The Manifestation of Affective Polarization on Social Media

Keywords: affective polarization, BERT, mass polarization, social media, supervised machine learning

Extended Abstract

Political polarization is rising around the world, not least in the United States. Although several forms of polarization exist, affective polarization, or people's negative feelings toward their political opponents (Iyengar et al., 2012), has increased substantially troublingly (Iyengar et al., 2019). Though there is much work on social media's contribution to affective polarization, little is known about how this type of polarization manifests differently across digital spaces. Recent advances have made some progress (e.g., Marchal, 2021; Mentzer et al., 2020; Rathje et al., 2021; Yarchi et al., 2021) but relied on methodologies such as sentiment analysis and bag-of-words approaches, which have limited abilities to capture the nuances of affective polarization. The current research addresses these limitations by using supervised machine learning to build a classifier that can detect expressions of affective polarization in social media (Facebook, Twitter) content. We focus on COVID-19 posts/tweets from the first half year of the pandemic (n = 8,603,695), which provides a glimpse not only into how people talk about contentious issues on social media but also into how affective polarization changes as issues grow increasingly partisan. Our findings make several contributions to the growing literature on affective polarization, and especially have important implications for those seeking to capture affective polarization in text, or understand how affective polarization manifests on social media.

One key contribution of our study is the introduction and validation of a classifier that can accurately capture affective polarization in social media texts. Prior efforts have grouped different types of polarization together or relied on reductive measures (e.g., dictionary approaches and/or sentiment analysis), limiting researchers' ability to capture the theoretical conceptualizations of these constructs. Our first step was to explicate what it means for a social media text to express affective polarization. We conceptualized affective polarization in social media content as *expressions of dislike or negativity toward those the author of a social media post disagrees with about politics.*

With this starting point, we conducted a content analysis to capture this concept in tweets and Facebook posts. Two research assistants were trained based on a detailed codebook to reliably identify this concept in tweets and Facebook posts (Krippendorff's alpha = 0.774), and then labeled a set of posts (n = 3,194). With this labeled dataset, we trained a model, leveraging a fine-tuned version of BERT (Bidirectional Encoder Representations from Transformers; Devlin et al. 2018), to identify affective polarization in social media texts at scale. Validation showed that this model performed well, with *accuracy*, *precision*, *recall*, and *F-score* all exceeding 0.80. This stands in contrast to prior attempts which have not used human validation to check model performance.

Conceptual and methodological contributions aside, the current research also speaks to how affective polarization manifests on social media, particularly around unfamiliar issues that transform into hot-button topics. To this end, we focused on social media discourse about COVID-19 during the first six months of the pandemic in 2020. We find that affective polarization followed largely similar trajectories across Facebook and Twitter. This was confirmed using time series analyses. On both platforms, affective polarization was relatively

low in January and February, although at this point polarization was higher on Twitter than on Facebook. Affective polarization then rose, following very similar trajectories across both platforms, in two waves. These similarities may suggest that we are observing trends that generalize to social media in a broad sense, and are not unique to one specific platform.

Using time series analyses, we also identified the interplay between affective polarization and virality and the interplay between virality on the two platforms. First, our time series analyses suggest that affective polarization precedes viral engagement on Facebook, but not on Twitter. There may be several explanations for this, including potential differences in the user base (perhaps Facebook users are more likely to engage with affective polarization) or differences in the platform's infrastructure (such as the way in which algorithms on each respective platform treat retweets and shares). Second, we find that virality on Facebook on a given day was predictive of virality on Twitter on the following day. This diffusion of virality from one platform to another suggests some linking of engagement across platforms—not necessarily causal, but temporally predictable. Perhaps popularity on one platform may motivate others to spread that content on other platforms.

In sum, this paper introduced and validated an improved way of measuring affective polarization in natural language. In doing so, we investigated how COVID-19 discourse on social media grew increasingly polarized as the pandemic unfolded, and we also shed new light on cross-platform dynamics and the interplay between polarization and virality. Besides adding theoretical insight to the growing literature on affective polarization, we believe these efforts can serve as a building block for future attempts to capture affective polarization in text. Given the complexities of affective polarization, we caution scholars to use careful human validation of automated computational ways of capturing this rather nuanced construct. By better understanding how polarization manifests on social media, we hope that scholars and digital architects will be better able to understand the interplay between polarization and social media, and better equipped to build constructive digital spaces.

References

Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., & Westwood, S. J. (2019). The origins and consequences of affective polarization in the United States. *Annual Review of Political Science*, *22*, 129–146. https://doi.org/gfv79s

Iyengar, S., Sood, G., & Lelkes, Y. (2012). Affect, not ideology: A social identity perspective on polarization. *Public Opinion Quarterly*, *76*(3), 405–431. https://doi.org/bsck

Marchal, N. (2021). "Be Nice or Leave Me Alone": An Intergroup Perspective on Affective Polarization in Online Political Discussions. *Communication Research*, 00936502211042516. https://doi.org/10.1177/00936502211042516

Mentzer, K., Fallon, K., Prichard, J., & Yates, D. (2020). Measuring and Unpacking Affective Polarization on Twitter: The Role of Party and Gender in the 2018 Senate Races. *Hawaii International Conference on System Sciences 2020 (HICSS-53)*. https://aisel.aisnet.org/hicss-53/dsm/data_mining/4

Rathje, S., Bavel, J. J. V., & Linden, S. van der. (2021). Out-group animosity drives engagement on social media. *Proceedings of the National Academy of Sciences*, *118*(26). https://doi.org/10.1073/pnas.2024292118

Yarchi, M., Baden, C., & Kligler-Vilenchik, N. (2021). Political Polarization on the Digital Sphere: A Cross-platform, Over-time Analysis of Interactional, Positional, and Affective Polarization on Social Media. *Political Communication*, *38*(1–2), 98–139. https://doi.org/10.1080/10584609.2020.1785067

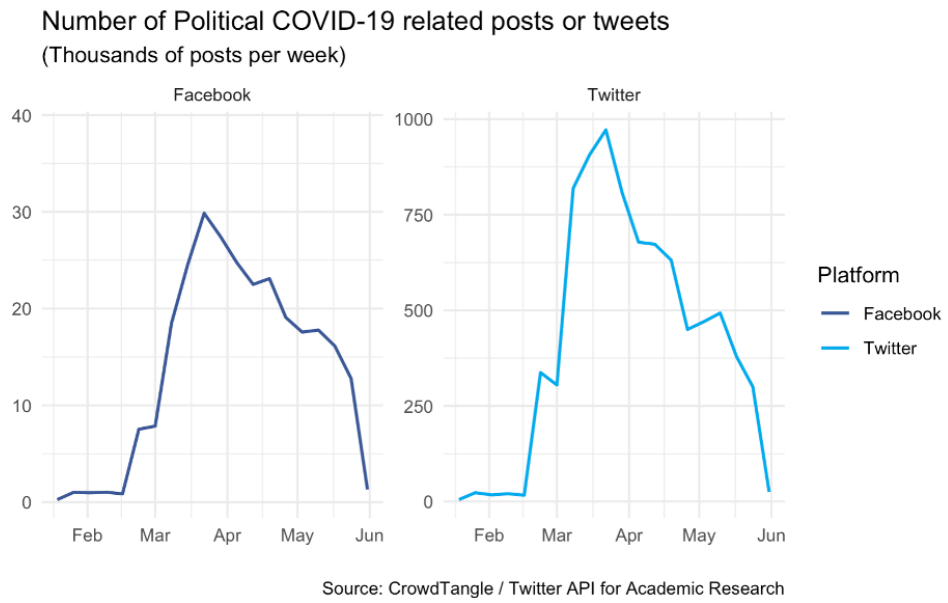

Figure 1. Post volume.

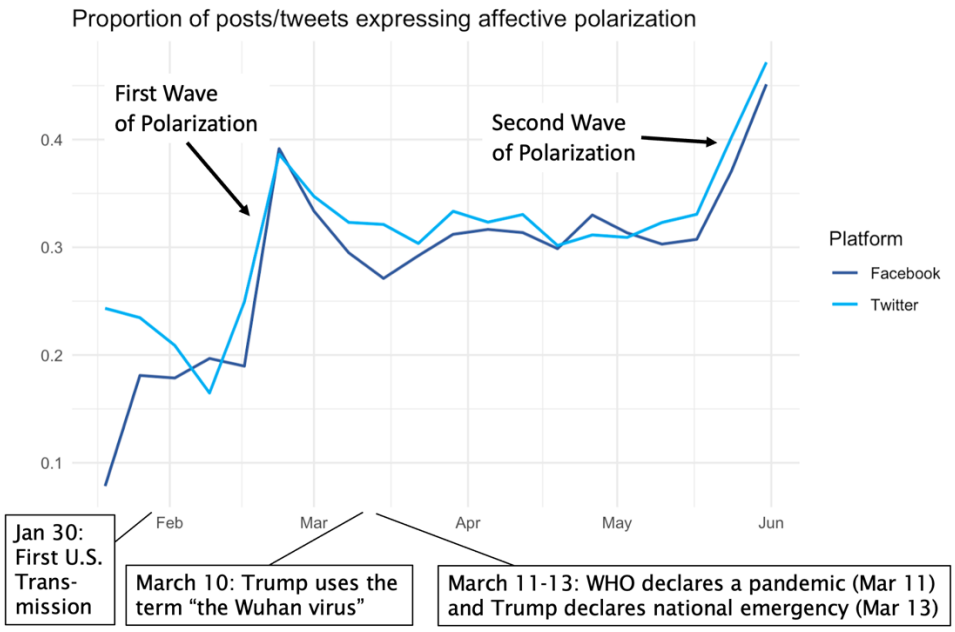

Figure 2. Affective polarization on Facebook and Twitter.

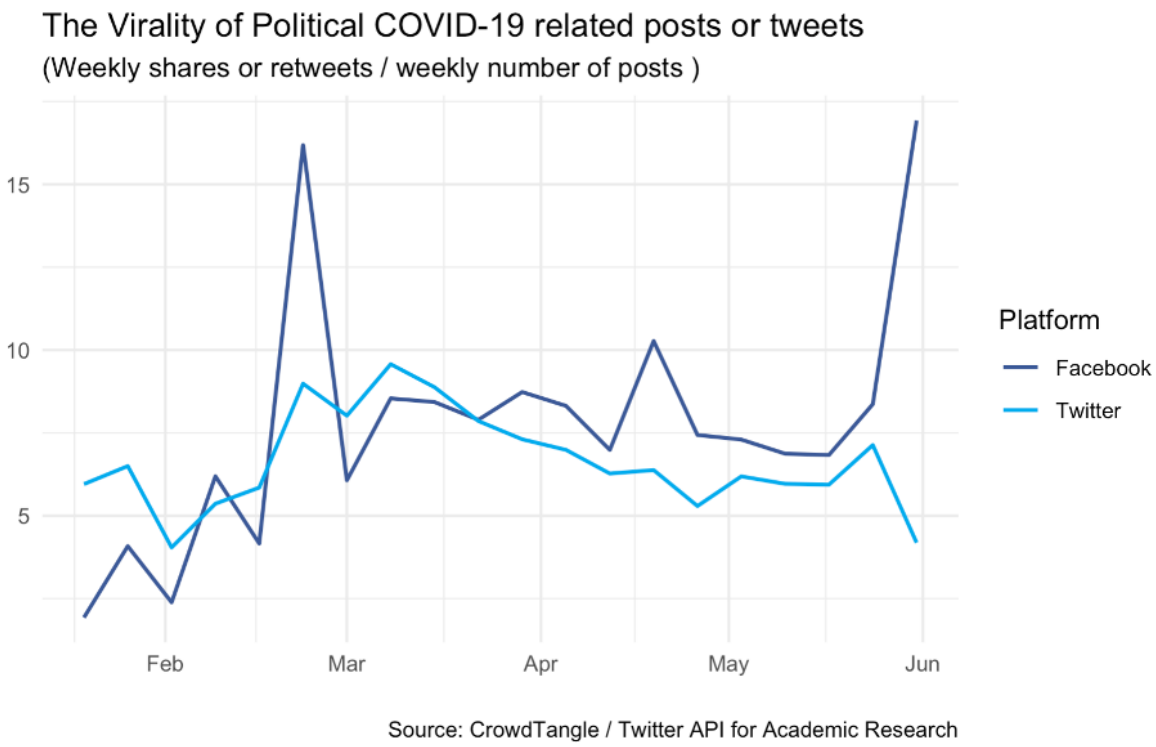

Figure 3. Virality on Facebook and Twitter

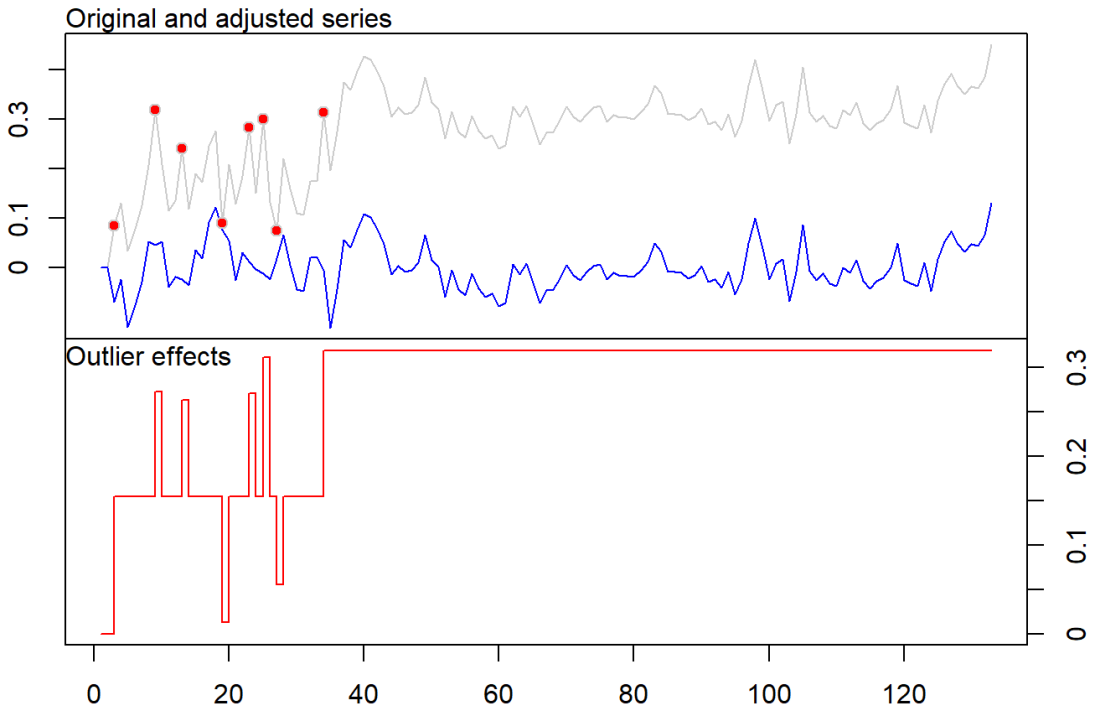

Figure 4. Outlier analysis for Facebook (Note: The blue line represents the data as predicted by the ARIMA model and the grey line represents the actual line data, with red dots indicating the time point of the outlier. The red line below represents the type of outlier: additive outliers, transient change, or level shifts.)

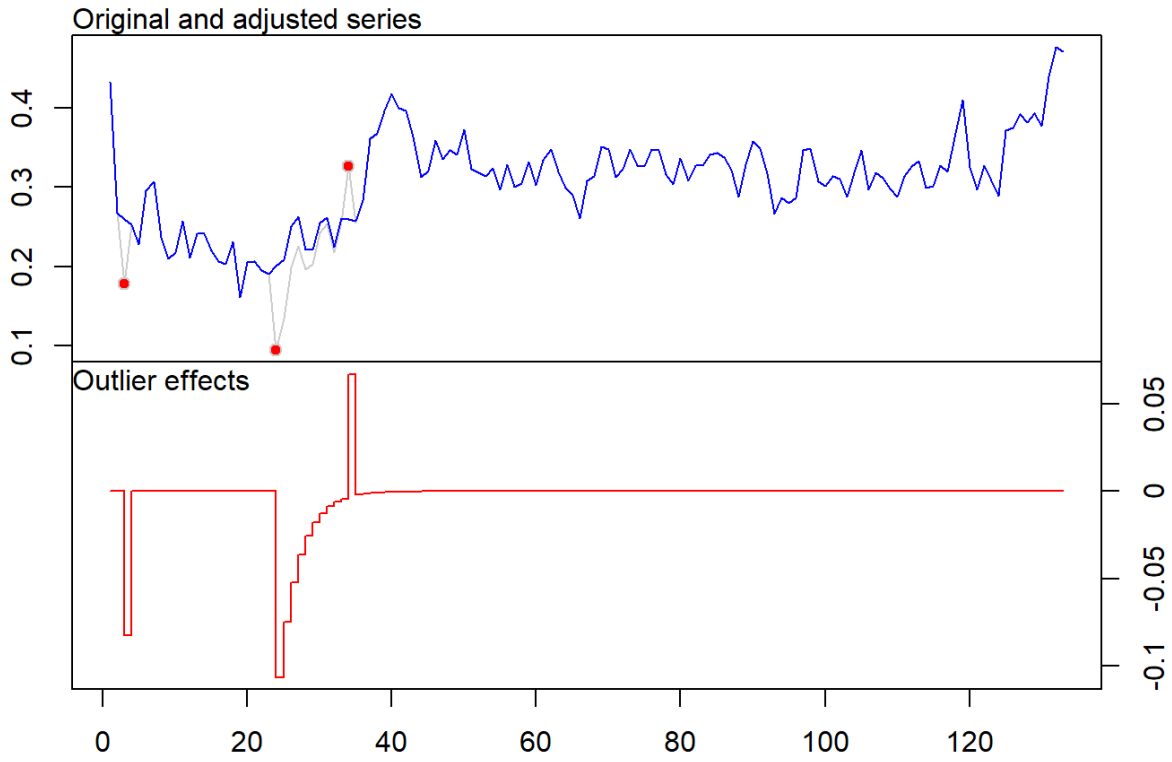

Figure 5. Outlier analysis for Twitter