Title: Manual validation of toxicity detection 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 H: Manual validation of toxicity detection

To evaluate the validity of toxicity scores used in our study, and to substantiate the claim that full video transcripts better capture toxic content than captions alone, we conducted two manual validation tasks using a sample of TikTok videos from our dataset. Two annotators independently labeled each sample for both tasks, and final labels were decided by mutual discussion.

Validation of perspective API toxicity scores

We randomly sampled 100 TikTok videos and manually labeled each as toxic or non-toxic, based on the primary verbal content in the transcript. Since our study employs the continuous toxicity scores from the Perspective API, we binarized the scores using the median toxicity score in our dataset as the cutoff point to assess agreement with human annotations. The validation yielded an accuracy of 92%, precision 0.92, recall 0.92, and F1 score 0.92. The inter-rater Cohen’s Kappa agreement is 0.79 (substantial agreement). These results indicate high agreement between human judgment and Perspective API predictions, justifying its use for toxicity inference in our analysis.

Validation of transcript vs. caption toxicity comparison

To test our claim in Finding 5—that full video transcripts capture more toxicity than video captions—we sampled 100 random video pairs, each consisting of a caption and a transcript for the same video. A pair was labeled as positive by human annotators if the transcript conveyed more toxic language or sentiment than the caption. This was compared against the Perspective API scores, where a difference of ≥ 0.1 between transcript and caption scores (transcript > caption) was treated as a positive case. The validation yielded an accuracy of 90%, precision 0.89, recall 0.89, and F1 score 0.89. The inter-rater Cohen’s Kappa agreement is 0.81 (near perfect agreement).

Manual inspection of the captions revealed that many consist solely of hashtags, often generic and non-toxic (e.g., #usa, #vote2024, #biden, #maga). These captions lack the linguistic content needed to detect nuanced or hostile rhetoric. In contrast, the transcripts frequently contain substantive political commentary, including personal attacks, emotionally charged statements, and explicit toxicity, which are not reflected in the caption alone.