Assessing Affective Polarization Using the Text, Audio, and Video from In-Person, Telephone and Online Interviews

Bryce J. Dietrich,[1*] Jeffery J. Mondak,[2] and Tarah Williams[3]

[1]Department of Political Science, University of Iowa, bryce-dietrich@uiowa.edu

[2]Department of Political Science, University of Illinois, jmondak@illinois.edu

[3]Department of Political Science, Allegheny College, twilliams@allegheny.edu

[*]To whom correspondence should be addressed; E-mail: bryce-dietrich@uiowa.edu

Abstract

Despite the widespread use of telephone surveys, the audio from these surveys has yet to be used for social science research. The same can be said for in-person and online interviews in which answers are often recorded, but no methodology has been developed to assess the way those answers are delivered. In this study, we develop the first automatic emotional speech recognition (AESR) system which can effectively identify the emotional intensity associated with responses obtained from in-person, online and telephone interviews. Using our system and the audio and video from three surveys, we find our multimodal measure of intensity is a statistically significant predictor of vote choice even when additional controls are included. Ultimately, our study dramatically expands the scope of survey research and gives researchers the tools necessary to better use the audio and video from survey responses to answer political questions, laying an important foundation for future research.

Introduction

The image of two individuals sitting silently next to one another, texting rather than talking, has gained iconic status. In a relatively short span of time, exchanges that take place via the written word have come to characterize much of interpersonal communication. People send one another texts and emails, and they post comments on platforms such as Twitter, Instagram, and Facebook. Although this increase in brief written communication brings undeniable efficiency, we all also recognize, whether implicitly or explicitly, that something is lost when we move from talking to texting. Readers of Twitter and Facebook posts often fail to capture the intended meanings of those messages because the written words alone do not convey the writer's underlying state of sarcasm, anger, or elation. Nonverbal expressions, like a change in the tone of a voice or eyes gazing downward when something is said, often signal the communicator's emotions in a manner the written word, even with the slap-dash aid of emojis, rarely can match.

The absence of the spoken word or simply observing how someone acts when they speak may deny us insights regarding a person's emotional state, but the reality is that social scientists interested in the political significance of emotion rarely have capitalized on the richness of nonverbal expressions in their research. We may want citizens to "sound off" so that their "opinions are heard," but we most often measure emotions through text – text that inherently implicates cognitive processes, and omits any meaning communicated through vocal tone or facial expressions. Survey respondents are asked to ponder, recall, and report what emotions they have experienced, a measurement strategy that produces data in which representations of respondents' emotions are mediated by those same respondents' cognitions. People's feelings of happiness, fear, or annoyance then are reduced to a few self-reported numbers on five- or seven-point scales which often fail to capture emotions that occur prior to conscious awareness (McDermott 2007) and can be sometimes compromised by other factors, like one's partisanship (Lodge and Taber 2013) and social desirability bias (Soubelet and Salthouse 2011). However, this is how emotions are typically measured in survey research, and how they necessarily are measured on internet surveys – surveys

on which no one speaks, and no one listens.

The alternate scenario we envision, and that we seek to explore in the present paper, is that survey respondents' politically-significant emotional states can be measured validly, efficiently, and unobtrusively through analysis of those respondents' spoken words. In this study, we use as our raw material audio and video recordings from in-person, telephone and online surveys. In each, respondents are asked to explain what they like and dislike about various presidential candidates, but, unlike previous efforts, we actually listen to and watch the way those words are spoken. This underscores our central methodological contribution – an automatic emotional speech recognition protocol (AESR) which can be more generally employed to analyze audio and video recordings of in-person, telephone and online survey interviews. We then employ this technology to understand how affective polarization influences word choice and subsequent voting decisions through a process called "spreading activation" (Anderson 1983). The convergence of multiple factors related to political and psychological theories of emotion, technological innovations in speech science, and salient features of the contemporary political arena make such an inquiry particularly salient.

The Physiology of Emotion

In research on politics and emotions, information on several aspects of emotional response may be enlightening. First, and most fundamentally, it would be useful to be able to distinguish between an individual's neutral states and states of emotional activation, something previous scholars have shown is often indicative of underlying political opinions (Marcus, Neuman and MacKuen 2000) and predictive of behavior (Dietrich, Hayes and O'Brien 2019). Second, emotional responses should be subject to some basic form of categorization, like a response's positive or negative valence. Third, information about the temporal aspects of emotional expressions is needed if the analyst is to differentiate among a fleeting response, an enduring state (i.e., a mood) and a chronic disposition (i.e., a trait). Fourth, if possible, it would be beneficial to identify the specific emotions

activated under various circumstances and in response to various prompts. It is our contention that all of these ends can be achieved with a focus on the physiology of emotion, and particularly the nonverbal expression of emotion as measured by AESR.

The plausibility of AESR in research on political behavior hinges on whether voice and face signals provide valid and reliable information about emotions and whether means can be devised to enable the systematic measurement of that information. To understand why they do requires that we step back and consider the bases of emotional response. Our perspective emphasizes the role of biology, and especially the thesis that evolutionary forces have given rise to emotional responses that are purposive. This view draws on the foundation established by leading contemporary scholars on politics and emotion, and especially the work of George Marcus and his colleagues (e.g., Marcus, Neuman and MacKuen 2000; Neuman et al. 2007). The Marcus et al. theory of affective intelligence holds that multiple systems of emotion function to direct or manage learning, and to control attention, such as in response to threat. We add one critical point to this view, which is that the motor expression of emotion – i.e., communication via facial expressions, voice, and gestures – also is of adaptive benefit. Indeed, the significance of communication of emotions via facial expressions (e.g., Stewart, Waller and Schubert 2009), vocal expressions (e.g., Scherer 2003), and a combination of both (e.g., Owren and Bachorowski 2007) has been studied extensively outside of political science, but has received less attention within the discipline.

Research in neuroscience establishes that the activation of emotion systems is preconscious (for review, see Posner, Russell and Peterson 2005). Indeed, it is precisely because of preconscious response that emotions serve to control attention. This leads to two practical questions. First, given the neurological basis of emotional response, are there alternate physiological measures that might be preferable to AESR for applied research on political behavior? Second, to what extent, if any, is motor expression of emotion the involuntary consequence of antecedent neurological/physiological processes?

Aspects of social and political judgment, often with focus on emotion, have been studied in

recent years using standard techniques in neuroscience (for review, see Spezio and Adolphs 2007). These include measures of brain function such as functional Magnetic Resonance Imaging (e.g., Greene et al. 2001), Event-Related Potentials (e.g., Amodio et al. 2007), and even variants of the lesion method (e.g., Knoch et al. 2006). Also, several recent studies have examined emotions and politics via physiological measures of blink amplitude and skin conductance (e.g., Smith et al. 2011; Hibbing, Smith and Alford 2014; Fournier, Soroka and Nir 2020; Oxley et al. 2008). We see value in these approaches, but we also view AESR as an important complement. A focus on vocal and facial expression of emotions potentially offers an unobtrusive and efficient means to measure a socially-discernible manifestation of emotional response and to do so in a manner that does not require in-person contact with the survey respondent. These features–that it is unobtrusive, efficient, and that it centers on observable motor expression–make AESR a highly–desirable addition to survey-based data acquisition.

From the perspective of the survey respondent, AESR is invisible, and thus fully unobtrusive. This contrasts starkly with the techniques noted above. With fMRI, the respondent is instructed to remain extremely still, and is posed questions while encased in a confined horizontal space, and while subjected to the deafening roar of the machine's magnet. ERPs are measured with electrodes placed at multiple locations on the respondent's scalp, and the respondent is required to minimize blinking during the procedure. The traditional lesion method and its newer variants require either the availability of patients who have suffered actual damage to a particular brain area or the local disruption of brain function. Blink amplitude is measured with electrodes placed just below a person's eyes, and skin conductance is measured with sensors attached to the respondent's fingers. These techniques all instill a high level of artificiality to the data acquisition process. Further, because all of these procedures except for the traditional lesion method make use of specialized equipment, their application requires that the respondent be brought to a laboratory. Data acquisition is costly and inefficient, making these approaches infeasible at present for large N studies such as telephone surveys and difficult for studies that require respondents from outside of the researcher's locale.

Measures of skin conductance and blink amplitude capture physiological responses over which the individual has little or no control. But is the same true of motor expression? To some extent, people do regulate their facial expressions, voices and gestures. Nonetheless, the emotional cues transmitted through motor expression–and measured with techniques such as AESR and facial recognition applications–emerge primarily as the consequence of physiological changes that are beyond the individual's control, and often beyond the person's awareness. Russell (2003) equates this process to one's body temperature. Even though your body temperature is always present and you can note it whenever you want, only extreme changes become noticeable. However, regardless of the magnitude, changes exist prior to the conscious salience of words such as "hot" or "cold." Russell (2003) argues emotions work in a similar way and can affect behavior prior to conscious awareness by changing the way we process new and existing information.

Darwin (1872) first noted that vocal cues and facial expressions signal the activation of particular emotions. Darwin's focus was on the adaptive roles of these somatic changes. For instance, vocal outbursts reflective of fear tend to be loud, enabling such expressions to serve a socially-beneficial warning function. Initially gaining prominence in the 1930s (see Fairbanks and Pronovost 1939), research on emotional non-verbal cues, like changes in vocal intonations, has made tremendous progress since then both in specifying how emotions trigger physiological changes that ultimately influence nonverbal features and in identifying the specific prosodic qualities associated with various emotional responses. On the first of these points, the process by which emotions influence nonverbal expressions is best exemplified by considering the human voice. Johnstone and Scherer (2000, 222) summarize current understanding:

> (E)motions are accompanied by various adaptive responses in the autonomic and somatic nervous systems. These responses will lead to changes in the functioning of parts of the speech production system, such as respiration, vocal fold vibration, and articulation. For example, with a highly aroused emotion such as rage, increased tension in the laryngeal musculature coupled with raised subglottal pressure will provoke

a change in the production of sound at the glottis, and hence a change in voice quality.

Across scores of studies, convergent evidence has accumulated regarding the patterns in vocal expression associated with both the general activation of emotions and the presence of specific emotional responses. Although dozens of discrete emotions have been examined, the literature is most clear on the prosodic markers that differentiate neutral states from states of emotional activation, and on the vocal characteristics associated with a handful of basic, archetypal emotions. Given the consistent use of vocal pitch as a measure of emotional arousal in the psychology and computational linguistics literature, it has been used the most in the social sciences. For example, drawing from earlier work in social psychology (e.g., Tigue et al. 2012) Klofstad and co-authors (e.g., Klofstad and Anderson 2018; Klofstad 2016) have demonstrated pitch changes can influence vote choices. Similarly, using large-N studies of elite speech, Dietrich and co-authors (e.g., Dietrich, Hayes and O'Brien 2019; Dietrich, Enos and Sen 2019; Dietrich and Juelich 2018) have used vocal pitch as a measure of emotional activation or intensity.

Similar relationships have been found for facial expressions. Due to biological and social pressures (Susskind et al. 2008), facial expressions comprise specific movements (Jack et al. 2012) in a signaling/decoding effort (Schyns, Petro and Smith 2009). Although facial expressions have been linked to specific emotional states (Scherer and Grandjean 2008), considerable evidence has also been found relating facial expressions to emotional activation (e.g., Sato and Yoshikawa 2007). For example, several studies have measured the intensity of emotional expressions using facial features (e.g., Adolph and Alpers 2010). The most notable of this work is on the various types of smiles that people use to convey their underlying emotional states (e.g., Ambadar, Cohn and Reed 2009). In political science, this latter work has made an appearance in the presidential studies literature (e.g., Stewart and Ford Dowe 2013). Here, scholars have found that voters are not only keenly aware of subtle changes in facial expressions, but those changes can also influence their evaluations (e.g., Stewart, Waller and Schubert 2009).

Noticeably lacking from this literature is a multimodal approach that combines text, audio

and image data to classify emotional expressions. We seek to rectify this in the following pages. More specifically, our two-part objective is to (1) develop and implement an AESR procedure that automatically combines these related data streams into a common classifier which can be implemented during mass opinion surveys, regardless of modality. We then (2) demonstrate the utility of our approach through a familiar application, affective polarization and the presidential vote choice. We now turn to the description of procedures we employed, the data we acquired, and the technology developed for the present study.

Data

Data were collected from in-person, telephone and online interviews. Our in-person interviews used a convenience sample of 252 respondents from a mid-sized county in Kentucky and were conducted on the campus of a regional university between 11/11/2012 and 11/15/2012. Respondents were recruited through advertisements both on campus (for these, university staff were specifically targeted; see Kam et al. 2007) and in the local community. We recorded the interviews using digital recorders and Plantronics Supra Binaural headsets which exceed our minimum audio quality requirements.

Two telephone interviews were also conducted, the first of which roughly co-occurred with the 2012 in-person interviews. That phone survey was conducted by a call center on the same campus as the in-person interviews and consisted of 234 respondents. All phone numbers were drawn from the aforementioned Kentucky county, and were randomly dialed. The phone interviews began on 11/11/2012 and ended on 11/30/2012. These phone interviews were recorded on computers equipped with the DLI Personal Logger system, which feeds audio signals directly to the computer's sound card.

Our second telephone survey was conducted by a call center at a large public university between 10/28/2019 and 11/10/2019. All phone numbers where drawn from a mid-sized Iowa county and

were randomly dialed. The purpose of the phone interview was to gauge sentiment towards the 2020 Iowa caucuses, so Democrats were over-sampled. A portion of the sample was then randomly assigned to have their interviews recorded using computers equipped with a Computer Assisted Telephone Interviewing (CATI) system. In total, this sample yielded 183 respondents for whom audio recordings could be obtained.

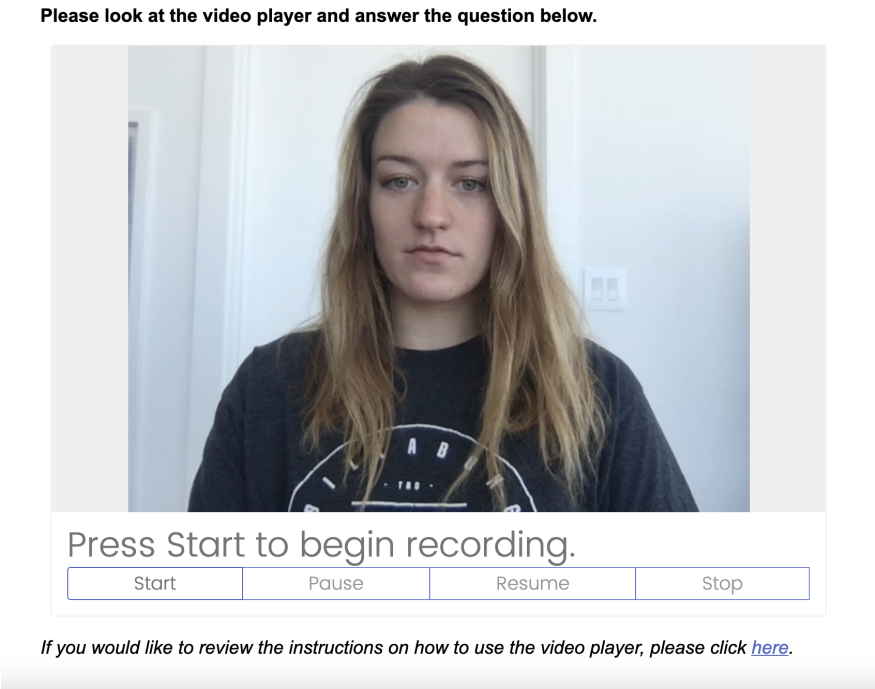
Lastly, video data were obtained from a nationally-representative online survey conducted by Qualtrics surrounding the 2020 Presidential election. The first response was recorded on 10/26/2020 and the last was recorded on 11/24/2020. All responses were obtained using a custom Qualtrics plugin written (in Java) for the purposes of this study, an example of which is shown in Figure 1. Here, respondents were asked to record themselves answering open-ended questions about the main presidential candidates. In total, this sample yielded 332 respondents for whom video recordings could be obtained.[1]

Automatic Emotional Speech Recognition

The AESR system has four main steps: speaker diarization, labeling training data, data augmentation, and fitting a multiple input neural network, each of which is described below. We apply this system to all candidate like/dislike statements as an initial proof of concept and to better understand the relationship between affective polarization and vote choice, something we discuss in the next section.

[1]This study received Institutional Review Board (IRB) approval from three institutions and follows the American Political Science Association (APSA) Council's "Principles of Guidance for Human Subjects Research." Informed consent was obtained for all participants. As explained in the Supplemental Information, numerous precautions were also taken to protect the identities of all participants, like destroying all audio and video recordings upon publication.

Figure 1: Example of Qualtrics Plug-in



Speaker Diarization

Speaker diarization involves partitioning an input audio stream into homogenous segments corresponding to a speaker's identity (Knox and Lucas 2021). In this study, segmenting the speakers was only necessary for the telephone surveys. There, timestamps were provided for the start of the question and the end of the response, meaning the interviewer and interviewee needed to be separated before analysis could begin. For the in-person interviews, segmentation was done manually using Audacity (`http://audacityteam.org/`). For the online survey, respondents recorded themselves, so only one speaker was present in each file.

For this study, we used a modified version of the "call home" model developed by Snyder et al. (2018) which was trained using phone conversations between two people. Prior to fitting the model, a pre-trained support vector machine was used to identify voiced samples (Giannakopoulos 2015). Text was then extracted from the voiced samples using Google's Cloud Speech-to-Text API (for review, see Ziman et al. 2018). Any voiced samples in which the API either (1) did not return

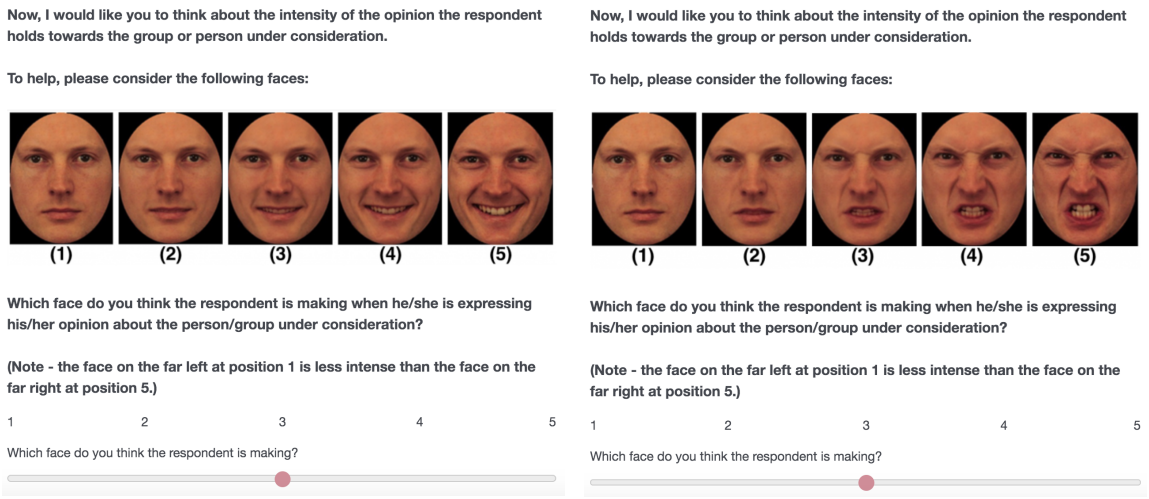
any text or (2) the text that was returned was a one word answer (e.g., "Yes" or "No"), are too short to train the model in Kaldi (Povey et al. 2011). We validated our approach using a random sample of 50 like/dislike statements which yielded 114 audio segments that were voiced and more than a 1-word answer. We correctly separated the interviewer from the respondent in 100 of those audio segments, meaning our algorithm returned an accurate speaker label 87.72 percent of the time. For the analyses below, we combined the segments in which our algorithm identified the respondent as the speaker into a single audio file, yielding one audio file for each like/dislike statement.

Labeling Training Data

Given the relationship between nonverbal expressions and emotional activation, we asked undergraduate coders to label the "intensity" of the response. Not only is this concept closely related to emotional activation (Ethofer et al. 2006), but we also found it was easier for our coders to understand. For each dataset, a 25 percent random sample was obtained and labeled using the prompts shown in Figure 2. Depending on whether the coder thought the respondent was saying something positive or negative about the candidate, they received either smiling (see Panel A) or frowning (see Panel B) faces. They were then asked "Which face do you think the respondent is making when he/she is expressing his/her opinion about the person/group under consideration?" with faces on the the left being less "intense" than the faces on the right. One hundred training files were randomly drawn from each dataset and labeled by all of our coders. Intercoder reliability was assessed using the Interclass Correlation Coefficient (ICC). Since we ultimately randomly assigned two coders to all the files in our larger training set and then took the average, we set $k = 2$ and used a random effects model. This was done using the `ICC` function from the `psych` library in the `R` statistical software language. Ultimately, the ICC was 0.75, 0.78 and 0.85 for our Qualtrics, Kentucky and Iowa data, respectively. All of these results would be described as "average" or "good" reliability (Koo and Li 2016).

Our analysis focuses on candidate like/dislike statements, which are commonly used by schol-

Figure 2: Audio File Annotation



Note: To annotate the audio files, we first asked our coders to assess whether the respondent expressed a favorable or unfavorable opinion towards the person or group described in the audio file. If they said the opinion was favorable, then they received the "happy" scale on the left. If they said the opinion was unfavorable, they received the "angry" scale on the right. The slider at the bottom could take on any value between 1 and 5 which serves as our main variable of interest.

ars to study vote choice (e.g., Zaller et al. 1992), affective polarization (e.g., Levendusky 2018) and online processing (e.g., Lodge, McGraw and Stroh 1989). In our 2012 in-person and telephone surveys, we had 942 and 1,047 of these files. In our 2019 telephone and 2020 online surveys, we had 1,087 and 1,706 like/dislike statements. This resulted in 4,782 files for the purpose of our analysis, of which 1,316 (or 28%) were labeled for training purposes. Two coders were randomly assigned to each file and the file label was the average of their two scores, with the minimum score being 1 (least intense) and the maximum being 5 (most intense).

Multiple Input Neural Network

A detailed outline of the neural network trained for this study can be found in the Supporting Information (SI). Generally speaking, we estimated a multi-input neural network where the text, audio and image data streams were used in conjunction to predict the intensity scores described in the previous section. Prior to a pooling step, separate convolutional neural networks were built for the audio/image data and a recurrent neural network was created for the text data. Although such

an approach is common in computer science, we are unaware of any study in political science that have combined text, audio and video data in a similar way.

Feature Extraction

For the audio model, twenty Mel Frequency Cepstral Coefficients were extracted from each audio file. Speech can be analyzed at the frame- and trend-level. The Mel Frequency Cepstral Coefficients (MFCCs) are an example of the latter since they provide a summary of the energy distribution at specific frequencies for the entire speech signal. Ultimately, they are returned in the Mel scale which relates the perceived frequency (or pitch) to the actual measured frequency. Since the vocal tract is manipulated to change the perceived frequency (or pitch), the MFCCs essentially capture the shape of the vocal tract which is why they are frequently used for audio classification tasks (for review, see Desai, Dhameliya and Desai 2013).

For the text models, we first generated automated transcripts using the aforementioned Google Cloud Speech-to-Text API (for review, see Bokhove and Downey 2018). Similar to what was done in our diarization model, these transcripts were first used to filter out one-word answers. Once this was done, word embeddings were then created from the text of each statement. Word embeddings are commonly used for text classification, especially in political science (e.g., Rheault et al. 2016). For example, Rheault and Cochrane (2020) use word embeddings with augmented political data to produce ideological scalings for members of the British, Canadian and American legislatures. Similar to these authors, we implemented our model using the default hyperparameters proposed by Mikolov et al. (2013). The only departure is the choice of window size which we set at ± 10 words due to some relatively short like/dislike responses.

For the image models, images must be taken from each video. This was done using key frame extraction, which is a process by which a video is summarized using some number of frames. Generally speaking, motion detection (Liu, Zhang and Qi 2003) and visual descriptors (Gianluigi and Raimondo 2006) are the main ways key frames are extracted. We use an open source tool

Figure 3: Examples of Key Frame Extraction and Facial Cropping with Haar Cascade

(a) Key Frames

(b) Key Frames Cropped

Note: This plot shows how we converted the Qualtrics videos into storyboards for the purpose of analysis. In Panel A, we provide example output from the Katna algorithm, implemented in Python. Facial cropping is shown in Panel B. This was done in OpenCV using a pre-trained Haar Cascade.

implemented in Python called Katna,[2] which combines these approaches with another clustering algorithm. Using OpenCV[3] and a pre-trained Haar Cascade[4], the resulting key frames were then cropped to focus on the faces of each respondent. Figure 3 shows an example of the final 2x2 image that was ultimately used for classification. Additional details can be found in Section S2.1 of the SI.

Data Augmentation

Similar to previous studies in computer science (for review, see Perez and Wang 2017), we augmented our training data to improve our out-of-sample performance. For the audio data, we created a version of the training data that included random white noise which aimed to simulate differences

[2] https://pypi.org/project/katna/

[3] https://docs.opencv.org/4.x/index.html

[4] https://github.com/opencv/opencv/blob/master/data/haarcascades/haarcascade_frontalface_alt2.xml

in phone quality. For the same reason, another version of the training data was created where the volume was randomly increased and decreased. The image data were augmented in a similar way, with one augmented dataset randomly increasing and decreasing image quality, while the other increased/decreased the brightness. Finally, our text data were augmented using WordNet and Word2Vec. More specifically, we know words are often textually different, but semantically the same. For example, the statement "I hate football" is synonymous with the statement "I despise football," but textually they are distinct. Given that, we created two additional versions of our training data where semantically similar words (excluding stop words, like "is" and "the") were imputed using distance metrics calculated from both WordNet (Miller 1995) and Word2Vec (Goldberg and Levy 2014), the latter of which was trained using the Google News archive.[5]

Model Performance

Our model was trained for 2,500 epochs using a batch size of 8. In machine learning, the number of epochs controls the number of times the algorithm passes through the entire training dataset. Generally speaking, there is no optimal number of epochs, instead researchers should track the validation error. If it continues to go down (see Figure 4), then allowing the algorithm to train for a large number of epochs is acceptable (Williams, Casas and Wilkerson 2020). The batch size is the number of training examples the model uses to update the internal parameters. Again, there is no optimal batch size, but smaller batch sizes helps prevent overfitting (Wilson and Martinez 2003), which is why we used 8 for the purposes of our study.

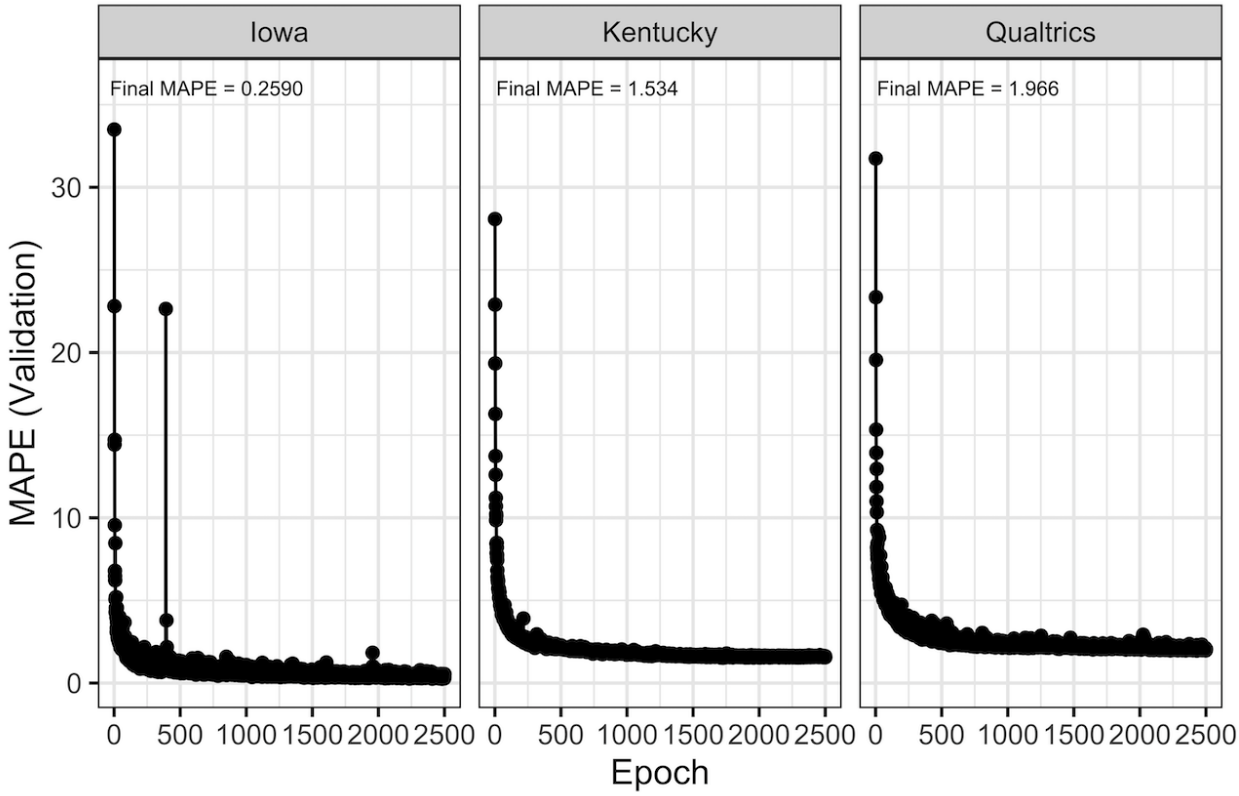
With these hyperparameters defined, we trained our model using the mean absolute percentage error (MAPE) as our loss function, which is defined as:

$$\texttt{MAPE} = \frac{100}{n} \sum_{t=1}^{n} \left| \frac{A_t - P_t}{A_t} \right| \tag{1}$$

where A_t is the actual value and P_t is the predicted value. Their difference is divided by A_t, then

[5] https://code.google.com/archive/p/word2vec/

Figure 4: Model Performance



summed for all fitted points (n). This result is then multiplied by 100 over n, yielding a measure of average model performance with *higher* values indicating *lower* performance.

MAPE is one of the most popular accuracy measures for continuous variables and has been used to assess emergency room admissions (Boyle et al. 2012), wind speeds (Prema and Rao 2015), and, perhaps most importantly, online sentiment (Bollen, Mao and Zeng 2011). MAPE is often used in practice because it is very intuitive and adaptable to a number of applications (De Myttenaere et al. 2016). It is also scale-independent making it useful when comparing models from different datasets (Byrne 2012). Although alternatives have been proposed (for review, see Davydenko and Fildes 2016), MAPE is still recommended in most statistical textbooks (e.g., Bowerman, O'Connell and Koehler 2005). Finally, one of the main criticisms to MAPE – its difficulty in predicting values that approach zero (Kim and Kim 2016) – does not apply to our data since the minimum label is 1, making this measure particularly useful for the present study.

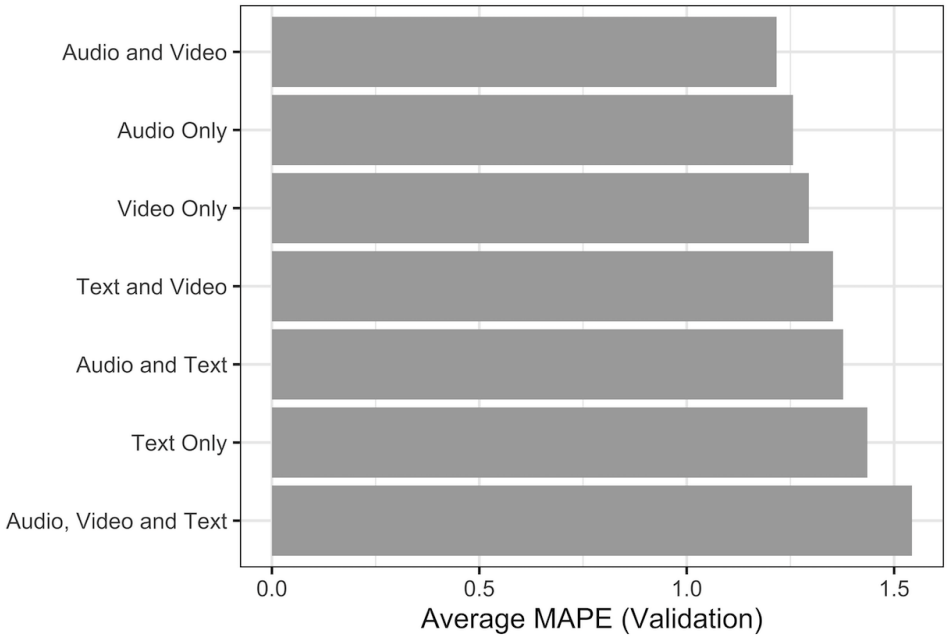
Figure 4 shows the validation MAPE for each epoch of training. Ultimately, the model performed best using the Iowa data and worst for the Qualtrics data, although there was only a slight difference between the performance of the Qualtrics and Kentucky models. For example, the final MAPE for our Iowa data was 0.259, meaning that, on average, whatever score the model returned would be less than ± 1 percent different than the actual score. Thus, if the actual intensity score was 1, then the model would return (on average) a value between 0.997 and 1.003.[6] Similarly, if the actual intensity score was 5, the model would return a value between 4.987 and 5.013. For our Qualtrics model (our worst performing model), if the actual intensity score was 1, then the model would return (on average) a value between 0.980 and 1.020. And, if the actual intensity score was 5, then the model would return (on average) a value between 4.902 and 5.098, suggesting all of our models performed very well.

Variable Importance

Since our Qualtrics survey is the only one to include audio, images and text, we use this data to assess variable importance by estimating models with different variable combinations. For example, the "audio only" model was only trained using the audio data we collected. Results are reported in Figure 5. To account for randomness during optimization, we report the average validation MAPE from five runs of the corresponding models with *lower* values implying *better* performance. Beginning with the audio, text and video data in isolation, we find the audio model performed the best with an average validation MAPE of 1.26, followed by the video model (1.29) and ending with the text model (1.44). A similar result is found when the text data is used in conjunction with either the audio or video data. More specifically, when text is added to the audio data the average validation MAPE increases to 1.38 as compared to when the audio data is used in isolation. A similar decrease in performance is found when text is added to the video data (average validation MAPE increases to 1.35). Finally, the average validation MAPE is lowest when the audio

[6]These numbers were derived by dividing 0.259 by 100 and multiplying it by the indicated value (1), then adding and subtracting that result from that same value (e.g., $1 - 1(\frac{0.259}{100}) = 0.997$).

Figure 5: Model Performance Declines When Text Data is Included



Note: On the x-axis, we report the average validation MAPE for five runs of the models listed on the y-axis. For each of these models, we list the data that was used for estimation. For example, the "Audio and Video" model uses both the audio and video data from the Qualtrics survey, whereas the "Text Only" model only uses the text data.

and video data is used without the text data (1.22), suggesting the latter is least important to our model's performance.

Given the better performance of the audio and video data, we next determine what portions of these data were most influential. Said differently, when our model looked at the video frames, were certain portions of the frames considered more than others? Similarly, were some parts of the audio signal given greater weight? To gain traction on these questions, we use Gradient-weighted Class Activation Mapping (Grad-CAM). Originally developed by (Selvaraju et al. 2017), Grad-CAM imports the gradients of a given label into the final convolutional layer to produce a localized map highlighting the portions of the image that are important for predicting the concept. Such mapping is not only useful to verify that the fitted model is "looking" at the portions of the image we would expect based on theory, but it can also yield insights into what image features are important for classification. Figure 6 shows one of these maps for our video only model. Here, the region that is considered the most in fitting this model is centered around the respondent's face which is

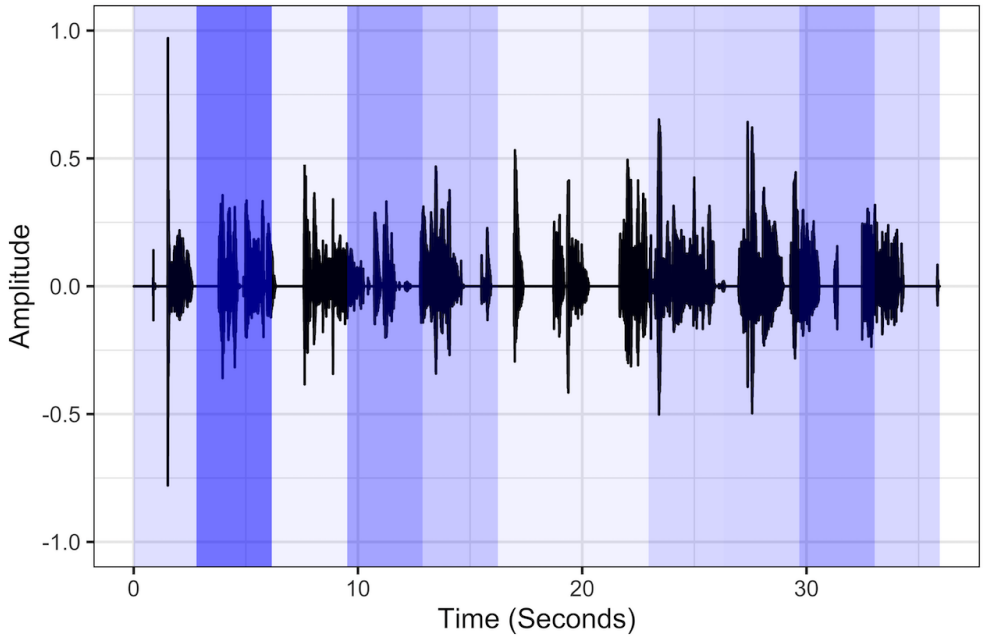
Figure 6: Gradient-Weighted Class Activation Map (Video Data)

Note: This figure shows the results from our Gradient-Weighted Class Activation Mapping. Brighter colors (e.g.,) imply that the video model is focusing more on that region when estimating, whereas darker colors (e.g.,) mean that region is less important to estimation.

consistent with previous literature (e.g., Bucy 2000). We also find that within the facial region, the respondent's eyes seem to be relevant, which is consistent with literature on gaze direction (e.g., Adams Jr and Kleck 2003). Finally, the lip region also seems to be somewhat relevant, which supports previous literature on the importance of smiles to emotional classification (e.g., Stewart and Ford Dowe 2013).

For the audio only model, we create a similar mapping. This is shown in Figure 7. Here, darker regions imply that portion of the audio signal is more influential when the model is being fit. However, this shading is indicative of specific time stamps, rather than the features themselves. For example, the first light blue region refers to the 0-2.80 seconds of the audio data. It does *not* refer to the corresponding change in amplitude. With that said, our mapping shows the model tends to focus on a beginning, middle and ending section with the beginning section being the most important for evaluation. This is consistent with the way humans process auditory signals (for review, see Szabó, Denham and Winkler 2016). Generally speaking, this is broken into two stages: (1) incoming sounds are grouped into general categories and then (2) within those categories further

Figure 7: Gradient-Weighted Class Activation Map (Audio Data)



Note: This figure shows the results from our Gradient-Weighted Class Activation Mapping. Darker (e.g.,) and lighter blues (e.g.,) imply some portions of the audio stream were weighted more and less during fitting, respectively. On the y-axis, we report the amplitude, whereas the x-axis reports the time in seconds. The amplitude is not considered in our mapping. Rather, the mapping can only show which portions of the audio (in seconds) are more or less relevant for classification purposes.

revisions are made. The pattern revealed by our mapping seems to follow such a process, with our model initially deciding whether an audio signal is high/low and then placing the audio signal within that broader demarcation.

An Application: Spreading Affective Polarization and Presidential Vote Choice

Background and Theoretical Expectations

Research has increasingly shown that "Democrats and Republicans both say that the other party's members are hypocritical, selfish, and close-minded, and they are unwilling to socialize across

party lines" (Iyengar et al. 2019, 130). This affective polarization has been shown to impact interpersonal relationships (Iyengar, Sood and Lelkes 2012), where people choose to live (Gimpel and Hui 2015), perceptions of the economy (Healy and Malhotra 2013), health behavior (Bavel et al. 2020), political participation (Huddy, Mason and Aarøe 2015; Iyengar and Krupenkin 2018) and voting (Ahler and Sood 2018). In this application, we expand on this literature by utilizing the theory of spreading activation to understand how affective polarization may permeate through a respondent's answer and help explain their vote choice.

Collins and Loftus (1975) proposed a model of spreading activation in semantic networks stating that information (e.g., raspberries) is stored in memory within larger conceptual networks (e.g., fruit) which are tied by bi-directional links. The strength of the associations varies between nodes within a conceptual network, with some connections being quite strong (e.g., raspberry and blackberry) while others are quite weak (e.g., raspberry and automobile). Activation of any given node then will spread along associative links to other related nodes, revealing the larger conceptual network. Thus, the speed of spreading activation is determined by the strength of the associative links between nodes and the strength of the initial activation node.

Both neuroimaging (De Zubicaray et al. 2001) and electrophysiological (Hirschfeld et al. 2008) studies have found that the process of spreading activation, especially in relation to speech production (Kormos 2014), is a biophysical response in which certain concepts are primed, making neighboring concepts more accessible in memory (for review, see Nozari and Pinet 2020). Although this is similar to how previous scholars in political science have conceptualized emotion (e.g., Marcus, Neuman and MacKuen 2000), it is distinct in that emotions are not considered static. Instead, during the course of a response, an individual's initial reaction (and resulting word choice) influences subsequent reactions, ultimately producing the response that is uttered (Hatfield, Cacioppo and Rapson 1993).

Using this model, we will determine when spreading activation occurs while respondents are explaining what they like/dislike about the major presidential candidates in their respective elec-

tions. Although there are a number of ways to operationalize spreading activation, we consider the extent to which the intensity of one word influences the intensity of another. To generate these measures, we used a word-level version of the aforementioned neural network in which the audio files of each response were first aligned with the text using the Ochshorn-Hawkins algorithm (Ochshorn and Hawkins 2017). Once done, we then extracted audio and images for each individual word, yielding word-level intensity scores for all words uttered in every like/dislike response.

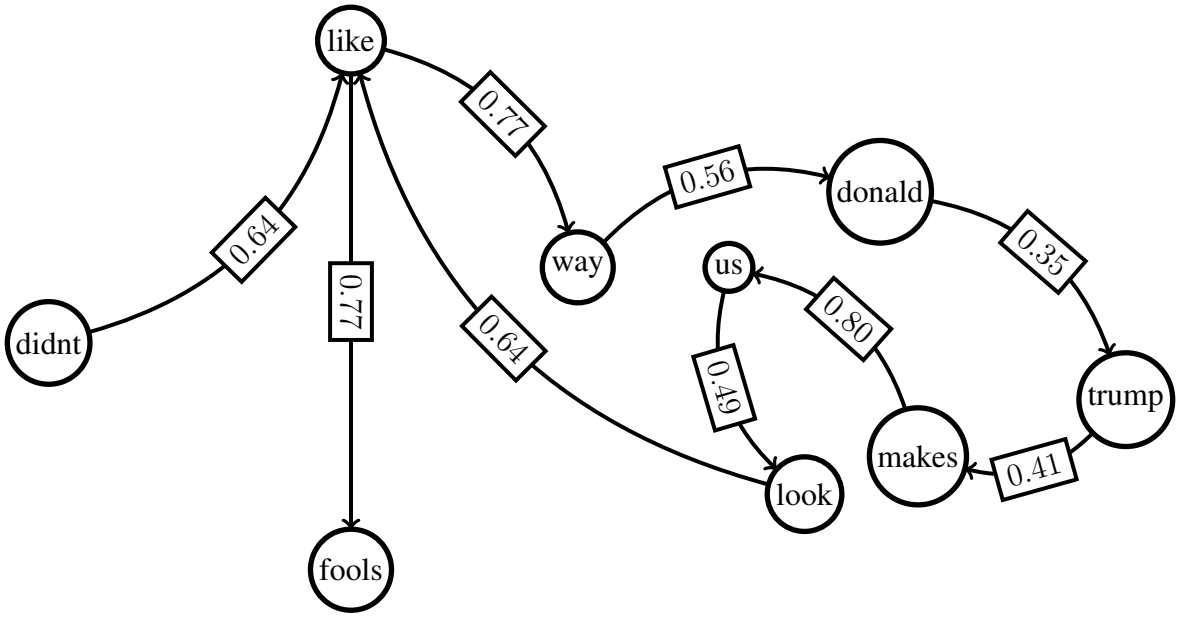
Using these word-level measures, we then test three main theoretical expectations. First, we expect activation to spread faster when partisans are discussing what they dislike about the opposing party as opposed to what they like. This is because the semantic network surrounding their articulation of out-party dislikes is more likely to be densely connected. Second, we expect the inverse to be true when partisans are discussing what they like about an opposing candidate. Since this concept is unlikely to be central in their semantic network, it will be less accessible, making it more difficult for spreading activation to occur. Finally, these patterns will be predictive of vote choice. For example, respondents who display the most spreading activation while explaining what they like about an opposition candidate are precisely the respondents who are most likely to vote against their own party.

Operationalizing Spreading Activation

We use social network analysis to measure the extent to which intensity spreads during the course of a respondent's answer. Figure 8 provides an example of one of the graphs we used to create our measures. Here, we provide the first ten words of a response (excluding stop words, like "is" and "the") to the question "What do you dislike about Donald Trump?" Using the raw intensity scores associated with each word, we first calculated the extent to which intensity increased from one word to the next, represented as a percent increase. For example, the intensity score associated with the word "fool" (5.17) is around 12 percent higher than the intensity score associated with the word "like" (4.61). To ensure all weights are greater than zero, we next standardized these percents

to range from the minimum (0) to maximum (1) for each response. In Figure 8, these standardized weights are shown for each bigram edge, with values greater and less than .50 being associated with intensity increases and decreases, respectively. Additional details are provided in the SI.

Figure 8: Example of Weighted Directed Graph Created for Each Response



Note: Each node is one of the first ten words in a response (excluding stop words) to the question "What do you dislike about Donald Trump?" The initial response was "I didn't like the way Donald Trump makes us look like fools." Edges are directed with the arrow indicating word order. Standardized weights are printed on each edge. These weights can range from 0 to 1 with values greater than .5 indicating intensity increased from one word to the next, whereas values less than .5 indicating an intensity decrease.

Using these graphs, we calculated the shortest path between all weighted edges using Dijkstra's algorithm (Dijkstra et al. 1959), with larger values implying greater distances. After all the paths were calculated, we then took the average for each response graph, yielding a measure that is higher when the word-level intensity tends to increase from one word to the next. Previous literature suggests such a pattern is indicative of spreading activation since is implies that the word-level intensity is permeating throughout the network. Conversely, spreading activation is less likely to occur when a respondent begins their answer with very little intensity and that does not change as they explain more about what they like or dislike about the candidate in question. In this instance, the average path length will be the shortest since the word-level intensity starts relatively low and

remains constant for the duration of the response.

Results

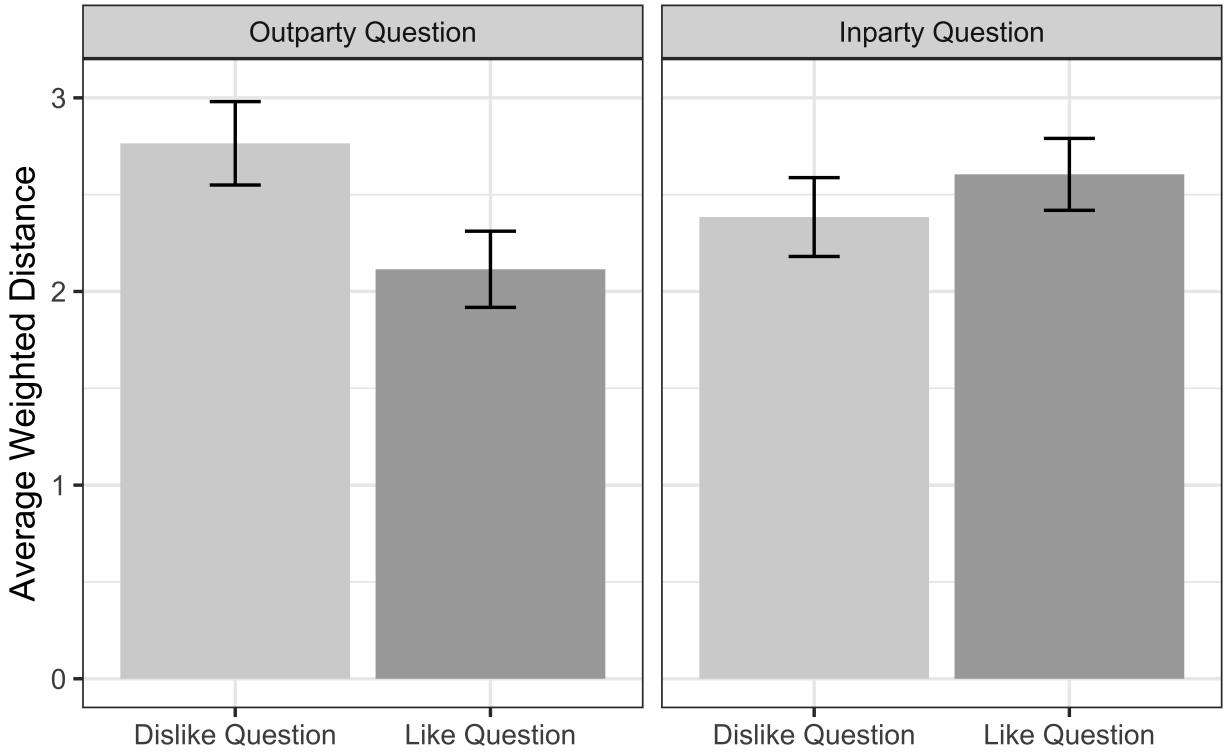
How Does Spreading Activation Influence The Structure Of A Response?

We begin by determining when and where spreading activation is most likely to occur when respondents are explaining what they like or dislike about presidential candidates. By definition, affective polarization centers around negative outparty sentiment (Druckman and Levendusky 2019), so we expect to find more evidence of spreading activation when respondents are explaining what they dislike about opposing party candidate, as opposed to what the like. More specifically, we should see the *highest* amount of interconnectivity among respondents' answers when they are explaining what they dislike about the opposing party's candidate, and we should see the *lowest* amount of interconnectivity among respondents' answers when they are explaining what they like about that same candidate.

In Figure 9, we test this expectation by calculating the average weighted distance between words when respondents are explaining what they dislike about the opposing party's candidate as compared to what they like. Here, we find the average weighted distance to be highest for outparty dislike responses (2.77), suggesting that intensity is most likely to increase from one word to the next when respondents are expressing what they dislike about the opposing candidate. We also find the average weighted distance is lowest when respondents are explaining what they liked about the opposing party candidates (2.11). Consistent with our first two expectations, this difference is statistically significant at the 0.001 level ($t = 4.31, df = 609, p < 0.001$).

To help interpret these results, we re-calculated the average weighted distance using 100 bootstrapped graphs for each response. These results are reported in the SI. When random graphs are used, the average weighted distance for outparty dislike responses was 1.67, meaning what is reported in the first panel of Figure 9 is 65.87 percent higher than what we would expect based

Figure 9: Spreading Activation is Highest When Respondents Are Explaining What They Dislike About Opposing Party Candidates



Note: This figure plots the average weighted distance between bigrams with higher values implying word-level intensity increased from one word to the next, a pattern consistent with spreading activation. In the left panel, we show these averages when respondents are answering questions about the opposing party candidate. The right panel shows these averages for answers about their own party's candidate. Darker bars indicate the respondent was explaining things they liked, whereas lighter bars indicate they were talking about what they disliked. Vertical lines represent 95-percent confidence intervals.

on chance. A similar result is found for outparty like responses. Here, when random graphs are used, the average weighted distance is 1.21, which is, again, noticeably lower than what we actually observed. More specifically, what we report in the second panel of Figure 9 is 74.38 percent higher than what we found using random graphs, suggesting that our initial results cannot easily be attributed to chance alone.

Is Spreading Activation Indicative of Vote Choice?

We have, so far, found results consistent with our expectations: spreading activation is most likely to occur when respondents are discussing what they dislike about opposing party candidates and least likely to occur when they are discussing what they like about those same candidates. These findings are consistent with previous studies on affective polarization. However, the main question – and the one supporting the use of AESR – is whether spreading activation is indicative of vote choice.

Table 1 determines whether this is the case using respondents who identify as being members of the Republican or Democratic party (including leaners). Among these respondents, we identify those whose average weighted distance (see Figure 9) is higher when they are explaining what they dislike about the opposing candidate as compared to their own. In the *Outparty Dislike Intensity* variable, these individuals are given a 1 and all other respondents are given a 0. Said differently, *Outparty Dislike Intensity* returns a 1 when spreading activation is *higher* for the opposing party candidate (as compared to the respondent's own party's candidate) dislike statements. We constructed a similar variable for outparty like statements. More specifically, if the average weighted distance was higher when the respondent was explaining what they liked about the opposing party candidate (as compared to the respondent's own party's candidate), then the *Outparty Like Intensity* variable would be coded as a 1 and all other cases would be coded as 0. Both of these variables are then used to predict whether (1) or not (0) the respondent cast a vote for their own party's candidate (see *Inparty Vote*).[7]

Beginning with Table 1, Model 1, we find spreading activation is highly predictive of a respondent's vote choice. More specifically, if intensity tends to increase at a greater rate from one word

[7]A number of control variables were also included. `Strong Partisan` is simply a dummy variable capturing whether the respondent identified themselves as being a strong member of their party. Dummy variables were also used for gender and race. The `Male` variable records whether the respondents identified themselves as male (1). We also include a control for whether the respondent did (1) or did not (0) consider themselves Caucasian/White (`White`). Finally, we included fixed-effects for each survey.

Table 1: Inparty Votes Are Less Likely When Spreading Activation is Higher for Outparty Like Questions

	Dependent variable:	
	Inparty Vote	
	(1)	(2)
Constant	1.921***	1.500***
	(0.257)	(0.401)
Outparty Dislike Intensity	1.491***	1.468***
	(0.513)	(0.518)
Outparty Like Intensity	−0.942**	−0.960**
	(0.401)	(0.416)
Strong Partisan		1.831***
		(0.288)
Male		0.189
		(0.229)
White		−0.401
		(0.299)
Survey Fixed Effects	✓	✓
N	734	734
Log Likelihood	−275.885	−248.868
AIC	561.769	513.736

Note: In all models, the dependent variable equals 1 when respondents voted for their own party's candidate. These models report the results from Firth logistic regressions. More specifically, estimate were derived using the `brglm` function from the `brglm` library in the `R` statistical software language. All variables are described on page 25. Checkmarks (✓) indicate fixed effects are included for each survey. Levels of significance are reported as follows: $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$. Standard errors are reported in parentheses.

to the next when respondents are explaining what they dislike about an opposing party's candidate as compared to their own, then those respondents are less likely to vote against their own party. The inverse is true when respondents are explaining what they like about an opposing party's candidate. Here, when intensity increases at a greater rate from one word to the next, then respondents are more likely to vote for a candidate outside their own party. Not only are these results consistent with affective polarization, but they hold when additional controls are included in Model 2. In the SI, we also replicate these results using differences in intensity, instead of the dummy variables outlined above. The results reported in Table 1 hold when this alternative operationalization is utilized.

Table 2: Variables Associated with Spreading Activation Have A Higher Average Marginal Effect Than Strength of Partisanship

Variable	AME	z	p	95% Conf. Int.
Outparty Dislike Intensity	0.17	2.87	0.00	$[0.05, 0.28]$
Outparty Like Intensity	-0.10	-2.24	0.03	$[-0.19, -0.01]$
Strong Partisan	0.08	2.89	0.00	$[0.03, 0.13]$
Male	0.00	0.18	0.86	$[-0.04, 0.05]$
White	-0.04	-1.05	0.29	$[-0.10, 0.03]$

Note: Average marginal effects (AME) for Table 1, Model 2. These were estimated using the `margins_summary` function from the `margins` library in the `R` statistical software language.

Table 2 reports the average marginal effects (AME) for all variables reported in Model 2. These were calculated using the `margins_summary` function from the `margins` library in the `R` statistical software language. Here, we find the AME for both *Outparty Dislike Intensity* and *Outparty Like Intensity* is higher than the AME for our control variables, including the dummy variable associated with strength of partisanship. For example, if intensity tends to increase at a greater rate from one word to the next when respondents are stating what they dislike about an opposing party's candidate as compared to their own (see *Outparty Dislike Intensity*), then those respondents are 17 percent *more* likely to vote with their own party. Similarly, respondents whose intensity increases more when they are explaining what they like about an opposing party candidate as compared to their own (see *Outparty Like Intensity*) are 10 percent *less* likely to vote with their own party. By

comparison, respondents who indicate they are strong partisans are only 8 percent *more* likely to vote with their own party, as compared to those who do not. Thus, the AME for *Outparty Dislike Intensity* and *Outparty Like Intensity* is around 113 and 25 percent higher than the AME for *Strong Partisanship*, respectively.

Whether the topic is affective polarization (Iyengar, Sood and Lelkes 2012; Mason 2016), candidate evaluations (Lodge, McGraw and Stroh 1989), voter ambivalence (Basinger and Lavine 2005) or gauging opinions on important public policy topics (Soss and Schram 2007), asking voters what they like and dislike is ubiquitous with the study of American politics. Although the words used in these open-ended responses are undoubtedly important, Tables 1 and 2 demonstrate that the nonverbal cues given during those responses also carry considerable weight. Indeed, both *Outparty Dislike Intensity* and *Outparty Like Intensity* are derived using the audio and video feeds from respondents answering questions political scientists have been asking for decades. This is not to say text-based measures or simply counting the number of likes and dislikes have no place in survey research, but our results suggest nonverbal expressions may serve as important complements to these efforts.

We further underline this point with two robustness checks. In our first robustness check, we consider whether our results hold when the number of like and dislike statements – measures used by previous scholars to assess affective polarization (Levendusky 2018; Levendusky and Malhotra 2016) – are included as controls. These results are reported Table 3, Model 1. Here, we include the number of dislikes listed towards the opposing party candidate minus the number of dislikes listed towards the respondent's party's candidate. We also include a similar measure for like statements, where instead of differencing the number of dislikes, we difference the number of likes. Given that the dependent variable is still whether respondents vote for their party's candidate (*Inparty Vote*), the dislike and like differences should be a positive and negative predictor, respectively, which is what we find. However, we also find that the variables created for this study (see *Outparty Dislike Intensity* and *Outparty Like Intensity*) are still highly significant predictors. This is perhaps not too surprising since the correlation between *Outparty Dislike Intensity* and the difference in the

Table 3: The Effects of Spreading Activation Cannot Be Easily Attributed To Like/Dislike Counts or the Number of Positive/Negative Words

	Dependent variable:	
	Inparty Vote	
	(1)	(2)
Constant	1.376^{***}	1.653^{***}
	(0.421)	(0.419)
Outparty Dislike Intensity	1.521^{***}	1.382^{***}
	(0.520)	(0.511)
Outparty Like Intensity	-0.940^{**}	-0.807^{**}
	(0.410)	(0.408)
Strong Partisan	0.985^{***}	0.738^{***}
	(0.267)	(0.244)
Male	0.324	0.103
	(0.239)	(0.223)
White	-0.133	-0.273
	(0.309)	(0.294)
# of Outparty Dislikes − # of Inparty Dislikes	0.646^{***}	
	(0.124)	
# of Outparty Likes− # of Inparty Likes	-0.555^{***}	
	(0.114)	
Outparty Dislike Text-Based Intensity		0.512^{**}
		(0.223)
Outparty Like Text-Based Intensity		-0.358
		(0.227)
Survey Fixed Effects	✓	✓
N	734	734
Log Likelihood	-237.496	-266.591
AIC	494.993	553.182

Note: In all models, the dependent variable equals 1 when respondents voted for their own party's candidate. These models report the results from Firth logistic regressions. More specifically, estimate were derived using the `brglm` function from the `brglm` library in the `R` statistical software language. All variables are described on pages 25 and 30. Checkmarks (✓) indicate fixed effects are included for each survey. Levels of significance are reported as follows: $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$. Standard errors are reported in parentheses.

number of outparty and inparty dislike statements is only 0.10, suggesting that what people say is not necessarily the same as how people say it. Indeed, there is even a smaller correlation (0.08) between *Outparty Like Intensity* and the difference between the number of outparty and inparty like statements, suggesting AESR can provides insights into how respondent's vote that cannot be easily captured by simply counting the number of likes and dislikes.

In our second robustness check, we consider whether our results hold when text-based measures of intensity are included as controls. These, too, have been used by previous scholars to assess affective polarization (Rathje, Van Bavel and van der Linden 2021; Stapleton and Dawkins 2021). More specifically, in Model 2 we include two variables derived from the National Research Council (NRC) Valence-Arousal-Dominance (VAD) dictionary (Mohammad 2018). Similar to existing VAD lexicons (Bradley and Lang 1999; Warriner, Kuperman and Brysbaert 2013), the NRC dictionary includes over 20,000 English-language words scored on three dimensions: valence, arousal and dominance. As explained above, the measures we introduce in this study are most closely associated with the arousal dimension. Given that, we used the NRC arousal category to score all words in the respondent like/dislike statements from 0 (low arousal) to 1 (high arousal). We then used the average of these scores for each response to create dummy variables that are similar to our main independent variables. The first of these variables – called *Outparty Dislike Text-Based Intensity* – equals 1 when respondents use words with higher NRC arousal scores when they are explaining what they dislike about the opposing party candidate as compared to their own. Similarly, when respondents use words with higher NRC arousal scores when they are explaining what they like about an opposing party candidate as compared to their own, then the *Outparty Like Text-Based Intensity* variable would be coded as a 1 and all other cases would be coded as 0. Given that the dependent variable is still whether respondents vote for their party's candidate (*Inparty Vote*), we expect the coefficients associated with *Outparty Dislike Text-Based Intensity* and *Outparty Like Text-Based Intensity* to be positive and negative, respectively, which is what we find.

In Table 3, we also find that both *Outparty Dislike Intensity* and *Outparty Like Intensity* are

Table 4: Variables Associated with Spreading Activation Have A Higher Average Marginal Effect Than Traditional Text-Based Measures

Variable	AME	z	p	95% Conf. Int.
Outparty Dislike Intensity	0.16	2.69	0.01	$[0.04, 0.27]$
Outparty Like Intensity	-0.09	-1.98	0.05	$[-0.18, -0.00]$
Outparty Dislike Text-Based Intensity	0.06	2.31	0.02	$[0.01, 0.11]$
Outparty Like Text-Based Intensity	-0.04	-1.58	0.11	$[-0.09, 0.01]$
Strong Partisan	0.08	3.05	0.00	$[0.03, 0.14]$
Male	0.00	0.07	0.94	$[-0.05, 0.05]$
White	-0.03	-0.93	0.35	$[-0.10, 0.03]$

Note: Average marginal effects (AME) for Table 3, Model 2. These were estimated using the `margins_summary` function from the `margins` library in the `R` statistical software language.

still highly significant predictors.[8] Moreover, we found the correlation between the former and the associated text-based measure (*Outparty Dislike Text-Based Intensity*) was only 0.12, suggesting simply determining the extent to which words are indicative of arousal fails to capture the influence of nonverbal cues on the vote. The same can be said for *Outparty Like Intensity* where a low correlation was also found (0.15) between this variable and the one derived from the NRC dictionary (*Outparty Like Text-Based Intensity*). To further underline this point, Table 4 shows the AME for the variables associated with spreading activation is higher than the comparable AME from our text-based measures. More specifically, the AME for *Outparty Dislike Intensity* is 167 percent higher than the AME for *Outparty Dislike Text-Based Intensity* and the AME for *Outparty Like Intensity* is 125 percent higher than the AME for *Outparty Like Text-Based Intensity*. Not only does this emphasize the substantive importance of the variables introduced in this study, but it also shows that the measures we derived from AESR provide important complements to traditional measures, like those derived from the number of likes and dislikes offered by respondents, and others which use dictionary-based methods.

[8]In the SI, we also find the same results when differences in intensity are used instead of these dummy variables.

Discussion and Conclusion

Despite the widespread use of telephone surveys for decades, the audio from these common data streams have received scant attention from political scientists. Similarly, when we conduct in-person interviews, the way respondents behave rarely enters into our assessments of their responses. Yet, these data can yield important insights into the intensity with which respondents hold their opinions. The same can be said for online interviews where responses are becoming easier to record. However, to our knowledge no prior study has used the data from these audio and video recordings to gain insights regarding respondents' emotions. We develop the first Automatic Emotional Speech Recognition (AESR) system which can be used by future scholars to automatically measure the emotional intensity of respondents during in-person, telephone and online surveys. In doing so, we demonstrate how audio and image data can be extended beyond elite rhetoric and media portrayals to understand the mass public in real-time.

We argue emotions begin below conscious awareness, which is why scholars increasingly rely on physiological output to measure them. However, physiological measures are very intrusive and cannot be used for large-N studies, especially those conducted over a telephone or the internet. With fMRI, respondents are encased in a confined horizontal space. When ERPs are used electrodes are placed at multiple locations on the respondent's scalp. Blink amplitude is measured with electrodes placed just below a person's eyes, and skin conductance is measured with sensors attached to the respondent's fingers. Not only do these techniques require highly artificial settings, but all require expensive equipment and a laboratory to house it. Our AESR system can be used with any audio and video data of sufficient quality obtained from in-person, online and telephone surveys, which greatly reduces costs and make large scale data acquisition possible.

Our central finding suggests that the emotional intensity of respondent answers – as derived from a novel machine learning algorithm – is not only a significant predictor of vote choice, but can also give us insights into how candidate opinions may be stored in memory. We use these insights to learn about how partisans think about the opposing party, with a focus on how the

intensity of one word influences another. Taken together, these results, in combination with the performance metrics associated with the machine learning models we estimated for this study, serve as an important proof of concept. Indeed, we find our automated measures are stronger predictors of vote choice than measures derived from dictionary-based methods, as well as strength of partisanship. Moreover, additional metrics show that audio and image data are most important to assessments of respondent intensity. Collectively, these results suggest that the non-verbal content associated with different response patterns should be actively considered when conducting survey research.

We also note the substantive importance of the findings we present. Spreading activation is a theory that has been used extensively in psychology, but has gained less traction in political science. In our study, we show that respondents are becoming more and less intense from one word to the next, and that these patterns are indicative of how they will eventually vote. If this is the case, then it suggests when we ask questions like "What do you dislike about Joe Biden?", that respondents are placing different cognitive weight on the words they are speaking. For example, we find that spreading activation is more likely to occur when respondents are explaining what they dislike about opposing party candidates. Although we are not the first to say that voters tend to think about the opposing party differently (Cassese 2021), we provide some insights into the *way* that information is stored and later accessed. Given that the purpose of this analysis was to prove the many ways AESR can be used, we will leave it to future scholars to explore the broader implications of our findings, but we think they have illuminated potentially a new way to think about how affective information is processed and later used to make important political decisions, like who to vote for.

Regardless of how future scholars use the tools created in this study, it is clear that respondents seem to deliver some responses with greater emotional intensity. Often, we think of survey responses as series of numbers, but current results show that the *way* those responses are delivered is also important. Nonverbal cues carry meaning. The present study suggests that we can extract politically-relevant information from nonverbal data conveyed by respondents to in-person, online,

and telephone interviews. This information, in turn, greatly expands the capacity of scholars to explore the consequences of emotion in political behavior.

References

Adams Jr, Reginald B and Robert E Kleck. 2003. "Perceived gaze direction and the processing of facial displays of emotion." *Psychological science* 14(6):644–647.

Adolph, Dirk and Georg W. Alpers. 2010. "Valence and arousal: a comparison of two sets of emotional facial expressions." *American Journal of Psychology* 123(2):209–219.

Ahler, Douglas J and Gaurav Sood. 2018. "The parties in our heads: Misperceptions about party composition and their consequences." *The Journal of Politics* 80(3):964–981.

Ambadar, Zara, Jeffrey F Cohn and Lawrence Ian Reed. 2009. "All smiles are not created equal: Morphology and timing of smiles perceived as amused, polite, and embarrassed/nervous." *Journal of nonverbal behavior* 33(1):17–34.

Amodio, David M, John T Jost, Sarah L Master and Cindy M Yee. 2007. "Neurocognitive correlates of liberalism and conservatism." *Nature neuroscience* 10(10):1246.

Anderson, John R. 1983. "A spreading activation theory of memory." *Journal of verbal learning and verbal behavior* 22(3):261–295.

Basinger, Scott J and Howard Lavine. 2005. "Ambivalence, information, and electoral choice." *American Political Science Review* 99(2):169–184.

Bavel, Jay J Van, Katherine Baicker, Paulo S Boggio, Valerio Capraro, Aleksandra Cichocka, Mina Cikara, Molly J Crockett, Alia J Crum, Karen M Douglas, James N Druckman et al. 2020. "Using social and behavioural science to support COVID-19 pandemic response." *Nature human behaviour* 4(5):460–471.

Bokhove, Christian and Christopher Downey. 2018. "Automated generation of 'good enough' transcripts as a first step to transcription of audio-recorded data." *Methodological Innovations* 11(2):2059799118790743.

Bollen, Johan, Huina Mao and Xiaojun Zeng. 2011. "Twitter mood predicts the stock market." *Journal of computational science* 2(1):1–8.

Bowerman, Bruce L, Richard T O'Connell and Anne B Koehler. 2005. *Forecasting, time series, and regression: an applied approach*. Vol. 4 South-Western Pub.

Boyle, Justin, Melanie Jessup, Julia Crilly, David Green, James Lind, Marianne Wallis, Peter Miller and Gerard Fitzgerald. 2012. "Predicting emergency department admissions." *Emergency Medicine Journal* 29(5):358–365.

Bradley, Margaret M and Peter J Lang. 1999. Affective norms for English words (ANEW): Instruction manual and affective ratings. Technical report The center for research in psychophysiology.

Bucy, Erik P. 2000. "Emotional and evaluative consequences of inappropriate leader displays." *Communication Research* 27(2):194–226.

Byrne, Robert F. 2012. "Beyond Traditional Time-Series: Using Demand Sensing to Improve Forecasts in Volatile Times." *Journal of Business Forecasting* 31(2).

Cassese, Erin C. 2021. "Partisan dehumanization in American politics." *Political Behavior* 43(1):29–50.

Collins, Allan M and Elizabeth F Loftus. 1975. "A spreading-activation theory of semantic processing." *Psychological review* 82(6):407.

Darwin, Charles. 1872. *The expression of the emotions in man and animals*. John Murray.

Davydenko, Andrey and Robert Fildes. 2016. "Forecast error measures: critical review and practical recommendations." *Business forecasting: Practical problems and solutions* 34.

De Myttenaere, Arnaud, Boris Golden, Bénédicte Le Grand and Fabrice Rossi. 2016. "Mean absolute percentage error for regression models." *Neurocomputing* 192:38–48.

De Zubicaray, Greig I, Stephen J Wilson, Katie L McMahon and Santhi Muthiah. 2001. "The semantic interference effect in the picture-word paradigm: An event-related fMRI study employing overt responses." *Human brain mapping* 14(4):218–227.

Desai, Nidhi, Kinnal Dhameliya and Vijayendra Desai. 2013. "Feature extraction and classification techniques for speech recognition: A review." *International Journal of Emerging Technology and Advanced Engineering* 3(12):367–371.

Dietrich, Bryce J, Matthew Hayes and Diana Z O'Brien. 2019. "Pitch perfect: Vocal pitch and the emotional intensity of congressional speech." *American Political Science Review* 113(4):941–962.

Dietrich, Bryce J, Ryan D Enos and Maya Sen. 2019. "Emotional arousal predicts voting on the US supreme court." *Political Analysis* 27(2):237–243.

Dietrich, Bryce Jensen and Courtney L Juelich. 2018. "When presidential candidates voice party issues, does Twitter listen?" *Journal of Elections, Public Opinion and Parties* 28(2):208–224.

Dijkstra, Edsger W et al. 1959. "A note on two problems in connexion with graphs." *Numerische mathematik* 1(1):269–271.

Druckman, James N and Matthew S Levendusky. 2019. "What do we measure when we measure affective polarization?" *Public Opinion Quarterly* 83(1):114–122.

Ethofer, Thomas, Silke Anders, Sarah Wiethoff, Michael Erb, Cornelia Herbert, Ralf Saur, Wolfgang Grodd and Dirk Wildgruber. 2006. "Effects of prosodic emotional intensity on activation of associative auditory cortex." *Neuroreport* 17(3):249–253.

Fairbanks, Grant and Wilbert Pronovost. 1939. "An experimental study of the pitch characteristics of the voice during the expression of emotion." *Communications Monographs* 6(1):87–104.

Fournier, Patrick, Stuart Soroka and Lilach Nir. 2020. "Negativity biases and political ideology: A comparative test across 17 countries." *American Political Science Review* 114(3):775–791.

Gianluigi, Ciocca and Schettini Raimondo. 2006. "An innovative algorithm for key frame extraction in video summarization." *Journal of Real-Time Image Processing* 1(1):69–88.

Giannakopoulos, Theodoros. 2015. "pyaudioanalysis: An open-source python library for audio signal analysis." *PloS one* 10(12):e0144610.

Gimpel, James G and Iris S Hui. 2015. "Seeking politically compatible neighbors? The role of neighborhood partisan composition in residential sorting." *Political Geography* 48:130–142.

Goldberg, Yoav and Omer Levy. 2014. "word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method." *arXiv preprint arXiv:1402.3722* .

Greene, Joshua D, R Brian Sommerville, Leigh E Nystrom, John M Darley and Jonathan D Cohen. 2001. "An fMRI investigation of emotional engagement in moral judgment." *Science* 293(5537):2105–2108.

Hatfield, Elaine, John T Cacioppo and Richard L Rapson. 1993. "Emotional contagion." *Current directions in psychological science* 2(3):96–100.

Healy, Andrew and Neil Malhotra. 2013. "Retrospective voting reconsidered." *Annual Review of Political Science* 16:285–306.

Hibbing, John R, Kevin B Smith and John R Alford. 2014. "Differences in negativity bias underlie variations in political ideology." *Behavioral and brain sciences* 37:297–307.

Hirschfeld, Gerrit, Bernadette Jansma, Jens Bölte and Pienie Zwitserlood. 2008. "Interference and facilitation in overt speech production investigated with event-related potentials." *Neuroreport* 19(12):1227–1230.

Huddy, Leonie, Lilliana Mason and Lene Aarøe. 2015. "Expressive partisanship: Campaign involvement, political emotion, and partisan identity." *American Political Science Review* 109(1):1–17.

Iyengar, Shanto, Gaurav Sood and Yphtach Lelkes. 2012. "Affect, not ideology social identity perspective on polarization." *Public opinion quarterly* 76(3):405–431.

Iyengar, Shanto and Masha Krupenkin. 2018. "The strengthening of partisan affect." *Political Psychology* 39:201–218.

Iyengar, Shanto, Yphtach Lelkes, Matthew Levendusky, Neil Malhotra and Sean J Westwood. 2019. "The origins and consequences of affective polarization in the United States." *Annual Review of Political Science* 22:129–146.

Jack, Rachael E, Oliver GB Garrod, Hui Yu, Roberto Caldara and Philippe G Schyns. 2012. "Facial expressions of emotion are not culturally universal." *Proceedings of the National Academy of Sciences* 109(19):7241–7244.

Johnstone, Tom and Klaus R Scherer. 2000. "Vocal communication of emotion." *Handbook of emotions* 2:220–235.

Kim, Sungil and Heeyoung Kim. 2016. "A new metric of absolute percentage error for intermittent demand forecasts." *International Journal of Forecasting* 32(3):669–679.

Klofstad, Casey A. 2016. "Candidate voice pitch influences election outcomes." *Political Psychology* 37(5):725–738.

Klofstad, Casey A and Rindy C Anderson. 2018. "Voice pitch predicts electability, but does not signal leadership ability." *Evolution and human behavior* 39(3):349–354.

Knoch, Daria, Alvaro Pascual-Leone, Kaspar Meyer, Valerie Treyer and Ernst Fehr. 2006. "Diminishing reciprocal fairness by disrupting the right prefrontal cortex." *science* 314(5800):829–832.

Knox, Dean and Christopher Lucas. 2021. "A dynamic model of speech for the social sciences." *American Political Science Review* 115(2):649–666.

Koo, Terry K and Mae Y Li. 2016. "A guideline of selecting and reporting intraclass correlation coefficients for reliability research." *Journal of chiropractic medicine* 15(2):155–163.

Kormos, Judit. 2014. *Speech production and second language acquisition*. Routledge.

Levendusky, Matthew and Neil Malhotra. 2016. "Does media coverage of partisan polarization affect political attitudes?" *Political Communication* 33(2):283–301.

Levendusky, Matthew S. 2018. "Americans, not partisans: Can priming American national identity reduce affective polarization?" *The Journal of Politics* 80(1):59–70.

Liu, Tianming, Hong-Jiang Zhang and Feihu Qi. 2003. "A novel video key-frame-extraction algorithm based on perceived motion energy model." *IEEE transactions on circuits and systems for video technology* 13(10):1006–1013.

Lodge, Milton and Charles S Taber. 2013. *The rationalizing voter*. Cambridge University Press.

Lodge, Milton, Kathleen M McGraw and Patrick Stroh. 1989. "An impression-driven model of candidate evaluation." *American Political Science Review* 83(2):399–419.

Marcus, George E, W Russell Neuman and Michael MacKuen. 2000. *Affective intelligence and political judgment*. University of Chicago Press.

Mason, Lilliana. 2016. "A cross-cutting calm: How social sorting drives affective polarization." *Public Opinion Quarterly* 80(S1):351–377.

McDermott, Rose. 2007. "Cognitive Neuroscience and Politics Next Steps." *The Affect Effect. Dynamics of Emotion in Political Thinking and Behavior* pp. 375–397.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. pp. 3111–3119.

Miller, George A. 1995. "WordNet: a lexical database for English." *Communications of the ACM* 38(11):39–41.

Mohammad, Saif. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. pp. 174–184.

Neuman, W. R., G. E. Marcus, A. N. Crigler and M. MacKuen. 2007. *The affect effect: Dynamics of emotion in political thinking and behavior*. University of Chicago Press.

Nozari, Nazbanou and Svetlana Pinet. 2020. "A critical review of the behavioral, neuroimaging, and electrophysiological studies of co-activation of representations during word production." *Journal of Neurolinguistics* 53:100875.

Ochshorn, Robert and Max Hawkins. 2017. "Gentle: A robust yet lenient forced aligner built on Kaldi." `https://lowerquality.com/gentle/`. Accessed: 2022-01-12.

Owren, Michael J and Jo-Anne Bachorowski. 2007. "Measuring emotion-related vocal acoustics." *Handbook of emotion elicitation and assessment* pp. 239–266.

Oxley, Douglas R, Kevin B Smith, John R Alford, Matthew V Hibbing, Jennifer L Miller, Mario Scalora, Peter K Hatemi and John R Hibbing. 2008. "Political attitudes vary with physiological traits." *science* 321(5896):1667–1670.

Perez, Luis and Jason Wang. 2017. "The effectiveness of data augmentation in image classification using deep learning." *arXiv preprint arXiv:1712.04621* .

Posner, Jonathan, James A. Russell and Bradley S. Peterson. 2005. "The Circumplex Model of

Affect: An Integrative Approach to Affective Neuroscience, Cognitive development, and Psychopathology." *Development and Psychopathology* 17:715–734.

Povey, Daniel, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz et al. 2011. The Kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*. Number CONF IEEE Signal Processing Society.

Prema, V and K Uma Rao. 2015. "Time series decomposition model for accurate wind speed forecast." *Renewables: Wind, Water, and Solar* 2(1):1–11.

Rathje, Steve, Jay J Van Bavel and Sander van der Linden. 2021. "Out-group animosity drives engagement on social media." *Proceedings of the National Academy of Sciences* 118(26).

Rheault, Ludovic and Christopher Cochrane. 2020. "Word embeddings for the analysis of ideological placement in parliamentary corpora." *Political Analysis* 28(1):112–133.

Rheault, Ludovic, Kaspar Beelen, Christopher Cochrane and Graeme Hirst. 2016. "Measuring emotion in parliamentary debates with automated textual analysis." *PloS one* 11(12):e0168843.

Russell, James A. 2003. "Core Affect and the Psychological Construction of Emotion." *Psychological Review* 110:145–172.

Sato, Wataru and Sakiko Yoshikawa. 2007. "Enhanced experience of emotional arousal in response to dynamic facial expressions." *Journal of Nonverbal Behavior* 31(2):119–135.

Scherer, Klaus R. 2003. "Vocal communication of emotion: A review of research paradigms." *Speech communication* 40(1-2):227–256.

Scherer, Klaus R and Didier Grandjean. 2008. "Facial expressions allow inference of both emotions and their components." *Cognition and Emotion* 22(5):789–801.

Schyns, Philippe G, Lucy S Petro and Marie L Smith. 2009. "Transmission of facial expressions of emotion co-evolved with their efficient decoding in the brain: behavioral and brain evidence." *Plos one* 4(5):e5625.

Selvaraju, Ramprasaath R, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*. pp. 618–626.

Smith, Kevin B, Douglas Oxley, Matthew V Hibbing, John R Alford and John R Hibbing. 2011. "Disgust sensitivity and the neurophysiology of left-right political orientations." *PloS one* 6(10):e25552.

Snyder, D., D. Garcia-Romero, G. Sell, D. Povey and S. Khudanpur. 2018. X-vectors: Robust DNN Embeddings for Speaker Recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Soss, Joe and Sanford F Schram. 2007. "A public transformed? Welfare reform as policy feedback." *American Political Science Review* 101(1):111–127.

Soubelet, Andrea and Timothy A Salthouse. 2011. "Influence of social desirability on age differences in self-reports of mood and personality." *Journal of personality* 79(4):741–762.

Spezio, Michael L and Ralph Adolphs. 2007. "Emotional processing and political judgment: Toward integrating political psychology and decision neuroscience." *The affect effect: Dynamics of emotion in political thinking and behavior* pp. 71–95.

Stapleton, Carey E and Ryan Dawkins. 2021. "Catching My Anger: How Political Elites Create Angrier Citizens." *Political Research Quarterly* p. 10659129211026972.

Stewart, Patrick A, Bridget M Waller and James N Schubert. 2009. "Presidential speechmak-

ing style: Emotional response to micro-expressions of facial affect." *Motivation and Emotion* 33(2):125.

Stewart, Patrick A and Pearl K Ford Dowe. 2013. "Interpreting President Barack Obama's facial displays of emotion: Revisiting the Dartmouth group." *Political Psychology* 34(3):369–385.

Susskind, Joshua M, Daniel H Lee, Andrée Cusi, Roman Feiman, Wojtek Grabski and Adam K Anderson. 2008. "Expressing fear enhances sensory acquisition." *Nature neuroscience* 11(7):843–850.

Szabó, Beáta T, Susan L Denham and István Winkler. 2016. "Computational models of auditory scene analysis: a review." *Frontiers in Neuroscience* 10:524.

Tigue, Cara C, Diana J Borak, Jillian JM O'Connor, Charles Schandl and David R Feinberg. 2012. "Voice pitch influences voting behavior." *Evolution and Human Behavior* 33(3):210–216.

Warriner, Amy Beth, Victor Kuperman and Marc Brysbaert. 2013. "Norms of valence, arousal, and dominance for 13,915 English lemmas." *Behavior research methods* 45(4):1191–1207.

Williams, Nora Webb, Andreu Casas and John D Wilkerson. 2020. *Images as data for social science research: An introduction to convolutional neural nets for image classification*. Cambridge University Press.

Wilson, D Randall and Tony R Martinez. 2003. "The general inefficiency of batch training for gradient descent learning." *Neural networks* 16(10):1429–1451.

Zaller, John R et al. 1992. *The nature and origins of mass opinion*. Cambridge university press.

Ziman, Kirsten, Andrew C Heusser, Paxton C Fitzpatrick, Campbell E Field and Jeremy R Manning. 2018. "Is automatic speech-to-text transcription ready for use in psychological experiments." *Behavior research methods* pp. 1–9.