Title: Data collection & feature extraction appendix for "Toxic politics and TikTok engagement in the 2024 U.S. election"
Authors: Ahana Biswas (1), Alireza Javadian Sabet (1), Yu-Ru Lin (1)
Date: August 20th, 2025
Note: The material contained herein is supplementary to the article named in the title and published in the Harvard Kennedy School (HKS) Misinformation Review.

Appendix A: Data collection & feature extraction

Data collection

To construct a comprehensive dataset of political TikTok content, we employed a multi-step data collection process focused on identifying politically active users and their engagement patterns during the 2024 U.S. presidential election. We first collected TikTok posts originating from the United States that contained the keyword "election" between January and March 2024, yielding over 361,000 posts by around 110,000 users. To refine our dataset, we identified relevant hashtags associated with U.S. politics, particularly those related to the presidential election, as described next. Then, we constructed a hashtag co-occurrence network (based on co-occurrence in video captions) and selected hashtags that appeared most frequently together using network backbone extraction. Next, to ensure a balanced representation of partisan content, we selected nine general election-related (#usa, #election2024, #election year2024, #electionyear, #america, #election2024prediction, #unitedstates, #election, #2024election), eight Democratic-leaning (#biden, #democrat, #biden2024, #voteblue2024, #joebiden, #joe, #biden2020, #democrats), and ten Republican-leaning (#donaldtrump, #trump2024, #trump, #maga, #republican, #trumptrain, #trump2020, #election2024trump, #electionfraud, #gop) hashtags that were among the most frequent in the co-occurrence network. Note that Biden was still the Democratic presidential candidate during this period. Using these hashtags, we identified politically active users (seed users) who post political content—that is, users who had at least three posts each containing any two of the hashtags.

We detected the partisan leaning of the seed users to monitor and limit partisan imbalance across iterative data collection rounds. This helped ensure that the final dataset was not disproportionately skewed in either direction due to our snowball sampling process. To determine the partisan alignment of these seed users, we employed FLAN-T5 large (Chung et al., 2024) to classify three of their posts that included voice-to-text transcriptions. The prompt was to simply classify a post as Republican/Democrat/Neither based on the stance expressed, and we use majority vote (out of 3 posts) to determine the leaning of the user. We were unable to identify the leaning of users for whom the voice-to-text transcriptions (15% of posts had transcriptions provided by API) were unavailable or who had fewer than three posts with transcriptions. Overall, we labeled over 11,000 posts by 3,826 (3.5%) users. Note that these seed users were responsible for around 27% posts in the dataset. To validate the accuracy of our labeling, we manually labeled 100 accounts, achieving a Cohen's Kappa of 0.53, indicating moderate agreement. Initially, we intended to collect videos posted by these politically active users. However, due to limitations in the TikTok API, we were unable to access their original posts. Instead, we collected videos that these users liked (1000 per user due to API restrictions), allowing us to analyze content that politically active users engaged with. To expand the seed users, we identified additional politically active users from the authors of these liked videos. This process was iteratively repeated three times, with diminishing returns in new user identification, suggesting saturation—3.5% new users in the third round compared to 47.5% after the second round.

Our final dataset comprised 4,624 labeled users, with the majority exhibiting a tendency to lean Republican (69%)—reflecting the initial data trends. Due to API constraints, we retrieved liked videos for 683 users (65% Republican-leaning, similar to the initial user leaning distribution), yielding 1,004,654 videos collected from January 1 to November 7, 2024. Using the previously mentioned hashtags—and additional tags like #kamala, #harris, #kamalaharris, #jdvance, #vance, #timwalz, and #walz—we

identified 111,114 political videos containing at least one of these hashtags (11.06% of total liked videos collected), 51,680 (46.5% of political videos) of which were downloadable using a third-party library Pyktok (the remaining videos were likely removed by the user or platform). The mean number of posts during a week (i.e., 7 days) in our dataset was 1,055.8. Note, we were unable to identify the exact reason for the removal of the remaining videos; nevertheless, our dataset reflects the naturally occurring set of publicly accessible political videos on TikTok during the collection period. Despite initial user imbalances and API limitations, our multi-stage, hashtag-seeded snowball sampling method produced a comprehensive video dataset that is balanced in partisan leanings, with downloadable content reflecting an approximately equal split between Republican- and Democratic-leaning videos (39.5% vs 37.1%) and included posts across a wide range of salient political topics and account demographics.

Video feature extraction

We extracted keyframes[1] from videos to capture the most visually representative moments. This is crucial because TikTok videos rely heavily on visual elements to convey implicit messages, evoke emotions, and engage audiences. We inferred user demographics, including age, gender, and ethnicity, and facial emotions such as happiness, anger, fear, and sadness from keyframes, using DeepFace (see Appendix G for manual validation), and RGB color saturation from each keyframe using OpenCV. We aggregated these features at the video level, such as calculating the mean RGB values or the proportions of various emotions and demographics featured, to provide insights into the overarching visual and emotional tone of each video. Figure A1 shows the demographic and video duration distribution in our dataset.

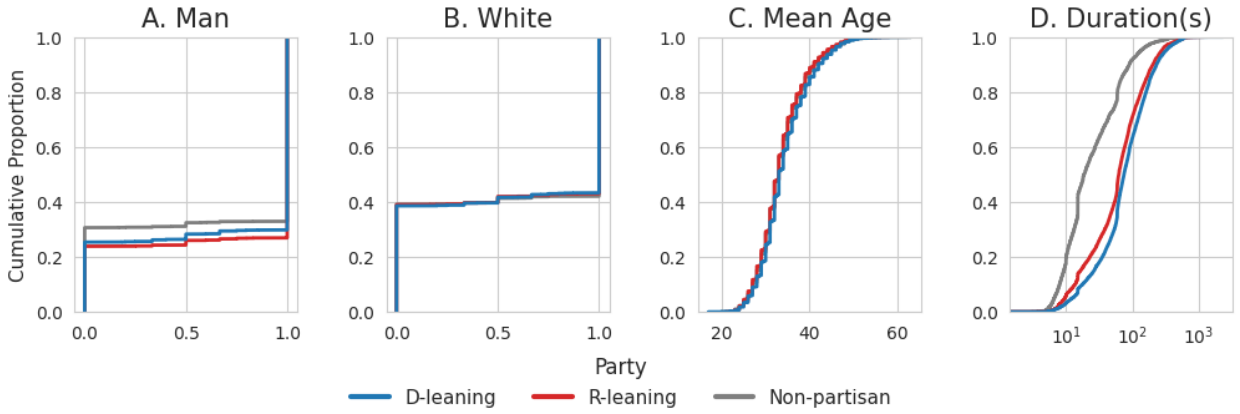

Figure A1. Demographic and video duration distribution for political TikTok videos. *Panel A shows proportion of man, panel B shows proportion of White, panel C shows mean age, as averaged over video keyframes. Our dataset is dominated by videos featuring men and White users, with an average age of 33.9. Panel D shows video duration in seconds, with an average duration of 89.8s.*

Furthermore, we extracted a variety of linguistic features from the video transcripts to analyze how language influences engagement. These features include markers of generalization (Somasundaran et al., 2007) and causation (Boyd et al., 2022) to assess logical structuring, hedges (Islam et al., 2020) and subjectivity (Wilson et al., 2005) to gauge certainty and bias, and emotion-related indicators such as positive emotions, anger, and anxiety (Boyd et al., 2022). These linguistic elements provide insights into

[1] Keyframes refer to a subset of still frames extracted from each video that represent significant visual transitions or scenes. These are used to summarize visual content and enable analysis of attributes like color saturation, facial expressions, and visual composition without processing every frame of the video.

the emotional and rhetorical strategies employed by content creators to engage or persuade their audiences.