When the Picture is not Complete: Decoding Visual Sentiment of Political Imagery

Abstract

What does it mean to define visual sentiment—the emotional resonance conveyed by images—when viewers consistently perceive things differently, especially when their political beliefs are involved? This study introduces a novel approach to visual sentiment analysis that directly addresses these perceptual differences in sentiment classification. In order to achieve this, we developed a dataset reflecting political divisions by curating images on a polarizing topic, annotated by individuals from distinct political affiliations. Using this dataset, we trained a deep learning multi-task, multi-class model to predict visual sentiment from different ideological viewpoints. By incorporating these diverse perspectives into the labeling and model training, our approach improves the accuracy of visual sentiment predictions and better mirrors human judgment. Ultimately, this study advocates for a paradigm shift in visual sentiment decoding, urging a move beyond traditional image-focused approaches to develop classifiers that more accurately capture the complexity of human sentiment.

Key words: visual data, sentiment analysis, annotators bias, visual data labeling, deep learning.

Word count: 7249

1 Introduction

The rapid advance of visual social media platforms like Instagram, TikTok, and X[1], signals a decisive shift toward a visual-centric culture where images crucially direct our attention (see, e.g., Domke, Perlmutter and Spratt 2002; Grabe and Bucy 2014; Schill 2012; Chavez 2023; Webb Williams 2023), shape attitudes and sentiments (Anastasopoulos et al., 2024; Bossetta and Schmøkel, 2023; Casas and Williams, 2019), and reinforce stereotypes (Carpinella and Bauer, 2021; Domke, Perlmutter and Spratt, 2002). This effect is especially strong in politics, where the way events are visually depicted can influence public reaction (Grabe and Bucy, 2009; Casas and Williams, 2019), particularly during pivotal times like elections (Grabe and Bucy, 2014) or when tensions around certain political issues rise (de Lima-Santos et al., 2023; Casas and Williams, 2019).

Despite significant efforts to understand the political effects of sustained exposure to politically charged visuals, many of these approaches may rest on shaky assumptions. For instance, prior research has often assumed a uniform reaction to certain political images—such as those depicting large crowds of immigrants—on public sentiment (Farris and Silber Mohamed, 2018; Haynes, Merolla and Ramakrishnan, 2016; Webb Williams et al., 2023). However, some recent findings, for instance, by Madrigal and Soroka (2023) challenge this view, suggesting that such imagery sometimes does not invariably amplify anti-immigrant sentiments. This indicates that the relationship between visual content and public sentiment is more complex and context-dependent than previously thought.

To further illustrate this point consider two images—A and B in Figure 1—commonly featured in news coverage of migrant caravans. Image A shows a young boy sitting on a man's shoulders, likely his father, as they walk peacefully together, while Image B shows a large, determined crowd resembling protesters. Although both images are sourced from official U.S. media accounts and accompany similar news stories, the sentiments they evoke—and, more crucially, their interpretations during coding—can differ significantly

[1]Formerly known as Twitter.

Figure 1: How Would You Label Sentiment of These Images?

Image A

Published by OANN X (Twitter) account (Jan 29, 2019).

Image B

Published by National Review ((Oct 26, 2018). Photo by Reuters/Carlos Garcia Rawlins.

depending on the viewer's preexisting attitudes toward immigration. This variability in interpretation exposes a critical gap in current methodologies for visual sentiment analysis, particularly for political or other sensitive or dividing images.

Prominent computer vision research has attempted to fill this gap by focusing on content-driven approaches to automated visual sentiment analysis (Joo and Steinert-Threlkeld, 2022), either by mainly interpreting images holistically (see e.g., Kossaifi et al. 2019; Ortis, Farinella and Battiato 2020; Torres 2018; Webb Williams 2023) or by identifying "affective regions" within images—specific objects that are believed to evoke strong emotions, like "smiling people" or "weapons" (see e.g., Torres 2024; Yang et al. 2018; Yang and Newsam 2010; Zhao et al. 2021). However, these approaches often overlook the subjective nature of image interpretation, particularly the influence of cultural, social, and political contexts

on viewer perceptions (Webb Williams et al., 2023).

In response to these limitations, we propose a novel approach that integrates attitudinal differences into training visual sentiment classifiers for political images. This method acknowledges image interpretation's subjective and context-dependent nature, aiming to produce more meaningful interpretations of visual sentiment in politically charged contexts. To illustrate our approach and its advantages we propose the following workflow:

(1) Identifying an Attitudinal Cleavage: First, we examine whether visual labeling reflects a stable societal gap, such as a political divide in the USA, which must be embedded in the model training. However, getting there is a challenge. Creating one label that averages sentiments for divisive visuals could result in a "neutral" label for these images, canceling out the sentiments of strong pro-issue and against-issue people among coders. This approach does not reflect the *true* sentiments that people actually hold.[2] Using separate models for different attitude groups can lead to fragmented analysis and could miss out on the informative contradictory nature of visual sentiment in an image. Additionally, this approach may complicate the interpretation of results.

To address these limitations, we propose a single multi-task multi-class classification model that predicts multiple sentiment labels for disagreeing groups of coders. This approach helps mitigate the potential bias from individuals with different attitudinal priors assigning opposing sentiment labels.

(2) Creating a Dataset of Sentiment Labels: We illustrate our approach using the topic of immigration, a politically polarizing topic where Democrats and Republicans typically disagree, representing a stable cleavage in political attitudes. We

[2]Of course, there is no such thing as a "true" sentiment. However, the problem we are addressing here is how to tailor computer vision strategies to account for significant, non-negligible variations in sentiment, to produce more accurate classifications. Labeling at scale, as proposed by Benoit et al. (2016), is considered one of the most promising solutions to this problem, at least with text; however, this approach may exacerbate the issue by averaging out polarized sentiments, thus masking the true diversity of opinions.

created a dataset where immigration-related images were labeled with sentiment scores. These scores were based on individual attitudes toward the images. After that, the scores were grouped by image, and then two sentiment labels were generated: one reflecting the sentiment of Democrats and another for Republicans, forming a pair for each image[3].

Importantly, what is known as *sentiment* in computational social sciences encompasses multiple related constructs in political and social psychology. By refining sentiment identification, we demonstrate that some perceptual divides lead to disagreements in sentiment classification, while others do not. Our empirical analysis identifies instances where specific sentiment representations significantly impact the accuracy of visual sentiment labeling, necessitating their inclusion in model training and labeling. Conversely, when these representations show no difference between coder groups, they can be omitted without compromising accuracy, allowing us to use a single sentiment label for these groups.

(3) Training a Multi-task Multi-class Classifier: We use paired sentiment labels, each representing the sentiment of coder groups (Democrats and Republicans), to develop visual sentiment classification models through a transfer learning approach. Specifically, we fine-tune and adapt established deep neural networks (ResNet50V2, DenseNet-121 and DenseNet-169) using our labeled data. In a typical multi-class image classification task, images serve as inputs, and label classes are the outputs of the classification model. Our approach extends this by using a multi-task, multi-class classification model to predict sentiment for two distinct partisan groups, Democrats and Republicans, within a unified framework. Here we conduct two exercises: (1) a multi-class classification, categorizing sentiment as negative, neutral, or positive for each group, and (2) a linear regression prediction, rating sentiment on a continuous

[3]We asked participants to self-identify as Democrats, Republicans, Independent, or Other, but only used responses from Democrats and Republicans for our analysis. For each image we averaged the scores across each of the partisan groups. On average each image was labeled by 32 Democrats and 33 Republicans.

scale [1,7] for each group. By designing the model to predict distinct sentiment labels for Democrats and Republicans, we capture group-specific sentiment nuances in one model.

(4) Verification: We evaluate the quality of our classification models using a test set that was not involved in the training or fine-tuning process.

(5) Practical Implications: Using our test data, we first demonstrate that averaging across sentiment labels can lead to poorly tailored results, as it aggregates sentiments that should not be combined, resulting in significant mislabeling and bias in sentiment identification. Second, by applying our approach to re-evaluate the visual sentiment of images from Madrigal and Soroka's (2023) study, we show that our classification method effectively captures the heterogeneity in visual sentiment among partisans and reflects variations in perceived vulnerability (being subjects of harm)—key aspects that the original authors also emphasized.

Ultimately, our study advocates for a shift in how computer vision approaches training classifiers to predict sentiment and other attitudinal outcomes at scale. Simply averaging labels to produce an *overall* sentiment is problematic for two key reasons: first, sentiment is inherently subjective, and, second, this subjectivity becomes even more pronounced when dealing with images that depict politically polarizing topics. Using a single, averaged label in such cases risks generating inaccurate or meaningless classifications. Instead, we emphasis the necessity to account for diverse perspectives, particularly those driven by partisan and ideological differences, as these cleavages often *stably* shape how people interpret political imagery. However, we also demonstrate that certain attitudes about visuals are less divisive, allowing for accurate labeling through averaging across coders in these cases.

From the perspective of policy implications, if certain images depicting polarizing topics systematically appear on social media, it is valuable to understand how they are

received by partisans with different ideological positions. Suppose a campaign manager wants to gauge how an ad targeting immigration resonates with both parties. A single averaged sentiment label might indicate the ad is viewed neutrally, hiding the fact that Democrats see it as compassionate while Republicans view it as irresponsible. This kind of oversight can prevent political strategists from crafting messages that effectively speak to their base. Our classifier helps quickly determine whether these images elicit similar associations or potentially exacerbate partisan divides. Our approach provides a straightforward, hands-on method for gaining this information.

Concluding, while partisanship is a significant dividing line for certain political issues—such as in the U.S., where strong and persistent political polarization shapes public opinion—other sensitive divides, both political and non-political, should also be considered in model training. Accounting for these divides is essential to capture the stable ways in which different societal cleavages influence people's reactions to sensitive visuals. This requires a robust theoretical understanding of the relevant social divides. Without incorporating this understanding into the labeling and model training processes, attempts to predict "as-if" human sentiments about visuals may fail to accurately reflect the real expression of these sentiments.

2 Measuring Sentiment

In social psychology, sentiments or attitudes are essentially *perspectives*—evaluations of a person, object, or concept—typically assessed along dimensions such as good versus bad, harmful versus beneficial, or likable versus unlikable (Ajzen and Fishbein, 2000; Ajzen, 2001; Eagly and Chaiken, 1993; Petty, Wegener and Fabrigar, 1997). Attitudes are not static; they are shaped by both past experiences and immediate context, shifting depending on what feels relevant at the moment (Liberman and Chaiken, 1996). This flexibility is driven by a mix of situational cues and deeply ingrained mental associations (Calanchini

et al., 2013; Van Bavel, Jenny Xiao and Cunningham, 2012) meaning that beliefs about an object appear automatically, but only the ones that are top of mind influence our attitudes at any given time (Feather, 1985). However, when attitudinal divisions are stable, we can expect more consistent judgments about the same objects from individuals within those divided groups.

When it comes to politically charged visuals, people's ideological leanings drive visuals' interpretations in the first place (Leong et al., 2019), and these interpretations tend to remain consistent as long as their political beliefs are stable. Partisanship, a core framework for how individuals understand politics in certain contexts (Campbell, 1960), heavily shapes the way people view these visuals. For example, Republicans may be more likely to see immigration-related imagery—like crowds or border crossings—as more alarming. This fits their more restrictive stance on immigration, leading them to quickly form negative attitudes toward the subjects in the image as soon as they understand the image scene. Democrats, on the other hand, may view the same scene as far less threatening, aligned with their more inclusive perspective, which should define the diversity in visual sentiment interpretations and labeling.

In this study, we propose a scalable solution that leverages human attitudes toward political images, aligned with predictable partisan divides. While not universally applicable, this approach is particularly well-suited for analyzing visual sentiment in contexts with stable attitudinal (e.g., ideological) cleavages—such as U.S. political imagery, where mass polarization between Democrats and Republicans creates clear, measurable differences in how politically divisive issues are perceived.

But how deep should we go in accounting for diverse perspectives? While sentiment variation can also be meaningful within parties (e.g., among Republicans) or along other dimensions (such as personal relevance to the topic), our approach suggests two guiding factors: (1) the degree of sentiment variation *one aims to capture*, and (2) whether empirical evidence supports meaningful differences in attitudes on specific issues across the groups

of interest.

In the workflow below, we expose individuals to images depicting a politically polarizing issue, anticipating a clear divide in their sentiment judgments. Rather than smoothing over these differences, we propose training a model that explicitly captures and incorporates these divisions to reveal the attitudinal distinctions each group expresses.

3 Empirical Case

3.1 Data Acquisition and Labeling

We choose immigration, as a political issue that systematically polarizes attitudes between Democrats and Republicans, as an application of our approach. Unlike other polarizing topics such as gun control, which often involve explicit affective objects (e.g., guns), immigration as a visual subject does not have universally recognizable visual cues that strongly signal emotional reactions. This makes it particularly valuable for examining how visuals can elicit diverse sentiments based solely on contextual understanding rather than overtly affective objects.

We collected images of real-live events accessed through publicly available social media accounts and stock image sources (such as Getty) that often serve as a primary source of imagery for many media outlets covering news on immigration. In total, we collected 832 images: 315 images are coming from the X (formerly Twitter) accounts of U.S. media outlets[4] and 517 images were collected from stock image sources.

We asked respondents standard socio-demographic questions, self-identify as either Democrats or Republicans[5]. For each participant, we randomly select 10 images from the full pool of images and then present these 10 images in a random order. Participants eval-

[4]We used official X (formerly Twitter) handles of 393 U.S. media outlets to query all tweets containing the term "migrant caravan" between December 2017 and October 2021. Focusing on tweets with images, we created a subset of over 2,000 tweets (see Author (year) for a detailed description of the dataset).

[5]Respondents who self-identified as Independents or from other parties were excluded.

uate each image based on their perception of its sentiment, using the following variables.

1. *"Sentiment"*: "Would you say that this image portrays the subject(s) or objects(s) in this picture in a positive or negative light? '1' stands for negative, '4' is neutral, and '7' stands for positive."

2. *"Subject of harm"*: "In your opinion, the subject(s) who is (are) portrayed in this picture is (are) more likely to be dangerous or harmless? '1' stands for dangerous, '4' is the middle ground, and '7' stands for harmless."[6]

3. *"Object of harm"*: "In your opinion, the subject(s) who is (are) portrayed in this picture is (are) more likely to be vulnerable or safe? '1' stands for vulnerable, '4' is the middle ground, and '7' stands for safe."

4. *"Accuracy"*: "Do you think that this image is a faulty or accurate representation of the story that actually occurred? '1' stands for faulty, '4' is the middle ground, '7' stands for accurate."[7]

3.2 Cleavages in Labeling and Image-Level Labels

We anticipate that visual sentiment labeling will vary across party lines.[8] To test our expectations, we generate image-level labels using the strategy outlined below and then examine whether sentiment labeling differs significantly between the partisan groups.

[6]Republicans should be generally more likely to perceive certain immigrants as a potential threat, especially if the visual frame aligns with this perception. For instance, Republicans should be more likely to perceive a crowd of men as dangerous compared to women with children. In contrast, Democrats will view the same subjects as less threatening and more harmless, which aligns with their more lenient views on immigration and social inclusion. The perceived threat from the subjects in immigration-related images is an important component of the overall sentiment people attach to these images (Madrigal and Soroka, 2023). By using the 'Subject of Harm' variable, we can effectively measure this aspect of visual sentiment that is relevant specifically to this topic.

[7]The perception of accuracy is tied to the level of trust in the image as a source of information about the event that occurred. Partisans are expected to judge the accuracy of an image based on their perception of norms regarding how immigration issues *should* be depicted to illustrate the topic accurately (Fahmy et al., 2006).

[8]However, in tasks such as assigning sentiment to images of politically sensitive topics, labels can be influenced by broader social and political disagreements, including fundamental characteristics like gender, age, or race (Webb Williams et al., 2023).

We use respondents' partisanship, coded as Democrats (1) and Republicans (0)[9], to create labels for each image. So each image receives three scores: 1) the overall average score from all participants; 2) the average score from Democratic participants; and 3) the average score from Republican participants. This procedure is repeated to assess visual sentiment for four outcomes of interest—sentiment, subject of harm, object of harm, and accuracy.

Figure 2 presents density plots of average evaluation scores for the four sentiment measures across Democrats (blue-shaded plot), Republicans (red-shaded plot), and the overall average (gray-shaded plot). Democrats and Republicans give similar evaluations for image accuracy and the portrayal of people as objects of harm, suggesting that averaging scores across all respondents does not introduce significant bias for these measures. However, there are clear differences in the average scores for 'Sentiment' and 'Subjects of Harm' between the two groups. This divergence suggests that using a single label for these variables could introduce considerable bias in large-scale labeling.[10]

In Appendix D, we provide density plots showing individual partisan evaluations for the images with the most and least polarizing average scores. This demonstrates that the cleavages at the image level are not simply a result of aggregating individual scores, but instead reflect genuine attitudinal divides at the individual level.

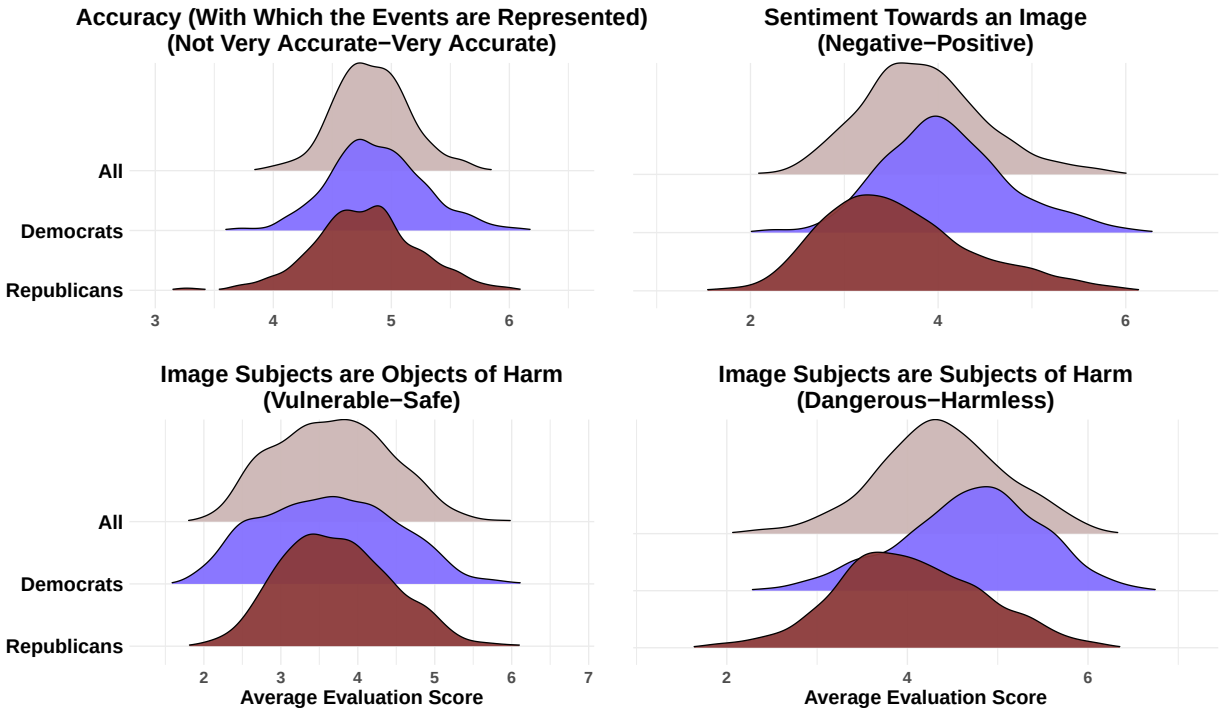
3.3 Label Classes

After gathering individual evaluations, we averaged them to produce a single score for each image on a 1 to 7 interval scale. While we use these interval-scale scores for linear predictions in our analysis, we also convert them into conventional sentiment labels. Specifically, we categorize the interval scores into three sentiment groups—negative, neu-

[9]Respondents self-identified as either Democrats or Republicans in the survey; others were excluded.

[10]In Figure C.1 of Appendix C, we explore additional heterogeneity in labeling and show that, for instance, gender does not result in attitudinal differences for political imagery on immigration.

Figure 2: Distribution of Image Evaluation Scores by Party



Note: The plots show the density distribution of average evaluation scores across 832 images. The grey-shaded plots represent the average scores calculated for each image across all respondents. The blue-shaded plot shows the distribution of average evaluation scores for each image based solely on evaluations from respondents who self-identified as Democrats. The red-shaded density plot shows the distribution of average evaluation scores for each image based on evaluations from respondents who self-identified as Republicans.

tral, and positive—for both 'Sentiment' and 'Subject of Harm.'[11] To assign these categorical labels to the interval average evaluation scores (AES), we partition the range AES $\in [1, 7]$ into three segments:

$$\text{Categorical Label} = \begin{cases} \text{negative}, & \text{if AES} \leq 3 \\ \text{neutral}, & 3 < \text{if AES} < 5 \\ \text{positive}, & \text{if AES} \geq 5 \end{cases} \tag{1}$$

[11]Since 'Subject of Harm' was initially measured on a scale from dangerous to harmless, we assume that images perceived as more dangerous correspond to a negative sentiment, while those perceived as more harmless correspond to a positive sentiment.

Table 1 presents the distribution of images categorized by sentiment labels for Democrats and Republicans. The table shows that most images are labeled with a neutral sentiment. However, images rated by Democrats tend to receive more positive sentiment labels compared to those rated by Republicans. This difference points to the sentiment heterogeneity across partisan groups.

Table 1: Distribution of Labeled Images Across Sentiment Categories

	Sentiment		Subject of Harm	
	Democrats	Republicans	Democrats	Republicans
Negative	32	212	20	77
Neutral	713	563	499	645
Positive	87	57	313	110

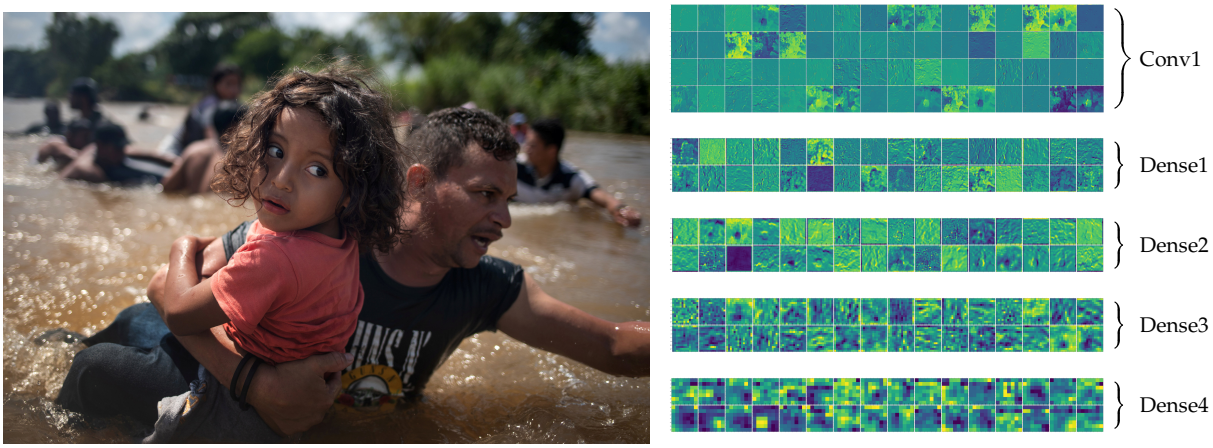
4 Visual Sentiment Prediction Using a Deep Learning Approach

Given the relatively small size of our image dataset, we use transfer learning[12] by fine-tuning pre-trained deep convolutional neural networks (CNN). Figure 3 illustrates feature maps at various levels of the CNN, showing how the image transforms as it progresses through the network's depth. We tested three well-established convolutional neural networks commonly used for visual sentiment analysis: ResNet50V2 (He et al., 2016), DenseNet-121, and DenseNet-169 (Huang et al., 2017), all pre-trained on the large ImageNet dataset. During fine-tuning, we adapted these models to our specific task by gradually 'unfreezing' and retraining several of the last layers or blocks of layers.

For each network, we implemented three fine-tuning strategies: 1) using the weights from the entire baseline model and only training the final fully connected layers responsible for classification (Version 1); 2) retraining one of the last convolutional blocks with ad-

[12]See the full description of the transfer learning approach and convolutional networks in Appendix E.

Figure 3: Feature Maps Visualization Heatmap (example with DenseNet-169)



Note: The left side displays the original image that was input into the neural network, while the right side illustrates the image transformations within the network. Each block represents the resulting image transformation at the first convolutional layer and at four dense blocks of the DenseNet-169. Specifically, Conv1 corresponds to the second layer of the original network, and Dense1 through Dense4 corresponds to layers 12, 58, 146, and 374 of the original network, respectively.

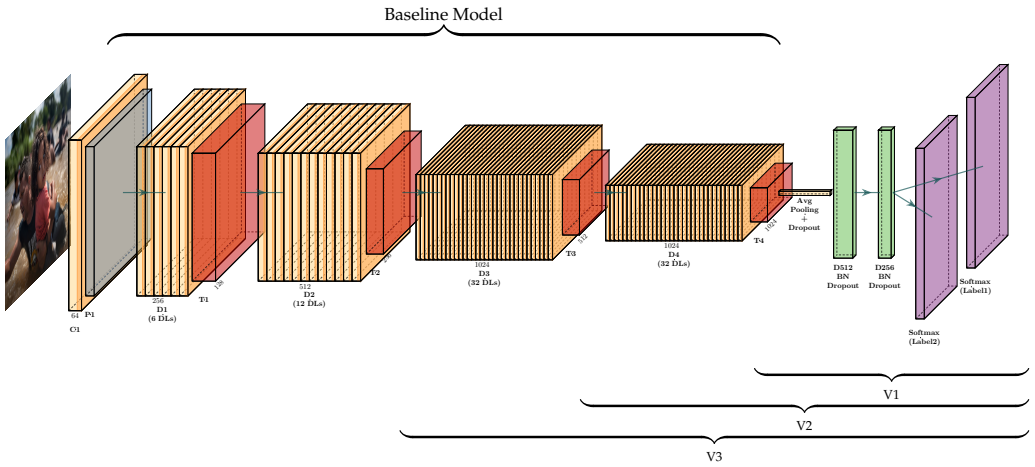
ditional pooling layers (or a dense block with an additional transition layer for DenseNets) and adding newly trained fully connected layers for classification (Version 2); 3) retraining two of the last convolutional blocks with additional pooling layers (or dense blocks with additional transition layers for DenseNet networks) and including newly trained fully connected layers for classification (Version 3) (see an example of the fine-tuning versions of the network architecture in Figure 4).

In each fine-tuning scenario, we incorporated Batch Normalization layers[13] to stabilize and accelerate training, and Dropout layers[14] to prevent overfitting.

[13] A batch normalization layer is a layer that normalizes the activations (output of a neuron after applying a specific activation function) of the previous layer for each mini-batch. This normalization is done by scaling and shifting the activations to have a mean of zero and a standard deviation of one The primary purpose of batch normalization is to stabilize and accelerate the training process by reducing the internal covariate shift, which helps in making the network less sensitive to the initialization of weights and learning rates.

[14] A dropout layer works by randomly setting a fraction of the input units to zero at each update during the training. Each time a batch of data is passed through the network, different sets of neurons are "dropped out" (ignored) and do not contribute to the forward pass or backpropagation. This reduces the likelihood of the network becoming overly dependent on specific neurons, leading to better generalization to out-of-sample data.

Figure 4: Model Architecture Indicating Fine-Tuning Versions for DenseNet-169



Note: The images illustrate the neural network architectures of the adapted convolutional neural networks (CNNs). Each model includes a baseline configuration, initially downloaded using Keras libraries and pre-trained on the ImageNet dataset. The versions Version (V1), Version 2 (V2), and Version 3 (V3) represent different re-training exercises for the CNN, showing the parts of the base model that have been updated and highlighting additional layers introduced between the base model and the final classification layer. While the schematic visualization depicts a softmax layer as the final classification layer, it is replaced with a linear activation function for the linear prediction task.

Each network was adapted for two main tasks: 1) multi-class classification with three categories (negative, neutral, and positive), and 2) linear prediction with sentiment labels measured on an interval scale [1,7]. We fine-tuned the models using the Python-based Keras API[15] with our dataset of 832 labeled images, which were divided into training, validation, and test sets with a 76.5/13.5/10 ratio taking into account an imbalanced distribution of labels across classes.[16] We used data augmentation[17] to increase the variation of the image data in the train set. The training was conducted using the Adam optimizer[18] with a learning rate of 0.001[19] over 50 epochs[20] with a batch size of 32[21] and included an early stopping mechanism with a patience parameter of 10 epochs.[22]

4.1 Dual-Task Classification

The key contribution of our approach is leveraging systematic attitudinal differences among respondents to generate multiple labels for each image. Hence, we derive two sen-

[15]Keras is an open-source, high-level neural networks API written in Python.

[16]Train set is the portion of data that is used to train the model, and on which the model learns the underlying patterns and relationships within this data by adjusting its parameters. The train set is usually the largest portion of the data, as the model needs a substantial amount of information to learn effectively. Validation set is used to fine-tune the model and make decisions about hyperparameters, such as learning rate or any performance with any additional layers to the model. The validation set helps in evaluating the model's performance during training without affecting the training data. By monitoring performance on the validation set, we may detect such issues as overfitting, when the model learns the train set too well and doesn't generalize effectively to new data. Test set is the final subset is an out-of-sample data that is kept completely separate from both the training and validation sets. After the model has been trained and tuned, it is evaluated on the test set to measure its performance on unseen data. This gives a clear, unbiased indication of how well the model generalizes to new data.

[17]We use ImageDataGenerator of Keras API to augment data on the fly, meaning that images are being augmented during process rather than saved into memory as a novel data, thus, each epoch can see slightly different variations of the data. To create a variability of images in the training process we used random rotation, width and height shifts, shear transformation, zoom transformation, and random horizontal flips.

[18]Adaptive Moment Estimation (Adam) adjusts the weights of the neural network to minimize the loss function, which measures the discrepancy between the network's predictions and the actual data.

[19]The learning rate is a hyperparameter that controls the size of the steps taken to update the model's weights during training.

[20]An epoch is one complete pass through the entire training dataset. Training over multiple epochs allows the model to refine its weights and improve performance.

[21]Batch size refers to the number of training samples used to update the model's weights in each epoch.

[22]Early stopping prevents overfitting by halting training when the validation loss does not improve for a specified number of epochs. The patience parameter determines how many epochs to wait for an improvement before stopping the training.

timent labels for each image based on separate evaluations from Democrats and Republicans–L_D and L_R, respectively–where L_D reflects the label from Democrats and L_R from Republicans.

Our approach centers on multi-task learning, where a single model is trained to handle multiple tasks simultaneously. The model's early layers learn shared features from input images, while later layers specialize for each task. Each task has a distinct loss function, and the overall loss is a weighted sum of these losses. This shared representation boosts task performance compared to training separate models (Crawshaw, 2020).

The key benefit is computational efficiency. By learning features relevant to multiple tasks concurrently, it avoids the need to train separate models, despite the added architectural complexity.

5 Results

The main result of our analysis is a classification model that predicts image sentiment separately for each party group. We find that the best performing models based on the prediction quality metrics for test set (out-of-sample data) are a fine-tuned DenseNet-169 Version 1 for predicting "Sentiment" and a fine-tuned ResNet50V2 Version 3 for "Subject of Harm". You may refer to Tables A.1 and A.2 in Appendix A, where we provide performance metrics for all trained and fine-tuned model specifications.

To evaluate and compare model performance in multi-class classification, we used two metrics: the *weighted F1* score[23] and the *accuracy* score, both of which can range up to a maximum of 1 (or 100% for accuracy).[24] Generally, the closer these values are to 1 on the test set, the better the classification model performs.

[23]The weighted F1 score for multi-class classification is calculated by taking the F1 score of each class, weighting it by the number of true instances in that class (support), and averaging these scores. To calculate the F1 score for each class, we first calculate Precision: $\text{Precision} = \frac{\text{True Positives}}{\text{True Positives}+\text{False Positives}}$ and Recall: $\text{Recall} = \frac{\text{True Positives}}{\text{True Positives}+\text{False Negatives}}$

[24]See Appendix A for a more detailed discussion.

The results presented in Table A.1 in Appendix show that for the "Sentiment" outcome, the DenseNet-169 model was the best performer. For the multi-class classification task, DenseNet-169 with all layers of the base model frozen achieved a weighted F1 score of 0.81 and 83% test set accuracy for Democrats, while it achieved an F1 score of 0.6 and 62% test set accuracy for Republicans. In the linear prediction task, DenseNet-169 with all layers frozen up to dense block 4 on the test set achieved a mean absolute error of 0.62 for Democrats and 0.62 for Republicans.[25]

For the "Subject of Harm" (see Table A.2 in Appendix for full results), the ResNet50V2 model with layers frozen up to convolution block 4 achieved a weighted F1 score of 0.91 and a test accuracy of 92% for Democrats, and a weighted F1 score of 0.85 with 86% test accuracy for Republicans. For linear prediction performance, the DenseNet169 model, with layers frozen up to dense block 4, produced mean absolute errors of 0.57 for Democrats and 0.65 for Republicans on the test set.[26]

These results suggest that our model predicts sentiment labels for Democrats more accurately than for Republicans across both sentiment outcomes.

To visually present the performance of the multi-class classification model, we plot confusion matrices, which evaluate the model by comparing its predictions with actual values. These matrices provide a detailed breakdown of correct and incorrect predictions for each class. Ideally, a well-performing model would have all observations concentrated along the upper-left to lower-right diagonal of the matrix, indicating perfect class prediction. Figures 5-6 show confusion matrices for the "Sentiment" and "Subject of Harm" outcomes, respectively.

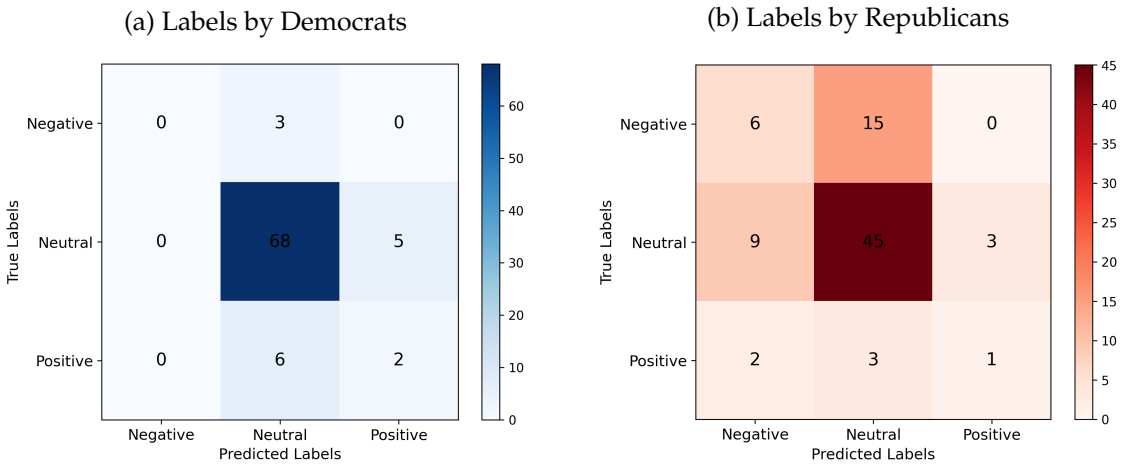
While F1 and accuracy scores offer an overall assessment of model quality, the confusion matrix illustrates which specific labels are challenging for the model to identify. Figures 5a-5b show that for the "Sentiment", our best-performing model (DenseNet-169

[25]See Figure F.1 in Appendix F. It plots actual vs. predicted labels of the linear predictions and the distribution of residuals.

[26]See Figure F.2 in Appendix F for actual vs. predicted labels and residual distributions.

V1) often confuses negative and neutral labels for Republicans (Figure 5b) and neutral and positive labels for Democrats (Figure 5a). For Democrats, the model frequently predicts only neutral, leading to consistent misclassifications of positive and negative labels.

Figure 5: Confusion Matrices: "Sentiment"



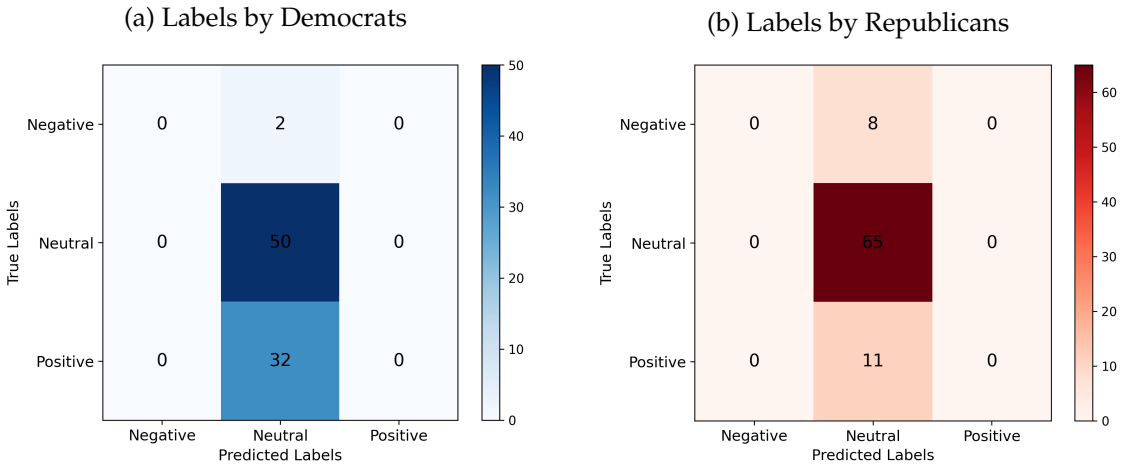
Note: Confusion matrices were constructed using the best-performing model for the "Sentiment" variable outcome - DenseNet-169 with additional dropout, batch normalization, and Dense layers, which was determined based on F1 score and accuracy on test set.

Figures 6a-6b present confusion matrices for the "Subject of Harm". Here, we observe a similar pattern, with the model consistently predicting neutral labels for all images, resulting in frequent confusing positive and neutral labels for Democrats (Figure 6a) and neutral with both positive and negative for Republicans (Figure 6b).

This over-prediction of neutral labels for both Democrats and Republicans is largely due to an imbalance in the labeled data, with a high prevalence of neutral labels (see Table 1). To improve model quality and address this imbalance, we focused on a subset of labeled images where Democrat and Republican labels differ. This subset included 292 images for 'Sentiment' and 328 images for 'Subject of Harm.'

Table 2 shows the class distribution, where neutral labels remain the majority but the distribution appears more balanced. Although this adjustment significantly reduced the size of our training set, we again applied a transfer learning approach to train and fine-

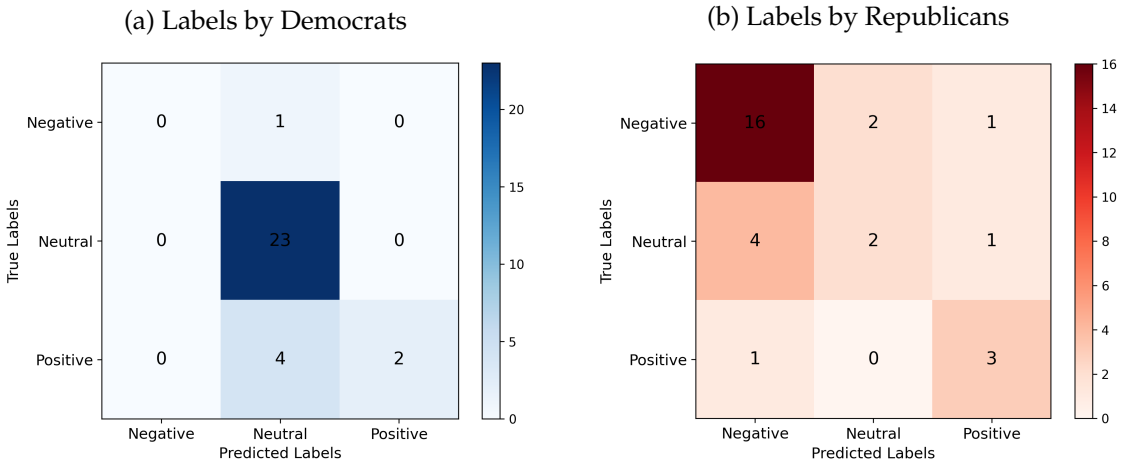
Figure 6: Confusion Matrices: "Subject of Harm"



Note: Confusion matrices are built based on the best-performing model according to the F1 score and accuracy on test set for the "Subject of Harm" outcome - ResNet50V2 model with retrained last two convolutional blocks and additional dropout, batch normalization, and Dense layers.

tune the same multi-class, multi-task classification models as in the baseline results.

Figure 7: Confusion Matrices: "Sentiment" (on a subsample)



Note: Confusion matrices were constructed using the best-performing model for the "Sentiment" variable outcome - DenseNet-169 (with pretrained layers frozen) with additional dropout, batch normalization, and Dense layers.

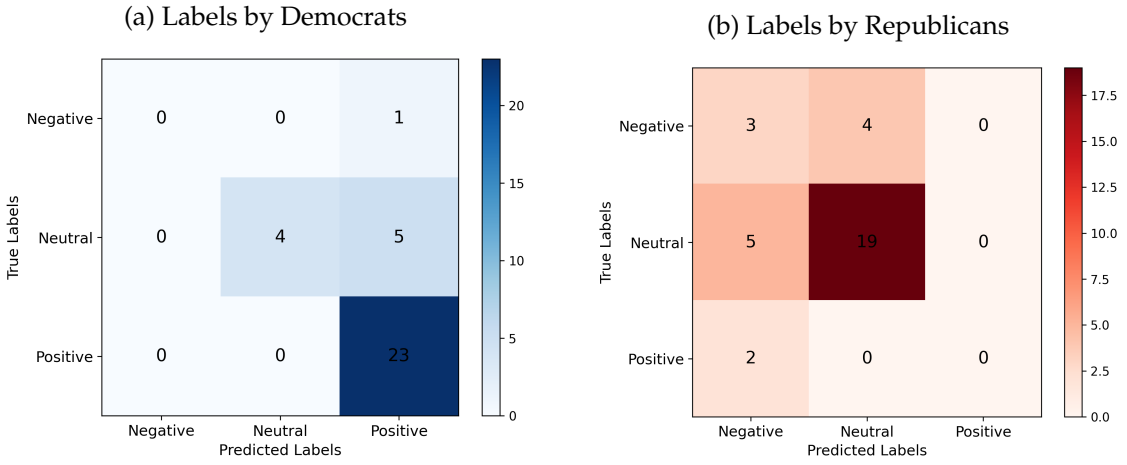
Results for all trained model specifications are shown in Tables B.3-B.4 in Appendix B. Among these, DenseNet169 Version 1 (with pretrained layers frozen) achieved the best performance for both "Sentiment" (Table B.3) and "Subject of Harm" (Table B.4).

Figures 7-8 present the confusion matrices for these top-performing models, showing

Table 2: Distribution of Labeled Images Across Sentiment Categories for Images where Democrats and Republicans Disagree

	Sentiment		Subject of Harm	
	Democrats	Republicans	Democrats	Republicans
Negative	7	187	11	68
Neutral	220	70	90	236
Positive	65	35	227	24

Figure 8: Confusion Matrices: "Subject of Harm" (on a subsample)



Note: Confusion matrices are built based on the best-performing model according to the F1 score and accuracy for the "Subject of Harm" outcome - DenseNet-169 (with pretrained layers frozen) with additional dropout, batch normalization, and Dense layers.

improved label predictions. While the models may still sometimes misclassify neutral labels as positive for Democrats or as negative for Republicans, they no longer predominantly predict neutral labels for most or all test set examples, as seen in the main results. Instead, these models now provide more balanced and accurate predictions across all labels.

6 Model Testing and Discussion

So what does our approach offer that traditional visual sentiment classifiers do not? The key takeaway from our research is the recognition, incorporation, and addressing of the systematic influence of coders' social characteristics, such as political affiliations, on the labeling of visual sentiment, especially *when these characteristics cannot be disregarded*.

Traditional visual sentiment analysis typically treats sentiment as a fixed attribute inherent to an image, determined solely by its visual content. Our study challenges this view by presenting both theoretical and empirical evidence that visual sentiment should be understood as an interplay between individual perceptions and attitudes and the content of the image itself, as both contribute to the perception of visual sentiment. Ignoring empirically consistent and theoretically significant sources of mislabeling—such as the effect of partisanship on how people react to visual depictions of politically polarizing topics—exacerbates bias and spreads misunderstanding about what sentiment is, how it is reflected, and, ultimately, how visual sentiment is generated when we scale up.

To illustrate the empirical usefulness of our approach, we conduct two exercises.

(1) First, we pass the images of our study through two of our classification models that best performed on the subsample of images evoking disagreement among partisans (DenseNet169 Version 1 from Tables B.3-B.4) and plot the results in Figure 9. Using the partisan-averaged true labels approach, which averages all coders' evaluations, we would assign a 'Neutral' label to 'Sentiment' (average score of 4.86) and 'Positive' label to 'Subject of Harm' (average score of 5.3) for image A, and a 'Neutral' label to 'Sentiment' (average score of 4.8) and a 'Positive' label to 'Subject of Harm' (average score of 5.6) for image B. However when we pass these images through our model, we predict that Democrats would assign a positive sentiment to the image of a man carrying a child, while Republicans evaluate this image neutrally. Conversely, Republicans would view a crowd marching as neutral, while Democrats see it as harmless (positive), as shown in Figure 9.

Figure 9: Predicted Sentiment Labels

Image A

"Sentiment" Dem: [Positive], Rep: [Neutral]
"Subject of Harm" Dem: [Positive], Rep: [Neutral]

Partisan-Averaged True Labels:
"Sentiment": [Neutral],
"Subject of Harm": [Positive]

Image B

"Sentiment" Dem: [Neutral], Rep: [Neutral]
"Subject of Harm" Dem: [Positive], Rep: [Neutral]

Partisan-Averaged True Labels:
"Sentiment": [Neutral], "Subject of Harm": [Positive]

Figure 10: Predicted Sentiment Labels (of Madrigal and Soroka (2023))

Image A

"Sentiment" Dem: [Neutral], Rep: [Negative]
"Subject of Harm" Dem: [Positive], Rep: [Neutral]

Partisan-Averaged True Labels:
"Sentiment": [Neutral], "Subject of Harm": [Neutral]

Image B

"Sentiment" Dem: [Positive], Rep: [Positive]
"Subject of Harm" Dem: [Positive], Rep: [Neutral]

(2) Second, we use images from the study by Madrigal and Soroka (2023) and include one of the images from their study in our sample, allowing us to know the actual label of this image. They argue that people's attitudes toward immigration, as depicted in images, are moderated by "threat sensitivity" (see p. 53). Specifically, predispositional threat sensitivity suggests that photos showing large groups of immigrants are likely to evoke feelings of threat, which can directly influence attitudes toward immigration. This effect is expected to be strongest among individuals who are naturally more sensitive to perceived threats. Feeding the images from their study into our classifiers confirm this expectation, as shown in Figures 10: there is a clear variation in sentiment between how Democrats and Republicans perceive images about immigration. The difference in sentiment aligns with how much individuals perceive the subjects in the images as potential threats, exactly as the authors found in their study while leveraging across responses produces a a misleading label. This indicates that our method predicts empirically mean-

ingful variation in visual sentiment and is generalizable.

Altogether, our approach is most effectively applied in contexts where variations in perceived sentiment towards visuals are likely driven by stable divisions, such as partisanship or ideology, as demonstrated in our study. The level of analytical granularity, however, can be adjusted to align with the specific goals and practical considerations of the researcher.

While this approach may not be essential for all tasks, it illustrates the importance of accounting for such divisions, particularly when analyzing politically sensitive content. Using our model, researchers can assess whether partisan differences influence the sentiment evoked by an image, facilitating more accurate and contextually informed interpretations of visual sentiment.

References

Ajzen, Icek. 2001. "Nature and operation of attitudes." *Annual review of psychology* 52(1):27–58.

Ajzen, Icek and Martin Fishbein. 2000. "Attitudes and the attitude-behavior relation: Reasoned and automatic processes." *European review of social psychology* 11(1):1–33.

Anastasopoulos, L Jason, Dhruvil Badani, Shiry Ginosar and Jake Ryland Williams. 2024. "Visible home style." *Electoral Studies* 90:102794.

Benoit, Kenneth, Drew Conway, Benjamin E Lauderdale, Michael Laver and Slava Mikhaylov. 2016. "Crowd-sourced text analysis: Reproducible and agile production of political data." *American Political Science Review* 110(2):278–295.

Bossetta, Michael and Rasmus Schmøkel. 2023. "Cross-platform emotions and audience engagement in social media political campaigning: Comparing candidates' Facebook and Instagram images in the 2020 US election." *Political Communication* 40(1):48–68.

Calanchini, Jimmy, Karen Gonsalkorale, Jeffrey W Sherman and Karl Christoph Klauer. 2013. "Counter-prejudicial training reduces activation of biased associations and enhances response monitoring." *European Journal of Social Psychology* 43(5):321–325.

Campbell, Angus. 1960. *The american voter*. University of Chicago Press.

Carpinella, Colleen and Nichole M Bauer. 2021. "A visual analysis of gender stereotypes in campaign advertising." *Politics, Groups, and Identities* 9(2):369–386.

Casas, Andreu and Nora Webb Williams. 2019. "Images that matter: Online protests and the mobilizing role of pictures." *Political Research Quarterly* 72(2):360–375.

Chavez, Leo R. 2023. *Covering immigration: Popular images and the politics of the nation*. Univ of California Press.

Crawshaw, Michael. 2020. "Multi-task learning with deep neural networks: A survey." *arXiv preprint arXiv:2009.09796* .

de Lima-Santos, Mathias-Felipe, Isabella Gonçalves, Marcos G Quiles, Lucia Mesquita and Wilson Ceron. 2023. "Visual Political Communication in a Polarized Society: A Longitudinal Study of Brazilian Presidential Elections on Instagram." *arXiv preprint arXiv:2310.00349* .

Domke, David, David Perlmutter and Meg Spratt. 2002. "The primes of our times? An examination of the 'power'of visual images." *Journalism* 3(2):131–159.

Eagly, Alice H. and Shelly Chaiken. 1993. *The psychology of attitudes*. Fort Worth, TX: Harcourt Brace Jovanovich.

Fahmy, Shahira, Sooyoung Cho, Wayne Wanta and Yonghoi Song. 2006. "Visual agenda-setting after 9/11: Individuals' emotions, image recall, and concern with terrorism." *Visual Communication Quarterly* 13(1):4–15.

Farris, Emily M and Heather Silber Mohamed. 2018. "Picturing immigration: How the media criminalizes immigrants." *Politics, Groups, and Identities* 6(4):814–824.

Feather, Norman T. 1985. "Attitudes, values, and attributions: Explanations of unemployment." *Journal of Personality and Social Psychology* 48(4):876.

Grabe, Maria Elizabeth and Erik P Bucy. 2014. Image bite analysis of political visuals: Understanding the visual framing process in election news. In *Sourcebook for Political Communication Research*. Routledge pp. 209–237.

Grabe, Maria Elizabeth and Erik Page Bucy. 2009. *Image bite politics: News and the visual framing of elections*. Oxford University Press.

Haynes, Chris, Jennifer Merolla and S Karthick Ramakrishnan. 2016. *Framing immigrants: News coverage, public opinion, and policy*. Russell Sage Foundation.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren and Jian Sun. 2016. Identity mappings in deep residual networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*. Springer pp. 630–645.

Huang, Gao, Zhuang Liu, Laurens Van Der Maaten and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 4700–4708.

Joo, Jungseock and Zachary C Steinert-Threlkeld. 2022. "Image as data: Automated content analysis for visual presentations of political actors and events." *Computational Communication Research* 4(1).

Kossaifi, Jean, Robert Walecki, Yannis Panagakis, Jie Shen, Maximilian Schmitt, Fabien Ringeval, Jing Han, Vedhas Pandit, Antoine Toisoul, Björn Schuller et al. 2019. "Sewa db: A rich database for audio-visual emotion and sentiment research in the wild." *IEEE transactions on pattern analysis and machine intelligence* 43(3):1022–1040.

Leong, Yuan Chang, Brent L Hughes, Yiyu Wang and Jamil Zaki. 2019. "Neurocomputational mechanisms underlying motivated seeing." *Nature human behaviour* 3(9):962–973.

Liberman, Akiva and Shelly Chaiken. 1996. "The direct effect of personal relevance on attitudes." *Personality and Social Psychology Bulletin* 22(3):269–279.

Madrigal, Guadalupe and Stuart Soroka. 2023. "Migrants, caravans, and the impact of news photos on immigration attitudes." *The International Journal of Press/Politics* 28(1):49–69.

Ortis, Alessandro, Giovanni Maria Farinella and Sebastiano Battiato. 2020. "Survey on visual sentiment analysis." *IET Image Processing* 14(8):1440–1456.

Petty, Richard E, Duane T Wegener and Leandre R Fabrigar. 1997. "Attitudes and attitude change." *Annual review of psychology* 48(1):609–647.

Schill, Dan. 2012. "The visual image and the political image: A review of visual communication research in the field of political communication." *Review of communication* 12(2):118–142.

Torres, Michelle. 2018. Framing a Protest: Determinants and Effects of Visual Frames. Technical report Working Paper.

Torres, Michelle. 2024. "A framework for the unsupervised and semi-supervised analysis of visual frames." *Political Analysis* 32(2):199–220.

Van Bavel, Jay J, Yi Jenny Xiao and William A Cunningham. 2012. "Evaluation is a dynamic process: Moving beyond dual system models." *Social and Personality Psychology Compass* 6(6):438–454.

Webb Williams, Nora. 2023. "What Type of Data Are Images?" *Webb Williams, Nora,'What Type of Data Are Images* .

Webb Williams, Nora, Andreu Casas, Kevin Aslett and John D. Wilkerson. 2023. "When Conservatives See Red but Liberals Feel Blue: Why Labeler-Characteristic Bias Matters for Data Annotation." *Available at SSRN* .

Yang, Jufeng, Dongyu She, Ming Sun, Ming-Ming Cheng, Paul L Rosin and Liang Wang. 2018. "Visual sentiment prediction based on automatic discovery of affective regions." *IEEE Transactions on Multimedia* 20(9):2513–2525.

Yang, Yi and Shawn Newsam. 2010. Bag-of-visual-words and spatial extensions for land-use classification. In *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*. pp. 270–279.

Zhao, Sicheng, Xingxu Yao, Jufeng Yang, Guoli Jia, Guiguang Ding, Tat-Seng Chua, Bjoern W Schuller and Kurt Keutzer. 2021. "Affective image content analysis: Two decades review and new perspectives." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44(10):6729–6751.