

## Building an ICCN Multimodal Classifier of Aggressive Political Debate Style: Towards a Computational Understanding of Candidate Performance Over Time

Dhavan V. Shah, Zhongkai Sun, Erik P. Bucy, Sang Jung Kim, Yibing Sun, Mengyu Li & William Sethares

To cite this article: Dhavan V. Shah, Zhongkai Sun, Erik P. Bucy, Sang Jung Kim, Yibing Sun, Mengyu Li & William Sethares (2023): Building an ICCN Multimodal Classifier of Aggressive Political Debate Style: Towards a Computational Understanding of Candidate Performance Over Time, *Communication Methods and Measures*, DOI: [10.1080/19312458.2023.2227093](https://doi.org/10.1080/19312458.2023.2227093)

To link to this article: <https://doi.org/10.1080/19312458.2023.2227093>



Published online: 21 Jun 2023.



Submit your article to this journal [↗](#)



View related articles [↗](#)



View Crossmark data [↗](#)



# Building an ICCN Multimodal Classifier of Aggressive Political Debate Style: Towards a Computational Understanding of Candidate Performance Over Time

Dhavan V. Shah<sup>a</sup>, Zhongkai Sun<sup>b</sup>, Erik P. Bucy<sup>c</sup>, Sang Jung Kim<sup>d</sup>, Yibing Sun<sup>a</sup>, Mengyu Li<sup>a</sup>, and William Sethares<sup>a</sup>

<sup>a</sup>University of Wisconsin-Madison, Madison, WI, USA; <sup>b</sup>Amazon Alexa AI, Seattle, WA, USA; <sup>c</sup>Texas Tech University, Lubbock, TX, USA; <sup>d</sup>University of Iowa, Iowa City, IA, USA

## ABSTRACT



Understanding the implications of aggressive political debate style amid corrosive modes of campaign politics requires fine-grained analyses of political performance, attending to multiple communication modalities. Politicians' facial expressions, emotional tone, and speech content can all independently convey aggression and dominance, and often work in combination for purposes of emphasis. Yet micro-coding individual visual, tonal, and verbal features across more than a handful of debate segments becomes extremely labor intensive, hampering research, especially historical, longitudinal, and cross-cultural work. To address this limitation, we develop a novel multimodal classifier using an Interaction Canonical Correlation Network (ICCN) that incorporates video and audio features with speech coding of candidate debate performance, trained on a 20% sample of 10-second segments from each of the first televised U.S. presidential debates between 1980 and 2020. In the analysis, we demonstrate this classifier can accurately detect aggression by political candidates in U.S. debates. We sharpen its performance by distinguishing between debate eras characterized by lower and higher levels of aggression and validate the approach by comparing the performance of unimodal with multimodal classification. This classifier opens new avenues for computational social science research, including explaining candidate behavior within debates at a larger scale and across different eras.

## KEYWORDS

Computer vision; Dominance displays; Interaction Canonical Correlation Network; multimodal classification; political performance; Trump; presidential debates

Studies of campaign debates and other televised political events have long recognized the power of visuals to influence viewers evaluations of candidates extending back to the original Nixon-Kennedy debates of 1960 (Bucy, 2016; Druckman, 2003; Kenski & Stroud, 2005; Lang & Lang, 1961). Of course, these nonverbal cues work in combination with candidate characteristics, especially candidate gender, to shape voter reactions to facial and vocal features of debate performance (Boussalis et al., 2021). As prior work has shown, on television, candidate movement and behavior, including facial expressions, bodily gestures, level of physical activity, and paralinguistic cues like voice tone and interruptions, become salient, forming a candidate's nonverbal style (Bucy & Stewart, 2018).

Recent behavioral analysis of candidate performances during debates points to social dominance as a key outcome of effective debate performances (Bucy, 2016; Bucy & Gong, 2018; Seiter & Weger, 2020). However, this appears less true for female candidates, who are punished by voters for displays of anger, a clear dominance marker (Boussalis et al., 2021), and embracing a more

**CONTACT** Dhavan V. Shah  [dshah@wisc.edu](mailto:dshah@wisc.edu)  University of Wisconsin-Madison, 5152 Vilas Communication Hall 821 University Avenue, Madison, WI 53706, USA

The first two authors contributed equally to this paper.

Work done while at UW-Madison.

© 2023 Taylor & Francis Group, LLC

physically assertive or “agentic” style of behavior (Everitt et al., 2016). Whereas an effective performance enhances candidate stature, a weak or ineffectual performance diminishes it. A growing body of work documents the influence of candidate performance, demonstrating the importance of considering the full repertoire of communicative channels, including verbal, non-verbal, and tonal markers of candidate behavior, over any one channel in isolation. Multimodal influence not only has the capacity to attract and sustain voter attention but also serves to entertain and distract, emotionally arouse, and generate commentary on social media during real-time events like debates (Shah et al., 2016). As a substantial body of research has shown, nonverbal communication, in particular, conveys social dominance and transmits social cues important to leader evaluation by voters (Seiter & Weger, 2020).

Over the 60 years between the first televised U.S. presidential debates and the most recent encounters between Joe Biden and Donald Trump, a sea change in performative style and candidate aggression has occurred (Alexander, 2011). Changes on stage parallel larger trends in American and European politics toward a more confrontational, “in your face” style of partisan engagement (Mutz, 2015; Norris & Inglehart, 2019). Whereas previous explications of negativity in politics (e.g., Geer, 2006; Herbst, 2010; Jamieson et al., 2017) were able to explain away acerbic rhetorical attacks as a normatively undesirable but attention-getting technique for making issue-based claims, and differentiating and mobilizing partisans, the new negativity is much more polarizing, personal, and confrontational (Brubaker, 2017; Bucy et al., 2020).

Understanding the implications of growing candidate aggression, especially during debates, requires fine-grained analysis of political performance, attending to expressions, voice tone, and content of speech. Despite the contributions of studies that analyze text, voice, or visuals in isolation, multimodal approaches will yield more explained variance than a unimodal approach (Bucy et al., 2020; Shah et al., 2016). Comparing the public conduct of candidates over time and across contents also requires examination of several distinct characteristics at a high degree of temporal precision, given the dynamic nature of political confrontation. Such large-scale coding of hundreds of hours of political debates on a detailed, second-by-second basis calls for leveraging computer vision and multimodal classification techniques that consider visual, audio, and features of speech when gauging the presence of candidate aggression.

Along these lines, Wu and Mebane (2022) present a deep learning framework for constructing multimodal representations for vision and language tasks that can overcome missing observations of text or image features, using modality translation to reduce pretraining demands. We build on this multimodal approach to consider the relationships between semantic, acoustic, and visual properties of aggression using deep canonical correlation analysis (DCCA). Given that semantic features are often derived from validated language processing systems, trained word embeddings, or advanced language models, whereas audio and video features are human-engineered and comparatively underdeveloped, semantic features often outperform non-semantic features in emotion classification tasks (Sun et al., 2019). Yet different communication modalities reveal features of the same moment in different ways. Consequently, we propose a novel computational model, an Interaction Canonical Correlation Network (ICCN) that learns the (hidden) correlations between features extracted from semantic coding and audio features (i.e., text-based audio) and semantic coding and video features (i.e., text-based video).

Moreover, our analytical approach attempts to overcome limitations of current debate analysis, including limits in the amount of coding that can be accomplished manually and the ability to code debates for candidate behaviors or other forms of politically aggressive displays by politicians on a moment-to-moment basis. Thus, this paper advances a deep learning multimodal classifier trained on a large sample of debate segments—20% of each first U.S. presidential debate from 1980 to 2020 coded for visual, tonal, and semantic elements at 10-second increments – and then compare the performance of traditional classification (which uses only a single modality) with the multimodal approach. In the process, we also confirm the sharp rise in multimodal displays of political aggression over time and demonstrate how models can be fine-tuned to address changing debate styles observed

in these data by comparing a one-step (full historical period) with a two-step (models distinguishing historical periods) hierarchical training strategy.

### Aggressive debate performance and multimodal triggers

While popular and even academic accounts of political debates traditionally assumed they were venues for showcasing differing policy views first, and candidate personality and presence second (see McKinney & Warner, 2013; Schroeder, 2008), a close reading of recent research in political nonverbal communication recasts televised and parliamentary debates as competitive contexts in which candidates vie for dominance largely through the use of emotional displays, nonverbal behavior, and strategic use of the body to attract and sustain viewer attention (see Bucy & Stewart, 2018; Koppensteiner & Grammer, 2010; Koppensteiner et al., 2016; Seiter & Weger, 2020). Although carefully staged, the 60- to 90-minute televised debate format reveals how candidates hold up under extended (albeit controlled) questioning and how they perform relative to other contenders. Recently, the split-screen format of debate telecasts has allowed viewers to continuously monitor the arguments, gestures, and reactions of each candidate, heightening access to unfolding drama (Cho et al., 2009; Wicks et al., 2017).

Research has long sought to identify the factors that matter most in citizens' responses to debate content, including what candidates say (Boydston et al., 2013; Cheng, 2020; Stromer-Galley & Bryant, 2011), how they say it (Shah et al., 2016), and the gestures and facial expressions used in communicating intent (Bucy et al., 2020; Druckman, 2003). Recent work extends these insights to examine how presidential candidates' verbal, tonal, and nonverbal behaviors correlate with and predict viewers' "second screen" responses – their use of web-connected devices to express real-time reactions to the debate experience (Chen, 2021). Shah et al. (2016) find that candidates' nonverbal behaviors – their expressions, gestures, and blink rate – are consistent, robust, and significant predictors of the volume and valence of social media expression during debates, especially angry or threatening expressions and defiance gestures. Notably, the only verbal and tonal elements that mattered also reflect an aggressive style, namely rhetorical attacks and an angry or threatening tone, suggesting "subsequent analyses should consider accumulated or conditional effects, such as how an angry or threatening tone works alongside corresponding facial expressions" (Shah et al., 2016, 1838).

A more aggressive pattern of performative style was observed in 2016 (Bucy et al., 2020), where Trump's violation of normative boundaries, particularly those related to protocol and politeness, and open displays of frustration and anger, resonated in the form of social media responsiveness significantly more than Clinton's more controlled approach.<sup>1</sup> Again, the primacy of visual markers of this aggressive style were noted for their power in shaping public responses, though closer examination finds that the tonal and verbal markers of this transgressive style also spurred real-time reactions, further suggesting that visual, tonal, and semantic features work in tandem to signal aggression and dominance.

From the standpoint of gaining attention, Trump's bombastic and unrepentant approach worked. As noted by a growing number of scholars, with the shift to a political style that foregrounds performativity, scholarship must move beyond semantics and rhetorical arguments to assess the influence of gestural meanings and nonverbal communication (Alexander, 2011; Hall et al., 2016; Seiter & Weger, 2020). Trump's wanton disregard for norms of civility and polite exchange was

---

<sup>1</sup>To verify that Clinton's lower level of expressiveness compared to Trump was not an artifact of coding bias or a misapplication of coding criteria due to gender differences between the candidates, we compared our earlier coding of the 2012 presidential debates between Barack Obama and Mitt Romney against the expressiveness of Trump in 2016. To make valid comparisons across election years, behavior measures for Trump (coded at 10-second intervals in 2016) were aggregated to 30-second intervals – the unit of time analyzed for our 2012 coding. Comparisons between Trump and both Obama and Romney show the same pattern of pronounced expressiveness by Trump relative to both male candidates (see Bucy et al., 2020, p. 645), with Clinton largely mirroring the norms of recent Republican and Democratic contenders. Thus, the differences observed in 2016 were due to a sharp increase in expressiveness by Trump rather than a muted performance by Clinton due to gender dynamics on stage.

amplified by active displays of anger, regular interruptions, character attacks, blame language, and attempts to intimidate (Bucy et al., 2020).

### Multimodal coding of aggressive markers

Given the changing combative terrain of contemporary politics and political performance, particularly in U.S., it is important to track markers of this aggressive style and understand its implications for social dominance and audience attention. Content analysis by human coders documenting the visual, tonal, and verbal features of presidential candidates' performance provides insights into understanding how presidential candidates enact an aggressive style through expressions, vocal tone, and specific language cues (Bucy et al., 2020). Yet the time and labor required to produce human-coded data capturing multimodal features of debate performances at scale are daunting, especially visual and auditory characteristics (Joo et al., 2019). Developing machine learning (ML) classifiers from human-coded training sets addresses this issue by 1) reliably classifying the features of debate performances in an automated fashion, and 2) scaling the content analysis findings from the human-coded dataset. When developed and applied consistently across candidates, ML classifiers permit accurate documentation of multimodal markers from candidate performances in debates.

Automated classification of multimodal features in debates allows (a) closer observation of presidential debates on a moment-to-moment basis (e.g., second-by-second visual and audio recognition) rather than at 10-second increments, and (b) deeper examination of a much larger corpus of data (e.g., all 90 minutes of a full-length debate for as many events as occur at the presidential and vice-presidential level in a given election cycle versus the typical isolated case study debate that most studies of debates analyze). Moreover, detection of multimodal markers at each second of a debate uncovers unique patterns of performance that are not easily detectable by human observation (Dietrich et al., 2019; Joo et al., 2019; Kang et al., 2020). When combined with time-series of various forms of continuous response measures that gauge a person's immediate, ongoing reactions to a stimulus such as dial-o-meter response systems, eye-tracking measures, electroencephalography (EEG), and other psychophysiological measures (see Bucy, 2022), automated classification techniques open new vistas for research.

Social media data offer another avenue for fine-grained analysis of candidate behavior, enabling detailed prediction of the influence of aggressive political performances on audience reactions and evaluations. Such approaches entail linking micro-coding of candidate behavior to specific patterns of language production, emotional expression, or discursive shifts to examine real-time dynamics at scale as compared with less consistent and non-scalable human coding of these features (Lukito et al., 2021; Shah et al., 2016). Multimodal machine learning methods are thus needed to further unpack and expand the implications of aggressive performative style in candidate debates and document moment-by-moment responses, whether in the laboratory or on social media.

Computational studies applying machine learning classifiers to automatically capture candidates' performances in debates have thus far mainly focused on analyzing visual markers of behavior, such as facial expressions or body gestures (Boussalis & Coan, 2021; Joo et al., 2019; Kang et al., 2020). These studies have taken different computational approaches to classify facial expressions and body gestures. Kang et al. (2020) trained the classifier to an existing dataset of images with facial expressions (i.e., Expression-in-the-Wild dataset) to classify emotions expressed through candidates' facial expressions. For body gestures, Kang et al. (2020) used OpenPose, an open-source library. Joo et al. (2019) used two approaches to classify facial and body features: 1) relying on the open-source libraries OpenFace and OpenPose to extract facial and body feature vectors, and 2) training a classifier based on human-coded facial and bodily expressions in the 2016 presidential debate. Boussalis and Coan (2021) utilized Microsoft Face API, a commercial machine learning classifier, to extract candidates' facial expressions, and compared the output with human-coding of candidates' expressions (Boussalis & Coan, 2021). The pace of development in this domain means open source (e.g., Dlib, TensorFlow, Caffe, OpenCV) and commercial (e.g.,

Amazon Rekognition, Microsoft Azure Face API, Google Cloud Vision API, Bodytrak) tools innovate and proliferate rapidly.

Automated tools have fostered a growth in scholarship in the computational field toward ML classification of nonverbal performances of presidential candidates – a marked shift from text-oriented computational analysis. However, most studies do not consider the multimodal nature of candidates' performances in debates, limiting their analytical scope to the visual or nonverbal dimension, hindering our understanding of the full impact of the interplay of these elements on candidate performance (cf. Wu & Mebane, 2022). As Liebenthal et al. (2016, p. 7.) note, “compared to written words, spoken words contain additional nonverbal emotional information (i.e., emotional prosody) that is physically and perceptually intertwined with the verbal information.” Such observations highlight the need to consider verbal features in combination with nonverbal features in emotion classification tasks (Sun et al., 2019).

Aggressive political style, in particular, can only be accurately assessed by understanding the interplay between multimodal markers of aggression that together construct these performances. This paper therefore proposes an automated multimodal classification model of aggressive political style in presidential debates that considers the relationships between semantic, acoustic, and visual properties of aggression using deep canonical correlation analysis (DCCA), thereby contributing to the expansion of computational and political communication scholarship. Previous studies have shown the utility of a multimodal approach to investigating visual, tonal, and verbal markers of aggressive political styles (e.g., Bucy et al., 2020; Shah et al., 2016). From a technical perspective, such multimodal language analysis can be framed as a classification problem. Features extracted from each of the three modalities are combined and used as input to a model; the output may be a measure of the degree of aggressiveness of the subject, or it may be binarized into a label that indicates either aggressive or nonaggressive behavior. Recent developments in multimodal language analysis allow researchers to improve classification by learning hidden correlations between text-, audio, and video. We build on this multimodal approach using a novel model, the Interaction Canonical Correlation Network (ICCN), that learns the hidden correlations between extracted features based upon Deep Canonical Correlation Analysis (DCCA) and the proposed embeddings (Andrew et al., 2013; Sun et al., 2019). The approach advanced by Sun et al. (2019) extracts the interaction features of a Canonical Correlation Analysis (CCA) in a DCCA-based network, learning the inherent correlations among visual, tonal, and verbal markers focusing on text and audio (i.e., text-based audio) and text and video (i.e., text-based video).

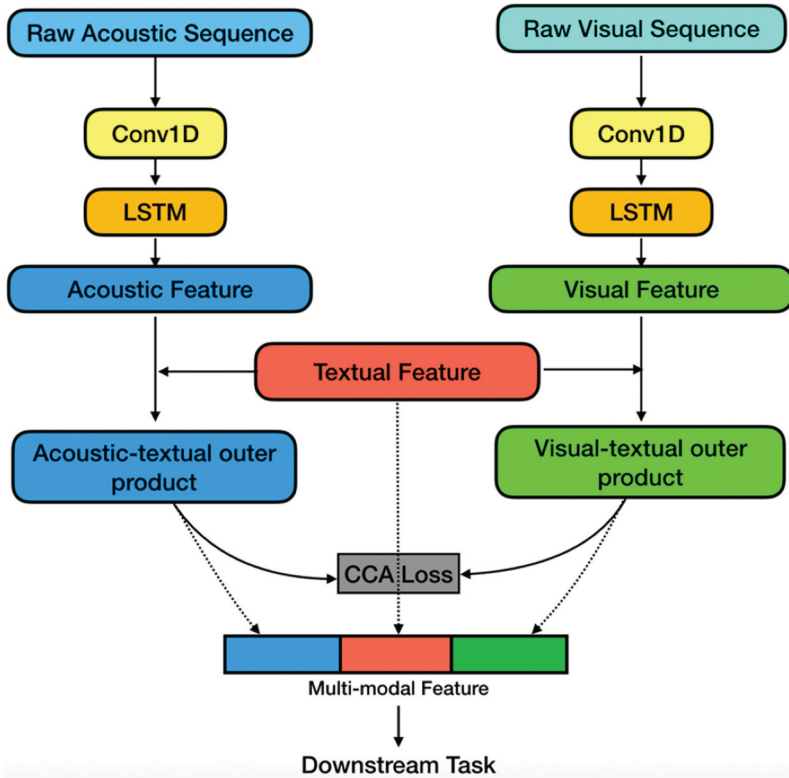
Hidden correlations refer to relationships or connections between variables that are not immediately apparent or easily observable, which only become apparent when more sophisticated techniques such as DCCA and other machine learning algorithms are used to analyze the data (Andrew et al., 2013). Techniques such as feature engineering and regularization used to identify and account for hidden correlations significantly improve the accuracy of an algorithm's predictions. Manual coding of verbal, tonal, and nonverbal aggression within election debates can be used for training. Once trained, the ICCN machine learning classifier can identify aggressive expressions with a high degree of accuracy.

Figure 1 provides a high-level description of our procedure for merging information from the three communication modalities. To start, the audio and video sequences are extracted independently, each with a dedicated convolutional layer aggregated by a Long Short-Term Memory (LSTM), a type of recurrent neural network capable of classifying, processing, and making predictions based on time series data. The output of the LSTMs are the Acoustic Features and the Visual Features, which are combined separately (each in its own path) with the textual sentence embedding through the outer product.<sup>2</sup> These outer products form two-dimensional image-like representations that are each input into a Convolutional Neural Net (CNN). The outer-product matrices of text-audio and text-video are input into the Deep CCA network, that is, the coefficients of the CNNs are trained using the Canonical

---

<sup>2</sup>The LSTM transforms a time series of either raw audio or raw video features gathered over the 10-sec. time interval into a single fixed-length feature vector. A simpler alternative would be to concatenate the raw features into a high dimensional vector for use in the classifiers, as concatenation has a fixed (and finite) memory, while the LSTM is a recursive network that can retain information over a longer period. A comparison between ICCN and the different baseline methods can be found in Sun et al. (2020).



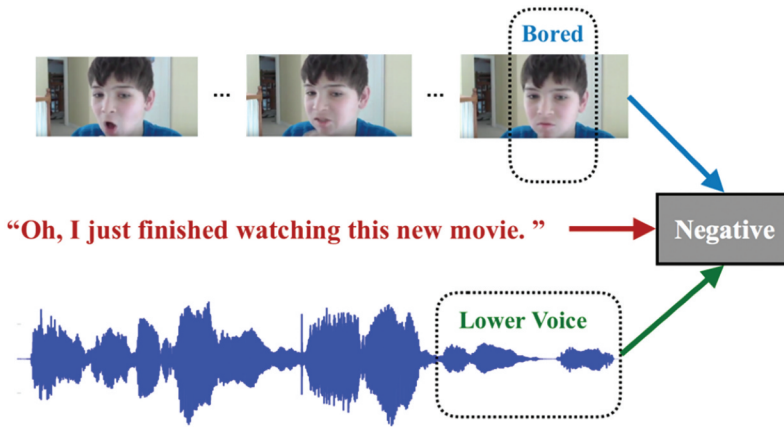


**Figure 1.** ICCN text-based audio feature and text-based video feature alignment.

Correlation loss function to learn a new space where the input representations are highly correlated. Once trained, the outputs of the CNNs can be thought of as “text-aligned audio features” and “text-aligned video features.” These are then concatenated with the sentence embedding to form the full feature vector representing all three modalities of interest. The full feature representation can then be used in downstream tasks such as classification of aggressive behaviors. Code implementing the method described above (will be made available) at the GitHub repository.

Figure 2 provides an illustration of the strength of the method in combining the three modalities. Not relying on any one modality for expressive value provides a robust way to detect behavior. Even if one or more of the modes may be ambiguous (and hence useless in classification), another mode may contain sufficient evidence to allow correct classification. Semantic characteristics may require tonal or visual features for proper classification. The lower voice and bored expression coupled with the addition of the “movie” provides the evaluative information of a negative evaluation that would be missed without all three. Even with the primacy of text in the example, the nonverbal characteristics are essential for classification.

The ICCN approach used here advances work on debate aggression in three ways. First, it takes the interrelationships between text and audio (i.e. text-based audio) and text and video (i.e., text-based video) into consideration and therefore has higher accuracy than a visuals-only, audio-only, or text-only approach. Second, once built, the validated classifier can be used to track and detect aggression throughout a much wider range of election debates – and at a much finer level of detail than human coding. Substantively, such computational tools can help the scholarly community across a range of disciplines better understand the inherent nature of aggressive leadership styles and the mechanisms underlying their appeal.



**Figure 2.** Illustration of features drawn from three modalities for classification.

## Methods

To analyze change in candidate aggression spanning the four-decade period from 1980 to 2020, 20% of each first debate for each election year was coded by human annotators, with the 2016 election debate oversampled due to its higher volume of aggressive expressions, a point we return to below. For this coding, debate videos were first divided into 10-second segments. Within each segment, three key behaviors signaling aggression were documented: angry/threatening facial expressions, angry/threatening tone of voice, and verbal put downs or character attacks. Each candidate was coded separately to isolate their unique behavioral repertoires. The coding scheme for aggression, adapted from previous bibehavioral analysis (see Bucy et al., 2020; Masters et al., 1986; Shah et al., 2016), is summarized in Table 1.

For each of the 11 debates included in this over-time analysis, two continuous 9-minute windows (each representing 10% of the total debate) were selected for coding. To evenly distribute possible sources of influence such as candidate fatigue or debate strategy, these sampling windows were randomly drawn from the early, middle and later parts of the debates. Specific time points for each coded debate window are available from the authors. Altogether, 18 minutes of each 90-minute debate (or about 198 minutes total) were coded for training set data.

Intercoder reliability was established using our in-depth coding of the first 2016 presidential debate. Two trained graduate student coders followed a detailed codebook with variable definitions to

**Table 1.** Aggression Coding for Facial, Tonal and Verbal Expressions.

Category	Description	Operationalization
Facial Aggressiveness	Angry/threatening facial expressions	<ul style="list-style-type: none"> <li>– Lowered eyebrows</li> <li>– A staring gaze</li> <li>– Lower teeth showing with fixed stare</li> <li>– Lowered mouth corners</li> <li>– Lips pressed firmly together</li> <li>– Facial rigidity associated with little to no movement</li> <li>– Overall sentiment expressed is negative and has a hostile feeling</li> </ul>
Tonal Aggressiveness	Angry/threatening voice tone	<ul style="list-style-type: none"> <li>– The tone has a menacing, accusatory, or hostile feel</li> <li>– Impolitely disagree or challenges his/her rival</li> <li>– Making dire predictions about will happen to the country if the opponent is elected</li> <li>– Tone reveals a desire to do political battle or contest action by the opponent.</li> <li>– The overall tone is enraged, feisty bold, and/or aggressive.</li> </ul>
Verbal Aggressiveness	Verbal put-downs or character attacks	<ul style="list-style-type: none"> <li>– Calling the opponent forgetful, unqualified, not having the right temperament, not having enough energy or stamina for the job, or assailing other personal qualities.</li> </ul>



document the presence or absence (1 = present, 0 = absent) of each defined category for each candidate, over each 10-second segment. One coder (who was male) specialized in the verbal variables, while the other coder (who was female) focused on nonverbal and tonal variables.<sup>3</sup> For intercoder reliability, 69 individual segments, or 13% of the analyzed content from the first 2016 debate, were randomly selected at 9 different time points and assessed by a third coder (who was male). Levels of agreement are reported below.

Because the variables were nominal, manifest, and non-normally distributed (showing low variability), percent agreement is reported instead of alpha reliability scores (see Feng, 2015). Although percent agreement does not make allowances for chance agreement, it is appropriate for nominally scaled coding under these conditions. Coding of all three variables used in the analysis showed an acceptable to high level of agreement. For anger/threat displays, agreement was 91.3% for Clinton and 89.9% for Trump. For use of an angry/threatening tone, agreement was 79.7% for both Clinton and Trump. Because character attacks and put downs occurred far less frequently than nonverbal behaviors, every segment featuring a verbal attack was double-coded. Percent agreement was initially a bit low—78.7% for Clinton and 74.1% for Trump. After instances of disagreement were reviewed and discussed between coders, agreement rose to 82.2% for Clinton and 89.6% for Trump. With intercoder reliability established, 20% of all the debate content, as described above, was then manually annotated by the third coder who was now fully trained to identify and accurately document a range of variables. For quality control, 10% of this over-time coding (about 2 minutes per debate) was spot-checked to ensure consistency. No serious inconsistencies were observed.

If aggression was identified across any of the three modalities in any individual 10-second segment, the segment was labeled “1” (aggression); otherwise it was labeled “0” (no aggression). This approach allows plotting the degree of aggression across the elections analyzed. Within the 40-year span, the frequency of aggressive segments is around 0.2 (or 20%) for most election debates, except for 1984 (Mondale/Reagan), 2016 (Clinton/Trump), and 2020 (Biden/Trump). As shown in [Figure 3](#), the small peak in aggression in 1984 is dwarfed by 2016 and 2020. The two most recent elections have witnessed a sharp increase in the expression of aggression in presidential debates to a point where now 0.8 of all coded segments, or around 80% of segments, are identified as containing an aggressive element by human coders.

To construct the multimodal classifier, features are extracted from video and audio files and correlated transcripts, individually coded as containing aggressive features. Facial expressions and audio signals were processed from the video files and the text layer from the correlated transcripts of the debate. [Figure 4](#) illustrates this process of feature extraction from the sampled debate segments using a 2016 exemplar.

For the textual layer, transcripts of segments are input into a pre-trained BERT lower-case model to extract textual features. For each segment, a 768-dimension feature vector is extracted to represent the text. For the acoustic layer, TorchAudio powered by PyTorch is used to extract MFCC and FBANK coefficients as acoustic features. For each segment, the audio is divided into 440 frames, and 36 features are extracted for each frame, resulting in a vector of size  $440 \times 36$ . For the facial features, OpenFace is used to extract facial action features including facial landmarks and facial action units. For each segment, the video is divided into 300 frames, and 35 features are extracted for each of the frames, resulting in a vector of size  $300 \times 35$ . Features generated from different layers are used to construct a multimodal vector used to classify candidates' aggressive behaviors. The clearest feature resulting from this coding is the distinct increase of aggression for debates after 2016, reflecting the same pattern observed in the simply dichotomized plotting of debate segments over time in [Figure 3](#).

---

<sup>3</sup>The codebook was developed for a more in-depth times-series analysis of candidate behavior in presidential debates (see Bucy et al., 2020). Only selected variables are used here.

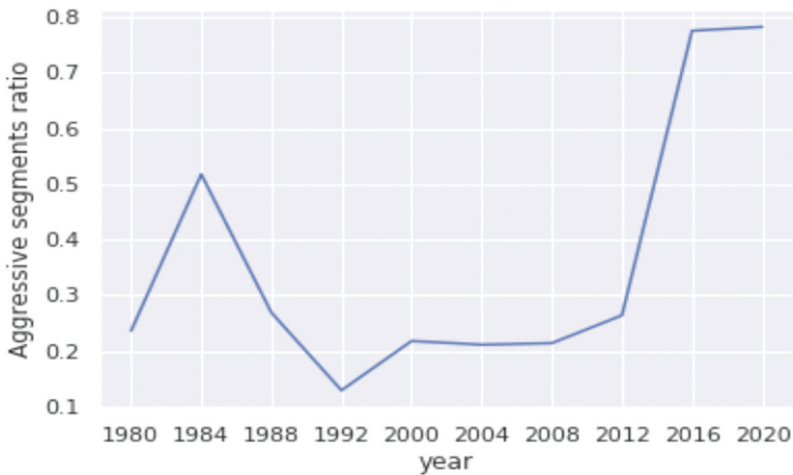


Figure 3. The proportion of aggressive segments in each debate.

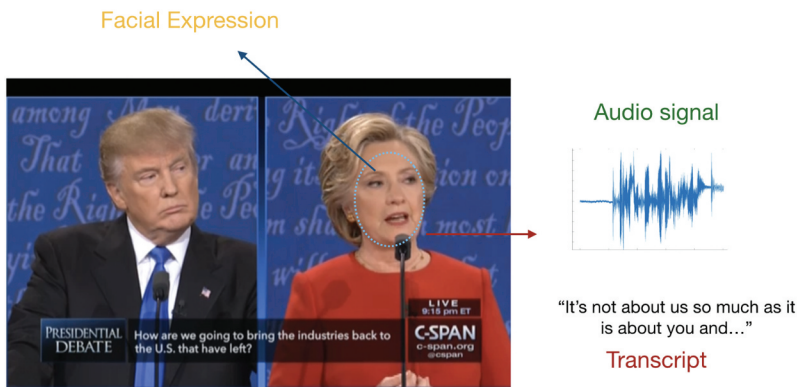


Figure 4. Feature extraction from video files and transcript of debate segment.

Given the observable shift in the degree of aggression in the debates that include Trump, who introduced a new pugilistic televised debate style to American politics – greatly amping up characteristics last seen in much milder form during the Reagan-Mondale debate over three decades earlier – we considered the need for a two-step classification strategy to account for this historic change. A two-step strategy distinguishing unique eras over a singular approach to modeling the entire period has a number of advantages including: (1) allowing researchers to consider disruptive events that subsequently alter the data landscape when working across long time spans, (2) permitting analysts to incorporate non-stationary data, whose statistical properties change over time, into model building, and (3) allowing oversampling of certain classes (in this case, the post 2016 debates) without biasing the classification in the undersampled era. Accordingly, to capture the differences, a hierarchical “two-step” strategy is used to generate parallel models for debates before and after 2016 in an effort to exploit these clear differences, allowing us to contrast a “one-step” or singular approach to modeling the whole period.

In the two-step strategy, the first step builds an algorithm to capture and divide the data into two eras (2–1 and 2–2), representing debates with relatively low and relatively high levels of aggression. The second step is used to predict the aggressiveness of each expression within each era. This two-step

model can then be compared with a one-step model where a single classifier is used throughout for benchmarking purposes, as diagrammed in Figure 5. The two-step approach can be thought of as a way of injecting prior knowledge of the kinds of distinctions likely to shape the classification directly into the network structure, in this case, a shift in the degree of aggressiveness beginning in 2016. Notably, the two-step method is used only for the selection of the network structure and during the training of the networks; it is not used in the evaluation of their performance, providing a conservative test of model comparison.

The hierarchical training strategy shown in Figure 5 begins with the basic multimodal model 1, which is trained to predict whether the segment is from a prior-2016 debate or not. The basic multimodal model 2-1 is trained on the segments prior to 2016 (i.e., 1980 to 2012) in the training dataset, while the basic multimodal model 2-2 is trained on the 2016 and 2020 segments in the training dataset. The coding of all debate segments in 2016 allowed us to compensate for the smaller pool of training data based on the 20% sample. This multimodal classification process is repeated for performance testing.

Figure 6 illustrates the basic multimodal classification model. Audio and video feature sequences are processed by a LSTM while the text vector is processed by a hidden layer. The multimodal vector is formed by the processed individual modalities for input in the final classifier – a pipeline for using ICCN to process multimodal features for the classification.

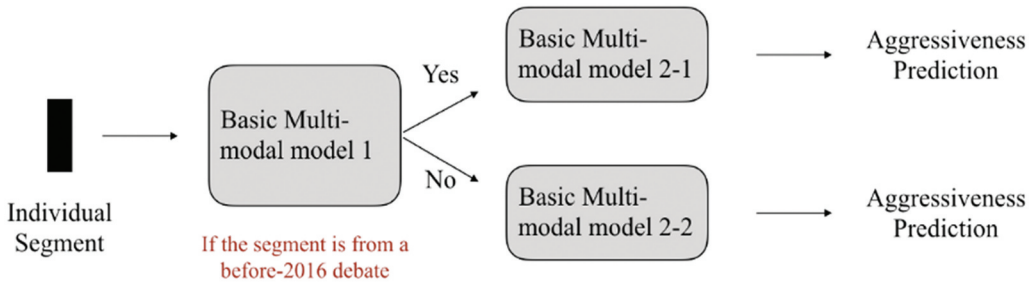


Figure 5. Hierarchical training strategy comparing classification models.

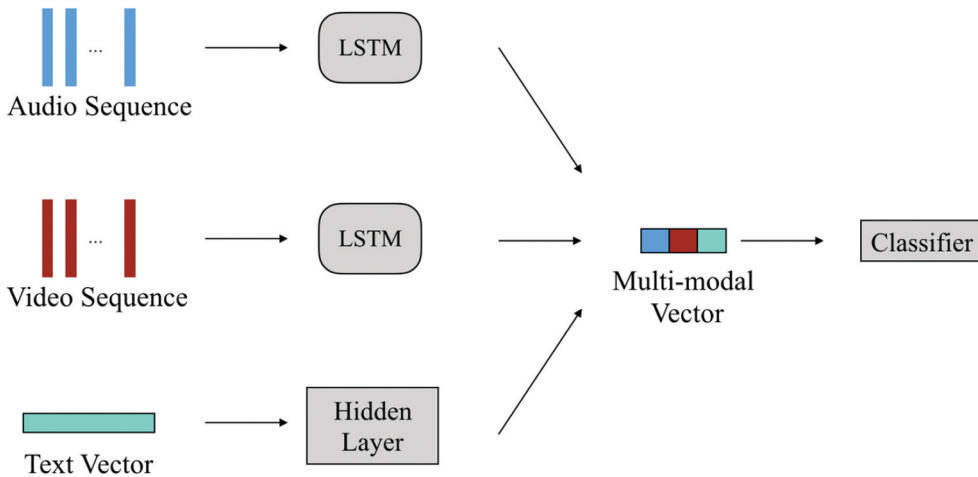


Figure 6. Basic multimodal classification model.

## Results

Findings from the analysis are discussed in two parts. First, we compare the performance of the one-step with the two-step strategy, then the accuracy of the unimodal classification with the multimodal ICCN approach, and discuss the implications for future analysis. Second, we compare the performance of traditional classification (which uses only a single modality) with the multimodal approach. The results show that the hierarchical two-step training strategy is efficient and effective, confirming the shift in aggressive political debate style since 2016 and providing a framework for research spanning long, disrupted or non-stationary data. As we detail below, the Interaction Canonical Correlation Network (ICCN) that captures inherent interrelationships of textual features with accompanying audio and video characteristics performs more accurately than any single modality, improving the overall performance of classification of aggressive behavior beyond other models. When attempting to detect candidate behavior with only a single modality, classification with text alone performed worst, followed by audio and then video. Two findings are especially notable here: (1) that text performed less prominently as a feature predicting aggressiveness than expected, especially prior to 2016; and (2) that even using this approach, the importance of visuals still dominates, generally providing better predictions than the other modalities on their own.

The first model utilizes a basic multimodal approach to predict whether the segment is from a debate before 2016. The performance of the first step is high, with 93.5% accuracy. This shows that expressive styles shift dramatically before and after the 2016 election debates. The aggressive political style that appeared in the two most recent U.S. election cycles is a marked departure from past debates. The first model can also be used in testing future election debates to assess whether aggressive expressions are limited to debates with Trump in them or whether there is a consistent pattern after 2016 and 2020. Both models have the same overall structure: an individual segment is inputted into the classifier and the model predicts aggressiveness as defined by facial, tonal and verbal expressions – features extracted from text and audio (i.e. text-based audio) and text and video (i.e., text-based video). The main difference is that the two-step model allows separate training of the before/after portion of the classifier.

Table 2 helps to summarize and compare the overall performance of the one-step strategy and two-step strategy. Five-fold cross-validation is used to avoid overfitting. The results show that the one-step training strategy achieves a 0.83 aggression prediction F1 score across the whole test set. However, the model performs poorly (aggression prediction's F1 = 0.31, aggression prediction's precision = 0.26, aggression prediction's recall = 0.39) for predicting aggression before 2016. Using the two-step strategy improves the performance. The F1 score increases from 0.31 to 0.66 for the data prior to 2016, with sharp rises in precision (0.26 to 0.72) and recall (0.39 to 0.61). The two-step strategy also improves the prediction performance after 2016 with the aggression F1 increasing from 0.83 to 0.85.

To compare the single-modality classifications to the multimodal approach and analyze which modality contributes most to aggression detection, an ablation study is done that uses the three modalities separately in aggression classification (see Table 3). Results suggest that tonal and visual markers are important to aggression detection in political debates, where verbal control and practiced public presentation center the performance. Combining all the signals and considering the inherent relationships among them improves performance, raising the F1 for the classification of aggression from a low of 0.66 for text-only classification to 0.85 for the multi-modal approach. Moreover, we find

**Table 2.** Performance Summary Using One-step and Two-step Strategy.

	Precision aggression		Recall aggression		F1 aggression	
	<i>1-step</i>	<i>2-step</i>	<i>1-step</i>	<i>2-step</i>	<i>1-step</i>	<i>2-step</i>
Full Period	0.82	0.89	0.84	0.81	0.83	0.85
Prior-2016	0.26	0.72	0.39	0.61	0.31	0.66
2016 & 2020	0.92	0.92	0.84	0.88	0.88	0.90

**Table 3.** Performance Summary Using Unimodal and Multimodal Approach (Two-Step Strategy).

	Text			Audio			Video			Multimodal		
	Prec.	Recall	F1	Prec.	Recall	F1	Prec.	Recall	F1	Prec.	Recall	F1
Full Period	0.59	0.74	0.66	0.80	0.82	0.81	0.73	0.81	0.77	0.82	0.88	0.85
Prior to 2016	0.45	0.56	0.50	0.59	0.72	0.65	0.56	0.69	0.62	0.60	0.74	0.66
2016 & 2020	0.66	0.74	0.70	0.84	0.92	0.88	0.81	0.85	0.83	0.87	0.93	0.90

that textual signals are less useful than audio and video signals, particularly in presidential debates prior to 2016, when decorum demanded a certain level of verbal civility and candidates were more likely to use sanctioned and accepted language during debates, limiting aggressive verbal content. The F1 score for the classification of aggression increases from 0.50 for text-only classification to 0.66 for the multi-modal approach. Throughout both periods, nonverbal features from presidential debates perform well in conveying aggressiveness, likely because such displays are an acceptable aspect of candidate performance during U.S. presidential debates, though the classifier performs especially well after 2016, with the F1 score jumping to 0.90 for the multimodal classifier. The results affirm the importance of the audio and video modalities in the analysis of election debates, especially recent ones.

## Discussion

The implications of this work for improving our understanding of aggression in politics are threefold. First, we have developed a multimodal classifier of aggressive political style incorporating visual, audio, and textual features using a novel ICCN classifier of candidate behavior from U.S. presidential debates that can serve as a baseline for future development of computational tools for political communication research. Second, we have provided evidence of the need to simultaneously attend to visual, tonal, and verbal markers of candidate communication to optimally classify aggressive political style within debate content, with our model incorporating the hidden correlation in text-based audio and text-based video outperforming other models. Third, we have demonstrated the utility of a two-step strategy of model development recognizing an abrupt and fundamentally different pattern of aggression in the 2016 and 2020 debates, which facilitates higher classification accuracy across debate eras. Instead of relying on a single network to derive all relevant information directly from the input data, we segment the network into pieces to be trained independently, thus exploiting our knowledge about the structure of the data to help in the design of the network structure and in its training. The “two steps” can be viewed as an alternative network structure trained in an idiosyncratic manner; the improvements in classification performance clearly justify the approach and suggest future research may benefit from this strategy when dealing with abrupt changes over long time periods, or, more generally, when there are two classes of cases.

Developing this multimodal classifier and observing the patterns revealed in its development reinforces the need for a fine-grained analysis of political performance, one that attends, at a minimum, to candidate facial expressions, tone of voice, and verbal content. Another performative characteristic considered in prior manual coding but not incorporated into this classifier is candidate gestures, which can convey aggression and defiance through hand, arm, and other bodily movements that challenge or disregard authority, express belligerence, or dismiss an opponent (Bucy & Stewart, 2018; Bucy et al., 2020). Future research should incorporate gestural features, which can be extracted from video by pre-trained models like TensorFlow and BodyTrak.

Further analysis would also benefit from analyzing more debates at different jurisdictional levels (i.e., local, statewide, and national) and across cross-national contexts. The aggressive style of politics this classifier is able to capture in U.S. presidential debates is certainly not restricted to the executive level in the U.S., nor to televised debates. The sharp rise in aggression the two-step model reveals a potential stylistic shift toward populist politics increasingly evident in U.S. state-

level politics and across Europe (e.g., Hungary, Poland, Italy, and France), Latin America (e.g., Venezuela, Bolivia, and Ecuador), and Asia (especially India). Indeed, the trend toward populist candidates and their associated political style is a global phenomenon (Moffit, 2016; Norris & Inglehart, 2019), linked to bombast, emotionalization, simplification, and negativity (Engesser et al., 2017).

Whether tied to this rise of populist politics or longstanding features of debate clashes that are now expressing themselves more fully, research would benefit from a broadened understanding of political aggression during candidate debates and other moments of political confrontation. Most of what we know about televised political debates is centered on Western democracies and at the national level. Development of tools for computational classification allows research to expand and scale quickly, fostering the conditions for comparative cross-national and state-level research. Broadening the scope of analysis would also allow for the validation of computational tools in cross-cultural contexts, where there may be differences in how political style is manifested. Developed using open-source tools and benchmarking comparisons, an accurate classifier of political aggression has the potential to provide reliable coding of political behavior at speed and scale, hastening research, increasing the volume of analyzed content, avoiding many forms of investigator bias, and enhancing transparency.

Given the capabilities of the current model, researchers now have a building block for a larger-scale comparative debate analysis over time. Future studies could further aim to develop classifiers capable of analyzing televised debate videos in real time, similar to the continuous response measures that are already in use by political consultants. Of course, this classifier, developed with TorchAudio and OpenFace, is already outdated to the degree that commercial tools, such as Amazon Rekognition and Microsoft Azure Face API, employed within the same ICCN framework, would likely outperform the F1 scores generated using the open-source tools. Although beyond the scope of this research, studies building on this multimodal classification development should engage in a comparison of pre-trained models.

Future research based on this tool, or an improved multimodal classifier could also be used to contrast the expressive differences between conservatives and liberals, incumbents and challengers, male and female politicians, and among candidates discussing different issues, across different time periods and spanning different local, statewide and national contexts (see, for example, Hargrave & Langengen, 2021). Research could expand with the further development of multimodal classifiers designed to reflect more positive sides of candidate performance, including reassuring, affiliative, and affirming displays, which may generate capture when a more relaxed and agreeable style of politics is appropriate and efficacious, especially for female politicians, consistent with work by Boussalis et al. (2021).

Furthermore, our two-step strategy of content classification, which reliably categorized presidential debate behavior into two different eras, points the way toward improving model accuracy and providing temporal nuance in political analysis. With this baseline of comparison, future election debates can now be compared with our benchmark findings to assess how they cluster – either into the more aggressive or more polite era. We could feed the model with new data and test whether and how the aggression shift observed only exists in the two presidential elections involving Trump, or if the level of aggression we observed in 2016 and 2020 is a harbinger of things to come for U.S. politics, as Trump’s style becomes adopted by other candidates. Indeed, an early glimpse of this was seen in the 1984 Mondale-Reagan debate, though at a fraction of the level of aggressiveness on display in 2016 and 2020. More broadly, future analyses might find that the bombastic and aggressive political style documented here has become a global trend, or that the US is a trailing case, borrowing a populist style seen in other countries. The ability of the classifier to integrate visual, audio, and textual modalities creates broader opportunities to examine other forms of political performance, including public addresses, press conferences, ceremonial speeches, and legislative discussions.

The developed model also opens new avenues for continuous response measurement and “dual-screening” research. As mentioned, candidates’ aggressive verbal and nonverbal behaviors during election debates have an impact on viewers’ moment-by-moment responses as gauged via tools such as



dial-o-meters, eye-tracking, and EEG, or via their second-screen interactions on social media (Bucy, 2022; Bucy et al., 2020; Shah et al., 2016). Verbal, tonal, and especially visual markers especially seem to drive the attention economy on social media, increasing mentions of candidates and affecting sentiment. Future research could relate aggressive political style in election debates to deeper, potentially anti-democratic and antisocial effects on social platforms (Neudert et al., 2019). When candidates show aggression during the debate, do their supporters denigrate opponents online? Do aggressive political performances provoke stronger responses from conservatives or liberals? These and other questions can be addressed with the finer-grained analysis and scalable science this type of work advances.

To further open, accelerate, and enhance research across different areas, multimodal research that embraces the open science approach of sharing databases, stimuli for analysis, and analytical techniques should be embraced. As part of this broader project, we have secured a collaborative agreement, in principle, with the C-SPAN Archives at Purdue University to create the C-SPAN Video Library Data Co-op that will make selections of televised debate video, audio, transcripts, along with manual and machine coding of debate features freely available to the broader research community. With the formation of this C-SPAN Data Co-op, we plan to introduce a new interdisciplinary public resource for political scholarship to facilitate greater collaboration between computational and social scientists. The aim of this effort is to produce improved behavioral detection models at a level of precision and scope not possible in previous eras, and to generate new insights related to election debates and other political events that advance our understanding of the dynamics of political communication (Bucy et al., 2022).

Technologies of Natural Language Processing (NLP) are constantly evolving, and newer network structures, training strategies, cost functions, and aggregation techniques are regularly introduced. For example, one might consider replacing the LSTMs in Figures 1 and 3 (which aggregate time series features into a single representative feature vector) with BERT (Devlin et al., 2018) or other transformer-style encoders. The cost function defined by the CCA might be generalized directly to deep CCA (Sun et al., 2019) or changed to some other appropriate loss function entirely. Similarly, the CNNs of Figure 1 might be replaced or augmented by more complex network structures. Other kinds of generalizations of the basic technology suggested here might consider more than just the aggressive element of the discourse and behavior, perhaps generalizing to multi-emotion or other kinds of sentiment classification strategies. As computer vision (and audition) grow to incorporate improved methods of feature extraction, improved network structures, and improved algorithms, we would anticipate that the performance of multi-modal methods will increase. With the release of our curated data sets in collaboration with C-SPAN and with the availability of our code (see Github site for this paper), we hope that our work may provide a benchmark against which further progress can be measured.

Beyond improving the efficiency and accuracy of computer vision detection systems to enable rapid and near real-time analysis of political events by providing a “one stop shop” for researchers interested in multimodal analysis of presidential debates, there are many other benefits to forming a data co-op that an open science approach to research brings. Such benefits include harnessing tacit knowledge that exists within the research community by enabling many minds to engage with similar questions at low cost; improving research practices that lead to difficulties with replication and confirmation of previous results; enabling research areas to grow faster and more efficiently by reducing the amount of duplicate work and wasted efforts that reinvent the wheel; and, fostering a more inclusive and democratic research environment by expanding access to information and lowering barriers to entry (Engzell & Rohrer, 2021).

Over time, the development of a presidential debate Data Co-op could be the beginning of a much broader collaborative research effort. As of this writing, the C-SPAN Archives contain over 271,000 hours of political video including not only coverage of U.S. presidential campaigns, elections, and administrations but also extensive video records of all three branches of American government, special hearings and investigations, impeachments, foreign leader addresses to

Congress, prime minister's questions from the UK, historical documentaries, panels and discussions, specials on First Ladies, African-American History, student leaders, the Civil War, and others covering the gamut of political culture. Over time, the C-SPAN data co-op idea could extend to these and other areas of American political life and bring the open science framework to the analysis of video from different perspectives and approaches at scale.

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the National Research Foundation of Korea Grant funded by the Korean Government, NRF-2016S1A3A2925033, and by the Vice Chancellor for Research and Graduate Education at the University of Wisconsin–Madison through funds provided to D.V.S. Additional support was provided by funding to E.P.B. from the Marshall and Sharleen Formby Regents Endowed Professorship in the College of Media and Communication at Texas Tech University.

## Notes on contributors

*Dhavan V. Shah* is the Jack M. McLeod Professor of Communication Research and Louis A. & Mary E. Maier-Bascom Chair at the University of Wisconsin, where he is Director of the Mass Communication Research Center (MCRC) and Research Director of the Center for Communication and Civic Renewal (CCCR), both in the School of Journalism and Mass Communication. His research centers on (1) the influence of message construction and processing in social evaluations and behaviors, (2) the capacity of mass and interpersonal communication, primarily through online networks, to shape civic engagement, participation, and trust, and (3) the effects of computer-mediated interactions, particularly support expression, on the management of cancer, aging, and addiction.

*Zhongkai Sun* is currently working as an Applied Scientist at Amazon Alexa AI, where his research interest is leveraging large language models to advance conversational AI. His research explores aspects such as reinforcement learning with human feedback for controllable text generation, and the integration of multimodal and knowledge graphs with large language models. Before joining Amazon, Zhongkai Sun received his Ph.D. from the University of Wisconsin-Madison with his dissertation on multimodal language analysis.

*Erik P. Bucy* is the Marshall and Sharleen Formby Regents Professor of Strategic Communication in the College of Media and Communication at Texas Tech University, where he teaches and conducts research on misinformation, visual and nonverbal communication, and public opinion about the press. He is the co-author of *Image Bite Politics: News and the Visual Framing of Elections* (with Maria Elizabeth Grabe) and past editor of *Politics and the Life Sciences*, an interdisciplinary journal published by Cambridge. Bucy is a US-UK Fulbright Scholar and Honorary Fellow of the Mass Communication Research Center at the University of Wisconsin-Madison.

*Sang Jung Kim* is an assistant professor at the University of Iowa in the School of Journalism and Mass Communication. She studies the interaction between technology, politics, and social identity, with particular attention to the mediating role of social media platforms and the spread of information to the public. She explores the identities of message creators and message receivers on social media platforms—including racial identity, gender identity, and political identity—and utilizes both experimental methods and computational approaches to understand how consumers and creators of such content introduce and are impacted by biases.

*Yibing Sun* is a Ph.D. student in the School of Journalism and Mass Communication at the University of Wisconsin-Madison. Her research is centered on the production and effects of visual content within the networked communication environment. Specifically, she seeks to unravel the intricate relationship between textual and visual content and explore how memes are disseminated online using computational methods and experiments.

*Mengyu Li* is a Ph.D. student in the School of Journalism and Mass Communication at the University of Wisconsin-Madison. Using computational and experimental methods, she studies the impact of multi-modal, multi-affordance, and multi-platform social media in shaping (mis)perceptions and driving action in the realms of gender politics and public health.

*William A. Sethares* is a Professor in the Department of Electrical and Computer Engineering at the University of Wisconsin in Madison. He has been a scientific researcher at the Rijksmuseum in Amsterdam and is the Honorary International Chair Professor at the National Taipei University of Technology. His research interests include adaptation and learning in image and signal processing with a special focus on applications of natural language processing to problems in communications, the social sciences, and the arts.

## Data availability statement

For Data Coding, Codebooks and Computer Code:

- **Data Coding and Codebooks of 1976 through 2020 Presidential Debate Videos**

BOX DATA ARCHIVE: <https://uwmadison.box.com/s/kuqqypkcgcwr5miaog5iu855upooqty>

- **Computer Code for the Interaction Canonical Correlation Network Method**

GITHUB REPOSITORY:  
<https://github.com/zsun227/ICCN>

## References

- Alexander, J. C. (2011). *Performance and power*. Polity Press.
- Andrew, G., Arora, R., Bilmes, J., & Livescu, K. (2013). Deep canonical correlation analysis. *Proceedings of Machine Learning Research* 28(3):1247–1255. <https://dl.acm.org/doi/10.5555/3042817.3043076>
- Boussalis, C., & Coan, T. G. (2021). Facing the electorate: Computational approaches to the study of nonverbal communication and voter impression formation. *Political Communication*, 38(1–2), 75–97. <https://doi.org/10.1080/10584609.2020.1784327>
- Boussalis, C., Coan, T. G., Holman, M. R., & Müller, S. (2021). Gender, candidate emotional expression, and voter reactions during televised debates. *American Political Science Review*, 115(4), 1242–1257. <https://doi.org/10.1017/S0003055421000666>
- Boydston, A. E., Glazier, R. A., & Pietryka, M. T. (2013). Playing to the crowd: Agenda control in presidential debates. *Political Communication*, 30(2), 254–277. <https://doi.org/10.1080/10584609.2012.737423>
- Brubaker, R. (2017). Why populism? *Theory and Society*, 46(5), 357–385. <https://doi.org/10.1007/s11186-017-9301-7>
- Bucy, E. P. (2016). The look of losing, then and now: Nixon, Obama, and nonverbal indicators of opportunity lost. *American Behavioral Scientist*, 60(14), 1772–1798. <https://doi.org/10.1177/0002764216678279>
- Bucy, E. P. (2022). Embodied politics and emotional expression in the populist era: Research advances amid a disruptive decade. In K. Döveling & E. Konijn (Eds.), *Routledge international handbook of emotions and media, 2e* (pp. 247–266). Routledge. <https://doi.org/10.4324/9780429465758-16>
- Bucy, E. P., Foley, J. M., Lukito, J., Doroshenko, L., Shah, D. V., Pevehouse, J. C., & Wells, C. (2020). Performing populism: Trump’s transgressive debate style and the dynamics of Twitter response. *New Media and Society*, 22(4), 634–658. <https://doi.org/10.1177/1461444819893984>
- Bucy, E. P., & Gong, Z. H. (2018). In/Appropriate aggression in presidential debate: How Trump’s nonverbal displays intensified verbal norm violations in 2016. In C. Senior (Ed.), *The facial displays of leaders* (pp. 73–95). Palgrave MacMillan. [https://doi.org/10.1007/978-3-319-94535-4\\_4](https://doi.org/10.1007/978-3-319-94535-4_4)
- Bucy, E. P., Shah, D. V., Sun, Z., Sethares, W., Borah, P., Kim, S. J., & Duan, Z. (2022). Detecting Nonverbal Aggression in Presidential Debate: A Demonstration and Rationale for a CCSE Presidential Debate Data Co-op. In R. X. Browning, (Ed.), *Democracy and the Media: The Year in C-SPAN Archives Research* (Vol. 8, pp. 239–253). Purdue University Press. <https://doi.org/10.2307/j.ctv33t5gjm.15>
- Bucy, E. P., & Stewart, P. A. (2018). The personalization of campaigns: Nonverbal cues in presidential debates. In W. R. Thompson (Ed.), *Oxford Research Encyclopedia of Politics*. Oxford University Press. Gen. <https://doi.org/10.1093/acrefore/9780190228637.013.52>
- Chen, H. T. (2021). Second screening and the engaged public: The role of second screening for news and political expression in an OSROR model. *Journalism & Mass Communication Quarterly*, 98(2), 526–546. <https://doi.org/10.1177/1077699019866432>
- Cheng, M. (2020). Acclaims, attacks, defenses: Critical discourse analysis of Ma Ying-jeou’s 2012 Taiwan presidential debates discourse. *Discourse & Society*, 31(1), 19–43. <https://doi.org/10.1177/0957926519877696>
- Cho, J., Shah, D. V., Nah, S., & Brossard, D. (2009). “Split screens” and “spin rooms”: Debate modality, post-debate coverage, and the new videomalaise. *Journal of Broadcasting & Electronic Media*, 53(2), 242–261. <https://doi.org/10.1080/08838150902907827>

- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
- Dietrich, B. J., Hayes, M., & O'Brien, D. Z. (2019). Pitch perfect: Vocal pitch and the emotional intensity of congressional speech. *American Political Science Review*, 113(4), 941–962. <https://doi.org/10.1017/S0003055419000467>
- Druckman, J. (2003). The power of television images: The first Kennedy-Nixon debate revisited. *The Journal of Politics*, 65(2), 559–571. <https://doi.org/10.1111/1468-2508.t01-1-00015>
- Engesser, S., Fawzi, N., & Larsson, A. O. (2017). Populist online communication. *Information, Community & Society*, 20(9), 1279–1292. <https://doi.org/10.1080/1369118X.2017.1328525>
- Engzell, P., & Rohrer, J. (2021). Improving social science: Lessons from the open science movement. *PS: Political Science & Politics*, 54(2), 297–300. <https://doi.org/10.1017/S1049096520000967>
- Everitt, J., Best, L. A., & Gaudet, D. (2016). Candidate gender, behavioral style, and willingness to vote: Support for female candidates depends on conformity to gender norms. *American Behavioral Scientist*, 60(14), 1737–1755. <https://doi.org/10.1177/0002764216676244>
- Feng, G. C. (2015). Mistakes and how to avoid mistakes in using intercoder reliability indices. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 11(1), 13–22. <https://doi.org/10.1027/1614-2241/a000086>
- Geer, J. G. (2006). *In defense of negativity: Attack ads in presidential campaigns*. University of Chicago Press.
- Hall, K., Goldstein, D. M., & Ingram, M. B. (2016). The hands of Donald Trump: Entertainment, gesture, spectacle. *Hau: Journal of Ethnographic Theory*, 6(2), 71–100. <https://doi.org/10.14318/hau6.2.009>
- Hargrave, L., & Langengen, T. (2021). The gendered debate: Do men and women communicate differently in the house of commons? *Politics & Gender*, 17(4), 580–606. <https://doi.org/10.1017/S1743923X20000100>
- Herbst, S. (2010). *Rude democracy: Civility and incivility in American politics*. Temple University Press.
- Jamieson, K. H., Volinsky, A., Weitz, I., & Kenski, K. (2017). The political uses and abuses of civility and incivility. In *The Oxford Handbook of Political Communication* (pp. 205–218). Oxford University Press.
- Joo, J., Bucy, E. P., & Seidel, C. (2019). Automated coding of televised leader displays: Detecting nonverbal political behavior with computer vision. *International Journal of Communication*, 13, 4044–4066. <https://ijoc.org/index.php/ijoc/article/view/10725>
- Kang, Z., Induhara, C., Mahorker, K., Bucy, E. P., & Joo, J. (2020). Understanding political communication styles in televised debates via body movements. In A. Bartoli & A. Fusiello (Eds.), *Computer vision: ECCV 2020 workshops*, Lecture notes in Computer Science (Vol. 12535, pp. 788–793). Springer, Cham. [https://doi.org/10.1007/978-3-030-66415-2\\_55](https://doi.org/10.1007/978-3-030-66415-2_55)
- Kenski, K., & Stroud, N. J. (2005). Who watches presidential debates? A comparative analysis of presidential debate viewing in 2000 and 2004. *American Behavioral Scientist*, 49(2), 213–228. <https://doi.org/10.1177/0002764205279423>
- Koppensteiner, M., & Grammer, K. (2010). Motion patterns in political speech and their influence on personality ratings. *Journal of Research in Personality*, 44(3), 374–379. <https://doi.org/10.1016/j.jrp.2010.04.002>
- Koppensteiner, M., Stephan, P., & Jäschke, J. P. M. (2016). Moving speeches: Dominance, trustworthiness, and competence in body motion. *Personality and Individual Differences*, 94, 101–106. <https://doi.org/10.1016/j.paid.2016.01.013>
- Lang, K., & Lang, G. E. (1961). Ordeal by debate: Viewer reactions. *Public Opinion Quarterly*, 25(2), 277–288. <https://doi.org/10.1086/267020>
- Liebethal, E., Silbersweig, D. A., & Stern, E. (2016). The language, tone and prosody of emotions: Neural substrates and dynamics of spoken-word emotion perception. *Frontiers in Neuroscience*, 10, 506. <https://doi.org/10.3389/fnins.2016.00506>
- Lukito, J., Sarma, P., Foley, J., Abhishek, A., Bucy, E., Doroshenko, L., & Shah, D. (2021). Resonant moments in media events: Discursive shifts, agenda control, and twitter dynamics in the first Clinton-Trump debate. *Journal of Quantitative Description: Digital Media*, 1. <https://doi.org/10.51685/jqd.2021.019>
- Masters, R. D., Sullivan, D. G., Lanzetta, J. T., McHugo, G. J., & Englis, B. G. (1986). The facial displays of leaders: Toward an ethology of human politics. *Journal of Social and Biological Structures*, 9(4), 319–343. <https://doi.org/10.1016/S0140-17508690190-9>
- McKinney, M. S., & Warner, B. R. (2013). Do presidential debates matter? Examining a decade of campaign debate effects. *Argumentation & Advocacy*, 49(4), 238–258. <https://doi.org/10.1080/00028533.2013.11821800>
- Moffit, B. (2016). *The global rise of populism: Performance, political style, and representation*. Stanford University Press.
- Mutz, D. (2015). *In-your-face politics: The consequences of uncivil media*. Princeton University Press.
- Neudert, L. M., Howard, P., & Kollanyi, B. (2019). Sourcing and automation of political news and information during three European elections. *Social Media+ Society*, 5(3), 2056305119863147. <https://doi.org/10.1177/2056305119863147>
- Norris, P., & Inglehart, R. (2019). *Cultural backlash: Trump, Brexit, and authoritarian populism*. Cambridge University Press.
- Schroeder, A. (2008). *Presidential debates: Fifty years of high-risk TV*. Columbia University Press.
- Seiter, J. S., & Weger, H., Jr. (2020). *Nonverbal communication in political debates*. Lexington Books.

- Shah, D. V., Hanna, A., Bucy, E. P., Lassen, D. S., Van Thomme, J., Bialik, K., Yang, J., & Pevehouse, J. C. (2016). Dual screening during presidential debates: Political nonverbals and the volume and valence of online expression. *American Behavioral Scientist*, *60*(14), 1816–1843. <https://doi.org/10.1177/0002764216676245>
- Stromer-Galley, J., & Bryant, L. (2011). Agenda control in the 2008 CNN/YouTube debates. *Communication Quarterly*, *59*(5), 529–546. <https://doi.org/10.1080/01463373.2011.614212>
- Sun, Z., Sarma, P., Sethares, W., & Bucy, E. P. (2019, September). Multimodal Sentiment Analysis Using Deep Canonical Correlation Analysis. Paper presented to the International Speech Communication Association, INTERSPEECH 2019.
- Sun, Z., Sarma, P., Sethares, W., & Liang, Y. (2020, April). Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(5), 8992–8999. <https://doi.org/10.1609/aaai.v34i05.6431>
- Wicks, R. H., Stewart, P. A., Eubanks, A. D., Eidelman, S., & Dye, R. G. (2017). Visual presentation style 1: A test of visual presentation styles and candidate evaluation during the first 2016 presidential debate. *American Behavioral Scientist*, *61*(5), 533–544. <https://doi.org/10.1177/0002764217704317>
- Wu, P. Y., & Mebane, W. R., Jr. (2022). MARMOT: A deep learning framework for constructing multimodal representations for vision-and-language tasks. *Computational Communication Research*, *4*(1). <https://doi.org/10.5117/CCR2022.1.008.WU>