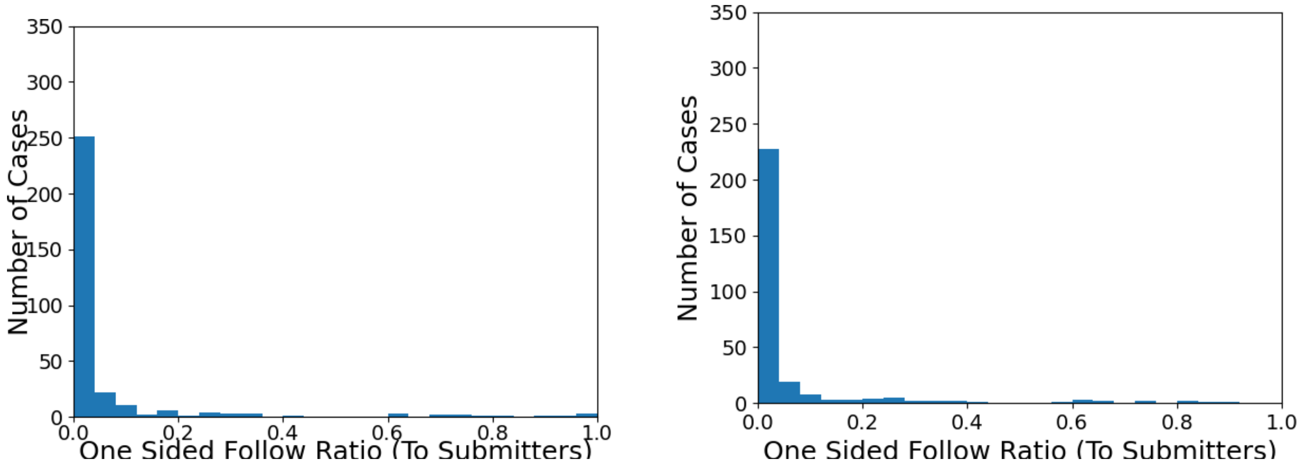
Figure 8 shows the distribution of the ratio of one sided follow relations (to submitters) between the submitters of the obtained



(a) the first day (Day 0)

(b) the last day (Day 6)

Figure 7. The histograms of the ratio of one sided follow relations (from submitters) between the submitters of the obtained 318 tweets and the users who gave likes to them on the first day (Day 0) and the last day (Day 6) of the investigation period.



(a) the first day (Day 0)

(b) the last day (Day 6)

Figure 8. The histograms of the ratio of one sided follow relations (to submitters) between the submitters of the obtained 318 tweets and the users who gave likes to them on the first day (Day 0) and the last day (Day 6) of the investigation period.

318 tweets and the users who gave likes to them. In most cases, the ratio was less than 0.2. However, we found 14 cases where the ratio was more than 0.6. In one case of them, we found that 20 users gave likes to a single tweet promising to disclose submitter's personal information and all of them had one sided follow relations (to submitters) with the submitter. Figure 8 shows that the number of users who had one sided follow relations (to submitters) with submitters did not decrease. It is probable that submitters were careful to follow unfamiliar users even if the users followed them and gave likes to their tweets.

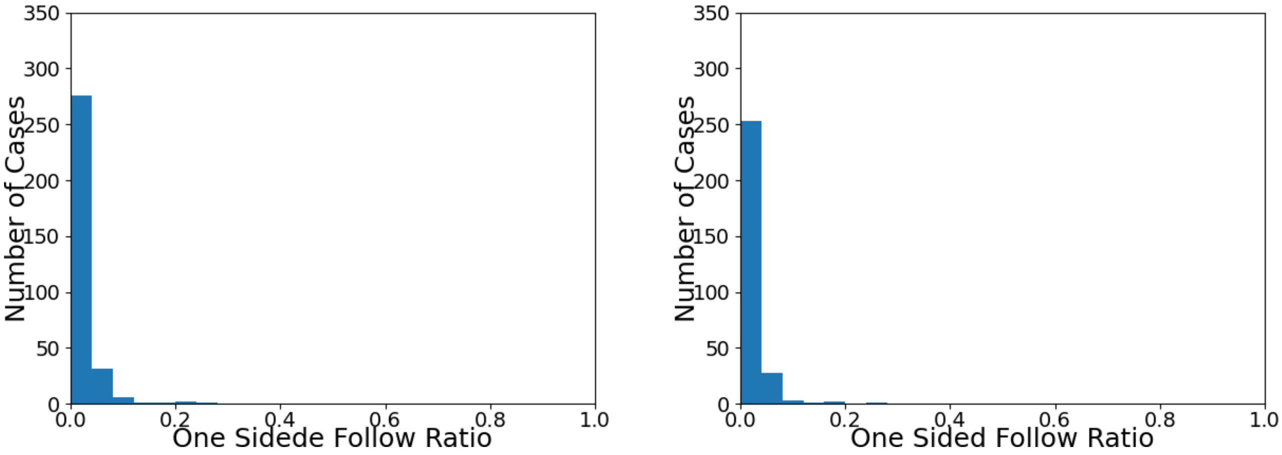
B. One Sided Follow relations among users who gave likes to submitters' tweets

We discuss the one sided follow relations among users who gave likes to tweets disclosing submitters' personal information. In order to discuss this problem, we introduce the ratio of one sided follow relations among users who gave likes to a tweet. Suppose that the number of users who gave likes to tweet t is n and there are m cases where one user of them follows another user but is not followed by the user. Then, the ratio of one sided follow relations among the users who gave likes to tweet t, $P_{OSFamongU}(t)$, is defined as follows:

$$P_{OSFamongU}(t) = \frac{m}{n(n-1)/2}$$

Figure 9 shows the distribution of the ratio of one sided follow relations among the users who gave likes to the obtained 318 tweets. In most cases, the ratio was less than 0.1. Figure 9 shows that the number of users who had one sided follow relations with other users did not decrease. It is probable that users were careful to follow unfamiliar users even if the users followed them and gave likes to the same tweets.

ISBN: ISBNFILL



(a) the first day (Day 0)

(b) the last day (Day 6)

Figure 9. The histograms of the ratio of one sided follow relations among the users who gave likes to the obtained 318 tweets on the first day (Day 0) and the last day (Day 6) of the investigation period.

V. Conclusion

In this paper, we investigated the one sided relations between submitters and users who gave likes to submitters' tweets promising to disclose their personal information. The results of our investigation show that giving likes to tweets promising to disclose submitter's personal information is not a sufficient trigger to get to follow users. Submitters were careful to follow unfamiliar users even if the users followed them and gave likes to their tweets. Also, users were careful to follow unfamiliar users even if the users followed them and gave likes to the same tweets.

References

[1] Y. Watanabe, T. Nakano, H. Nishimura, and Y. Okada, "An Investigation of Twitter Users Who Gave Likes to Tweets Disclosing Submitters' Personal Information," in Proceedings of the Eighth International Conference on Human and Social Analytics (HUSO 2022), May 2022, pp. 10–15. [Online]. Available: https://www.thinkmind.org/index.php?view=article&articleid=huso_2022_1_30_80026 [accessed: 2023-02-14]

[2] C. Johnson III, Safeguarding against and responding to the breach of personally identifiable information, Office of Management and Budget Memorandum, 2007. [Online]. Available: https://obamawhitehouse.archives.gov/sites/default/files/omb/assets/omb/memoranda/fy2007/m07-16.pdf [accessed: 2023-02-14]

[3] B. Krishnamurthy and C. E. Wills, "On the leakage of personally identifiable information via online social networks," Computer Communication Review, vol. 40, no. 1, 2010, pp. 112–117. [Online]. Available: https://doi.org/10.1145/1672308.1672328 [accessed: 2023-02-14]

[4] S. Fox et al., Trust and Privacy Online: Why Americans Want to Rewrite the Rules, The Pew Internet & American Life Project, 2000. [Online]. Available: http://www.pewinternet.org/2000/08/20/trust-and-privacy-online/ [accessed: 2023-02-14]

[5] A. Acquisti and R. Gross, "Imagined communities: Awareness, information sharing, and privacy on the facebook," in Proceedings of the 6th International Conference on Privacy Enhancing Technologies, ser. PET'06. Berlin, Heidelberg: Springer-Verlag, 2006, pp. 36–58. [Online]. Available: https://doi.org/10.1007/11957454_3 [accessed: 2023-02-14]

[6] S. B. Barnes, "A privacy paradox: Social networking in the United States." First Monday, vol. 11, no. 9, 2006. [Online]. Available: http://firstmonday.org/article/view/1394/1312 [accessed: 2023-02-14]

[7] S. Barth, M. D. de Jong, M. Junger, P. H. Hartel, and J. C. Roppelt, "Putting the privacy paradox to the test: Online privacy and security behaviors among users with technical knowledge, privacy awareness, and financial resources," Telematics and Informatics, vol. 41, 2019, pp. 55–69. [Online]. Available: https://doi.org/10.1016/j.tele.2019.03.003 [accessed: 2023-02-14]

[8] J. A. Obar and A. Oeldorf-Hirsch, "The biggest lie on the Internet: ignoring the privacy policies and terms of service policies of social networking services," Information, Communication & Society, vol. 23, no. 1, 2020, pp. 128–147. [Online]. Available: https://doi.org/10.1080/1369118X.2018.1486870 [accessed: 2023-02-14]

[9] A. Viseu, A. Clement, and J. Aspinall, "Situating privacy online: Complex perception and everyday practices," Information, Communication & Society, 2004, pp. 92–114. [Online]. Available: https://doi.org/10.1080/1369118042000208924 [accessed: 2023-02-14]

[10] S. Livingstone, "Taking risky opportunities in youthful content creation: teenagers' use of social networking sites for intimacy, privacy and self-expression." New Media & Society, vol. 10, no. 3, 2008, pp. 393–411. [Online]. Available: https://doi.org/10.1177/1461444808089415 [accessed: 2023-02-14]

[11] M. Alshaikh, M. Zohdy, R. Olawoyin, D. Debnath, Z. Gwarzo, and J. Alowibdi, "Social network analysis and mining: Privacy and security on twitter," in 10th Annual Computing and Communication Workshop and Conference (CCWC), 2020, pp. 0712–0718. [Online]. Available: https://ieeexplore.ieee.org/document/9031147 [accessed: 2023-02-14]

[12] A. N. Joinson, U.-D. Reips, T. Buchanan, and C. B. P. Schofield, "Privacy, trust, and self-disclosure online." Human-Computer Interaction, vol. 25, no. 1, 2010, pp. 1–24. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/07370020903586662 [accessed: 2023-02-14]

[13] Z. Tufekci, "Can You See Me Now? Audience and Disclosure Regulation in Online Social Network Sites," Bulletin of Science, Technology & Society, vol. 28, no. 1, 2008, pp. 20–36. [Online]. Available: https://journals.sagepub.com/doi/abs/10.1177/0270467607311484 [accessed: 2023-02-14]

[14] Y. Watanabe, H. Nishimura, Y. Chikuki, K. Nakajima, and Y. Okada, "An investigation of twitter users who disclosed their personal profile items in their tweets honestly," in Proceedings of the Sixth International Conference on Human and Social Analytics (HUSO 2020), Oct 2020, pp. 20–25. [Online]. Available: http://www.thinkmind.org/index.php?view=article&articleid=huso_2020_1_40_80035 [accessed: 2023-02-14]

[15] Twitter, Inc. Twitter API. [Online]. Available: https://developer.twitter.com/en/docs/twitter-api [accessed: 2023-02-14]

Examining Content and Emotion Bias in YouTube's Recommendation Algorithm

Obianuju Okeke, Mert Can Cakmak, Billy Spann, Nitin Agarwal
COSMOS Research Center
University of Arkansas at Little Rock (UALR)
Little Rock, USA
Email: {oiokeke, mccakmak, bxspann, nxagarwal}@ualr.edu

***Abstract*—Detection, characterization, and mitigation of bias in modern systems of automated and autonomous decisions is a growing interdisciplinary field. This study aims to explore YouTube's video recommendation bias to determine if an inherent bias has an unintended impact of occluding vulnerable communities and minority groups. Our findings suggest that the algorithm recommends videos evoking more positive emotions and higher user engagement. We also discovered that content related to our seed videos was filtered out in a systematic but gradual pendulum-like motion. This analysis of potential emergent biases will be applicable in analyzing the fairness of recommender systems, patterns of content consumption, information diffusion, echo-chamber formation, and other significant problems.**

***Index Terms*—Keywords-Recommender Systems; Recommendation Bias; YouTube; Topic Modeling; Emotion Modeling.**

I. Introduction

According to YouTube's Chief Product Officer Neal Mohan [3], around 70 percent of videos watched on YouTube are recommended videos, this means that an average of 7 out of 10 videos a user watches are recommended by YouTube. Although YouTube's goal of profit generation through increased watch-time is intended to be harmless and business-oriented, this pattern could have the unintended consequence of occluding vulnerable communities and crisis-torn societies. For our research, we studied the impact of the algorithm on videos related to the Uyghur group, a vulnerable community in the China-Uyghur crisis. According to the Council on Foreign Relations, more than a million Uyghurs - a Muslim, Turkish speaking ethnic group, have been detained since 2017 in the China Xinjiang region [15]. Platforms such as YouTube remain an indispensable outlet for such minority groups and vulnerable communities to spread awareness on important issues [21]. It also serves as a window to the world to receive vital information [19]. These groups depend on free and open platforms such as YouTube to vocalize the crisis they endure in their respective societies. According to Silverman, content evoking polarization is propagated faster than non-polarizing content [22]. We, therefore, expect content and emotions related to our seed videos to be propagated across recommendation depths.

II. Literature Review

In this section, we discuss research related to our study which includes previous works on topic shifting, emotion detection [1], and bias in recommender systems. Bias in recommendation engines has been extensively studied to understand its nature, structure, and effects, especially in the area of radicalization, polarization, and spread of misinformation [18]. These studies have described how homophilic communities are generated through recommended videos as well as factors which drive the creation of such interconnected communities, leading to filter bubble effects and echo-chambers [23]. Insights from such studies are crucial in identifying the emergence of homogeneity in recommender systems. Topic drift is a technique that has been used by many researchers in studying how content evolves. By studying content evolution, we are able to determine if content remains the same or changes relative to a standard metric. O' Hare et al. [7] analyzed sentiment-annotated corpus of textual data to determine topic drift among documents within a corpus. Liu et al. developed an LDA (Latent Dirichlet Allocation)-based method for topic drift detection in micro-blog posts [5] Topal et al. identified and quantitatively studied the effects of topic shift in social media comments [17]. Papakyriakopoulos et al., addressed hyperactive users and their effects on political discussion and recommender systems [13]. According to Papakyriakopoulos, recommendation algorithms favor the interest of hyperactive users, creating significant social influence bias and causing alterations in political opinions. By identifying inherent topics using topic modeling [4], [9], the authors classified content by topic to examine the activities of hyperactive users and determine if engagement distribution diverges. In this paper, we aim to identify inherent bias in YouTube's recommendation algorithm, and determine if the identified bias works to occlude videos related to vulnerable communities across recommendation depths. Some of the questions we hope to answer include:

- **RQ1**: How do we identify bias in content related to vulnerable communities?
- **RQ2**: What kind of videos drive recommendations on YouTube?
- **RQ3**: How do videos related to vulnerable communities change across recommendation depth?

Unlike other methodologies which have adopted a more manual approach through the use of raters in the content analysis process [16], we programmatically assign topic communities to videos across recommendation depths. Through our research, we also aim to track the evolution of content across recommendations for a detailed view on how content related to the Uyghur ethnic group is recommended on YouTube. In the next section, we present our data collection methodology.

III. Data Collection

To begin data collection, we conducted expert workshops to identify keywords related to the China-Uyghur issue. These keywords were used as search queries on YouTube's search engine to generate the 10 seed videos used in our research. Recommended videos were gathered using custom-made crawlers over "depths" of recommendations. The seed videos generated the 1st video depth, after which subsequent depths served as parent videos to generate the next sets of recommended videos. This process continued until recommended videos for 5 depths were generated. To prevent personalization in recommendations, we did not log into the account used for video collection. Also, a new browser instance was started and cookies from each previous recommendation depth were cleared to enable a fresh search of videos for the next depth crawl. This approach allowed us to generate a total of 38,970 videos across 5 depths. To focus our study, we filtered out duplicates and non-English videos which reduced our dataset to 14,332 videos, after which videos were categorized by depth. Video text data such as titles, descriptions and transcripts were used for this research.

The collection of video transcripts was divided into four sub-tasks.**Task 1:** Video transcripts were fetched using YouTube Transcript API [24]. 14,332 video ids were fed to the API and 12,611 transcripts were gathered. **Task 2:** We found that transcripts were disabled for 1,721 videos. For such videos, we used the OpenAI Whisper model [20] to extract the video transcripts. This led to an additional 1567 transcripts of which 154 videos were unavailable as they were identified as 'live shows', 'removed' or returned null in our script. **Task 3:** We identified and translated non-English transcripts to English transcripts using Google Translate API and removed transcripts which had less than 80% English content. **Task 4:** Lastly, the results were combined together and processed for analysis.

IV. Methodology

In this section, we discuss the techniques used in our study.

A. Emotion and Popularity Assessment Methodology

For this study, we analyzed emotions embedded in video text data (title, description and transcript) across 6 emotions: joy, anger, sadness, fear, surprise, love. We use emotion drift to identify emotion bias across depths of recommendations. The resulting emotion diversity in content were illustrated on a line graph with each depth representing a traversed hop of recommended videos. A fine-tuned version of transfer learning [10], T5-base-fine-tuned-emotion, was utilized for Natural Language Processing (NLP) tasks to ensure accuracy of results. To further understand the emotion drift pattern in recommended videos, we analyzed user engagement using engagement metrics of all videos such as likes and views. With the engagement metrics, we studied the change in metrics across depths, to determine if the user interaction supports the emotion drift pattern across recommendation depths.

B. Topic, Network and Content Analysis Methodology

Although previous research methodologies have concatenated video text information (video titles, video description, and video transcript) for video content analysis [18], this research analyzed these three components separately as well as in combination. By analyzing these components separately, we hoped to identify a variability in content concentration at varying levels of video text detail. The goal of topic drift detection is to investigate if recommendations stay on the topic of the Uyghur crisis as we move through recommended videos and by how much content diverges if drift is detected. To measure topic drift across depths, we computed topic similarity using Hellinger distance [11], [12], [29] and Jensen-Shannon divergence [26], [28]. Hellinger and Jensen-Shannon divergence are distance metrics used in estimating document similarity. Hellinger divergence is represented as the symmetric midpoint of Kullback–Leibler divergence [25], [30] while Jensen-Shannon divergence is a finite, smoothed version of Kullback–Leibler divergence [27]. These distance metrics calculate similarity within the range of 0 to 1, where values closer to 0 indicate a smaller distance and, therefore, larger similarity. We computed a final topic similarity score using the average of both scores across depths. Next, we analyzed the video recommendation network. Recommendation graphs for each depth consisting of video ids as nodes and recommendations as edges were generated and examined. The distribution of eigenvector centrality scores, which measure the influence a node has on a network of videos across depths was computed to determine if a sub-cluster of videos were highly influential (more recommended) compared to other videos. We then analyzed our data to determine the topic communities of videos in each recommendation depth. For this approach, we used the BertTopic model [14], a model which uses transformers and class-based term frequency-inverse document frequency (c-TF-IDF) to create dense clusters and produce interpretable topics, while maintaining important words in the topic description. By programmatically assigning each video to its respective topic community across depths. We were able to detect how influential videos evolved across recommendations.

V. Results

A. Topic Drift Analysis

As earlier discussed, the goal of topic drift detection is to determine if recommended videos stay on the topic of the China-Uyghur crisis or drift as users move through recommended videos. Distance metrics are often measured between 0 and 1, where scores closer to 0 depict high similarity (contents are similar) and scores closer to 1 depict low similarity (contents are different). For this research, drift is observed if there is an increase in the distance between depths resulting in decreased content similarity. This is seen as a rising trend in the distance metric line graph. Using video titles, video descriptions, video transcripts, and a concatenation of all text information to analyze topic drift, we compared the similarities to answer two questions;

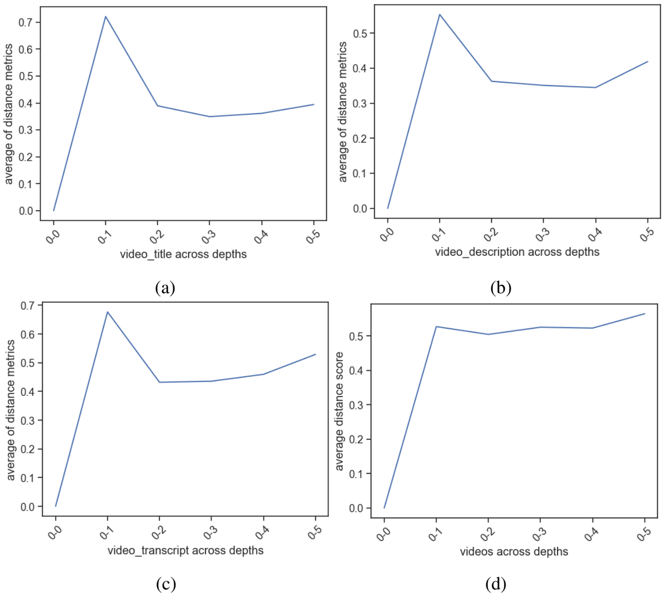
- **Are our seed videos different from videos across depths?**

This question was answered by analyzing the similarity between the seed videos and each depth of recommendation. The goal was to measure the video similarity between the seed and recommended videos in each depth of recommendation.

- **Do recommended videos become increasingly similar or different from each other?**

This question was answered by analyzing the similarity between adjacent depths of recommendation. The goal was to measure the video similarity across depths of recommended videos.

1) Similarities between the seed videos and subsequent depths of recommendation: In Fig. 1, we observe that the seed videos are significantly different from depth 1 videos. Once we approach depths 2 – 3, the videos increase in similarity to the seed videos compared to videos in depth 1, but remain significantly different in general. This result shows that, depth 1 recommendations were highly unrelated to the China-Uyghur crisis, but, as the users moves through depths 2 to 5, the videos become somewhat similar to our seed videos, but not to a relevant degree.



(a)

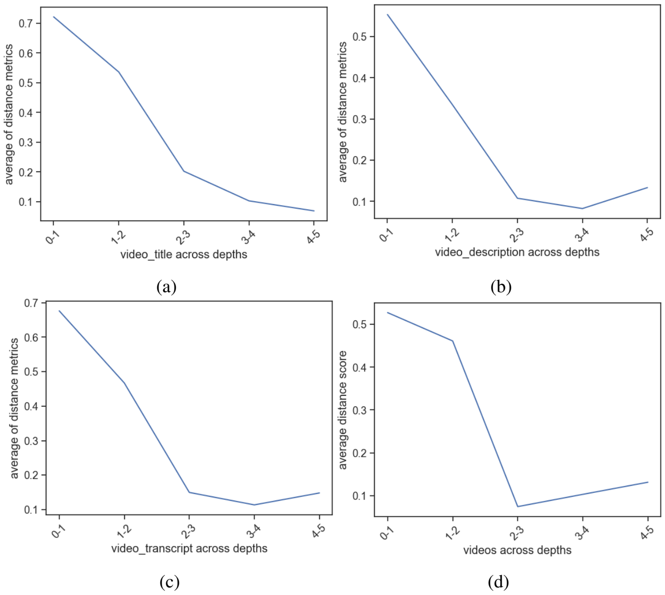
(b)

(c)

(d)

Fig. 1: Line graph showing how recommended videos become increasingly different from seed videos using (a) video titles (b) video descriptions (c) video transcript (d) all text information

2) Similarities between adjacent depths of recommendation: This question investigates how similar each depth of recommended videos is compared to its previous depth. From this analysis, we observe that as we move through recommended videos, each depth of videos becomes more similar to its previous depth, reaching maximum similarity between depths 3 and 4. Both results suggest that, although each depth of recommended videos becomes more different from our seed videos, each depth of videos also becomes more similar to its previous depth. With this pattern, the difference in recommended videos is not immediately noticed and the user is gently re-introduced to content unrelated to the seed videos.



(a)

(b)

(c)

(d)

Fig. 2: Line graph showing how recommended videos become more similar across depths using (a) video titles (b) video descriptions (c) video transcript (d) all text information

B. Network and Content Analysis

Next, network analysis was performed on each depth of recommended videos. For each depth, each video is ranked using its eigenvector centrality measure, to determine its influence in the network. For a given graph G:=(V,E) with —V— vertices, let A = (avt) be the adjacency matrix, i.e., avt = 1 if vertex v is linked to vertex t, and avt = 0 otherwise. The relative centrality score, Xv of vertex v can be defined as:

$$X_y = \frac{1}{\lambda}\Sigma_{t\epsilon M(v)} X_t = \frac{1}{\lambda}\Sigma_{t\epsilon v} a_{vt} X_t \qquad (1)$$

where M(v) is the set of neighbors of v and is a constant. With a small rearrangement, this can be rewritten in vector notation as the eigenvector equation.

$$Ax = \lambda x \qquad (2)$$

To find the most influential videos, we isolated and analyzed the top 10 videos with the highest eigenvector centrality score per depth. The mean eigenvector centrality score for the top 10 videos per depth was found and videos which had an eigenvector centrality score above the resulting mean were identified and categorized as ‘above-average’ / highly influential videos. Our results suggest that these ‘above-average’ influential videos act as attractors by driving the recommendations of videos and directing how the conversation evolves across depths. We also see that the top 10 videos in each depth fluctuate in the count of ‘above-average’ influential videos in each depth, as seen in Table I. As we move through the depths, we observe a steady increase in the number of ‘above-average’ influential videos until depth 3. Once we arrive at depth 3, the count of above-average influential videos began to steadily decrease. Also, content of these influential videos seem to drift from our seed videos as we approach depth 5. To visualize the content divergence of above-average videos from our seed videos after depth 3, we performed topic modelling on the seed videos and the whole dataset to generate the latent topics present in the recommendations and assign each video a topic community number.

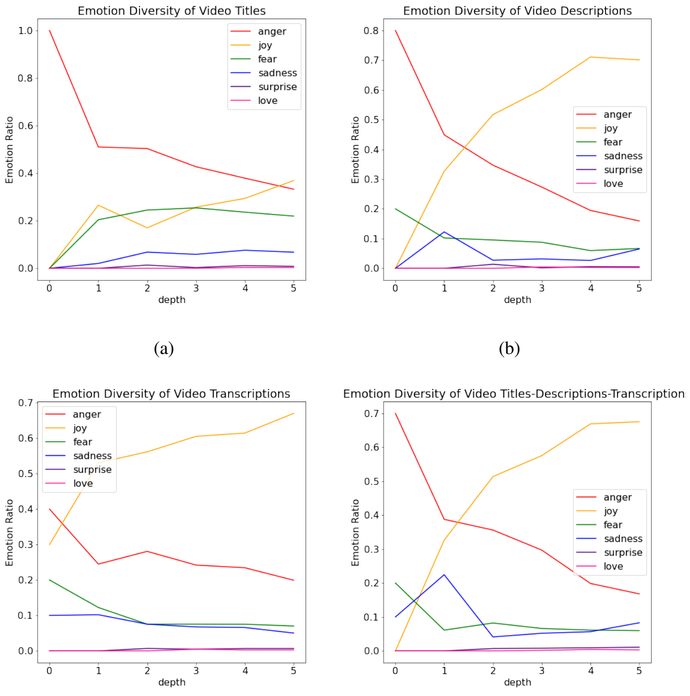
TABLE I: TOPIC COMMUNITIES OF HIGHLY-INFLUENTIAL VIDEOS PER DEPTH

Videos	Count of highly-influential videos	Topic communities
Seed	N/A	-1
Depth 1	1	1
Depth 2	4	-1, -1, -1, -1
Depth 3	4	1, 9, -1, -1
Depth 4	2	1, 9
Depth 5	3	1, 9, 13

Topic modelling was done using BERTopic, to identify the topic communities present in our seed videos and the topic communities of highly influential videos in each recommendation depth. By doing this, we were able to visualize the topical content of our ‘above-average’ influential videos and track the movement of content topically related to our seed videos as we moved across depths. We observed that all of our seed videos belonged to one topic community, -1, while the highly influential videos across depths belonged to a mix of topic communities. From Table I, we see that the highly influential video at depth 1 is introduced into the algorithm, and steers depth 1 away from the content on Uyghur crisis. Conversely, as we move to depth 2 the highly influential videos are fully turned back to topics related to our seed videos. At depth 3, the highly influential videos contain an equal mix of videos related and unrelated to our seed videos but once we arrive depth 4, our seed video content is filtered out once more from the list of highly influential videos. This result shows that, as we progress through the recommendations, videos related to our seed videos are filtered out from the recommendations in a pendulum-like motion. From Table I, we observe that the algorithm seems to swing back and forth from content related to the Uyghur crisis, reducing its influence with each motion until it is finally filtered out of the recommended videos. We are also able to see that content in depth 5 is topically unrelated to our seed videos as seen in the difference in topic communities from our seed videos in Table I.

C. Emotion and Popularity Analysis

1) Emotion Analysis: To study the pattern of emotion drift across depths, we considered video text data at 4 different levels; video titles, video description, video transcript and a combination of all texts. By doing this, we were able to apply emotion assessment and visualize emotion drift at different levels of video details, as seen in Fig. 3(a), 3(b), 3(c) and 3(d). The results show that the most dominant emotion in our seed videos was anger for all levels of video detail, as illustrated in the figures. As we traverse the recommendation depths, we see the positive emotion (joy) emerge for each depth in all emotion graphs and the negative emotions (anger, fear, and sadness) decrease significantly.



(a)

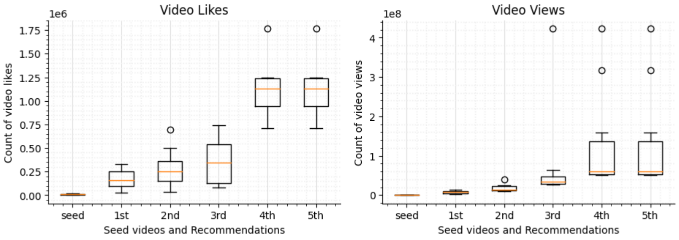
(b)

(c)

(d)

Fig. 3: Emotion assessment for video text data (a) titles only (b) descriptions only (c) transcripts only (d) all text information

2) Popularity Analysis: By analyzing the emotions of the video across depths using video text data, we discovered that there was a significant decrease in negative emotion (anger) and a significant increase in positive emotion (joy). To investigate the significance of this emotion drift pattern, we analyzed user interaction with the videos using engagement metrics across depths. This analysis was to determine if more popular videos were recommended across depths. For this experiment, a popular video is described as a video which has significantly high views and high positive engagement in the form of likes. As a result, the engagement metrics we considered were the views and likes of each video. On inspecting our seed videos, we found they all had a very high view count but a significantly low like count, suggesting that although our seed videos were widely watched, they did not elicit positive interaction from the audience. This is to be expected as the China-Uyghur crisis has been monitored internationally with the discourse being widely criticized. From the video like box-plot in Fig. 4(a), we see that, as we move through recommendation depths, the median likes of recommended videos are significantly higher compared to the seed videos and increase linearly until we hit depth 3, after which, there is an exponential increase in video likes by depth 4 and depth 5. Secondly, our video views box-plot in Fig. 4(b) shows that the views of recommended videos are higher compared to the seed videos but unlike video likes, we see a steady growth in view count across depths of recommended videos. The result of our popularity analysis shows that more popular videos are present in recommended videos, which further explains the high occurrence of positive emotions in higher depths of recommendations.



(a)

(b)

Fig. 4: The box-plot show the increasing median count of (a) video likes and (b) video views from seed videos to recommended videos.

VI. Discussion

RQ1: How do we identify bias on content related to vulnerable communities?

In examining the results from our emotion and popularity analysis, we observed that the anger emotion significantly decreases across depths, while there is a proportional increase in the joy emotion in videos after each recursive depth of recommendation. In addition, we see that the engagement metrics (views and likes) increase significantly as we move to higher depths of recommendation, suggesting increased user engagement with recommended videos. In summary, the algorithm seems to recommend more popular videos with positive emotions (joy) in an attempt to keep users engaged for longer periods of time. This pattern demonstrates recommender bias which steers users away from unpopular videos with negative emotions. This trend poses the risk of occluding content related to the China-Uyghur crisis.

RQ2: How do videos related to vulnerable communities change across recommendation depth?

Our topic drift analysis shows that as users watch recommended videos, the videos become increasingly different from our seed videos across recommendations. We also found that each depth of recommended videos became increasingly similar to its immediate previous depth suggesting that videos across recommendations are similar in content. These drift patterns show that the algorithm gently drifts from our seed videos by recommending videos that are increasingly different from our seed videos but similar to adjacent depths of recommendations until recommended videos significantly drift from content related to the China-Uyghur crisis at depth 5.

RQ3: What kind of videos drive recommendations in the context of this study?

Through our network analysis, we observe that each depth has a set of highly influential videos which act as attractors to drive video recommendations. The gradual shift in topics we observe from seed videos to depth 5 in Fig. 1 seems to be due to a *pendulum-like* motion of the algorithm. From Table I, our results show that depths 3, 4 and 5 show a back-and-forth swing of the algorithm. There is an alternate filtering and re-introduction of content related to our seed videos across depths, maintaining a steady plateau in similarity of depths 3 - 5 to our seed videos until the China-Uyghur crisis topics are filtered out of the recommendations.

VII. Conclusion and Future Works

For this research, we employed the use of drift analysis to identify bias across recommended videos. Our results showed that YouTube's recommendation system tends to lessen negative emotions such as anger and amplify positive emotions such as joy across recommended videos on the platform. We also see that highly influential videos at each depth act as attractors to gently draw recommendations away from content related to our seed videos in a pendulum-like motion. In future research, we plan to expand this research into exploring an alternate narrative which elicits a different emotion (e.g joy) and comparing the findings with those of our current research. We are also interested in developing a framework which serves to methodologically compare content across various discourse and exploring the effects of the YouTube algorithms on such datasets.

Acknowledgment

This research is funded in part by the U.S. National Science Foundation (OIA-1946391, OIA-1920920, IIS 1636933, ACI-1429160, and IIS-1110868), U.S. Office of the Under Secretary of Defense for Research and Engineering (FA9550-22-1-0332), U.S. Office of Naval Research (N00014-10-1-0091, N00014-14-1-0489, N00014-15-P-1187, N00014-16-1-2016, N00014-16-1-2412, N00014-17-1-2675, N00014-17-1-2605, N68335-19-C-0359, N00014-19-1-2336, N68335-20-C-0540, N00014-21-1-2121, N00014-21-1-2765, N00014-22-1-2318), U.S. Air Force Research Laboratory, U.S. Army Research Office (W911NF-20-1-0262, W911NF-16-1-0189, W911NF-23-1-0011), U.S. Defense Advanced Research Projects Agency (W31P4Q 17-C-0059), Arkansas Research Alliance, the Jerry L. Maulden/Entergy Endowment at the University of Arkansas at Little Rock, and the Australian Department of Defense Strategic Policy Grants Program (SPGP) (award number: 2020-106-094). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding organizations. The researchers gratefully acknowledge the support.

References

[1] S. N. Shivhare and S. Khethawat, "Emotion Detection from Text." arXiv, May 22, 2012. doi: 10.48550/arXiv.1205.4944.

[2] J. M. Garcia-Garcia, V. M. R. Penichet, and M. D. Lozano, “Emotion detection: a technology review,” in Proceedings of the XVIII International Conference on Human Computer Interaction, New York, NY, USA, Sep. 2017, pp. 1–8. doi: 10.1145/3123818.3123852.

[3] J. E. Solsman, “Ever get caught in an unexpected hourlong YouTube binge? Thank YouTube AI for that,” CNET, Jan. 10, 2018. https://www.cnet.com/tech/services-and-software/youtube-ces-2018-neal-mohan/ [Accessed Jan. 28, 2023].

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” J. Mach. Learn. Res., vol. 3, no. null, pp. 993–1022, Mar. 2003.

[5] Q. Liu, H. Huang, and C. Feng, “Micro-blog Post Topic Drift Detection Based on LDA Model,” in Behavior and Social Computing, Cham, 2013, pp. 106–118

[6] F. A. Acheampong, C. Wenyu, and H. Nunoo-Mensah, “Text-based emotion detection: Advances, challenges, and opportunities,” Engineering Reports, vol. 2, no. 7, p. e12189, 2020, doi: 10.1002/eng2.12189.

[7] N. O’Hare et al., “Topic-dependent sentiment analysis of financial blogs,” in Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion, New York, NY, USA, Nov. 2009, pp. 9–16. doi: 10.1145/1651461.1651464.

[8] M. Suhasini and S. Badugu, “Two Step Approach for Emotion Detection on Twitter Data,” International Journal of Computer Applications, vol. 179, pp. 12–19, Jun. 2018, doi: 10.5120/ijca2018917350.

[9] A. Murakami, P. Thompson, S. Hunston, and D. Vajn, “‘What is this corpus about?’: using topic modelling to explore a specialised corpus,” Corpora, vol. 12, no. 2, pp. 243–277, Aug. 2017, doi: 10.3366/cor.2017.0118.

[10] C. Raffel et al., “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer,” Journal of Machine Learning Research, vol. 21, no. 140, pp. 1–67, 2020.

[11] A. Shemyakin, “Hellinger Distance and Non-informative Priors,” Bayesian Analysis, vol. 9, no. 4, pp. 923–938, Dec. 2014, doi: 10.1214/14-BA881.

[12] G.-H. Fu, Y.-J. Wu, M.-J. Zong, and J. Pan, “Hellinger distance-based stable sparse feature selection for high-dimensional class-imbalanced data,” BMC Bioinformatics, vol. 21, no. 1, p. 121, Mar. 2020, doi: 10.1186/s12859-020-3411-3.

[13] O. Papakyriakopoulos, J. C. M. Serrano, and S. Hegelich, “Political communication on social media: A tale of hyperactive users and bias in recommender systems,” Online Social Networks and Media, vol. 15, p. 100058, Jan. 2020, doi: 10.1016/j.osnem.2019.100058.

[14] M. Grootendorst, “BERTopic,” Mar. 11, 2022. https://maartengr.github.io/BERTopic/ [Accessed Jan. 09, 2023].

[15] “China’s Repression of Uyghurs in Xinjiang — Council on Foreign Relations.” https://www.cfr.org/backgrounder/china-xinjiang-uyghurs-muslims-repression-genocide-human-rights [Accessed Jan. 09, 2023].

[16] H. Heuer, H. Hoch, A. Breiter, and Y. Theocharis, “Auditing the Biases Enacted by YouTube for Political Topics in Germany,” in Proceedings of Mensch und Computer 2021, New York, NY, USA, Sep. 2021, pp. 456–468. doi: 10.1145/3473856.3473864.

[17] K. Topal, M. Koyuturk, and G. Ozsoyoglu, “Emotion -and area-driven topic shift analysis in social media discussions,” in 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Aug. 2016, pp. 510–518. doi: 10.1109/ASONAM.2016.7752283.

[18] M. Faddoul, G. Chaslot, and H. Farid, “A Longitudinal Analysis of YouTube’s Promotion of Conspiracy Videos.” arXiv, Mar. 06, 2020. doi: 10.48550/arXiv.2003.03318.

[19] L. P. Goldsmith et al., “The use of social media platforms by migrant and ethnic minority populations during the COVID-19 pandemic: a systematic review.” medRxiv, p. 2022.02.07.22270579, Feb. 07, 2022. doi: 10.1101/2022.02.07.22270579.

[20] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust Speech Recognition via Large-Scale Weak Supervision.” arXiv, Dec. 06, 2022. doi: 10.48550/arXiv.2212.04356.

[21] B. Auxier, “Social media continue to be important political outlets for Black Americans,” Pew Research Center. https://www.pewresearch.org/fact-tank/2020/12/11/social-media-continue-to-be-important-political-outlets-for-black-americans/ [Accessed Jan. 09, 2023].

[22] C. Silverman, “This Analysis Shows How Viral Fake Election News Stories Outperformed Real News On Facebook,” BuzzFeed News. https://www.buzzfeednews.com/article/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook [Accessed Jan. 09, 2023].

[23] B. Kitchens, S. L. Johnson, and P. Gray, “Understanding Echo Chambers and Filter Bubbles: The Impact of Social Media on Diversification and Partisan Shifts in News Consumption,” Aug. 26, 2020. https://misq.umn.edu/understanding-echo-chambers-and-filter-bubbles-the-impact-of-social-media-on-diversification-and-partisan-shifts-in-news-consumption.html [Accessed Jan. 09, 2023].

[24] J. Depoix, “YouTube Transcript/Subtitle API (including automatically generated subtitles and subtitle translations).” Jan. 08, 2023. [Accessed: Jan. 09, 2023]. [Online]. Available: https://github.com/jdepoix/youtube-transcript-api

[25] H. Sengar, H. Wang, D. Wijesekera, and S. Jajodia, “Detecting VoIP Floods Using the Hellinger Distance,” IEEE Transactions on Parallel and Distributed Systems, vol. 19, no. 6, pp. 794–805, 2008, doi: 10.1109/TPDS.2007.70786.

[26] M. Jamaati and A. Mehri, “Text mining by Tsallis entropy,” Physica A: Statistical Mechanics and its Applications, vol. 490, pp. 1368–1376, Jan. 2018, doi: 10.1016/j.physa.2017.09.020.

[27] A. Mehri, M. Jamaati, and H. Mehri, “Word ranking in a single document by Jensen–Shannon divergence,” Physics Letters A, vol. 379, no. 28, pp. 1627–1632, Aug. 2015, doi: 10.1016/j.physleta.2015.04.030.

[28] D. M. Endres and J. E. Schindelin, “A new metric for probability distributions,” IEEE Transactions on Information Theory, vol. 49, no. 7, pp. 1858–1860, 2003, doi: 10.1109/TIT.2003.813506.

[29] S. Sohangir and D. Wang, “Improved sqrt-cosine similarity measurement,” Journal of Big Data, vol. 4, no. 1, p. 25, Jul. 2017, doi: 10.1186/s40537-017-0083-6.

[30] S. Zhu, L. Liu, and Y. Wang, “Information retrieval using Hellinger distance and sqrt-cos similarity,” in 2012 7th International Conference on Computer Science and Education (ICCSE), 2012, pp. 925–929. doi: 10.1109/ICCSE.2012.6295217.

Comparing Toxicity Across Social Media Platforms for COVID-19 Discourse

Nahiyan Bin Noor, Niloofar Yousefi, Billy Spann, Nitin Agarwal
COSMOS Research Center
University of Arkansas at Little Rock
Little Rock, AR 72204, USA
e-mail: nbnoor@ualr.edu, nyousefi@ualr.edu, bxspann@ualr.edu, nxagarwal@ualr.edu

***Abstract*— The emergence of toxic information on social networking sites, such as Twitter, Parler, and Reddit, has become a growing concern. Consequently, this study aims to assess the level of toxicity in COVID-19 discussions on Twitter, Parler, and Reddit. Using data analysis from January 1 through December 31, 2020, we examine the development of toxicity over time and compare the findings across the three platforms. The results indicate that Parler had lower toxicity levels than both Twitter and Reddit in discussions related to COVID-19. In contrast, Reddit showed the highest levels of toxicity, largely due to various anti-vaccine forums that spread misinformation about COVID-19 vaccines. Notably, our analysis of COVID-19 vaccination conversations on Twitter also revealed a significant presence of conspiracy theories among individuals with highly toxic attitudes. Our computational approach provides decision-makers with useful information about reducing the spread of toxicity within online communities. The study's findings highlight the importance of taking action to encourage more uplifting and productive online discourse across all platforms.**

Keywords-Toxicity analysis; social network analysis; COVID-19; Parler; Twitter; Reddit.

I. INTRODUCTION

The most widely used social media platforms, such as Facebook, Twitter, and YouTube, have established community guidelines and enforcement mechanisms to regulate harmful content and misinformation, but free-speech platforms like Parler have been more accommodating towards hate speech, conspiracy theories, and potentially harmful misinformation. Reddit is another social media platform that is basically discussion based; it is a free-speech platform like Parler. However, after the increase of misinformation and hate speech, the policymakers imposed several guidelines and banned some subreddits that spread misinformation, toxicity, and hate speech. Parler is a micro-blogging platform comparable to Twitter that, by design, lacks the content moderation rules and capabilities of the platform it emulates. Parler was created before the emergence of COVID-19, but it has since become an important vector for online misinformation, a place where users can spread COVID-19 misinformation without restrictions. Even though there are multiple guidelines and regulations on Twitter and Reddit to stop people from posting a toxic posts, hate speech, or misinformation, it is not possible to remove toxicity from these platforms.

Managing social media platforms' security is difficult, but examining this harmful content can assist in solving the problem. Our study adds to the current body of knowledge on social media safety.

This paper considers misinformation a claim that contradicts or distorts the common understanding of verifiable facts [1]. Formerly obscure, in 2020, Parler enjoyed a surge in popularity following a push by conservative pundits and politicians to move away from larger, more mainstream social media platforms due to the perception of bias and censorship against conservative viewpoints on those platforms. In 2020, as the COVID-19 pandemic spread worldwide, users of Twitter and the primarily far-right user base of Parler engaged in discussions. They posted content about the vaccination efforts to stop the spread of COVID-19. This work is a comparative analysis of the toxicity of COVID-19-related content on Twitter, Parler, and Reddit from January 1, 2020, through December 31, 2020. Within our text corpus of users' posts, we compared the evolution of the toxicity level over the time frame of analysis. We presented evidence that Reddit contained a higher level of toxicity regarding the COVID-19 discourse than did Twitter and Parler over the four COVID-19-related content datasets we analyzed. From Reddit, among the four COVID-19-related content, the vaccination-related contents are more toxic than any other topic, which makes Reddit the most toxic platform.

This work answers four research questions:

1) How do Twitter, Parler, and Reddit differ about the existence of toxicity within user-generated text content?

2) Of the three platforms, which one contains the highest overall level of toxicity?

3) How did the average toxicity level change over time within Twitter, Parler, and Reddit datasets?

4) Which COVID-19-related topic is the most toxic in which social media?

The remainder of this paper is organized as follows. In Section 2, the related work that has been published regarding toxicity on social media is presented. Section 3 describes the data collection process and the methodology used in this paper. Section 4 presents the highlights from our results and analysis. Finally, Section 5 concludes with the contributions of this work and presents our plans and ideas for future work.

II. RELATED WORKS

This section will briefly overview toxicity and its spread in social media. Currently, a massive volume of content in social media demands tools and methods to detect toxicity. It will help to prevent the spread of toxicity in social media. Some researchers focused on this domain, and some studies developed a new method for this aim.

Sahana et al. [2] proposed a binary classification for detecting toxic contents; the authors classify toxic comments from non-toxic comments regardless of the nature of the toxicity. A similar approach has been made by Taleb et al. [3] in their research studied of different approaches to detect toxic comments on social media. For this purpose, the authors perform a binary classification to indicate whether a comment is toxic. On the other hand, Kumar et al. [4] suggested classifying toxic comments into various categories; for this assignment, he performs multiple machine learning approaches such as Logistic Regression, K Nearest Neighbors., Bernoulli Naïve Bayes (NB), multinomial NB, Support Vector Machine (SVM), and Random Forest. Comparing these algorithms helps us identify which method performs better in detecting multiple toxicities. Watanabe et al., [5] detected toxicity and hate speech on Twitter, proposed an ML technique using sentiment and semantic-based features.

Gröndahl et al. [6] claimed that current hate-speech detection models are inaccurate for the contents that are changed with simple techniques. Cheng et al. [7] investigated by using text quality metrics if it is possible to identify antisocial users in their post history in online forums. A multi-label classifier trained by Gunasekara and Nejadgholi [8] for detecting toxicity in online conversational text, their result indicated that character-level text representation methods perform better than word-level representations. Hanu [9] developed a trained model to predict toxic contents named Detoxify. Detoxify method provides a toxicity score for each content to indicate whether the content is toxic or not and scores for different toxicity categories such as threats, obscenity, insults, and identity hate.

Prior works of some researchers indicate that they use different methods in various social media platforms to detect toxicity. DiCicco et al. [10] compared the toxicity between Parler and Twitter and analyzed the highly toxic users and their networks on these two platforms. Obadimu et al. [11] used an NMF method to predict commenter toxicity on YouTube. They claimed that the performance of the NMF model is more accurate than other models. Obadimu et al. [12], in their other study, focused on evaluating various forms of toxicity. They investigated their assumption on the YouTube comments posted on pro- and anti-NATO channels. In a similar study, Pascual-Ferrá et al. [13] evaluated the toxicity of Pro-mask and Anti-mask related to COVID-19 on Twitter. The finding indicated that Anti-mask hashtags are more toxic than Pro-mask.

Wallace Chipidza [14] discovered a network of content posted on 30 politically biased and two neutral subcommunities on Reddit. Related to COVID-19, his finding from graph modeling indicates that most highly toxic contents are likely to be in political subreddits. Rafal Urbaniak et al. [15] used algorithmic detection and Bayesian statistical methods, analyzed Reddit's contents to find the correlation between username toxicity and different types of that. On the other hand, Yun Yu Chong and Haewoon Kwak [16] discussed detecting toxicity triggers in an Asian online community and how they can differ from Western online communities. Hind Almerekhi et al. [17], in their study, investigated the detection of toxic contents and the source of the toxicity in the discussion on Reddit. For this aim, they propose an approach for toxic comment and toxicity trigger detection.

III. DATA COLLECTION

The data from Twitter, Parler, and Reddit analyzed in this work consisting of a corpus of user posts collected based on a list of seed hashtags related to COVID-19 from January 1, 2020, through December 31, 2020 (Table 1).

A total of twelve datasets were created, four for each platform with mirroring hashtags and keywords. An open dataset from the Parler social network was created by Aliapoulios et al. [18], a complete dataset of all Parler data from August 2018 to when Parler was shut down in January 2021. The data for this paper was filtered by the seed list of keywords (Table 1). The Twitter data was collected using the Twitter Developer API [19] for the hashtags in (Table 1) posthoc. Because of this, tweets and accounts removed from Twitter for being labeled misinformation were not collected. Finally, Reddit posts and comments were collected using Pushshift API [20][21]. The customized python code was developed to collect data containing specific keywords during a specific period using the PSAW library [22]. Reddit data were collected from the whole of Reddit. A total of 72,327 posts and comments were collected from the 7511 subreddit. According to Twitter, Parler, and Reddit data-sharing guidelines, data collected in the study will be made available upon request.

TABLE 1. KEYWORDS USED FOR DATA COLLECTION WITH MEAN AND STANDARD DEVIATION OF TOXICITY SCORE.

Categories	Social Media	Records	Mean Toxicity	SD
Covid	***Twitter***	28,131	0.234	0.388
	Parler	16361	0.294	0.402
	Reddit	24501	0.1959	0.314
Lockdown	***Twitter***	1472	0.326	0.406
	Parler	5965	0.176	0.361
	Reddit	4781	0.216	0.32
Mask	***Twitter***	2423	0.313	0.416
	Parler	26165	0.264	0.388
	Reddit	16086	0.23	0.348
Vaccine	***Twitter***	610	0.302	0.411
	Parler	5928	0.119	0.304
	Reddit	26959	0.81	0.25

IV. METHODOLOGY

Before executing the toxicity analysis, the seed keywords and hashtags from each record in the datasets were removed so their presence would not influence the calculated toxicity scores for the overall target corpus. After the toxicity analysis, non-English posts for Parler, Posts and comments for Reddit, tweets, and retweets for Twitter were removed as Detoxify Unified was only trying to support the English language. Because of this, the results in other languages could have been more accurate. There were some missing, deleted, removed, and duplicate posts and comments on Reddit. There were some duplicate posts and comments that contained multiple keywords that were being searched. So, every duplicate value was removed to ensure that all datasets contained the unique value. When the analysis was completed, we computed toxicity scores for each Parler post, Twitter tweet, and Reddit post and comment in the dataset using Detoxify. Detoxify, a model created by Unitary AI (https://github.com/unitaryai/detoxify), uses a Convolutional Neural Network. It is trained with word vector inputs to determine whether the text could be perceived as toxic to a discussion. Given a text input, the Detoxify API returns a probability score between 0 and 1, with higher values indicating a greater likelihood of the toxicity label being applied to the text. Since toxicity scores are based on a probability score of 0 to 1, toxicity scores of 0.5 or greater indicate a piece of text labeled as toxic. Detoxify returns seven categories of toxicity scores in terms of level and type 1) toxicity, which is the overall level of toxicity for a piece of text 2) severe toxicity 3) obscene 4) threat 5) insult 6) identity attack and 7) sexually explicit. Detoxify is used since it is an open-source comment detection python library that identifies harmful and inappropriate texts online. This multilingual model has been trained in English, French, Italian, Spanish, Russian, Turkish, and Portuguese. Even though it can predict toxicity by giving a score, it is not efficient, while some words related to swearing, insults, or profanity are present in the text. They may predict a non-toxic text as toxic if there are certain words. For comparison, we also explored using Google's Perspective API, a related model with similar outputs used for determining toxicity. Previous datasets for other research were analyzed using both tools to compute the toxicity scores, finding similar values for toxicity scores across the same dataset.

V. Analysis and Results

In this section, we present our analysis and results. First, we discuss the overall posting frequency of our seed hashtags (and keywords) and the results of our toxicity analysis for each platform, Twitter, Parler, and Reddit.

For the Twitter dataset, the seed hashtags used in this analysis first appeared in March 2020. Of all the Twitter datasets, COVID had the most posts from March through December 2020. There was a peak in mid-April and near the end of June, and then a significant rise in the number of tweets in mid-November.

For Parler, the seed keywords (to mirror the Twitter target hashtags) registered posting activity near the end of May. Interestingly, all Parler datasets simultaneously registered a huge spike that peaked and then fell in posting frequency during November. This is a curious result that may indicate inorganic behavior at first glance. Further inspection of the dataset revealed that Parler users often adopted the behavior of using all four seed hashtags within a single post, which was not the behavior of Twitter users.

For Reddit, the number of posts and comments started to show up in the early weeks of 2020, which is earlier than Twitter and Parler. This is because some subreddit named ‘r/worldnews' and 'r/China_Flu’ have started discussions about COVID-19 since it first spread in China in Late December March. Each keyword-related post peaked from late November to early December.

Although each keyword or hashtag containing posts, comments, and tweets follow almost the same weekly trend throughout the year, three different platforms have different trends for different keywords. For instance, Twitter datasets had more tweets related to the f*ckcovid hashtag, whereas Parler had more posts containing the keyword f*ckmask. On the other hand, if we consider Reddit posts and comments, the f*ckvaccine keyword containing posts and comments was in the lead.

Thus, Twitter is more toxic based on COVID-related tweets, and Parler is more toxic for mask-related posts. Finally, Reddits' toxicity is mostly based on vaccine-related posts and comments. Multiple subreddit like 'r/Nonewnormal’ got banned due to spreading misinformation about vaccination during that time. We have collected posts and comments from that subreddit if they contain those four keywords related to COVID-19. Even though some subreddit got banned due to violation of community guidelines on Reddit. The posts and comments are collected using Pushshift API and analyzed toxicity on those posts.

As mentioned above, before executing toxicity analysis, these seed hashtags (mirroring keywords) were removed from each data record to avoid influencing the calculated toxicity scores for the overall target corpus. Upon completing our toxicity analysis methodology, we discovered that Twitter, Parler and Reddit differed in the existence of toxicity within their respective user-generated text content (toxicity scores > 0.5) from January 1, 2020, through December 31, 2020. When breaking down the content containing toxicity on each platform, Reddit contained a higher overall percentage, around 37% for all datasets, compared to Twitter, with just above 30%, and Parler, with 21.83%. (Table 2).

Although Reddit has the highest percentage of toxic posts (Toxicity score > 0.5), Twitter has the highest number of toxic posts containing the f*cklockdown hashtag, with 34.31% of tweets. In addition, Parler has 30.51% of the toxic post containing the keyword f*ckcovid.

However, surprisingly Reddit has 86% of toxic posts and comments containing the keyword f*ckvacccine, which is the highest among all platforms and all other hashtags and keywords. This made Reddit more toxic than the other two platforms. It is because Reddit has some forums that talk most about anti-vaccine.

TABLE 2. NUMBER AND PERCENTAGE OF TOXIC POSTS ON TWITTER, PARLER AND REDDIT FOR ALL TWELVE DATASET.

		Total	Percentage of Post with Toxicity Score > 0.5			Percentage of Posts with Toxicity score > 0.7			Percentage of Posts with Toxicity Score > 0.9		
Dataset	**Platform**	**Tweets/ Posts**	**Toxicity**	**Obscene**	**Insult**	**Toxicity**	**Obscene**	**Insult**	**Toxicity**	**Obscene**	**Insult**
#f*ckcovid	Twitter	28131	24.08%	22.05%	10.93%	21.65%	19.89%	8.60%	17.00%	10.06%	6.93%
#f*cklockdown		1472	34.31%	28.60%	16.37%	27.45%	22.96%	11.35%	20.11%	14.54%	6.05%
#f*ckmask		2423	31.24%	23.15%	19.81%	27.90%	20.59%	16.05%	22.86%	15.44%	5.94%
#f*ckvaccine		610	30.98%	23.28%	19.51%	27.21%	19.84%	14.75%	20.98%	14.26%	7.21%
#f*ckcovid	Parler	16361	30.51%	20.61%	18.01%	29.08%	19.16%	15.82%	15.48%	10.43%	7.03%
#f*cklockdown		5956	18.11%	13.06%	12.14%	17.45%	12.98%	11.90%	13.73%	8.14%	8.04%
#f*ckmask		26165	26.80%	15.71%	17.12%	23.38%	13.28%	13.48%	14.64%	9.96%	7.18%
#f*ckvaccine		5928	11.93%	9.06%	5.36%	10.90%	8.92%	4.82%	10.37%	8.52%	4.28%
#f*ckcovid	Reddit	24501	18.41%	13.14%	6.53%	13.35%	8.63%	4.11%	7%	3.64%	2.11%
#f*cklockdown		4781	20.08%	13.77%	6.60%	13.99%	9.10%	3.88%	7.12%	3.34%	1.86%
#f*ckmask		16086	23.37%	16%	10.64%	18.17%	11.29%	7.60%	10.52%	4.97%	4.28%
#f*ckvaccine		26959	86%	81.50%	39.99%	77.67%	70.33%	29.94%	57.27%	44.42%	18.37%

There was a huge community that discussed the covid vaccine. These subreddits are responsible for spreading misinformation related to the Covid vaccine. Though this subreddit eventually got banned, we collected posts from those banned subreddit. Most of the Twitter content had a higher probability of being labeled as toxic than the Parler and Reddit content, except the f*ckvaccine keyword for Reddit. Surprisingly, for the overall toxicity category, the Twitter content for all datasets had a higher percentage of content with toxicity scores greater than 0.7 and greater than 0.9 than did the Parler content and Reddit content. Again, Parler only exceeded Twitter in the percentage of harmful content for the COVID dataset. In contrast, Reddit exceeded the other two platforms in the percentage of harmful content for the vaccine dataset.

This is an interesting result because we expected to see more harmful content on Parler due to the free-speech nature of the platform and how they tout their lack of censorship as a selling point for users. We also expected to see the highest toxicity on Reddit for the vaccine dataset. We also looked at the obscene and insult toxicity categories for each tweet and post for all twelve datasets. Of the seven categories of toxicity scores obtained from Detoxify, only three contained enough data to warrant inclusion in the discussion: toxicity (overall), obscene, and insult. More Twitter content fell into the obscene category than did Parler and Reddit content for all datasets except the vaccine dataset from Reddit, with the highest percentage being within the Lockdown dataset (28.6% for Twitter, 13.06% for Parler and 13.77% for Reddit) and vaccine dataset (23.28% for Twitter 9.06% for Parler and 81.50% for Reddit). However, more Parler content fell into the insult category than Twitter content and Reddit content for the COVID dataset (18.01% vs. 10.93% vs. 6.53%).

The percentage of harmful content (overall toxicity category) within the vaccine datasets varied considerably between platforms (30.98% for Twitter versus 11.93% for Parler versus 86% for Reddit). So, overall, the toxicity analysis revealed that Twitter was more toxic than Parler and Reddit in all, but one case, the COVID dataset and Reddit were more toxic than Parler and Twitter for the vaccine dataset. The toxic content was more obscene and insult type for both platforms. However, the harmful content on Twitter was obscener than that of Parler, especially within the Lockdown dataset. The toxic content on Parler was more of an insulting type within the COVID dataset. Finally, the vaccine dataset on Reddit contained the highest toxic, obscene, and insulting posts than the other two platforms.

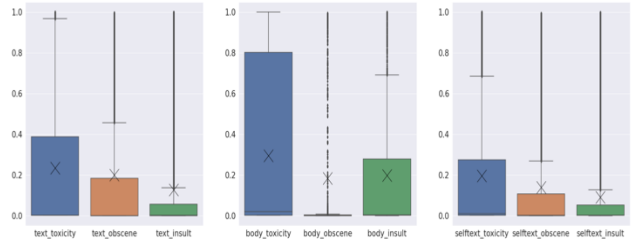


Figure 1. f*ckcovid hashtag for three classes (Toxicity, Obscene, Insult) for Twitter (left) vs Parler (middle) vs Reddit (right).

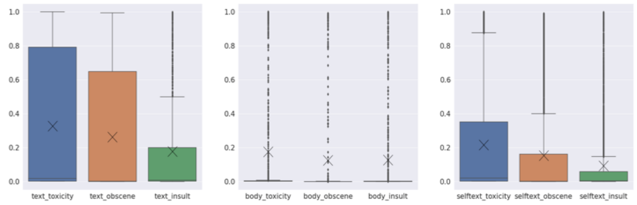


Figure 2. f*cklockdown hashtag for three classes (Toxicity, Obscene, Insult) for Twitter (left) vs Parler (middle) vs Reddit (right).

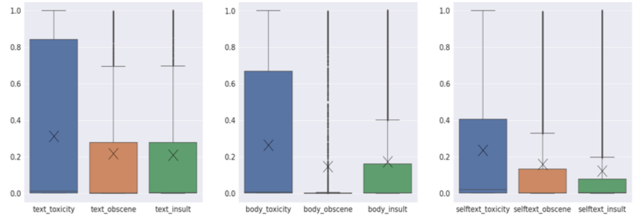


Figure 3. f*ckmask hashtag for three types of classes (Toxicity, Obscene, Insult) for Twitter (left) vs Parler (middle) vs Reddit (right).

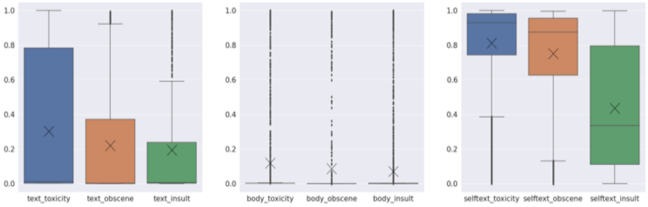


Figure 4. f*ckvaccine hashtag for three types of toxicity (Toxicity, Obscene, Insult) for Twitter (left) vs Parler (middle) vs Reddit (right).

The Twitter data, for example, shows that a few conversations are very toxic, and those few highly toxic conversations are driving up the overall toxicity level of the platform. The same goes for the Reddit vaccine dataset as well. This has important implications for platform administrators, who can significantly reduce the strongest drivers of toxicity by moderating the relatively few, highly toxic users rather than attempting larger platform-wide changes to all users. The toxicity standard deviation metrics revealed some unique contrasts between the platforms (Table 1). The standard deviation of toxicity values for content within the lockdown, mask, and vaccine categories are higher on Twitter than on Parler and Reddit, indicating that there is more variation in toxicity for these datasets. However, values were higher for Parler for content within the COVID category.

Figure 1 to Figure 4 illustrate that the term f*ckcovid on Parler is generally more toxic than on Twitter and Reddit. The mean toxicity is marked by a cross on each boxplot, slightly higher than Parler. However, for the other terms except for vaccine, Twitter is more toxic. All five points for the vaccine dataset on Reddit are the highest among all platforms. From the seven toxicity classes, we take three severe classes to compare in our statistical analysis. For F*Lockdown hashtags, Parler and Reddit are less toxic than Twitter if we consider the mean toxicity from the boxplot for both platforms. On the other hand, for f*ckcovid and f*ckmask hashtags, there is a significant increase in toxicity in Parler. On Twitter, the most toxic term f*ckmask whereas for Parler, it is f*ckcovid, and on Reddit, it is f*ckvaccine.

VI. CONCLUSIONS AND FUTURE WORKS

Twitter and Parler both experienced moderate levels of toxicity regarding COVID-19 content. However, Reddit had the highest toxicity related to the vaccine dataset, which is much higher than any other keywords or platforms. This paper compares and analyzes the toxicity of these three social media platforms in the same period. The methods were applied to different datasets for Twitter, Parler, and Reddit. The key finding of this research indicates Reddit is the most toxic social media platform among these three and Parler contained less toxicity compared to Twitter and Reddit regarding COVID-19 discourse.

Although the finding indicates toxicity levels were higher overall on Twitter for all datasets except for COVID-19 and vaccine, it was surprising to observe higher toxicity levels on Twitter since it is a moderated platform with clear guidelines for content posted, whereas Parler's guidelines emphasize a lack of moderation. Even though Reddit experienced the highest toxicity for the vaccine topic, the moderators took necessary steps to decrease toxicity by banning the anti-vaccine subreddit named 'r/Nonewnormal'. One possible explanation for the unexpectedly high toxicity on the Parler COVID dataset is that Twitter began removing users and posts sharing COVID-19 misinformation in April 2020, sparking anger and prompting many users to migrate to Parler instead [23]. In addition to being detrimental to the overall health of social networks, the moderate proportion of harmful content on these platforms surrounding COVID-19 topics may have affected users' perceptions of the effectiveness and importance of periodic lockdowns, wearing of face masks, and becoming vaccinated. The contributions of this work include evidence that 1) Twitter contained a higher level of toxicity regarding COVID-19 discourse than did Parler and Reddit; 2) Reddit contained the highest level of toxicity among all three social platforms for vaccine-related discussion. 3) Parler contained the highest level of toxicity among all three social platforms for COVID-related discussion.

A potential limitation of this paper is the methodology used to collect and analyze the data - the seed hashtag stem #f*ck can be used positively or negatively, depending on the context. The model used in this paper to classify content as toxic or not has difficulty distinguishing the semantic context of profanity and often classifies profane words as toxic, regardless of intent. We will keep this limitation in mind going forward in our future works. In future work, we plan to expand our keywords and collect more data from these three platforms, which are easy to get under their guidelines. We are working on other popular social media platforms like TikTok and Facebook. In addition, we will further explore the vaccine and lockdown topics due to their notably higher toxicity on Reddit.

ACKNOWLEDGEMENT

This research is funded in part by the U.S. National Science Foundation (OIA-1946391, OIA-1920920, IIS-1636933, ACI-1429160, and IIS-1110868), U.S. Office of the Under Secretary of Defense for Research and Engineering (FA9550-22-1-0332), U.S. Office of Naval Research (N00014-10-1-0091, N00014-14-1-0489, N00014-15-P-1187, N00014-16-1-2016, N00014-16-1-2412, N00014-17-1-2675, N00014-17-1-2605, N68335-19-C-0359, N00014-19-1-2336, N68335-20-C-0540, N00014-21-1-2121, N00014-21-1-2765, N00014-22-1-2318), U.S. Air Force Research Laboratory, U.S. Army Research Office (W911NF-20-1-0262, W911NF-16-1-0189, W911NF-23-1-0011), U.S. Defense Advanced Research Projects Agency (W31P4Q-17-C-0059), Arkansas Research Alliance, the Jerry L. Maulden/Entergy Endowment at the University of Arkansas at Little Rock, and the Australian Department of Defense Strategic Policy Grants Program (SPGP) (award number: 2020-106-094). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding organizations. The researchers gratefully acknowledge the support.

REFERENCES

[1] A.M. Guess, and B.A. Lyons, "Misinformation, disinformation, and online propaganda." Social media and democracy: The state of the field, prospects for reform 10 2020.

[2] B. S. Sahana, G. Sandhya, R. S. Tanuja, Sushma Ellur, and A. Ajina. "Towards a safer conversation space: detection of toxic content in social media (student consortium)." In 2020 IEEE Sixth International Conference on Multimedia Big Data (BigMM), pp. 297-301. IEEE, 2020.

[3] M. Taleb, A. Hamza, M. Zouitni, N. Burmani, S. Lafkiar, and N. En-Nahnahi. "Detection of toxicity in social media based on Natural Language Processing methods." In 2022 International Conference on Intelligent Systems and Computer Vision (ISCV), pp. 1-7. IEEE, 2022.

[4] A.K. Kumar and B. Kanisha. "Analysis of multiple toxicities using ML algorithms to detect toxic comments." In 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), pp. 1561-1566. IEEE, 2022.

[5] H. Watanabe, M. Bouazizi, and T. Ohtsuki. "Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection." IEEE access 6, 2018: 13825-13835.

[6] T. Gröndahl, L. Pajola, M. Juuti, M. Conti, and N. Asokan. "All you need is" love" evading hate speech detection." In Proceedings of the 11th ACM workshop on artificial intelligence and security, pp. 2-12. 2018.

[7] J. Cheng, C. Danescu-Niculescu-Mizil, and J. Leskovec. "Antisocial behavior in online discussion communities." In Proceedings of the international aaai conference on web and social media, vol. 9, no. 1, pp. 61-70. 2015.

[8] I. Gunasekara and I. Nejadgholi. "A review of standard text classification practices for multi-label toxicity identification of online content." In Proceedings of the 2nd workshop on abusive language online (ALW2), pp. 21-25. 2018.

[9] L. Hanu, (2020). Unitary team. Detoxify. Github.

[10] K. DiCicco, N. B. Noor, N. Yousefi, B. Spann, M. Maleki and N. Agarwal. “Toxicity and networks of COVID-19 discourse communities: A tale of two media platforms.” In The 3rd Workshop on Reducing Online Misinformation through Credible Information Retrieval, 2023, forthcoming.

[11] A. Obadimu, E. Mead, M. N. Hussain, and N. Agarwal. "Identifying toxicity within youtube video comment." In International conference on social computing, Behavioral-cultural modeling and prediction and behavior representation in modeling and simulation, pp. 214-223. Springer, Cham, 2019.

[12] A. Obadimu, E. Mead, and N. Agarwal, "Identifying latent toxic features on YouTube using non-negative matrix factorization." In The Ninth International Conference on Social Media Technologies, Communication, and Informatics, IEEE. 2019.

[13] P. Pascual-Ferrá, N. Alperstein, D. J. Barnett, and R. N. Rimal, "Toxicity and verbal aggression on social media: Polarized discourse on wearing face masks during the COVID-19 pandemic." Big Data & Society 8, no. 1, 2021: 20539517211023533.

[14] W. Chipidza, "The effect of toxicity on COVID-19 news network formation in political subcommunities on Reddit: An affiliation network approach." International Journal of Information Management 61, 2021: 102397.

[15] R. Urbaniak et al., "Namespotting: Username toxicity and actual toxic behavior on Reddit." Computers in Human Behavior 136, 2022: 107371.

[16] Y. Y. Chong, and H. Kwak. "Understanding toxicity triggers on Reddit in the context of Singapore." In Proceedings of the International AAAI Conference on Web and Social Media, vol. 16, pp. 1383-1387. 2022.

[17] H. Almerekhi, S.B.B.J. Jansen, and C.S.B.H. Kwak. "Investigating toxicity across multiple Reddit communities, users, and moderators." In Companion proceedings of the web conference 2020, pp. 294-298. 2020.

[18] M. Aliapoulios, E. Bevensee, J. Blackburn, B. Bradlyn, E. De Cristofaro, G. Stringhini, and S. Zannettou. "A large open dataset from the Parler social network." In ICWSM, pp. 943-951. 2021.

[19] https://developer.twitter.com/en/products/twitter-api/academic-research, last accessed March 2023.

[20] J. Baumgartner, S. Zannettou, B. Keegan, M. Squire, and J. Blackburn. "The pushshift reddit dataset." In Proceedings of the international AAAI conference on web and social media, vol. 14, pp. 830-839. 2020.

[21] https://github.com/pushshift/api, last accessed March 2023

[22] https://psaw.readthedocs.io/en/latest, last accessed March 2023

[23] L. König and P. Breves. "Providing health information via Twitter: professional background and message style influence source trustworthiness, message credibility and behavioral intentions." Journal of Science Communication 20, no. 4, 2021: A04.

Complex Behavior Vs. Design - Interpreting AI: Reminders from Synthetic Psychology

Elliot Swaim
CS. Grinnell College
Grinnell, USA
email swaimell@grinnell.edu

Fernanda Eliott
CS. Grinnell College
Grinnell, USA
email eliottfe@grinnell.edu

***Abstract*—Can a simple agent design (*i.e.*, that uses a small set of simple rules) trigger complex behavior? To investigate that question, we implemented Braitenberg vehicles in a Khepera robot simulator using the Java programming language. We decided to avoid a fancy look from popular simulators to prevent enhancing visual, unrelated sophistication to our experiments. We ran our Braintemberg-inspired Khepera robots, recorded the simulations, and watched the recordings. Our simulations provide interesting insights as we discuss a distinction between *interpreted behavior* and *embedded behavior*. Given the popularity of AI-powered (Artificial Intelligence) tools, we hope our discussion inspired by Braitenberg and synthetic psychology will provide fruitful reflections on the role of anthropomorphism in interpreting AI.**

Keywords*—*AI; anthropomorphism; behavior; Braitenberg vehicles; synthetic psychology.

I. Introduction

Valentino Braitenberg authored a book [1] that proposes thought experiments *via* vehicles (or robots) that embody human-like elements, such as love or aggression. The vehicles illustrate synthetic psychology, *i.e.*, the notion that we can investigate ourselves, biological creatures, through the development of machines embodied and observed in an environment [2]. Although the vehicles follow very simple rules, their actions may be interpreted as sophisticated behavior. From observing them, we may project meaning onto their actions; however, they are void of any true complexity. Despite the book being published in the ’80s, the context has never been as current as right now. For instance, consider current inquiries on AI-powered language models and sentience. What happens if a considerable number of people become convinced that an AI is sentient and should be protected? Would that make people more likely to protect a machine rather than an animal?

On one hand, one could try to approach the “sentience” question in regard to machines in the same way we do with other humans: driving inspirations from folk psychology, we could use our abilities to attribute mental states and do it toward machines (e.g., their beliefs, desires, intentions). According to Ratcliffe and Hutto [3], despite an intense debate on which cognitive processes support humans’ folk psychological abilities, there is a considerable consensus on what folk psychology is: the ability to attribute intentional states, beliefs, and desires to others to predict and explain behavior. In a similar vein, while comparing observable properties of an external system with the unobservable properties of an internal system, Caporael [4] ponders Turing (1950/1964) and a flavor of a solution: “inferring that others have thought, consciousness, minds, or feelings is by comparing their behavior with what we expect or know to be our own in similar circumstances.”

On the other hand, that approach is subject to anthropomorphic bias [4], or to attribute human-like characteristics to non-human creatures or things. We may have the inclination to infer complexity in a system beyond what can be validly deduced from the observable outcomes, especially when those outcomes provide human clues. For instance, in one study [5] where participants were asked to determine between text that was authored by a human and text that was generated by a machine, participants were more likely to guess that a human authored the text if the text was expressed aloud than if participants were only able to read it. Because human speech lends itself more to anthropomorphism than text alone, participants tended to infer more complexity from it. The authors discuss their findings’ implications in the case of human dehumanization in text-based media on the one hand and anthropomorphizing machines in speech-based media on the other.

Still, anthropomorphism helps us interact with Artificial Intelligence (AI) according to its intent (such as with self-driving vehicles) and develop trust in machines [6], which can lead people to have a false understanding of AI. Digging deeper into *sentience* goes beyond the scope of this work; however, we would like to point readers to [7], where DeGrazia distinguishes sentience (beings capable of having pleasant or unpleasant experiences) from consciousness (beings capable of having subjective experience) to investigate if conscious although not sentient creatures could have interests and moral status. DeGrazia [7] examines animals and insects and comments on the implications for autonomous machines.

A. Our Work and Contributions

Our research attempts to investigate if, even on a very basic level where there is little motivation for anthropomorphism, the observable outcomes of an artificial agent can still communicate more complexity than that which is embedded in the agent. (Interestingly, that approach could, at some degree and with caution, return back to humans, as there may be situations in which we attribute more complexity than that which is embedded in us.) With the hope that synthetic psychology can resourcefully illustrate that complex behaviors do not necessarily imply complex design, we adapted and implemented Braitenberg vehicles into a robot controller (that uses the Java programming language) and ran simulations in different environments.

Next, we recorded and watched the simulations to provide possible interpretations for the robot's behavior (see in Section II, our discussion on embedded behaviors *vs.* interpreted behaviors). It is not our claim that our interpretations are the only ones possible, and we also acknowledge that those are subject to biases, since we played a role in the entire process. **Still, our interpretation/study is important because we are equipped to discriminate between embedded and interpreted behaviors.** (Note that this is an initial phase of our project; in future work, we plan on running a pilot study involving interpretation derived from a group of human participants to continue our investigation.) Nevertheless, it is our claim that the combination agent in an environment can favor the interpretation of behavior as complex and that 'complexity' may lead to false assumptions toward the agent design - we believe that such an awareness is essential for the general population as personal assistants and AI-powered tools get more common.

It is also our claim that stronger efforts should be made to investigate ways of educating people to make a distinction between behavior and design so that we all are better equipped to make sense of AI technologies' impact on the world. We identified Braitenberg vehicles as an accessible way of creating educational materials (and accessible in terms of both needed technology and framework). Braitenberg vehicles provide so many fruitful applications that it has been explored in other disciplines as well, such as in neuroscience [8].

The vehicles do not explore language but rather acts in an environment and how an observer interprets those acts. Our goal is to use synthetic psychology to remind us of the dangers of anthropomorphism, as we use it to exemplify that very simple rules and frameworks can still suggest meaning or somewhat complex behavior. Our work shows that simple design can create visual patterns that foster interpreted behaviors.

Contributions. Our contributions are the adaptation and implementation of Braitenberg vehicles in a robot controller for Khepera simulation and a contextualized discussion on the distinction between design and behavior. Finally, we consider our framework to be accessible, and others can easily adapt it to use and spread awareness of AI.

This work is organized as follows: in Section I, we introduce our work and contributions. In Section II, we provide background information and more details about Braitenberg vehicles. In Section III, we describe our methods and experimental setup, followed by results and discussion in Section IV. Finally, we present our conclusions and suggestions for future work in Section V.

II. Background

Communication does not necessarily need words to occur; for example, when we join a queue at a store, we communicate that we aim to buy something once it is our turn; other customers respond by joining the queue behind us. According to Tversky [9], by using position, form, and movement in space, gestures, and actions convey a plentiful set of meanings. In that sense, differently from solely symbolic words, visual communication can directly convey content and structure (both literally and metaphorically). Although it may lack the rigorous definitions that words can offer, visual communication delivers both flexibility and suggestions for meanings. Such flexibility, in its turn, requires context and experience to interpret conveyed meanings.

Caporael [4] suggests that *anthropomorphism* results from a schema that we apply to *phenomena*, such as machines, while *mechanomorphism* would be the other way around or the attribution of machine-like attributes to humans. Focusing on three psychological determinants (1. the accessibility and applicability of anthropocentric knowledge, 2. the motivation to explain and understand the behavior of other agents, and 3. the desire for social contact and affiliation), Epley and colleagues [10] present a theory to explain when people are more likely to anthropomorphize. Taking into account ethical issues in AI, Salles and colleagues [11] discuss and examine anthropomorphism, as "It is a well-known fact that AI's functionalities and innovations are often anthropomorphized".

Braitenberg vehicles were conceived to demonstrate how complex behaviors can arise from simplistic concepts or rules and that we can seek to understand the complex behaviors we see in humans and animals by attempting to reconstruct those behaviors using simple concepts (a method called *Synthetic Psychology*).

> "Watching vehicles of brand 4a in a landscape of sources, you will be delighted by their complicated trajectories. And I am sure you will feel that their motives and tastes are much too varied and intricate to be understood by the observer. (...) You forget, of course, that we have ourselves designed these vehicles" [1].

Whereas the aim of Synthetic Psychology is to understand human or animal behavior through reconstruction, our aim is to use Braitenberg vehicles as inspiration to navigate the distinction between behavior and design in artificial agents. To that end, we distinguish between two types of behaviors: *interpreted behaviors* and *embedded behaviors*.

Embedded behaviors are patterns of actions that the agent actually follows. They are the behaviors coined into the agent's rule sets and are what result in the various series of actions an agent performs. (Note that we are not using any kind of learning in our experiments, just simple rules.)

Interpreted behaviors come from how an observer interprets the series of actions observed. They are patterns of action that exist in the observers' interpretation as a result of applying methods of interpretation to the series of actions they observe.

When attempting to understand artificial agents, it is important to investigate what they are in themselves in addition to what they are to us. It is essential to distinguish between behaviors that are embedded and coined to the agent from the behaviors which exist only as a pattern of action in our own interpretation.

Braitenberg's [1] **Vehicle 1** has only one sensor connected to a motor such that the stronger the activation of the sensor, the faster it goes. The sensor is tuned to a quality such as *light* or *temperature*, and this vehicle moves only forward in the absence of perturbations. The other vehicles are simple two-wheel objects, and both wheels are connected to sensors in simple ways so that the speed of each motor is correlated to the activation of the sensors. From these simple connections and rules, complex behaviors seem to arise.

Vehicle 2 has two sensors that are either parallel-connected (left sensor connected to the left motor and vice versa, Vehicle 2a), or cross-connected (left sensor connected to the right motor and vice versa, Vehicle 2b). Although both vehicles move faster in the presence of the source to which the sensors are tuned to, Vehicle 2a turns away from the source while Vehicle 2b turns toward it.

For Vehicle 2a, if the source is on one side of the vehicle, the corresponding sensor will have higher activation than the sensor on the other side. As a result, the wheel on the side of the source will move faster, causing it to turn away from the light. For Vehicle 2b, since the sensors are cross-connected to the motors, the motors on the opposite side move faster, causing it to turn toward it, perhaps even hitting the source. As the author points out, it may look like both vehicles "dislike" the source: 2a looks like a "coward" whereas 2b is "aggressive".

In **Vehicle 3**, the speed of the motors is inversely proportional to the sensor activation. Vehicle 3a is parallel-connected, and Vehicle 3b is cross-connected. Since higher sensor activation result in slower motor speeds, Vehicle 3a moves toward the source and rests in its vicinity. In contrast, Vehicle 3b comes to rest facing away from the source or even leaving as a result of a perturbation. This behavior makes it look like the vehicles "like" the source: Vehicle 3a "loves" it, while 3b acts as an "explorer": likes the source but is open to other sources as well.

In **Vehicle 4**, the speed of the motors is related to the sensor activation through an arbitrary activation function. The behaviors of these vehicles depend on the activation chosen. Vehicles 2 and 3 are both particular types of Vehicle 4. The book continues to introduce more vehicles with increasingly more complex rules and connections. However, we focus on the first four in our research.

III. Methods and Experimental Set Up

Seeking simplicity rather than a fancy look, we chose to implement and run the vehicles using the WSU (Wright State University) Khepera Simulator [12]. We aimed to avoid advanced features found in more recent simulators – which could elicit more sophistication in an observer's interpretation. In addition, the simulator provides noise in the sensor data, which helps introduce more random variation to each run of the experiments.

To simulate the two light or distance sensors on the Braitenberg vehicles with the eight light and distance sensors found on the Khepera robots, we averaged the sensor activation values from each of the four sensors on each side of the robot to approximate what a single sensor on each side of the robot might sense.

Directional vs Omnidirectional Sensors. Because of the nature of the Khepera robots and of the simulator we used, we had to adjust the vehicles accordingly. Whereas in Braitenberg vehicles the sensors are omnidirectional, each of the Khepera robot's sensors is directional. By averaging the activation values from each of the sensors on either side of the robot, we were able to somewhat reduce the impact of using directional sensors rather than omnidirectional sensors in our implementation of the vehicles.

In our implementation of the vehicles, only the sensors on the side facing the light detect the light, so the average of the sensors on the side opposite the light read zero. In addition, because there is a forward-facing, diagonally forward-facing, sideways, and backward-facing sensor on each side of the Khepera robot but no diagonally backward-facing sensor, the robot has a slight "blind spot" diagonally behind it where light can only be detected through the backward facing and side sensors, and since no sensor would detect the light straight-on, the detected brightness would be less than the theoretical brightness that an omnidirectional sensor would detect.

Sensor Activation Values. A second difference is due to Khepera's reading sensors. The light sensors range from 500-512 for no light and diminishing for full light exposure. We determined through experimentation that we could use the relationship between the value read by the light sensors and the distance to the light as $100{*}log_2(x)$ where x is the distance to the light source. Thus, before averaging the values from the various light sensors, we first calculated the distance from the value read by the sensor. Then, we calculated the brightness of the light using the inverse square law. Then the brightness of the light falling on each sensor on each side of the robot was averaged to get the brightness falling on each side of the robot. The distance sensors range from 0 when nothing is detected to 1023 right up next to something (wall, object, or obstacle). We simply used the values given by the robot's distance sensors.

Obstacle-avoiding. In the WSU simulator, if the robot crashes into a wall or light, the simulation halts. However, because many of the simplest Braitenberg vehicles do not have any obstacle-avoiding capabilities, to give us enough time to observe the vehicles and form interpreted behaviors based on their outcome, we embedded an obstacle-avoidance rules on top of Braitenberg's. Specifically, if the robot gets too close to an obstacle, it temporarily stops following the Braitenberg rules, turns approximately 180 degrees, and then continues following the Braitenberg rules.

Maps. The WSU simulator enables us to create maps consisting of walls (either vertical or horizontal) and light sources.

Those are considered the agent's (or robot's) environment.

Activation Function We used a bell-shaped activation function in vehicle 4. Small activation results in low speeds, medium activation in larger speeds, and high activation also in low speeds. In addition, we implemented four sub-types of vehicle 4. Our Vehicle 4a has two light sensors that are cross-connected to the motors; Vehicle 4b has two light sensors that are parallel connected; Vehicle 4c has two light sensors that are cross-connected and two distance sensors that are also cross-connected. Note that for Vehicle 4, we flipped the connections so that 4a and 4c are cross-connected and 4b is parallel connected.

IV. Results and Discussion

In Table I, we summarize our implementation of Braitenberg Vehicles. Using the WSU simulator, we designed eleven maps inspired by Braitenberg's descriptions while also aiming to trigger interesting behaviors. We implemented the vehicles 3a, 3b, 3c, 3d, 4a, 4b, and 4c. For each vehicle, we recorded five 1-minute runs per map. We defined short, 1-minute runs given our approach to simplicity. Although we experimented with various maps and vehicles, we present here only the vehicle/map combinations we saw as most significant for our discussion on complex behavior *vs.* design. In Figure 1, we show the four maps and respective interpreted behaviors for vehicles 3a, 4a, and 4c, followed by a discussion in Section IV-A.

TABLE I: OUR IMPLEMENTATION OF BRAITENBERG VEHICLES.

V#	Rules	Connection (Light)	Connection (Distance)
1	Proportional	A single sensor	Not Used
2a	Proportional	Parallel Connected	Not Used
2b	Proportional	Cross Connected	Not Used
3a	Inversely Proportional	Parallel Connected	Not Used
3b	Inversely Proportional	Cross Connected	Not Used
3c	Inversely Proportional	Parallel Connected	Cross Connected
3d	Inversely Proportional	Cross Connected	Cross Connected
4a	Activation Function	Cross-Connected	Not Used
4b	Activation Function	Parallel Connected	Not Used
4c	Activation Function	Cross-Connected	Cross-Connected

As we watched the runs, we collected our interpretations while still keeping in mind that we should prevent getting "trained" in watching the videos. We list in Figure 1 the behaviors as we interpreted what the robot was doing. This is a list of interpreted behaviors we saw in each of the selected vehicles and maps and in which run that behavior was seen (from 1 to 5). We also identify whether or not the behavior was due to the obstacle-avoidance rule we built on top of Braitenberg vehicles. Whether or not a behavior was due to the obstacle-avoiding rule was evident whenever the robot would rotate in place at a constant speed near an obstacle, as that should not happen while following the Braitenberg rules (to access the code, just email the authors) and, we provide links to our experiments' videos in the References [13]–[19].

A. Discussion

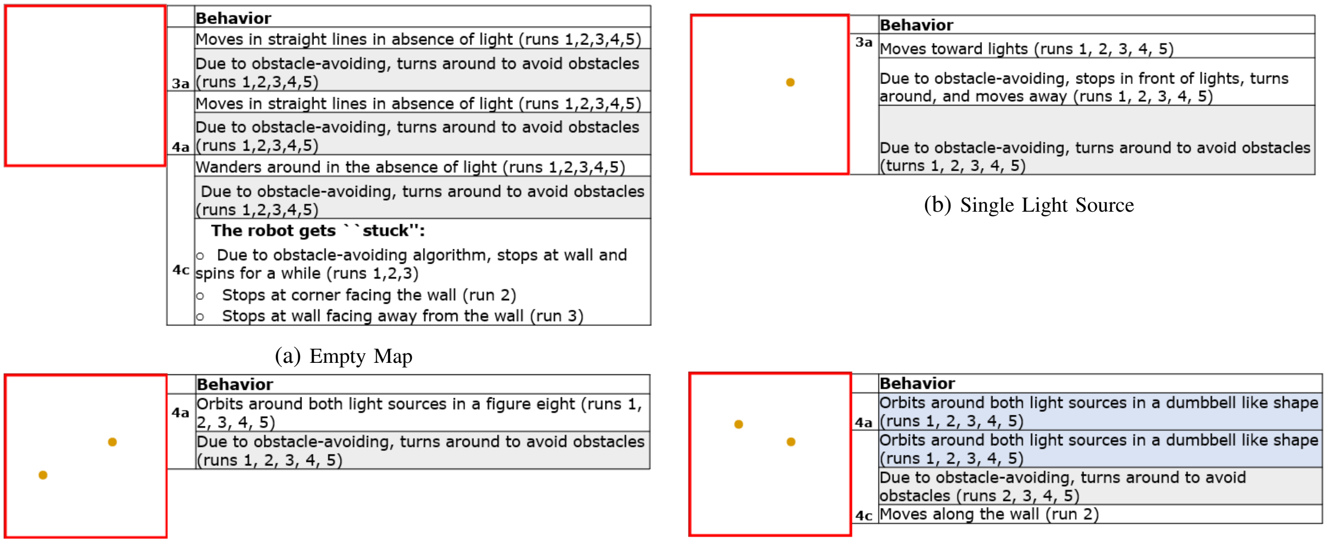
Braitenberg vehicles help to illustrate an important distinction when interpreting artificial agents: the distinction between interpreted behavior and embedded behavior. By observing the outcome of each robot in our experiments, we see a series of actions, e.g., it moves at such and such speed, turns by such and such amount, and speeds up or slows down at such and such times. While it may be that there is some intent or design behind the series of actions it performs ("this set of actions is Rule A", "that set is Rule B"), none of that is communicated to the observer by simply observing the series of actions it performs. In our experiments, the only thing an observer sees is the sum outcome of all the actions, not the rules or patterns that drove those actions.

Nevertheless, that does not stop us from trying to guess the patterns that may have driven the outcomes we see. As our experiments point out, the things we infer from observing the outcomes of an agent come from our interpretation of what the agent's embedded behavior may be, not necessarily the actual embedded behavior. Braitenberg vehicles illustrate this well because the embedded behaviors, the concepts or rules that each vehicle follows, are extremely simple, but they can result in seemingly complex interpreted behavior. In reality, the embedded behaviors of the vehicles are as simple as the rules that each vehicle follows. But the way we interpret the behaviors introduce far more complexity than what is actually coined to the vehicles.

For instance, some interpreted behaviors of Vehicle 4a are that it moves in straight lines in the absence of light, and that orbits around lights. However, the embedded behaviors are not the same as we may suppose; the embedded behaviors are merely that the left wheel moves at a speed related to the sensor activation of the left sensor, and likewise for the right wheel. It also has the added embedded behavior of turning around when it gets too close to an obstacle, such as a wall (obstacle avoidance). Those are the only rules the vehicle follows whereas, by describing that it orbits around lights, we are attaching significance to a certain series of actions the robot performed in certain runs that have no correspondence in the vehicle: **the interpreted behavior of orbiting around lights is not an embedded behavior**.

On the other hand, the interpreted behavior of turning around to avoid obstacles puts significance on another series of actions (stopping a certain distance from a wall, rotating in place, moving away from the wall), but this time that series of actions has a correspondence in the vehicle: the robot does indeed perform that specific series of actions in particular situations, and it is an embedded behavior.

The interpreted behaviors that we infer from observing the outcomes of artificial agents may or may not be the same as the embedded behaviors that it follows. If we want to



	Behavior
3a	Moves in straight lines in absence of light (runs 1,2,3,4,5)
	Due to obstacle-avoiding, turns around to avoid obstacles (runs 1,2,3,4,5)
4a	Moves in straight lines in absence of light (runs 1,2,3,4,5)
	Due to obstacle-avoiding, turns around to avoid obstacles (runs 1,2,3,4,5)
4c	Wanders around in the absence of light (runs 1,2,3,4,5)
	Due to obstacle-avoiding, turns around to avoid obstacles (runs 1,2,3,4,5)
	The robot gets ``stuck'': ○ Due to obstacle-avoiding algorithm, stops at wall and spins for a while (runs 1,2,3) ○ Stops at corner facing the wall (run 2) ○ Stops at wall facing away from the wall (run 3)

(a) Empty Map

	Behavior
3a	Moves toward lights (runs 1, 2, 3, 4, 5)
	Due to obstacle-avoiding, stops in front of lights, turns around, and moves away (runs 1, 2, 3, 4, 5)
	Due to obstacle-avoiding, turns around to avoid obstacles (turns 1, 2, 3, 4, 5)

(b) Single Light Source

	Behavior
4a	Orbits around both light sources in a figure eight (runs 1, 2, 3, 4, 5)
	Due to obstacle-avoiding, turns around to avoid obstacles (runs 1, 2, 3, 4, 5)

(c) Double Light Source at Distance

	Behavior
4a	Orbits around both light sources in a dumbbell like shape (runs 1, 2, 3, 4, 5)
	Orbits around both light sources in a dumbbell like shape (runs 1, 2, 3, 4, 5)
	Due to obstacle-avoiding, turns around to avoid obstacles (runs 2, 3, 4, 5)
4c	Moves along the wall (run 2)

(d) Double Light Source

Figure 1: Maps used to run implemented vehicles along with interpreted behaviors. We color-coded behaviors that we saw as the same. Yellow dots represent light sources, whereas red outlines the walls.

understand the agent in terms of the complexity intrinsic to it as opposed to the complexity we bring to it, **we must look beyond the outcome of the actions the agent performs and look additionally into the architecture of the agent to determine how those actions came about**.

B. Impact of Context on Interpretation

Another way our experiments demonstrate the difference between interpreted behavior and embedded behavior is by suggesting that interpreted behavior is contingent upon context (note our “Store” example in Section II). For example, in Vehicle 4c, there were several times that the robot would stop off walls or corners. Since the context of our experiments was that we were watching robots navigate various maps, we interpreted this as the robot getting “stuck” at a wall and considered it a bug rather than a feature, a failure rather than a behavior.

On the other hand, if we had been observing insect-like robots navigating a maze and happening to perform the exact same series of actions that our robots did, we would not be surprised about it temporarily stopping near a wall. We might think it is an interesting behavior when it would sometimes stop near a wall and spin in place. But because the context was that of robots navigating a map, we did not interpret these series of actions as behavior but rather as a bug.

Considering embedded behaviors, the time the robot spent stopped at a wall is no more significant than any other time the robot spent wandering about the map. While it was stopped at a wall, it continued following the same two behaviors it was always following: set the left motor’s speed according to the left sensor’s activation and set the right motor’s speed according to the right sensor’s activation. The only difference was that the result of the activation function applied to the sensor activation was zero, so the robot didn’t move.

As for when it would spin in place for a while near a wall, that seemed to be caused when the robot would, due to noise in the sensor activation values, get closer to the wall than what would normally be allowed before the obstacle-avoiding would kick in. As a result, once it did kick in after the robot turned 180 degrees and started moving away from the wall, it would still be close enough to the wall to trigger the obstacle-avoiding algorithm again. This would cause it to rotate again until random noise in the sensor activation would allow it to move away from the wall without triggering the obstacle-avoiding algorithm again. Thus, while the robot was spinning in place, it was still following the same behaviors it always did. What made it seem different than any prior set of actions was a function of how we interpret behaviors rather than a function of something coined to how the robot worked, and how we interpreted the behavior was a function of the context in which we observed the robot.

C. Anthropomorphic Language and Interpretation

Our experiments also help to point out how the use of anthropomorphic language can impact how we interpret the behavior of artificial agents. For instance, when examining Vehicle 3a, we noticed first that we found it easier to describe behaviors in anthropomorphic terms rather than through neutral language, and second that we both disagreed on how we anthropomorphically interpreted the robot’s behavior. Using the more neutral language we chose in the results listed in Figure 1, in the map with a single light source, vehicle 3a would move toward a light, stop in front of it, turn around, move away, and then eventually come back toward the light. But when describing it anthropomorphically, one of us described it as if it were a child excited to get a close look at the light only to quickly get bored and run off to find another light source. However, the other described it as if it were scared of the light, approaching it cautiously and then running away from it quickly. It was not difficult to recognize that the anthropomorphic language we used to describe the robots' behaviors was distinct from the embedded behaviors. When Braitenberg himself described the behaviors of some of the vehicles as symbolic of love, hatred, aggression, etc., it is clear that those are not a literal representation of the embedded behavior sets. However, we did find that it was easier to refer to specific interpreted behaviors through anthropomorphic language, and we considered that it would be far easier to communicate what kinds of outcomes we observed to someone inexperienced with robotics using anthropomorphic language than using a more neutral language.

However, this leads to two considerations: a) Even if it is clear that the anthropomorphic language is not literal, it could easily give the impression of far more complexity than what is embedded to the robot. And while even neutral language can suffer from the same problem, anthropomorphic language can amplify the issue. b) The same outcomes can be described through vastly different anthropomorphic descriptions.

Even if someone doesn't interpret the anthropomorphic language as literal, different descriptions may carry different connotations which color how one interprets the agent's behavior during any future interactions with the agent. And the entire lens, the entire framework through which all future observations or interactions with the agent are interpreted has nothing to do with the agent itself but only the description which happened to be attached with it. The same series of actions of the same agent can be interpreted in vastly different ways based on what kind of anthropomorphic framework is attached to it through anthropomorphic descriptions.

V. Conclusion

Thinking of a call for the AI community to serve the general population in educating people to make a distinction between behavior and design so that we all are better equipped to make sense of AI technologies, we identified Braitenberg vehicles as an accessible way of creating educational materials.

Here, we provided a framework to adapt Brainteberg vehicles into a Khepera simulator to examine the friction between behavior and design. We discussed the distinction between interpreted behavior and embedded behavior and the impact of context on interpretation, and anthropomorphic language on interpretation. In future work, we plan on conducting human studies and asking people from different backgrounds to interact with the simulator and watch the videos to investigate if *interpreted behaviors* will appear and how to improve our framework so that we can make it freely available to help the general population reflect on the distinction behavior *vs.* design in AI. Although we focused on visual communication, our approach can be extended to other types of communication; in addition, other connections are possible to explore using a khepera robot. Therefore, for future work, we suggest using more connections and activation functions and running human studies targeting the general population to check if this framework helps build AI literacy. Through these studies, a distinction between interpreted vs. embedded behaviors can be investigated, in addition to making a comparison with fancier robot simulators, to see what effect fancier features play in people's interpretation. Finally, participants may also observe robots in person to enable the comparison of results from participants that observed simulations with the ones that observed a robot.

Acknowledgment

The authors would like to thank Grinnell College's Mentored Advanced Projects (MAP program).

References

[1] V. Braitenberg, *Vehicles: Experiments in synthetic psychology*. MIT press, 1986.

[2] T. J. Prescott and D. Camilleri, "The synthetic psychology of the self," in *Cognitive architectures*. Springer, 2019, pp. 85–104.

[3] M. Ratcliffe and D. Hutto, *Folk psychology re-assessed*. Springer, 2007.

[4] L. R. Caporael, "Anthropomorphism and mechanomorphism: Two faces of the human machine," *Computers in human behavior*, vol. 2, no. 3, pp. 215–234, 1986.

[5] J. Schroeder and N. Epley, "Mistaking minds and machines: How speech affects dehumanization and anthropomorphism." *Journal of Experimental Psychology: General*, vol. 145, no. 11, p. 1427, 2016.

[6] A. Waytz, J. Heafner, and N. Epley, "The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle," *Journal of experimental social psychology*, vol. 52, pp. 113–117, 2014.

[7] D. DeGrazia, "Sentience and consciousness as bases for attributing interests and moral status: considering the evidence and speculating slightly beyond," in *Neuroethics and nonhuman animals*. Springer, 2020, pp. 17–31.

[8] D. Shaikh and I. Rañó, "Braitenberg vehicles as computational tools for research in neuroscience," *Frontiers in Bioengineering and Biotechnology*, vol. 8, p. 565963, 2020.

[9] B. Tversky, "Visualizing thought," in *Handbook of human centric visualization*. Springer, 2014, pp. 3–40.

[10] N. Epley, A. Waytz, and J. T. Cacioppo, "On seeing human: a three-factor theory of anthropomorphism." *Psychological review*, vol. 114, no. 4, p. 864, 2007.

[11] A. Salles, K. Evers, and M. Farisco, "Anthropomorphism in ai," *AJOB neuroscience*, vol. 11, no. 2, pp. 88–95, 2020.

[12] S. Perretta and J. Gallagher, "A portable mobile robot simulator for a world wide web robotics practicum," in *2003 Annual Conference, Web Systems and Web Services*, 2003, pp. 8–96.

[13] E. Swaim and F. Eliott, "Simulation videos. vehicle 3a, empty map," https://doi.org/10.6084/m9.figshare.21802581.v1, accessed: 2023-02-11.

[14] ——, "Simulation videos. vehicle 3a, single light source map," https://doi.org/10.6084/m9.figshare.21802575.v1, accessed: 2023-02-11.

[15] ——, "Simulation videos. vehicle 4a, empty map," https://doi.org/10.6084/m9.figshare.21802590.v1, accessed: 2023-02-11.

[16] ——, "Simulation videos. vehicle 4a, double light source," https://doi.org/10.6084/m9.figshare.21802593.v1, accessed: 2023-02-11.

[17] ——, "Simulation videos. vehicle 4a, double light source at a distance," https://doi.org/10.6084/m9.figshare.21802596.v1, accessed: 2023-02-11.

[18] ——, "Simulation videos. vehicle 4c, empty map," https://doi.org/10.6084/m9.figshare.21802599.v1, accessed: 2023-02-11.

[19] ——, "Simulation videos. vehicle 4c, double light source," https://doi.org/10.6084/m9.figshare.21802602.v1, accessed: 2023-02-11.