9^{th} International Conference on Computational Social Science IC^{2}S^2
July 17-20, 2023, Copenhagen, Denmark

Language Toxicity Patterns of Politically-Polarized Twitter Replies During the COVID-19 Pandemic

Keywords: language toxicity, URL domain bias, user engagement, political polarization, COVID-19

Extended Abstract

Toxic languages have emerged online since the breakout of COVID-19 in 2019. In most cases, any SNSs account can freely engage with another one with texts, which could lead to further negative and harmful online social behaviors, such as rounds of toxic replying. Some studies have looked at the characteristics of toxic language online. [1] discovered that anti-vaxxers are more aggressive in replying by analyzing toxic replies of English and Japanese tweets. [2] found that the ideological extremity is more associated with the conservatives than the liberals through network analysis. [3] identified toxic replies diffusing patterns on Twitter based on news outlets diffused on Twitter. However, the language toxicity patterns of politically polarized Twitter replies during the COVID-19 pandemic are still unclear. This study examined the patterns of language toxicity of "Right", "Center" and "Left" replies.

In this study, 542,212,429 English tweets were collected from February 20 2020 to May 30 2022 by querying COVID-19 related keywords: "corona virus", "coronavirus", "covid19", "2019-nCoV", "SARS-CoV-2", "wuhanpneumonia" using the Twitter Search API. 25,370,268 replies of the English tweets were used for this study. A politically-leaning URL domain list of news websites was then obtained by requesting from Allsides[1] for academic research purpose, which contains 160 "Left" and "Lean Left" URLs, 98 "Right" and "Lean Right" URLs and 180 "Center" URLs. By using the list, each reply was labeled as "Right" if its domain of the Twitter URL object was identified in the "Right" and "Lean Right" categories of the domain list; the other replies were labeled as "Left" and "Center", accordingly.

To examine the degree to which a user engages with labeled replies, we categorized users according to their replies' domain labels. For example, the "Right" user category includes users whose reply URL objects contained "Right" domains, exclusively. It happens that a reply does not contain any URLs. Please, keep in mind that this study only looked at replies that met two criteria: 1. The user to whom were replied ("in_reply_to_screen_name" in the standard Twitter object [2]) is not a "Null" value, and 2. The reply contains, at least, one domain in the Twitter URL object. Meanwhile, the study further considered the frequency with which each user was replied to in each politically-leaning category. For example, if "Left" domains occurred in a replied-to user's reply URL object three times without "Center" and "Right" domains occurring, this user was considered to be a three-time-replied-to user in the "Left" category and was called a three-time-replied-to "Left" category user. As a result, a user who were replied more frequently, in this study, is regarded as a more engaged user with a politically-leaning domain category. Short URLs were expanded them before the labeling. The language toxicities of replied texts were measured by Google's Perspective API [3]. Several other studies

[1] www.allsides.com

[2] `https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/object-model/tweet`

[3] `https://www.perspectiveapi.com/`

have demonstrated that its predictions are robust. The perspective API calculates a score for each text on a scale from 0 to 1.

The study found a negative correlation between maximum toxicity and the replied times. The maximum toxicity is the highest value of language toxicity of the category with specific replied times. Figure 1a illustrates the maximum toxicities of each category replied at different times, indicating that more-replied-to users were less likely to receive replies with high toxicity. The "Right" and "Center" categories replied-to users shared a similar maximum toxicity distribution (Kolmogorov-Smirnov test, $p > 0.05$), while the "Left" category showed a different distribution ($p < 0.05$). However, the statistical difference does not change the overall trend of the three categories (Figure 1a). In general, more frequently replied-to users shared lower maximum toxicities, regardless of user category. The "Left" category differs from the others, possibly due to the higher toxicity values of several outliers. For instance, some "Left" category outliers (indicated by arrows in Figure 1a) shared larger toxicities and some of them even reached more than 0.8. The outliers could be top toxic repliers. By contrast, the "Right" category users' maximum toxicities were less than 0.4, when they were replied more than, approximately 1,000 times. The maximum toxicity is compared between categories in Figure 1b. This reveals that the maximum toxicities of the "Left" category users are significantly higher than those of the other two categories (Mann-Whitney U test, $p < 0.005$). On the contrary, the three categories shared a similar distribution for median toxicity (Kolmogorov-Smirnov test, $p > 0.05$) (Figure 1c), and no significant median toxicity group was identified in the three categories (Figure 1d).

These results suggest that although there is no significant difference in the language toxicity across the replied-to users of "Right", "Left" and "Center" categories, the top toxic language repliers in each category can not be neglectable, especially for "Left" category. Top toxic repliers of the "Left" category could target popular replied-to users. Previous research confirmed that the right group was more distant from the neutral group than the left group [1]. However, this study found that "Right" category repliers were much more closed to the "Left" category than the "Center" category, in terms of maximum toxicities. The COVID-19 pandemic is "far from over"[4]. The language toxicity associated with COIVD-19, diffusing in the political community, should not be ignored. The SNSs are suggested to pay attention to the top toxic "Left" repliers and the users are advised to be careful while engaging with them, as as the authors of them might be emotional. Future work would be finding out the emotional patterns of the replies and the network features of the frequently replied-to users and discovering more solutions to combat the aggression of the toxic languages to keep our SNSs ecosystem healthier.

References

[1] Miyazaki, K., Uchiba, T., Tanaka, K. et al. Aggressive behaviour of anti-vaxxers and their toxic replies in English and Japanese. Humanit Soc Sci Commun 9, 229 (2022).

[2] Mosleh, M., Rand, D.G. Measuring exposure to misinformation from political elites on Twitter. Nat Commun 13, 7144 (2022).

[3] Martin Saveski, Brandon Roy, and Deb Roy. 2021. The Structure of Toxic Conversations on Twitter. In Proceedings of the Web Conference 2021 (WWW '21). Association for Computing Machinery, New York, NY, USA, 1086–1097.

[4] `https://news.un.org/en/story/2022/03/1113632`

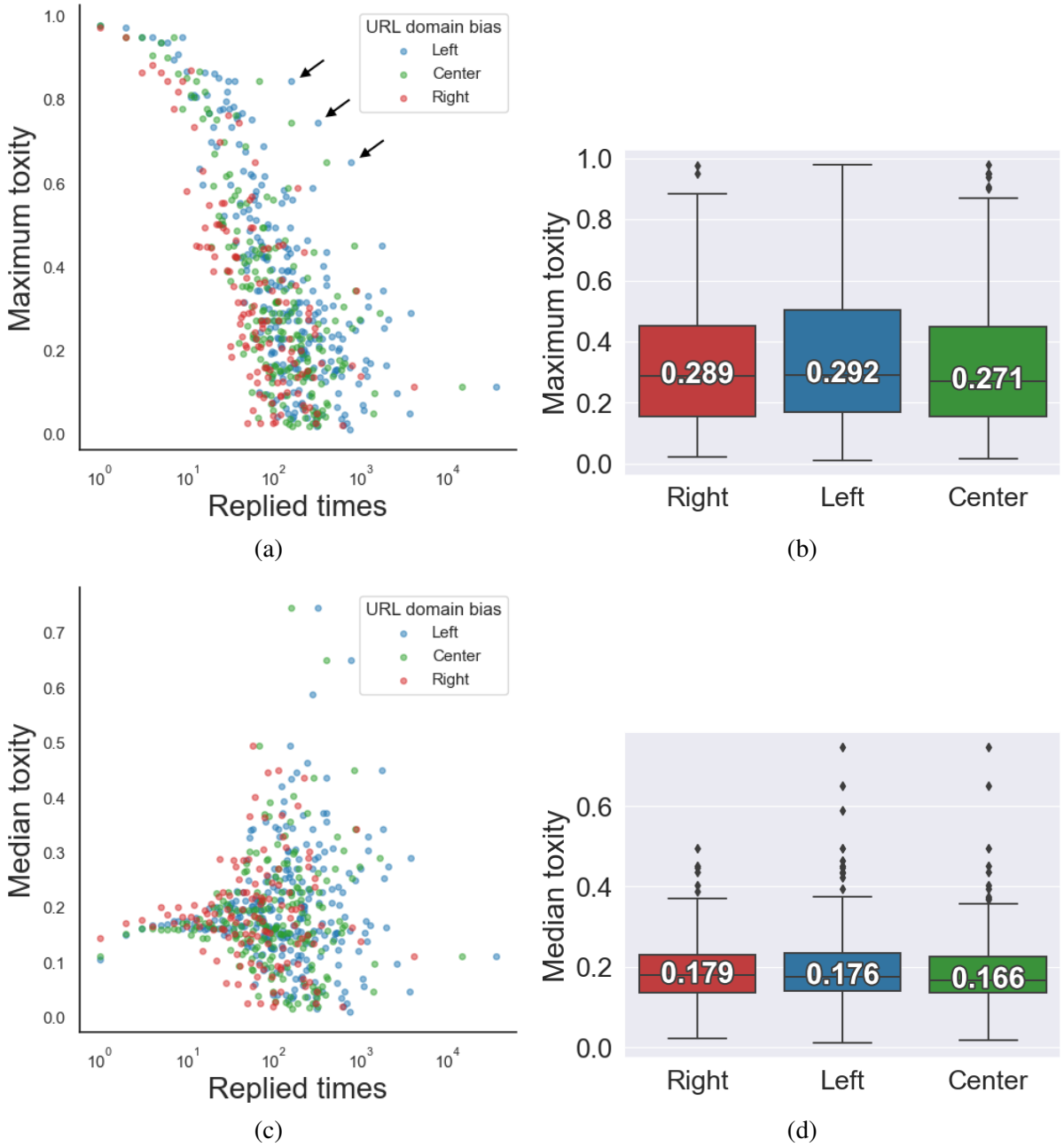

Figure 1: (a) The maximum toxicities of replies of each politically-leaning category of replied users received. The X-axis,repied times, is in log scale. The Y axis indicates the times each category of users were replied. Reds indicate users were engaging with "Right" domains, exclusively; Greens indicates users were engaging with "Center" domains, exclusively; Blues indicate uses were engaging with "Left" domains, exclusively. "Left" category outliers were indicated by arrows. Replies to the "Left" category users were significantly more toxic than the ones to the "Right" and "Center" category ($p < 0.005$ by Mann–Whitney U-test with a Bonferroni correction). (b) Boxplots represent the maximum toxicity of the replies of each user category. Each data point indicates the maximum toxicity of a user category. Each median score is annotated in the boxes. (c) and (d) are the median toxcity counterparts of (a) and (b), respectively.