

Measuring Polarization in Visual Content Using Counterfactual Image Generation *

Mohammad Mosaffa
Cornell University

Omid Rafieian
Cornell University

Hema Yoganarasimhan
University of Washington

Abstract

Political polarization is an increasingly significant issue in American politics, influencing public discourse, policy, and consumer behavior. While studies on polarization in news media have extensively focused on verbal content, non-verbal elements, particularly visual content, have received less attention due to the complexity and high dimensionality of image data. Traditional descriptive approaches often rely on feature extraction from images, leading to biased polarization estimates due to information loss. In this paper, we introduce the Polarization Measurement Using Counterfactual Image Generation (PM-CIG) method, which combines economic theory with generative models and multi-modal deep learning to fully utilize the richness of image data and provide a theoretically grounded measure of polarization in visual content. Applying this framework to a decade-long dataset featuring 30 prominent politicians across 20 major news outlets, we identify significant polarization in visual content, with notable variations across outlets and politicians. At the news outlet level, we observe that Republican-leaning outlets are more likely to use positive images of Republican politicians while Democratic-leaning outlets tend to favor the opposite. At the politician level, our results reveal considerable differences in the degree of polarization, with certain politicians consistently portrayed in a more ideologically slanted manner than others.

*Note: The most updated version of the paper can be found in the following link: <https://github.com/mmosaffa/Measuring-Polarization-in-Visual-Content-Using-Counterfactual-Image-Generation>

1 Introduction

Political polarization has emerged as a central issue in American politics over the past few decades. Three out of ten Americans now consider polarization one of the most significant challenges facing the country ([Skelley and Fuong, 2022](#)). When asked to describe the current political climate, the term “divisive” was the most frequent response, and similar terms like “polarized” and “partisan” were among the most common responses by Americans ([Doherty et al., 2023](#)). As individuals become more deeply rooted in their political identities, the potential for cross-party dialogue, compromise, and effective governance diminishes. The impact of polarization extends beyond politics, affecting public policy, social cohesion, and the overall functioning of democratic institutions. Thus, it is important to understand factors that reflect or intensify political polarization.

Given its role in shaping individuals’ political attitudes and beliefs, the news media landscape provides an ideal environment to study political polarization. In particular, several studies have focused on the verbal content in news articles as a rich source of information to propose measures for media bias and polarization and examine the demand-driven motives for news outlets to create content that matches the partisan preferences of their readers ([Groseclose and Milyo, 2005](#); [Gentzkow and Shapiro, 2010](#)). The research in this domain suggests that the choice of words and framing of issues can reveal accurate information about the speaker or the writer ([Gentzkow et al., 2019](#)).

The existence of ideological slanting in the choice of words and framing poses important questions about the non-verbal aspect of news content. The non-verbal content conveys meaning through channels other than language, such as facial expression and body language. In news articles, the non-verbal content is often communicated through visual content such as images. Editors often pay great attention to the choice of visual content, as visuals are more memorable, processed more rapidly, and elicit stronger emotional responses than text ([Sullivan and Masters, 1988](#); [Townsend and Kahn, 2014](#)). Further, new generations of consumers are increasingly interested in non-verbal forms of news content. Nevertheless, only a few studies have focused on the visual content to study polarization, and no prior work has utilized the richness of visual information ([Peng, 2018](#); [Boxell, 2021](#); [Caprini, 2023](#)).

In this paper, we bridge this gap and develop a framework to study polarization in visual content. In particular, we aim to answer the following questions:

1. How can we quantify political polarization in visual content, particularly in the images used in news articles?
2. What is the extent of visual bias and polarization in media?
3. How does the polarization in visual content vary over time and across politicians and news outlets?

To answer these questions satisfactorily, we face several challenges. First, we need to clearly define the parameter of interest that measures political polarization in visual content. To address this challenge, we turn to the very structure of the problem: the editor’s choice of visual content. We construct a flexible utility framework that models the editor’s choice of images from a set of available options. Using a utility framework allows us to compare the utility derived from two images that are as similar as possible but differ

in just one feature. This comparison enables us to build a polarization measure centered around that feature. For instance, consider two images of Donald Trump that differ only in whether he is smiling. For each outlet, we can define the difference in the utility from using the image with and without a smile. Intuitively, the greater the variability in this utility difference across outlets, the higher the level of political polarization in visual content.

Our second challenge originates from the high dimensionality of the visual content. The common approach in the literature is to find a mapping that extracts a certain feature (e.g., smile in an image) from the image (Peng, 2018; Boxell, 2021). However, this feature extraction approach relies solely on the selected feature and disregards other information in the image, which can introduce *omitted variable bias* and *extraction bias* in the analysis. We address this challenge by employing a generative approach that creates counterfactual versions of the same image with and without the feature of interest. This method enables us to retain all the information in the image and minimizes the degree to which other factors can confound our polarization measure.

The third challenge involves estimating our choice model for editorial image selection, which requires observing data on the news outlet’s choice between the two counterfactual versions of the image. To address this challenge, we exploit variations in image choices across outlets for similar events. Specifically, we estimate a choice model that predicts the likelihood of an outlet selecting a particular image based on all available information about the news article. We then link these estimates to our target parameter of interest for visual polarization, demonstrating how the variation in image choices across outlets can be used to identify the parameter.

Together, we build a unified framework that combines the economic structure of the problem with generative models to generate comparable counterfactual images and measure polarization in visual content. The key advantage of our framework in comparison to the traditional regression-based approaches is in overcoming the potential confounding and misestimation of polarization due to information loss. Additionally, our generative strategy extends beyond the study of polarization and can be applied to other contexts involving visual data. Another key benefit of our framework is its ability to provide individual-level measures, enabling us to examine the heterogeneity of polarization over time and across different politicians and news outlets.

We apply our framework to a comprehensive dataset comprising over 50,000 records of 30 prominent politicians from both the Republican and Democratic parties across 20 major news outlets over a 10-year span from 2011 to 2021. First, we utilize a state-of-the-art deep learning approach to predict the outlet of an image based on its visual information, the textual content of the article, and contextual data about the politician, year, and other relevant factors. We then employ Generative Adversarial Networks (GANs) to create sets of counterfactual images for all politicians with and without a smile. Using the model estimated in the initial step, we measure how the predicted probability of the image belonging to a particular outlet changes between these counterfactual images. Finally, we use these differences to quantify the extent of polarization at both the aggregate and individual levels.

Our results indicate that both Democratic- and Republican-leaning outlets ideologically slant their visual content. Specifically, we find that, compared to a neutral image of a Republican politician, a smiling image

increases the utility for Republican-leaning outlets but decreases it for Democratic-leaning outlets. Conversely, for Democratic politicians, a smiling counterfactual image generates lower utility for Republican-leaning outlets and higher utility for Democratic-leaning outlets compared to the neutral image. Notably, the degree of polarization among news outlets varies significantly. Among conservative outlets, *Fox News* and *Daily Mail* exhibit more ideological slanting in images compared to *Wall Street Journal*. Similarly, among liberal outlets, *Washington Post*, and *New York Times* show a higher degree of ideological slanting in visual content than *NBC News*.

We further document a high degree of heterogeneity in the extent of polarization across individual politicians. On the Republican side, Donald Trump appears to be the most polarizing figure. That is, the gap in the extent of ideological slanting is remarkably large, with the Republican-leaning outlets receiving higher utility from using a positive and smiley portrayal of him compared to Democratic-leaning outlets. On the Democratic side, we find Barack Obama and Kamala Harris to be among the most polarizing figures who are very differently portrayed in Democratic-leaning and Republican-leaning outlets.

In summary, our paper makes several contributions to the literature. Methodologically, we propose a framework that combines economic theory with generative models to provide robust measures of media bias and polarization in visual content. A key innovation of our framework is the use of Generative Adversarial Networks (GANs) in a way consistent with experimentation to generate counterfactual images based on a feature of interest, addressing the bias due to information loss present in social science studies that utilize image data. As such, our framework is general and applicable to a variety of settings where researchers want to study a feature in images. Substantively, our work demonstrates the existence of ideological slanting in the visual content used by media outlets, with substantial heterogeneity observed across time, politicians, and news outlets.

2 Related Literature

First, our paper relates to the study of political polarization in news media, a phenomenon well documented through the analysis of textual data. For instance, [Groseclose and Milyo \(2005\)](#) quantify media bias by examining the frequency with which different media outlets cite various think tanks, uncovering a persistent liberal bias through textual citation analysis. Similarly, [Gentzkow and Shapiro \(2010\)](#) investigate the factors driving media slant, highlighting the significant roles of consumer preferences and political affiliations by analyzing the alignment of newspaper language with political parties. [Jensen et al. \(2012\)](#) study the polarization of political discourse by analyzing records of Congressional speech and the Google Ngrams corpus, discovering a notable increase in discourse polarization since the late 1990s. Furthermore, [Gentzkow et al. \(2019\)](#) focus on the linguistic divide in Congressional speeches, showing how Democrats and Republicans increasingly use distinct vocabularies. While these studies predominantly focus on textual data, our work shifts the focus to polarization in visual content, specifically examining how news article images contribute to this phenomenon.

Research in this realm expands from textual to visual content recently due to advances in computer vision techniques. [Peng \(2018\)](#) use a two-stage approach with Azure Microsoft to extract facial expressions from 13,000 images of the 2016 election in the first stage and then analyze them through a regression model in the second stage, revealing bias toward politicians aligned with a news outlet's stance. [Boxell](#)

(2021) expand this by analyzing 70,000 images across more politicians and outlets. Caprini (2023) further extend this by generating and analyzing textual descriptions of images alongside news articles using Azure, applying Gentzkow et al. (2019) to show how the alignment of visual and textual bias amplifies polarization. Our study advances previous research in three significant ways. First, we enhance the traditional two-stage approach by addressing its inherent biases and issues with omitted variables, introducing a more robust methodology called Polarization Measurement using Counterfactual Image Generation (PMCIG) that uses the rich information in the image content and estimates a measure of polarization in visual content. Second, our improved method delivers clearer and more precise results, enabling us to quantify polarization at both the news outlet and politician levels and to track their evolution over the past decade. Third, we leverage a comprehensive dataset spanning from 2011 to 2021, which allows us to capture long-term trends in political coverage that earlier studies may have missed.

Lastly, from a methodological perspective, our work aligns with the growing trend in social science research that leverages computer vision techniques to analyze unstructured image data, a shift driven by the increasing availability of such data (Harbert, 2021). For example, marketing studies use image analysis to explore consumer behavior and brand perception (Hartmann et al., 2021; Lee, 2021; Liu et al., 2020; Peng et al., 2020). Traditionally, researchers employ a two-stage approach: first, extracting features from images to create structured data; second, applying statistical analysis to these features (Davenport, 2017). However, this approach has limitations, including potential information loss, endogeneity issues, and oversimplification due to parametric assumptions. To address these challenges, recent methodologies have emerged. Wei and Malik (2022) identify biases in econometric models using machine-learned variables from unstructured data and propose solutions to improve accuracy. Singh and Zheng (2023) introduce the RieszIV estimator, which incorporates high-dimensional unstructured data directly into causal analysis to manage endogeneity. Additionally, Luo and Toubia (2024) employ GANs for Controllable Stimuli Generation (CSG), enabling precise manipulation of image attributes to isolate causal effects. Building on these advancements, our PMCIG method combines GAN-based image manipulation with non-parametric deep learning models to more accurately quantify polarization in an image feature, holding all other features constant.

3 Setting and Data

We collect publicly available data on images of politicians from media websites spanning a 10-year period for the study. Below, we describe our data collection, cleaning, and labeling strategy.

3.1 Data Collection Strategy

We collect data on 30 politicians and 20 news outlets over a span of ten years, from 2011 to 2021. The set of politicians consists of those who ran for important public offices and/or held important national roles during the ten-year span of 2011–2021 and were commonly searched on Google (based on their popularity on Google searches using Google Trends data). The top 20 major news outlets are selected from the list presented in Flaxman et al. (2016). The details about the selection process for politicians and news outlets and their names are shown in Web Appendix A.1.

The data collection is done using an application referred to as SerpAPI. SerpAPI facilitates efficient large-scale image scraping by leveraging Google's image search to extract relevant images and metadata.

The process involves generating search queries based on specific criteria, such as targeted news outlets and date ranges, to extract images, article links, titles, and dates.



Figure 1: Search Results for "Joe Biden" site:"<https://www.cnbc.com>" before:2022 after:2020

Figure 1 presents an example of a query we use for data collection, where we use "site:" to restrict searches to specific news websites, "before:" and "after:" to filter results by publication date, and exact phrase searches to capture relevant content precisely. For each query, we aim to collect 80 news articles for each politician from each news outlet for the specified period.¹ This information for each image in the query is compiled into a structured data frame with the following fields:

- **Image:** A URL pointing to the image.
- **Alt:** A short description or alternate text corresponding to the image.
- **Href:** A hyperlink where the related news can be found.
- **Title:** The title of the news article or caption associated with the image.
- **Query Parameter:** A search query string utilized to retrieve the image and associated news, emphasizing the political figure and the source website along with a defined temporal span.

Overall, our data collection strategy provides us with a comprehensive data set of 287,275 politicians' images.

3.2 Data Cleaning and Identifying Politicians

After the collection of our dataset, a crucial step is data cleaning to ensure that images of a certain politician correctly show his/her face. In our original data set, we face the following challenges:

- Some news images do not feature the intended politician but instead capture relevant scenes or contexts of the news without the individual's presence.
- Certain images, although retrieved under a specific individual's search query (e.g., Joe Biden), may inadvertently include another politician (e.g., Donald Trump), introducing cross-representation.
- Several news images include multiple politicians, complicating the singular analysis of each politician's records.

¹Some queries may have fewer articles/images about specific politicians.

To address these challenges, we design a two-phase computer vision framework, as illustrated in Figure 2. In the first phase, we use a series of computer vision models to only keep images with one face presented in them. In the second phase, we want to ensure that the face in an image belongs to the intended politician. For that purpose, we use a sample of 20 high-quality face images from all 30 politicians in our sample as references for our face verification model. We then use that model for each instance in our one-face sample obtained from the first phase to verify that the predicted label for the face is the same as the intended politician. We present this two-phase approach in greater detail in Online Appendix A.2. The result of our two-phase data collection procedure in Figure 2 is a set of 53,563 images that only show a single face that belongs to the intended politician.

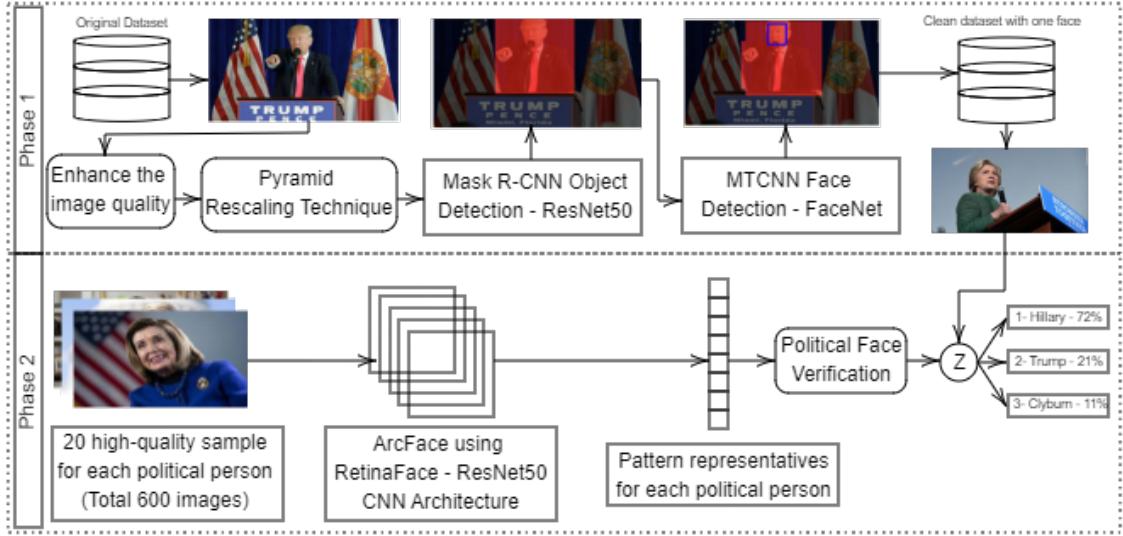


Figure 2: Two-phase computer vision framework for political person verification

4 Problem Definition

We formulate the problem from the perspective of a newsroom editor. As documented in earlier studies, images can have a significant impact on the extent to which readers engage with and click on news content (Matias et al., 2021). As such, editors must carefully select images for each article in order to maximize readers’ engagement with the article (Trilling et al., 2016). This involves choosing an image that appeals to the target audience and aligns with the editorial stance while ensuring that the visual elements complement the textual content. Figure 3 illustrates a sample news article, demonstrating the decision stage where an editor must decide to pick the most desired image for the given article.

CNN News	
Year:	2020
Politician:	Joe Biden
Politician's Side:	Democratic
Title:	Biden enters White House with slight rise in poll numbers

Figure 3: An editor of CNN selecting an image of Joe Biden for a given news article.

Consider a data set \mathcal{D} of news articles. Each news article i is characterized by a tuple (X_i, P_i, Y_i, Z_i) , where X_i represents the features associated with the key aspects of the article, such as its title, topic, text content, and publication date. P_i refers to the politician who is the focal subject of the article, Y_i is the news outlet that produces the article, and Z_i is the image and refers to the unstructured image data. The image Z_i can be interpreted as detailed pixel-level information capturing all the relevant aspects of the image, such as its background, brightness, other objects present, and the facial expression of the politician P_i .

The editor or news outlet receives utility U_i from producing article i with characteristics (X_i, P_i, Y_i, Z_i) , which is defined as follows:

$$U_i = u(Z_i, X_i, P_i, Y_i) + \varepsilon_i, \quad (1)$$

where $u(Z_i, X_i, P_i, Y_i)$ is the deterministic component and represents the editor’s expected utility from producing the article, and ε_i denotes the idiosyncratic error term, which follows an i.i.d. Type 1 Extreme Value distribution. This utility function captures all the main considerations of the editor when making image choice decisions, such as how well the image supports the outlet’s ideological stance (alignment with editorial policy, i.e., alignment of features of Z_i with Y_i), the image’s ability to attract and retain readers (reader-engagement), and the aesthetic quality and/or emotional impact of the image (visual appeal).

Intuitively, ideological slanting in images means that the outlet receives a higher (lower) utility from a more positive visual portrayal of a politician from the same (opposite) ideological side. The key challenge in developing a measure ideological slant in visual content comes from the ambiguity around the definition of a “positive” portrayal: an image is an unstructured and high-dimensional object, and there are presumably numerous ways for the outlet to choose a more or less positive image. As such, we need to define ideological slanting for a certain image feature. In our analysis, we focus on the presence of *smile* as the feature of interest because there is consensus that smile is a feature that contributes to a more positive portrayal ([Sülfow and Maurer, 2019](#)). However, our framework is not restricted to this particular feature, and can easily extend to other features or sets of features.

Let T_i denote whether the subject in image Z_i is smiling or not. To define the ideological slant measure for feature T_i , we need to compare the utility from two articles that are identical in all respects, except for the smile feature. Similar to the causal inference literature, we therefore define two counterfactual images

and denote them by $Z(T = 1)$ and $Z(T = 0)$, where $Z(T = 1)$ is the image with a smile and $Z(T = 0)$ is the same image without a smile. For ease of exposition, let $Z^{(-T)}$ denote all the information in image Z except the smile feature T . This implies that $Z(T = 1) \equiv (T = 1, Z^{(-T)})$ and $Z(T = 0) \equiv (T = 0, Z^{(-T)})$, i.e., the two counterfactual versions are identical in all features other than smile. Using this counterfactual set of images, we can define a measure of utility difference for a given news outlet y and politician p with respect to visual feature T as follows:

$$\Delta^T u(p, y) = \mathbb{E}_X [u(Z(T = 1), X, P, Y) - u(Z(T = 0), X, P, Y) \mid P = p, Y = y]. \quad (2)$$

This utility difference helps isolate the utility increase or decrease for a news outlet that solely comes from the presence of a smile in the image of any given politician. However, we cannot necessarily link this utility difference to ideological preferences, because there can be vertical preferences for the smile feature. For example, all news outlets may prefer a smiling image of a politician. What defines the polarization among news outlets is the difference in the utility difference $\Delta^T u(p, y)$ across news outlets. This helps cancel out the vertical preference for the feature and the remaining difference can be linked to the ideological slanting of the visual content. For example, we expect news outlets with higher conservative audience share (e.g., Fox News) to have a higher $\Delta^T u(p, y)$ for a conservative politician (e.g., Donald Trump) than a news outlet with lower conservative audience share (e.g., CNN). The difference in the utility difference across news outlets measures the polarization in the visual content as follows:

Definition 1. *For a given politician p and any pair of news outlets y_1 and y_2 , polarization in visual content with respect to feature T is defined as follows:*

$$\rho^T(p, y_1, y_2) = \Delta^T u(p, y_1) - \Delta^T u(p, y_2). \quad (3)$$

The proposed definition of polarization in visual content, $\rho^T(p, y_1, y_2)$, is flexible and allows for detailed analysis at the level of individual politicians and news outlets. For example, by examining the measure for a specific politician like Donald Trump across different outlets such as CNN and Fox News, we can observe how each outlet's editorial bias might influence the visual portrayal of the same politician. Fox News, which has a higher conservative audience share, might show a stronger preference for images where Trump is smiling, reflecting a more positive portrayal, while CNN, with a lower conservative audience share, might show a lesser preference for the same feature. This granularity in the analysis allows us to capture the varying degrees of ideological slant in visual content specific to each outlet's audience. Furthermore, by aggregating these measures across conservative outlets like Fox News and liberal outlets like CNN, we can assess the broader pattern of polarization in media representations, providing insights into how different ideological groups portray the same political figures.

5 Standard Reduced-Form Approach

In this section, we examine the standard reduced-form approach to quantifying polarization in visual content. In §5.1, we describe the two-step reduced-form approach and connect the estimated parameter under this approach to the measure of polarization in visual content as defined in Equation (3)). Next, we theoretically

examine the problem in §5.2 and present the drawbacks of this approach. Finally, in §5.3, we present the results from the reduced-form model and demonstrate the issues with this approach.

5.1 Two-step Model

Measuring $\rho^T(p, y_1, y_2)$ is inherently complex when dealing with unstructured image data. One approach is using a reduced-form model, such as regression, to study a function of $\rho^T(p, y_1, y_2)$. As in §4, we use data $\mathcal{D} = \{X_i, P_i, Y_i, Z_i\}$, where i is the index for each article, X_i is the set of non-visual features of the article, P_i is the politician presented in the article, Y_i is the outlet producing the article, and Z_i is the image shown in the article. As mentioned earlier, we want to measure polarization in visual content with respect to a focal feature T , such as whether the politician in the image is smiling or not. For each article i , we use T_i to denote this focal feature.

The key challenge is that we do not directly observe T_i . As a result, a common approach in social science settings is to use a *Two-step Approach*. This approach addresses the high dimensionality and complexity of unstructured data by first extracting meaningful features and then using these features in a structured econometric model. See [Davenport \(2017\)](#) for a general discussion of this approach in the broader social sciences literature, and [Boxell \(2021\)](#) and [Peng \(2018\)](#) for applications of this approach to the context of image polarization in media.² Methodologically, the two-step approach can be outlined as follows:

1. *Feature extraction using a machine learning model:* The first step involves using an off-the-shelf machine learning model, denoted as f_1 , to extract relevant feature(s) T from the unstructured data Z . This model $f_1 : Z \rightarrow T$ is used to derive the feature(s) T :

$$\hat{T} = f_1(Z) \quad (4)$$

2. *Econometric analysis:* After extracting features \hat{T} , the second step involves analyzing these features within a structured econometric model. There are two potential ways to relate \hat{T} and Y :

$$Y = f_2^{\text{T-independent}}(\hat{T}, X; \beta) + \varepsilon, \quad \text{or,} \quad (5a)$$

$$\hat{T} = f_2^{\text{T-dependent}}(Y, X; \beta) + \varepsilon, \quad (5b)$$

where $f_2^{\text{T-independent}}$ is the second-stage econometric model where extracted feature \hat{T} is used as an independent variable, and $f_2^{\text{T-dependent}}$ is the second-stage model where \hat{T} is used as dependent variable.

Social scientists typically prefer the second form because \hat{T} is an estimated variable that could include measurement error, and therefore using it as the independent variable is not appropriate ([Bollen and Davis, 2009](#)).

²Beyond the political polarization context, this two-step approach is now commonly used in marketing research involving images in other settings as well. For instance, unstructured data such as video and audio streams (Z) are analyzed to extract features like facial expressions (T), which are then used to study their impact on user engagement (Y) ([Lu et al., 2021](#)). In the online labor market, profile pictures (Z) are examined to identify features such as perceived race or attire (T) to assess job offer likelihood (Y), highlighting visual biases in employment opportunities ([Troncoso and Luo, 2022](#)). Additionally, research on social media posts about e-cigarettes (Z) extracts demographic features (T) to study tax policy compliance (Y), revealing demographic responses to legislative changes ([Anand and Kadiyali, 2024](#)).

We now describe how this approach can be used to quantify polarization in visual content with respect to feature T , defined as $\rho^T(p, y_1, y_2)$ in Equation (3). Consider a simple example where news outlets from Democratic-leaning outlets like the New York Times, CNN, and BBC (y_1) and Republican-leaning outlets like Fox News, Newsmax, and the Daily Mail (y_2) choose images for a politician, say Hillary Clinton (p). In this context, the first step involves estimating the binary indicator \hat{T} , which denotes whether the focal politician (Hillary Clinton) is smiling in the image ($\hat{T} = 1$) or not ($\hat{T} = 0$). This can be obtained from standard off-the-shelf emotion recognition models such as Face++³, Microsoft Azure⁴ or through in-house model training where we first label a sample of the images using human subjects and then train a model to predict the label given the image.

For the second step, we can use a logistic regression where we regress \hat{T}_i on characteristics X_i and Y_i . We want to see how our estimated parameters connect to the parameters of interest. For notation simplicity, consider the case where Y is a binary variable (e.g., $Y = 1$ for Democratic-leaning outlets, and $Y = 0$ for Republican-leaning outlets). We can write the log odds ratio for the resulting logistic regression as follows:

$$\log \left(\frac{\Pr(\hat{T} = 1 | Y, P, X)}{\Pr(\hat{T} = 0 | Y, P, X)} \right) = \alpha_0 + \alpha_1 X + \alpha_2 P + \beta Y \quad (6)$$

If the model in the equation above is well-specified, the log odds ratio characterizes the utility difference $\Delta^T u(p, y)$ between choosing an image with a smile versus one without a smile, as defined in Equation (2). In that event, we can connect the parameter β to the polarization measure defined in Equation (3) as follows:

$$\begin{aligned} \rho^T(p, Y = 1, Y = 0) &= \Delta^T u(p, Y = 1) - \Delta^T u(p, Y = 0) \\ &= (\alpha_0 + \alpha_1 X + \alpha_2 P + \beta) - (\alpha_0 + \alpha_1 X + \alpha_2 P) \\ &= \beta \end{aligned} \quad (7)$$

Thus, the coefficient β in the logistic regression model serves as an empirical estimate of the polarization measurement $\rho^T(p, Y = 1, Y = 0)$. If $\beta > 0$, it suggests that Democratic-leaning outlets ($Y = 1$) derive a higher utility from using a smiling image of Hillary Clinton compared to Republican-leaning outlets, implying a greater $\Delta^T u$ for the Democratic-leaning outlets. Conversely, if $\beta < 0$, Republican-leaning outlets ($Y = 0$) have a higher utility difference compared to Democratic-leaning outlets, indicating a greater $\Delta^T u$. Therefore, $\rho^T(p, Y = 1, Y = 0)$ can be approximated by β , and provides a quantitative measure of the ideological polarization across news outlets. This example is empirically investigated in detail in §5.3.

In linking the parameter estimate β to the polarization parameter of interest ρ^T , a key required assumption is that the logistic regression model is well-specified. That is, the estimates—particularly β —are not biased. In the following sections, we first theoretically examine the bias in β and then present empirical evidence in our setting to illustrate how this estimate could be biased.

³Face++ Emotion Recognition: <https://www.faceplusplus.com/emotion-recognition/>

⁴Microsoft Azure Vision: <https://azure.microsoft.com/en-us/services/cognitive-services/face/>

5.2 Drawbacks of This Approach

While the two-step model in §5.1 is easy to interpret and apply, it has two key drawbacks that can lead to incorrect inference: (1) *extraction bias*, and (2) *omitted variable bias*. In this section, we theoretically characterize these biases and discuss how they can appear in our problem.

We first discuss extraction bias, which occurs due to the imperfection of the feature extraction process (Wei and Malik, 2022). That is, the true label T may differ from the extracted label $\hat{T} = f_1(Z)$. This is because the first step in the two-step reduced form approach described in §5.1 usually contains some error in estimating the true label that can be characterized as ε_1 as follows:

$$T = f_1(Z) + \varepsilon_1 = \hat{T} + \varepsilon_1 \quad (8)$$

The error in Equation (8) can be viewed as a measurement error. From the econometrics literature, we know that the systematic dependence of this error to other relevant factors in the model can be consequential for the second step estimation. As such, we decompose this error into two parts, such that $\varepsilon_1 = \varepsilon_r + \varepsilon_e$, where ε_r is the random i.i.d. noise in the measurement process, and ε_e is the extraction error that can be correlated with the other relevant image features contained in $Z^{(-T)}$. For instance, a machine learning model designed to detect the facial expression of faces (e.g., whether a person is smiling) may also inadvertently capture brightness levels in the image, as brighter images are often associated with happier expressions. We rewrite the relationship between T and \hat{T} as follows:

$$T = \hat{T} + \varepsilon_r + \varepsilon_e \quad (9)$$

The second type of bias arises from relevant factors omitted from the second step, which can lead to the well-known *omitted variable bias*. This issue can arise even if we access the true labels T . To illustrate this point, we characterize the second-step relationship between the true label and covariates as follows:

$$T = f_2(Y, X) + \varepsilon_2, \quad (10)$$

where ε_2 is what is not explained by the observed covariates Y and X through the semi-parametric function f_2 . If this error term is independent of the covariates, we can identify function f_2 . However, this is often a very strong assumption given the amount of information in $Z^{(-T)}$ that is omitted from the model. For example, consider a collection of images by Donald Trump in articles by CNN and Fox News, where the smile labels are accurate, i.e., $\hat{T} = T$. Now, consider an image feature, such as the presence of the US flag. It is reasonable to see a higher likelihood of images with US flag in Fox News given their higher share of nationalist viewers. At the same time, it is more likely for a politician to smile in formal settings with flags present in the background. In this scenario, estimating the parameters of Equation (10) will lead to function f_2 picking up the association between the presence of flag and the smile in the image, because the presence of flag is omitted from the model. Given the high-dimensional and unstructured nature of the images, numerous other image features can result in omitted variable bias in our estimates. To characterize the

endogenous part of the error term ε_2 , we rewrite Equation (10) as follows:

$$T = f_2(Y, X) + \varepsilon_0 + \varepsilon_{ov}, \quad (11)$$

where ε_0 is the part that is independent of covariates, and ε_{ov} is the part that is correlated with Y or X , which can lead to omitted variable bias.

Together, if we want to estimate the second step of the two-step model in §5.1 by using \hat{T} as the outcome, the errors in Equation (9) also appear in the second step equation as follows:

$$\hat{T} = f_2(Y, X) + \varepsilon_0 + \varepsilon_{ov} + \varepsilon_r + \varepsilon_e, \quad (12)$$

where ε_{ov} leads to omitted variable bias, ε_e leads to extraction bias due to the use a machine learning model in the first step, and ε_r contributes to a higher uncertainty in model estimates. In Appendix B, we theoretically expand on these forms of bias that arise in our problem in a context with OLS estimator and characterize the bias.

5.3 Empirical Evidence

In this section, we empirically examine if the sources of biasedness discussed in §5.2 can bias the estimates of the two-step reduced form model outlined in 5.1. We focus on visual representations of Hillary Clinton as p , where y_1 corresponds to images published by Democratic-leaning news outlets such as *CNN*, *The New York Times*, and the *BBC*, and y_2 corresponds to those from Republican-leaning outlets like *Fox News*, *Newsmax*, and *The Daily Mail*. Our goal is to estimate the parameter β as characterized in Equation (6). For illustrative purposes, we examine a dataset comprising 1,021 articles, each containing images of Hillary Clinton. While this example provides a simplified analysis, we refer readers to Web Appendix C for a comprehensive examination of multiple politicians and a broader range of news outlets, where the primary insights remain consistent.

Following the two-step approach, we first employ the DeepFace model, proposed by Facebook (Taigman et al., 2014), as the machine learning algorithm f_1 to analyze the visual content Z . Specifically, DeepFace extracts facial expressions from images of Hillary Clinton, identifying whether she is smiling ($Z(\hat{T} = 1)$) or not ($Z(\hat{T} = 0)$).⁵ In the second step, we apply a logistic regression model f_2 to quantify the relationship between the estimated facial expression \hat{T} and the independent variable Y , which indicates the slant in using the smile for the politician of interest. The relationship is captured through the parameter β , where we model the probability of Clinton's happiness as a function of whether the image originated from a Republican or Democratic news source. The regression model is expressed as:

$$\log \left(\frac{\Pr(\hat{T} = 1 | Y)}{\Pr(\hat{T} = 0 | Y)} \right) = \alpha + \beta \cdot Y,$$

where $Y = 1$ represents Democratic outlets (*CNN*, *The New York Times*, or *BBC*) and $Y = 0$ corresponds to

⁵DeepFace utilizes a deep neural network with a nine-layer architecture and over 120 million parameters, including convolutional layers and a 3D alignment step. This model has been adapted for tasks such as facial expression and gender recognition and has been applied in various social science contexts (Dvorkin et al., 2021; Luca et al., 2022).

Republican outlets (Fox News, Newsmax, or *The Daily Mail*). The coefficients α (intercept) and β (slope) are estimated from the data. For simplicity, this example excludes article-level features X , although they can be incorporated, as shown in the comprehensive analysis in Web Appendix C.

This specification is a reduced-form model as it ignores the contextual information of a given image, represented by $Z^{(-T)}$. To account for the effect of information loss, consider the modified specification to include certain image characteristics of $Z^{(-T)}$. Specifically, we select three arbitrary image features: edge density, brightness, and contrast, which capture relevant aspects of the visual context. Thus, the extended specification of the logistic regression model becomes:

$$\log \left(\frac{\Pr(\hat{T} = 1 | Y, Z^{(-T)})}{\Pr(\hat{T} = 0 | Y, Z^{(-T)})} \right) = \alpha + \beta \cdot Y + \gamma_1 \cdot \text{EdgeDensity} + \gamma_2 \cdot \text{Brightness} + \gamma_3 \cdot \text{Contrast},$$

where $Z^{(-T)}$ includes the additional image characteristics, and γ_1 , γ_2 , and γ_3 are the coefficients that measure the effects of edge density, brightness, and contrast, respectively.

The results from both specifications are shown in Table 1. Column (1) presents the reduced-form model without image characteristics. In the reduced-form model, the *Democratic Outlets* variable is statistically significant ($p < 0.01$), with a positive coefficient of 0.369, indicating that Democratic news outlets are more likely to show Hillary Clinton smiling as a democratic politician. However, when we include contextual image characteristics—edge density, brightness, and contrast—in Column (2), the coefficient for *Democratic Outlets* decreases to 0.278 and loses statistical significance ($p < 0.10$). This reduction in both magnitude and significance suggests that part of the effect initially attributed to *Democratic Outlets* in the reduced-form model was actually capturing differences in how images were presented across outlets. The omission of these image characteristics led to biased estimates, as visual factors like contrast or edge density play an important role in shaping the portrayal of the subject, in this case, Hillary Clinton.

In Column (3), we find that the change in the coefficient for Democratic Outlets comes from the fact that the variable *Democratic Outlets* is correlated with some of the three image features excluded from the model in Column (1). It is crucial to emphasize that edge density, brightness, and contrast are just three arbitrary image characteristics selected from the much broader set of potential contextual features $Z^{(-T)}$. The visual content of an image encompasses a wide range of information, and the choice of these particular features is meant to demonstrate that even a limited selection of characteristics can significantly impact the results. The empirical evidence confirms the existence of extraction and/or omitted variable bias and highlights the crucial role it plays in inference when social scientists use reduced-form models. Failing to account for important contextual factors, such as image characteristics, can lead to biased estimates and incorrect conclusions about media bias and political polarization.

6 Algorithm: Polarization Measurement Using Counterfactual Image Generation

As discussed in §5, the feature extraction approach is subject to information loss as it ignores the high-dimensional information contained in the image. In this section, we develop an algorithm that recovers the polarization measure defined in §4, without incurring information loss. In this section, we introduce our algorithm: Polarization Measurement Using Counterfactual Image Generation (PMCIG).

Variable	Dependent Variable: \hat{T} (Happiness)			Dependent Variable: Y
	(1)	(2)	(3)	
Intercept	-1.170*** (0.104)	-1.809*** (0.257)	-1.604*** (0.235)	
Democratic Outlets (Y)	0.369** (0.141)	0.278† (0.145)		
Brightness Level		0.356 (0.561)	0.325 (0.512)	
Edge Density		0.878† (0.477)	2.312*** (0.456)	
Contrast		0.537 (0.561)	1.450** (0.513)	
No. of Obs.	1021	1021	1021	
Pseudo- R^2	0.005	0.013	0.040	
Log-Likelihood	-595.4	-591.3	-679.2	
Likelihood Ratio Test	6.87**	14.97**	56.94***	
<i>Note:</i>		†p<0.10; *p<0.05; **p<0.01; ***p<0.001		

Table 1: Results for logistic regressions. Columns 1-3 show results where T is the binary dependent variable (happiness), while column 4 shows results where Y (news outlet side) is the dependent variable where Republican outlets (Fox News, News Max, Daily Mail) are coded as $Y = 0$ and Democratic outlets (New York Times, BBC, CNN) are coded as $Y = 1$. Image characteristics include brightness level, edge density, and contrast.

Recall that the visual polarization measure is defined as:

$$\rho^T(p, y_1, y_2) = \Delta^T u(p, y_1) - \Delta^T u(p, y_2),$$

where $\Delta^T u(p, y) = \mathbb{E}_X [u(Z(T = 1), X, P, Y) - u(Z(T = 0), X, P, Y) | P = p, Y = y]$. Before outlining our approach to measuring visual polarization, we first present a thought experiment of what an ideal experiment to measure $\rho^T(p, y_1, y_2)$ would look like. Notice that in order to measure $\Delta^T u(p, y)$ for a given politician-outlet pair (p, y) for a feature/treatment T , we need to observe the difference in the editors' utilities (or choice probabilities) for two images that are exactly the same on all dimensions except T . Thus, an ideal experiment to measure $\rho^T(p, y_1, y_2)$ would involve presenting the editors of outlets y_1 and y_2 two images of politician p , that are exactly the same on all dimensions except the feature of interest, i.e., one where the politician is smiling (or the feature of interest T is turned on, i.e., $Z(T = 1) \equiv (T = 1, Z^{(-T)})$) and one where s/he is not smiling (i.e., $Z(T = 0) \equiv (T = 0, Z^{(-T)})$). Then, the observed difference in relative choice probabilities of the two images ($Z(T = 1)$ and $Z(T = 0)$) across the two outlets can be linked to polarization measure $\rho^T(p, y_1, y_2)$.

However, our data comes from an observational setting rather than this ideal experiment. Therefore, we only observe realized combinations of Z and y . For example, for an image Z with $T = 1$, we may see that it was chosen by outlet y_1 , as in the case where Biden's smiling image was chosen by CNN in Figure 3. However, we do not see what would be the probability of this image being chosen by y_1 if everything

about it was the same, but the feature/treatment of interest was turned off, i.e., $T = 0$. Similarly, we also do not see what would have been the likelihood of outlet y_2 choosing this image for the two cases: when $T = 1$ and when $T = 0$. Thus, we only observe one realized combination of $y-T$ for an image Z , as shown in Table 2. However, in order to quantify/measure polarization, we need to be able to reliably estimate/model the three other counterfactual outcomes for each image Z in the data. This limitation is similar in spirit to the well-established unobservability challenge in the potential outcomes framework ([Angrist and Pischke, 2009](#)).

	Treatment On: $Z(T = 1)$	Treatment off: $Z(T = 0)$
Outlet $y_1 = \text{CNN}$	✓	X
Outlet $y_2 = \text{FoxNews}$	X	X

Table 2: Factual and Counterfactual outcomes for the example from Figure 3, where CNN chose a smiling image of Biden.

As we can see from Table 2, there are two key challenges that we need to overcome to measure visual polarization.

- **Challenge 1: Counterfactual Image Generation**

Our polarization measure is defined for two counterfactual versions of an image that only differ in one feature. Using two counterfactual image versions addresses the issue of information loss in reduced-form approaches by using the rich information contained in images as opposed to mapping the image to a single feature. However, these counterfactual versions are not both observed in the data. Therefore, we need to employ a counterfactual image generation framework.

- **Challenge 2: Identification of the Polarization Parameter from Observed Data**

Second, even after we have two versions of each image ($Z(T = 1)$ and $Z(T = 0)$), we need to identify the polarization parameter $\rho^T(p, y_1, y_2)$ for any pair of news outlets, which is defined based on the news outlets' utility. However, identifying the utility function is challenging because we do not observe the news outlets' image selection process. Therefore, we need a systematic approach to use the observed data and identify the polarization parameter even when the identification of utility functions is not possible.

In this section, we first discuss our solution to both challenges and then present the unified algorithm.

6.1 Counterfactual Image Generation

As discussed in §5.2, standard reduced-form approaches that use off-the-shelf ML models to extract the feature/treatment of interest (T) suffer from omitted variable bias and extraction bias because the feature T is often correlated with other aspects of the image ($Z^{(-T)}$) that also affect an outlet's preference or likelihood of choosing that image. Therefore, it is important to isolate the feature of interest, such as the facial expression, without omitting the rich information contained in the image. To address this challenge, we leverage the recent developments in generative image models that allow us to manipulate images such that we can modify one specific aspect of an image (e.g., feature T) while keeping everything else constant. Generative models have recently been employed in several studies in visual/image analysis: [Athey et al. \(2022\)](#) use generative models to adjust features such as smiles in profile images to examine their effects on user preferences

and economic transactions in a micro-lending platform. [Ludwig and Mullainathan \(2024\)](#) utilizes these models to manipulate facial features, exploring how these alterations affect judicial decisions by modifying perceived attributes. Finally, in [Luo and Toubia \(2024\)](#), these models create facial images to assess the impact of perceived gender traits on discrimination within online marketplaces.

Specifically, we define two operators, π^1 and π^0 , which transform an image Z into either an image where the treatment/feature of interest is turned on $Z(T = 1)$ or one where it is turned off $Z(T = 0)$. The purpose of these operators is to create a set of control and treatment images such that the only difference between a pair of two images is the feature of interest (e.g., the facial expression or smile). The operators are defined as follows:

- π^1 : This operator ensures that the feature or emotion T is turned on. For example, when the treatment of interest is smile, then π^1 transforms an image of a politician into one where the politician is smiling, regardless of the initial state of T . Formally, applying π^1 to an image Z results in:

$$\pi^1(Z) = (T = 1, Z^{(-T)}) = Z^1. \quad (13)$$

- π^0 : This operator ensures that the feature or emotion T is turned off. For instance, π^0 transforms an image of a politician into one where the politician has a neutral expression, regardless of the initial state of T . Formally, applying π^0 to an image Z results in:

$$\pi^0(Z) = (T = 0, Z^{(-T)}) = Z^0. \quad (14)$$

We now discuss how we apply this idea to generate counterfactual images in our study. For each politician p , we select three neutral images $\tilde{Z}_{p1}^0, \tilde{Z}_{p2}^0, \tilde{Z}_{p3}^0$ to ensure a balanced and unbiased representation of the politician in a neutral manner. This is done to reduce potential biases that could arise from using a single image. To generate the counterfactual version of each image with a smile, we employ [AILabTools](#), which utilizes conditional Generative Adversarial Networks (cGANs) ([Mirza and Osindero, 2014](#)) for advanced image manipulation. This platform is chosen because it allows a precise modification of attributes, such as adding a smile while maintaining the authenticity of the other features in the image. The use of AILab ensures high-quality outputs, which is critical for our analysis of slant and bias. By applying the operator π^1 to these images, we generate corresponding smiling versions $\tilde{Z}_{p1}^1, \tilde{Z}_{p2}^1, \tilde{Z}_{p3}^1$, where the only difference is the added attribute, such as a smile. Figure 4 illustrates how we apply function π^1 to generate the counterfactual versions of Donald Trump's neutral images with a smile. The detailed process of how cGANs can be operationalized to provide desired images is supplied in Appendix E.1, following the GANs methodology introduced by [Goodfellow et al. \(2014\)](#).

We repeat this process for all politicians in our study to construct the set $\{\tilde{Z}_{p1}^0, \tilde{Z}_{p2}^0, \tilde{Z}_{p3}^0, \tilde{Z}_{p1}^1, \tilde{Z}_{p2}^1, \tilde{Z}_{p3}^1\}_p$. We then assess whether manipulating images by adding a smile affects other image characteristics, such as brightness and colorfulness. Our analysis, supported by histograms and statistical tests, confirms that the generated smiling images remain consistent with the original images in all aspects except the targeted attribute, thus validating the mitigation of extraction bias. Further details on the analysis can be found in the Appendix E.1.1.

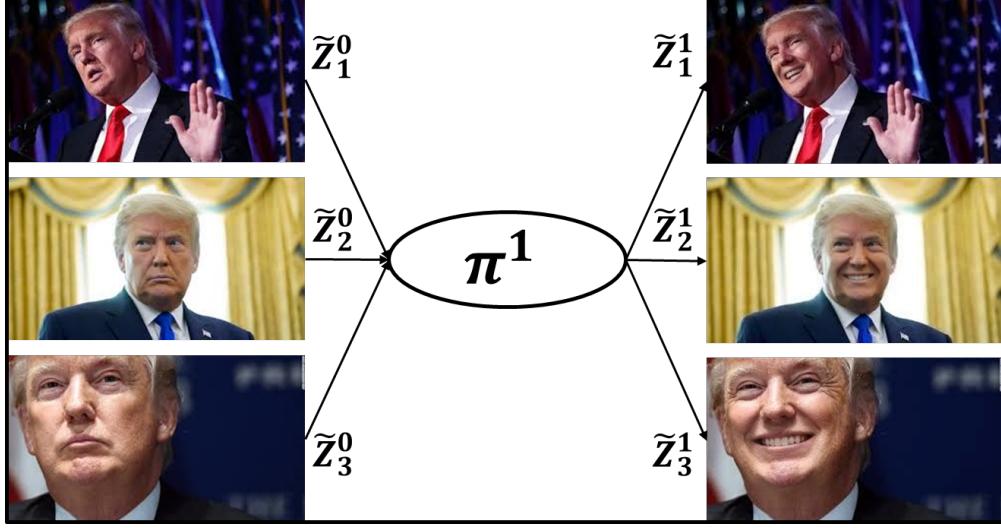


Figure 4: Three neutral selected images for Trump to generate a slanted version by having a smile.

6.2 Identification and Estimation of the Polarization Parameter

We now discuss the identification of our polarization parameter in Equation (3). Our measure is defined for a politician p and any pair of news outlets y_1 and y_2 . For a given pair of counterfactual images $\{\tilde{Z}_p^0, \tilde{Z}_p^1\}$, we can write the sample analogue estimator for the polarization parameter as follows:

$$\hat{\rho}(p, y_1, y_2) = \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} ([u(\tilde{Z}_p^1, X_i, p, y_1) - u(\tilde{Z}_p^0, X_i, p, y_1)] - [u(\tilde{Z}_p^1, X_i, p, y_2) - u(\tilde{Z}_p^0, X_i, p, y_2)]) \quad (15)$$

Clearly, if the utility function is known, we can directly estimate $\hat{\rho}(p, y_1, y_2)$ by integrating over articles. However, the key challenge is that we do not have the underlying utility function for news outlets because we do not observe their consideration sets when choosing the image to recover their utility parameters. In this section, we first discuss the identification of the polarization parameter given the data at hand in §6.2.1 and then discuss the estimation details in §6.2.2.

6.2.1 Identification: Linking Polarization to the News Outlet Prediction Problem

As discussed earlier, the main challenge in identifying the polarization parameter is that the elements of the sample analogue estimator are not directly identifiable because we do not observe news outlets' consideration set when choosing an image. Therefore, one solution is to link the elements of right-hand side (RHS) of Equation (15) to pieces identifiable from the data at hand. We can rewrite the sample analogue estimator

for the polarization parameter as follows:

$$\begin{aligned}
\hat{\rho}(p, y_1, y_2) &= \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} ([u(\tilde{Z}_p^1, X_i, p, y_1) - u(\tilde{Z}_p^0, X_i, p, y_1)] - [u(\tilde{Z}_p^1, X_i, p, y_2) - u(\tilde{Z}_p^0, X_i, p, y_2)]) \\
&= \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} ([u(\tilde{Z}_p^1, X_i, p, y_1) - u(\tilde{Z}_p^1, X_i, p, y_2)] - [u(\tilde{Z}_p^0, X_i, p, y_1) - u(\tilde{Z}_p^0, X_i, p, y_2)]) \\
&= \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} \left[\log \left(\frac{e^{u(\tilde{Z}_p^1, X_i, p, y_1)}}{e^{u(\tilde{Z}_p^1, X_i, p, y_2)}} \right) - \log \left(\frac{e^{u(\tilde{Z}_p^0, X_i, p, y_1)}}{e^{u(\tilde{Z}_p^0, X_i, p, y_2)}} \right) \right] \\
&= \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} \left[\log \left(\frac{\Pr(\tilde{Y} = y_1 | \tilde{Z}_p^1, X_i, p)}{\Pr(\tilde{Y} = y_2 | \tilde{Z}_p^1, X_i, p)} \right) - \log \left(\frac{\Pr(\tilde{Y} = y_1 | \tilde{Z}_p^0, X_i, p)}{\Pr(\tilde{Y} = y_2 | \tilde{Z}_p^0, X_i, p)} \right) \right], \tag{16}
\end{aligned}$$

where \tilde{Y} is a pseudo-variable that is equal to whichever news outlet who has a higher utility for producing an article with the fixed set of characteristics, including the image, content, and politician featured. In deriving Equation (17), we switch two elements (second line), apply $\log(\exp(\cdot))$ transformation (third line), and use a log-odds interpretation that connects the polarization parameter to a news outlet choice problem. As such, Equation (17) allows us to define the polarization parameter as the difference between the log-odds ratio of news outlet y_1 relative to y_2 in having a higher utility for the smiling image \tilde{Z}_p^1 and the log-odds ratio of news outlet y_1 relative to y_2 in having a higher utility the neutral image \tilde{Z}_p^0 .

We now link the identification of the polarization parameter to a news outlet prediction problem. This link depends on the link between the pseudo-variable \tilde{Y} and the actual Y variable that determines the news outlet for an article. The following proposition characterizes the assumption needed for our identification:

Proposition 1. *Suppose that we have data $\mathcal{D} = \{(X_i, P_i, Y_i, Z_i)\}_i$, where if an article i is produced by news outlet Y_i , then Y_i has higher utility from this article than other outlets. The identification of the predicted probabilities for news outlet prediction task results in the identification of the polarization parameter.*

The proof of this proposition is simple: under the assumption that the producing outlet has the highest utility, the identification of log-odds in Equation (17) becomes equivalent of log-odds for variable Y . Let $\hat{g}(y | z, x, p)$ is the conditional distribution that estimates the probability of an article with image z , characteristics x , and politician p is produced by news outlet y . We can then write the polarization parameter as follows:

$$\hat{\rho}(p, y_1, y_2) = \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} \left[\log \left(\frac{\hat{g}(y_1 | \tilde{Z}_p^1, X_i, p)}{\hat{g}(y_2 | \tilde{Z}_p^1, X_i, p)} \right) - \log \left(\frac{\hat{g}(y_1 | \tilde{Z}_p^0, X_i, p)}{\hat{g}(y_2 | \tilde{Z}_p^0, X_i, p)} \right) \right] \tag{17}$$

In light of Proposition 1, the critical question is in whether our identifying assumption is satisfied, and if so, where in the data. The variation we use in our data is the one in important events that are covered by all news outlets, such as important events like press conferences by the politician. For instance, consider different images of Donald Trump during the “Operation Warp Speed” speech in our dataset, as shown in Figure 5. Despite being captured from the same event, these images present Trump in varying ways, ranging from serious to expressive. We argue that in such events, our assumption that the producing outlet has the highest utility is more reasonable, because all outlets have access to the same set of images. In these events, the model’s observable choice over news outlets \mathcal{Y} can potentially mimic the true choice model over the set



Figure 5: Different images from Donald Trump’s “Operation Warp Speed” speech, illustrate how various news outlets may select images based on their editorial preferences.

of images \mathcal{L} . A notable feature of our data is the presence of such similar patterns that form a clustering structure, where we observe extensive choice variation by outlets within a cluster of very similar events. In the next section, we discuss how we design the structure of our news outlet prediction model such that it will be able to fully exploit the variation in similar events.

6.2.2 News Outlet Prediction: Beyond an Off-the-Shelf Application

As highlighted in Proposition 1, the news outlet prediction is at the core of estimating the polarization parameter. The prediction task is to use the training data \mathcal{D}_{train} to predict news outlet Y_i using the article-level information: (X_i, P_i, Z_i) . The resulting output in this step is model $g(y_i | Z_i, X_i, P_i; \theta)$ from the class of models \mathcal{G} . Given the multi-modal nature of inputs (e.g., text, image, categorical variables), we can use any flexible semi-parametric model that can capture complex patterns in the data. However, in learning a model for news outlet prediction, there are important challenges that we need to address. First, we need to ensure that the model utilizes the variation in similar events to satisfy the identifying assumption. Second, we need to ensure that our predictive model correctly identifies the link between the use of smile in an image and the outlet. For example, a purely loss-minimizing objective may end up estimating a model that underestimates the strength of the link between smile and the outlet by misattributing its link to features correlated with smile. Together, these challenges motivate us to go beyond the best off-the-shelf prediction model and impose some structure on the architecture of machine learning models we consider.

First, to fully exploit the presence of similar events, we use the contextual information in articles. The contextual information comprises textual data T_i from news titles, publication date D , politicians’ names P , and political affiliations A , and image data Z . We want our model architecture to fully capture the contextual information for two reasons. First, identifying clusters with diverse news outlets allows for identification of

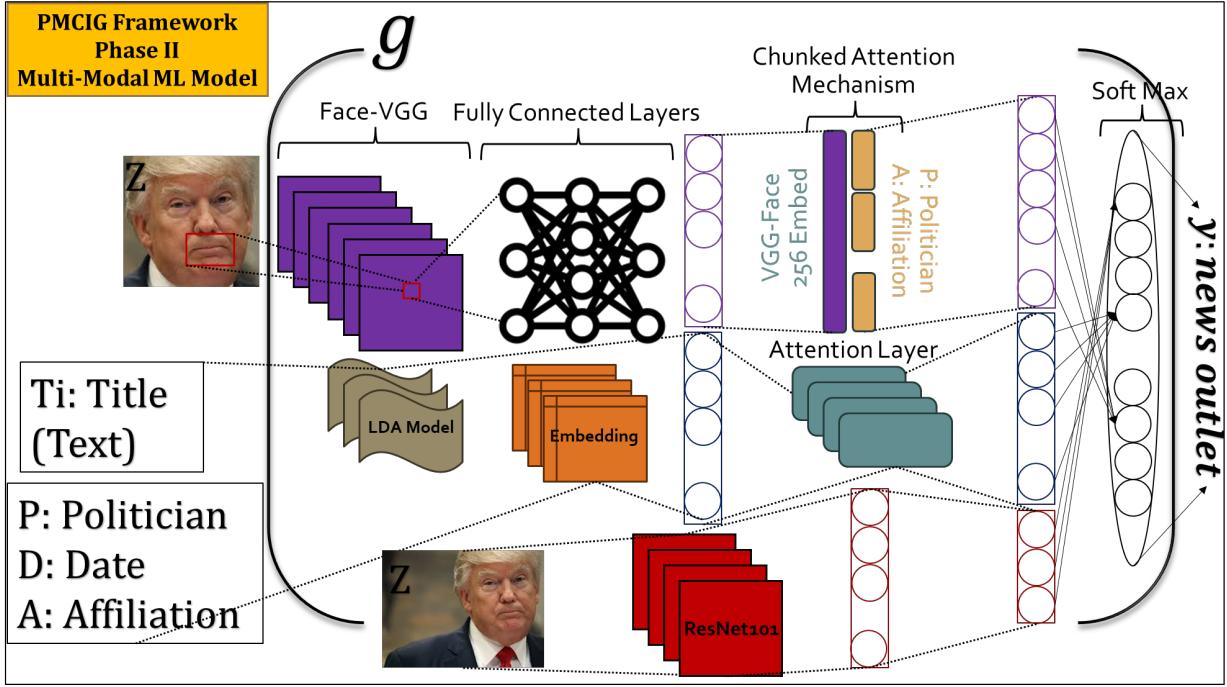


Figure 6: The multi-modal ML model

visual factors related to polarization. Second, identifying clusters with a single or only a few news outlets helps identify stylistic preferences of outlets, preventing those factors to play a role in determining the polarization parameter.

Figure 6 offers a visual overview of the model. To extract meaningful patterns in textual data T_i , we utilize Latent Dirichlet Allocation (LDA), which models the text data by identifying underlying topics across the articles. Although methods like BERT (Devlin et al., 2018) are known for their semantic understanding, LDA outperformed these models in our specific prediction task, particularly in identifying and quantifying topic distributions. Each article is thus represented as a 12-dimensional vector F_{LDA} , encapsulating the relevance of different topics within the text. The LDA processing steps, including pre-processing and model training, are detailed in Appendix E.2.1. Categorical data, including the publication date D , politicians' names P , and political affiliations A , provides additional context that enhances the model's ability to discern patterns relevant to news outlet classification. These features are represented using *embedding layers*, where each categorical value is mapped to a dense vector representation. Specifically, we use an embedding size of 4 for the publication dates D , 4 for the politicians' names P , and 2 for political affiliations A . For the image input, *ResNet-101* processes the entire image, leveraging its exceptional performance in general image recognition tasks to extract hierarchical and context-rich features from the broader visual content and scene information (He et al., 2016). To adapt these architectures to the specific requirements of our task, the last 10 layers of ResNet-101 and the last 5 layers of VGG-Face are fine-tuned, ensuring that both contextual and facial embeddings are optimized for our application.

The second challenge is related to correctly estimating the link between smile and news outlet prediction.

We use *MTCNN* architecture for face detection due to its ability to perform joint face detection and alignment with high accuracy, ensuring precise focus on facial regions (Zhang et al., 2016). Detected faces are then passed through the *VGG-Face* network, which is particularly well-suited for this task because it is pre-trained on facial expression data, making it highly effective at capturing facial attributes such as smiles (Parkhi et al., 2015). We design this part of the architecture to ensure that the predictive model accounts for the information in the politicians’ faces. Later in results, we show how adding this element to the structure allows the model to correctly identify the differences between the counterfactual image versions.

In summary, the integration of modalities occurs through specialized attention mechanisms:

- *Chunk attention* is applied to the VGG-Face embeddings, combining them with categorical data (politicians’ names P and affiliations A) to capture correlations between facial features and structured metadata related to the image. This fusion ensures the model can link specific facial attributes to political or identity-related information (Liang et al., 2024) (see Appendix E.2.2 for details).
- *Attention Mechanism* processes the structured data (categorical features, LDA topics), learning to prioritize the most relevant metadata features for the classification task (Vaswani, 2017) (see Appendix E.2.3 for details).
- The embeddings from ResNet-101 that are passed through fully connected layers, enriched with contextual image information, are directly incorporated into the final representation (see Appendix E.2.4 for details).

The outputs from these attention mechanisms and embeddings are used together in the final classification layer, which predicts the target news outlet. This design ensures that facial details, image context, and structured data interact effectively, enhancing the model’s performance in discerning patterns across modalities.

After concatenating the feature vectors, the output from this layer is then passed through a final softmax layer to produce the classification prediction. The complete model details are explained in Appendix E.2.5. The model is trained using the AdamW optimizer (Loshchilov and Hutter, 2017), which is particularly suited for large-scale data due to its adaptive learning rate and regularization through weight decay. The loss function used is weighted cross-entropy, which measures the discrepancy between the predicted probabilities and the true labels, optimizing the model to improve classification accuracy. The training process’s specifics, including the dataset’s division, hyperparameter tuning, and regularization techniques, are discussed in Appendix E.2.6.

With the model architecture defined, we can then maximize the following entropy over the training data:

$$\text{Entropy: } \max_{\theta} \mathcal{H}(\theta) = \max_{\theta} \left(- \sum_{i=1}^N \sum_{y \in \mathcal{Y}} g(Y_i = y | Z_i, X_i, P_i; \theta) \log g(Y_i = y | Z_i, X_i, P_i; \theta) \right) \quad (18)$$

The resulting model is what we use to estimate the polarization parameter using Equation (17).

6.3 Algorithm

We now present our algorithm *Polarization Measurement Using Counterfactual Image Generation (PM-CIG)*. We present the details of our unified PMCIG algorithm in Algorithm 1 in three steps. In the first step,

Algorithm 1 Polarization Measurement Using Counterfactual Image Generation (PMCIG)

1: Input:

Dataset $\mathcal{D} = \{(Z_i, X_i, P_i, y_i)\}_{i \in \mathcal{N}}$

Split \mathcal{D} into training set $\mathcal{D}_{\text{train}}$ and validation set \mathcal{D}_{val}

Selected images $\tilde{Z}_p \notin \mathcal{D}_{\text{train}}$ for each $p \in P$

Number of images for each politician used for counterfactual image generation S

News outlets $\mathcal{Y} = \{y^k\}_{k=1}^K$, with a baseline y^0 determined as the neutral outlet

2: Output:

Polarization Measurement $\hat{\rho}^T(p, y^k, y^0)$

3: Step 1: Apply Transformation π

4: For each $p \in \mathcal{P}$ and image index s ($1 \leq s \leq S$), apply π to obtain:

$$\tilde{Z}_{ps}^1 = \pi^1(\tilde{Z}_{ps}), \quad \tilde{Z}_{ps}^0 = \pi^0(\tilde{Z}_{ps}) \text{ for each } p \in \mathcal{P}$$

5: **Output:** $\mathcal{Z}_p = \{\tilde{Z}_{ps}^1, \tilde{Z}_{ps}^0\}_{s=1}^S$ for each $p \in \mathcal{P}$.

6: Step 2: Train Non/Semi-Parametric Multi-Modal Model g on $\mathcal{D}_{\text{train}}$

7: Given the training set $\mathcal{D}_{\text{train}} = \{(Z_i, X_i, P_i, y_i)\}_{i \in \mathcal{N}_{\text{train}}}$, train the multi-modal model $g : \mathcal{Y} \rightarrow [0, 1]$ to learn θ :

$$\max_{\theta} \mathcal{H}(\theta) = \max_{\theta} \left(- \sum_{i=1}^{\mathcal{N}_{\text{train}}} \sum_{y \in \mathcal{Y}} g(Y_i = y | Z_i, X_i, P_i; \theta) \log g(Y_i = y | Z_i, X_i, P_i; \theta) \right)$$

$$\text{where } g(Y_i = y | Z_i, X_i, P_i; \theta) = \frac{e^{u(Z_i, X_i, P_i, Y_i=y; \theta)}}{\sum_{y \in \mathcal{Y}} e^{u(Z_i, X_i, P_i, Y_i=y; \theta)}}$$

8: The trained model $\hat{g}(Y_i = y | Z_i, X_i, P_i; \hat{\theta})$ estimates the conditional probability of $y \in \mathcal{Y}$ given (Z, P, X) .

9: **Output:** Trained model $\hat{g}(Y_i = y | Z_i, X_i, P_i; \hat{\theta})$.

10: Step 3: Estimate Probabilities and Calculate $\rho^T(p, y^k, y^0)$ on \mathcal{D}_{val}

11: For each $j \in \mathcal{D}_{\text{val}}$, corresponding to $P_j = p$, compute:

$$\hat{P}(y^k | \tilde{Z}_{ps}^1, X_j, p) = \hat{g}(y^k | \tilde{Z}_{ps}^1, X_j, p; \hat{\theta}), \quad \hat{P}(y^0 | \tilde{Z}_{ps}^1, X_j, p) = \hat{g}(y^0 | \tilde{Z}_{ps}^1, X_j, p; \hat{\theta})$$

$$\hat{P}(y^k | \tilde{Z}_{ps}^0, X_j, p) = \hat{g}(y^k | \tilde{Z}_{ps}^0, X_j, p; \hat{\theta}), \quad \hat{P}(y^0 | \tilde{Z}_{ps}^0, X_j, p) = \hat{g}(y^0 | \tilde{Z}_{ps}^0, X_j, p; \hat{\theta})$$

12: For each $p \in \mathcal{P}$ and any y^k , calculate the article-level estimated Polarization Measurement for the article j as $\hat{\rho}_j^T(p, y^k, y^0)$ following:

$$\hat{\rho}_j(p, y^k, y^0) = \frac{1}{S} \sum_{s=1}^S \left[\log \left(\frac{\hat{g}(y^k | \tilde{Z}_{ps}^1, X_j, p)}{\hat{g}(y^0 | \tilde{Z}_{ps}^1, X_j, p)} \right) - \log \left(\frac{\hat{g}(y^k | \tilde{Z}_{ps}^0, X_j, p)}{\hat{g}(y^0 | \tilde{Z}_{ps}^0, X_j, p)} \right) \right]$$

13: Calculate the aggregated Polarization Measurement for a given politician p , for each news outlet $y^k \in \mathcal{Y}$ as:

$$\hat{\rho}^T(p, y^k, y^0) = \frac{\sum_{j:P_j=p} \hat{\rho}_j^T(p, y^k, y^0)}{N_p}$$

we generate the counterfactual image versions for each politician. To do so, we take three neutral images from each politician p and apply the function π^1 using the procedure outlined in §6.1. An important consideration in choosing the neutral images is that the images are within the joint distribution of the training

data. If the image is so different from the training data, our \hat{g} estimates can be inaccurate.

In the second step of our algorithm, we use the model architecture presented §6.2.2 to estimate the model \hat{g} using training data \mathcal{D}_{train} , which is a random subset of 85% of all articles in data \mathcal{D} . Finally, in the third step, we measure polarization on the validation data set \mathcal{D}_{val} , which is not used in the process of model building. This step incorporates the idea of cross-fitting presented in the literature on the intersection of machine learning and econometrics ensures that our estimates do not overfit (Gentzkow et al., 2019; Chernozhukov et al., 2018). In the final part of our algorithm, we aggregate over all articles and counterfactual image versions to measure polarization for each politician p in each outlet y^k relative to a baseline outlet y^0 . In our empirical application, we use Reuters as the baseline news outlet given its reputation to be a non-partisan news source. However, one could easily change that baseline to any other outlet to obtain polarization measures.

7 Empirical Evaluation and Findings

In this section, we present the results obtained from our empirical analysis using the PMCIG. We divide our results into two main parts. First, §7.1, we present results on the performance of our news outlet prediction model and how it examines the feature of interest. Next, in §7.2, we share our results on political polarization in visual content among news outlets, examining the findings at both the news outlet level and the political level.

7.1 Performance of the News Outlet Prediction Model

In this section, we present the performance of our multi-modal multi-class classification problem for news outlet prediction. As mentioned earlier, we use a 85%-15% split between training and validation data. For prediction accuracy measures, we use accuracy, precision, and recall in terms of percentages. Accuracy is simply measured as $\sum_{i=1}^N \mathbf{1}(Y_i = \text{argmax}_c \hat{Y}_{ic})/N$, where \hat{Y}_{ic} is our predicted probability for news outlet c producing article i . We further use Weighted Cross-Entropy (WCE) Loss as follows:

$$\text{Weighted Cross-Entropy} = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^C w_c (Y_{ic}(1 - \varepsilon) + \varepsilon/C) \log(\hat{Y}_{ic}), \quad (19)$$

where C is the number of classes, w_c represents the weight assigned to class c to handle class imbalance, and ε is the label smoothing factor applied to soften the target labels, encouraging the model to focus more evenly across all classes.

We use these prediction accuracy measures and evaluate three models on the validation data that progressively account for more information. The results are presented in Table 3. The first model only uses categorical inputs such as politician, affiliation, and date ($P + A + D$). We first add textual inputs in the second row, and then add image features in the final row.

Two key points emerge from Table 3. First, our final model achieves a remarkably great performance of approximately 44% correct classification of news outlets by using the information in the article, among the 20 news labels. It is worth emphasizing that a random benchmark only achieves a 5% accuracy. Second, we notice that the image information greatly contributes to the predictive accuracy of the model: while the non-image models achieve a maximum accuracy of 16.72%, the multi-modal model elevates accuracy to

Features Used	Accuracy (%)	Precision (%)	Recall (%)	WCE Loss
$P + A + D$	11.39	10.99	15.18	2.76
$P + A + D + Ti$	16.72	15.01	17.08	2.66
$P + A + D + Ti + Z$	43.81	43.65	41.40	2.29

Table 3: Performance Metrics for 20-Class Classification Task at Different Stages of Model Development for \mathcal{D}_{val}

43.81%. We present the performance measures across outlets in Appendix F.1. In the following section, we add more interpretability of our black-box predictive model by discussing how it captures the information in smile, and what happens to the predictions if we add a smile to an image.

7.1.1 How Does the Model Account for the Smile?

Understanding how a multi-modal model learns and utilizes visual features is essential for interpreting predictions, especially in complex tasks like news outlet classification. While high-dimensional embeddings capture rich information, their interpretability often remains limited (Ludwig and Mullainathan, 2024). Grad-CAM provides a valuable tool for visualizing the contribution of specific features to the model’s predictions, enabling us to understand the attention patterns of different components in the model. By examining the gradient of the model’s prediction function $g(\cdot)$ with respect to its inputs, Grad-CAM highlights the regions of an image that drive the decision-making process (Ludwig and Mullainathan, 2024). This technique is instrumental in our architecture, where two CNN branches—Face-VGG and ResNet101—process facial features and global image context, respectively.

To illustrate, we use Grad-CAM to analyze attention patterns in both CNN branches. Recall the example in Figure 5. Here, Figure 7 shows visualizations for two versions of the same image of Donald Trump from *Washington Post*: the original image (top row) and a smile-added version (bottom row) generated using the transformation function π^1 . The Grad-CAM visualizations illustrate two distinct patterns of model attention. First, the modified ResNet101 heatmaps predominantly focus on broader contextual features, such as the lectern, the logo, and background elements, including the event name (“Operation Warp Speed”). The gradients of $g(\cdot)$ with respect to the global image embeddings emphasize that ResNet101 captures cues related to the overall spatial and contextual layout of the image, such as positioning and visual saliency.

Secondly, the modified Face-VGG heatmaps display heightened sensitivity to facial features, with particular emphasis on the mouth region. For the original image, the gradient of $g(\cdot)$ with respect to the facial embeddings shows moderate activation around the mouth, corresponding to a neutral expression. When a smile is introduced, the activation in the mouth region intensifies and extends to other facial areas associated with smile dynamics. This indicates that Face-VGG excels at detecting smile-related features and subtle changes in facial expressions. Together, Figure 7 highlights the importance of accounting for the Face-VGG embeddings in the model architecture, which allows the model to learn the role of smile in news outlet prediction.

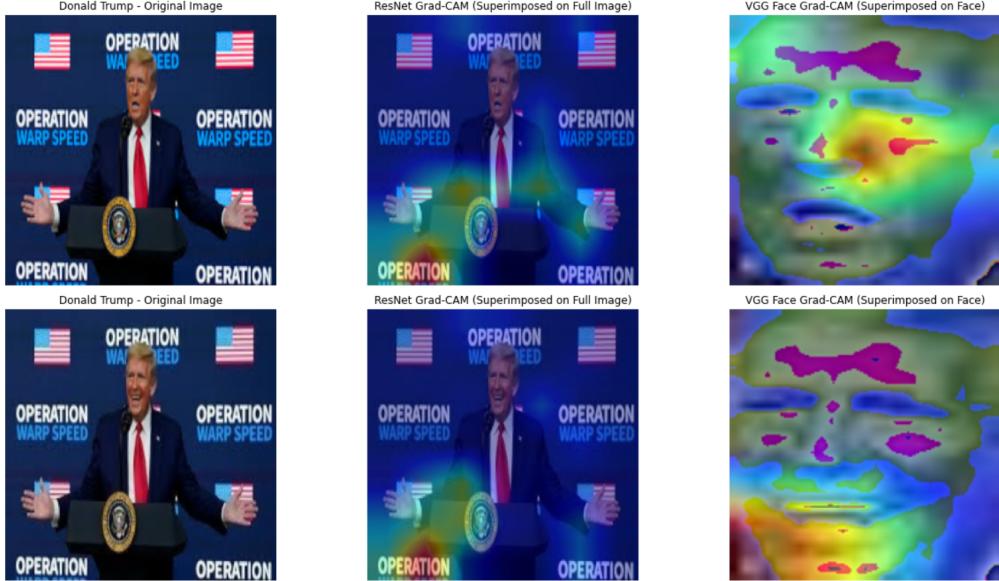


Figure 7: Grad-CAM Visualizations for Original and Smile-Added Versions of a *Washington Post* Image. The ResNet101 heatmaps (right column) capture information on the global image context, while the Face-VGG heatmaps (middle column) highlight the mouth region and facial expression changes.

7.1.2 How Does Adding a Smile Change Predictions?

In the previous section, we find that our model is able to capture the information in smile. In this section, we want to understand how adding a smile changes news outlet predictions. For example, for the images in Figure 7, we find that the adding a smile drops the probability of Washington Post as the label from 30.98% to 3.23%, and makes Daily Mail the predicted news outlet for the smiling version. As discussed in §6.2, our identifying variation comes from similar events covered by a diverse set of news outlets. Our goal in this section is to demonstrate how our multi-modal ML model g not only learns the structure of the comparable images for similar events but also captures the editorial preferences of different news outlets for smiles versus neutral expressions.

To do so, we turn to unsupervised learning techniques. First, we extract embeddings for all images using ResNet101. Each image is represented as a high-dimensional vector $\mathbf{e}_i \in \mathbb{R}^d$, where d denotes the dimensionality of the embedding space. These embeddings capture general visual features of the images, such as texture, color, and facial expressions, without being tied to specific news outlets. To make the embeddings interpretable and suitable for visualization, we reduce their dimensionality to two dimensions using Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE), followed by clustering to identify patterns. We present the details of PCA and t-SNE in our context in Appendix F.2.

Figure 8 shows the t-SNE projection for all images of Donald Trump, grouped into 20 clusters based on their visual similarity. Each cluster captures images with shared contextual or visual features and provides a conceptualization of “similar events” presented as the identifying variation. For instance, consider Cluster 18, as shown in the second Figure 9. Here, each point represents an image, and the color indicates the

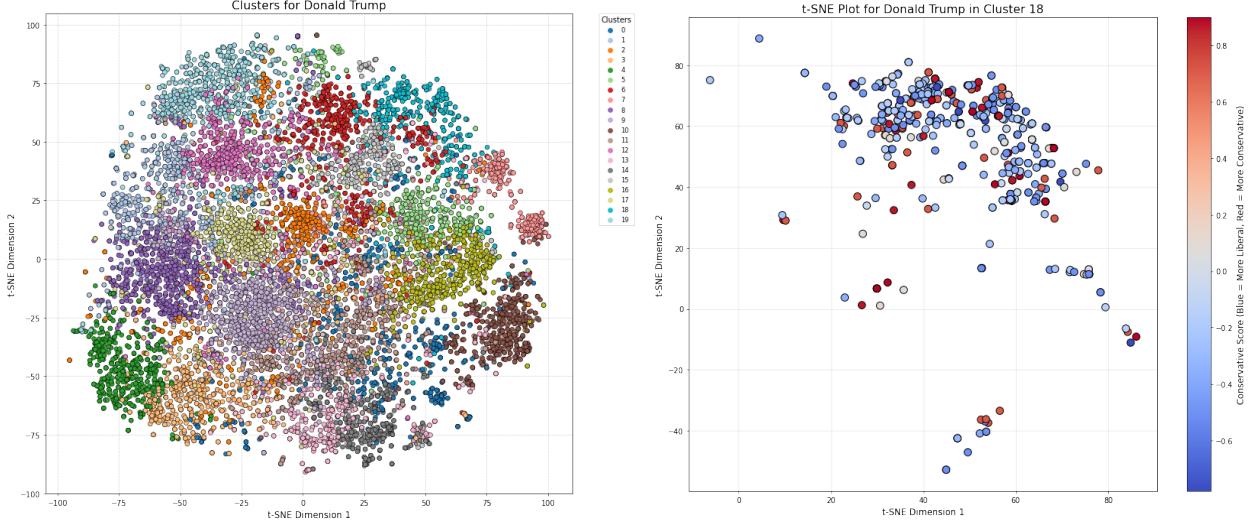


Figure 8: Clusters for Donald Trump: Images grouped into 20 clusters based on t-SNE projections.

Figure 9: t-SNE Plot for Cluster 18: Highlighting conservative scores for selected images.

conservative score of the news outlet that selected the image, ranging from liberal (blue) to conservative (red). The diversity in the news outlets in Figure 9 is helpful for us as it allows us to understand how different news outlets cover similar events and identify the role of subtle visual features such as a smile.

Figure 10 shows a few examples within Cluster 18 produced by different news outlets. Within this cluster, consider the image corresponding to ID 196491 in our dataset by the *Washington Post*, which is the same image of Donald Trump used in the previous section. In this image, Trump appears unhappy, leaning slightly towards an angry expression. As shown in Figure 10, this image is initially positioned on the top right-hand side of the plot. Next, we modify the original image (ID 196491) by applying the transformation function π^1 , which adds a smile to Donald Trump while preserving the contextual elements of the image. After computing the embedding for this modified image, its coordinates in the t-SNE space are updated, as illustrated by the gold dot in the scatter plot. Interestingly, the smile-added version of the original image shifts closer to images such as ID 50555 from *Fox News*. This demonstrates how the addition of a smile alters the embedding position to align with images that share similar facial expressions, even though the overall context remains unchanged.

Although the t-SNE coordinates provide interpretable insights into how our model works, they lose some of the complexity that our model \hat{g} captures. To examine how adding a smile changes the predicted labels, we present the model predictions in Table 4, which demonstrate substantial differences between the original image and the counterfactual version with the smile. Intuitively, the significant shift in predictions after adding a smile can be attributed to the movement of the modified image's embedding within the clustered space. Notably, we find that adding a smile substantially increases the predicted probability for *Daily Mail* (from 2.89% to 20.97%) and *News Max* (from 7.10% to 16.52%), slightly increases the predicted probability for *Fox News* (from 1.13% to 1.75%), slightly decreases the predicted probability for *ABS News* (from 2.59% to 1.70%), and substantially decreases the predicted probability for *Washington Post* (from 30.98% to 3.32%).

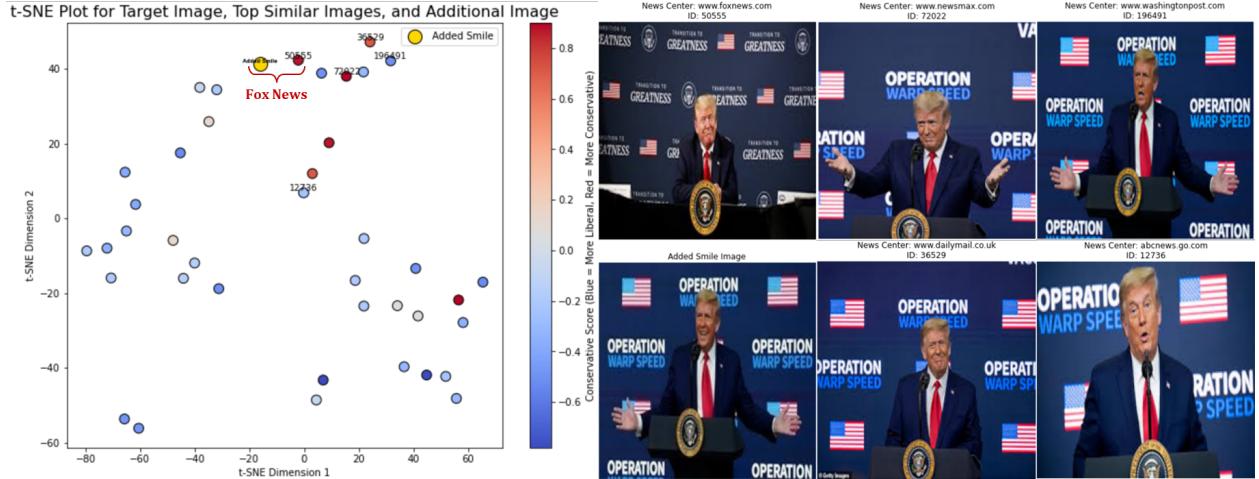


Figure 10: t-SNE visualization of a cluster for Donald Trump, showing the impact of adding a smile on the embedding position.

News Outlet	Original Image (%)	Smile-Added Image (%)
Daily Mail	2.89	20.97
Fox News	1.13	1.75
News Max	7.10	16.52
Washington Post	30.98	3.32
ABC News	2.59	1.70

Table 4: Change in Model Predictions for the Original and Smile-Added Versions of the Donald Trump Image.

In summary, we find an increase in the predicted probabilities for right-leaning news outlets when adding a smile to an image of Donald Trump, but a decrease in the predicted probabilities for left-leaning news outlets. However, this is just a single instance in the data that we use to illustrate the intuition behind how our algorithm works. In the next section, we present more systematic measures of visual polarization.

7.2 Visual Polarization

In this section, we present our main results on visual polarization. All results are direct applications of Algorithm 1 in §6.3. Before presenting the results, we review a few important considerations in applying our algorithm. First, we use a train-validation split and all the polarization measures are obtained on the validation data \mathcal{D}_{val} using the estimates obtained from the training data \mathcal{D}_{train} . Second, for each politician, we use three neutral images and generate three counterfactual versions. We only use these six images in our polarization measurement for each politician. Finally, our polarization parameter is defined based on two news outlets. We define Reuters as the baseline outlet y^0 and measure polarization relative to Reuters.

We start by demonstrating the overall distribution of visual polarization in §7.2.1. Next, in §7.2.2, we correlate our measure with existing measures of polarization at the news outlet level to examine whether there is any consistency. We then explore the extent of heterogeneity across news outlets in §7.2.3 to see which outlets exhibit higher levels of visual polarization. Finally, in §7.2.4, we document the overall degree

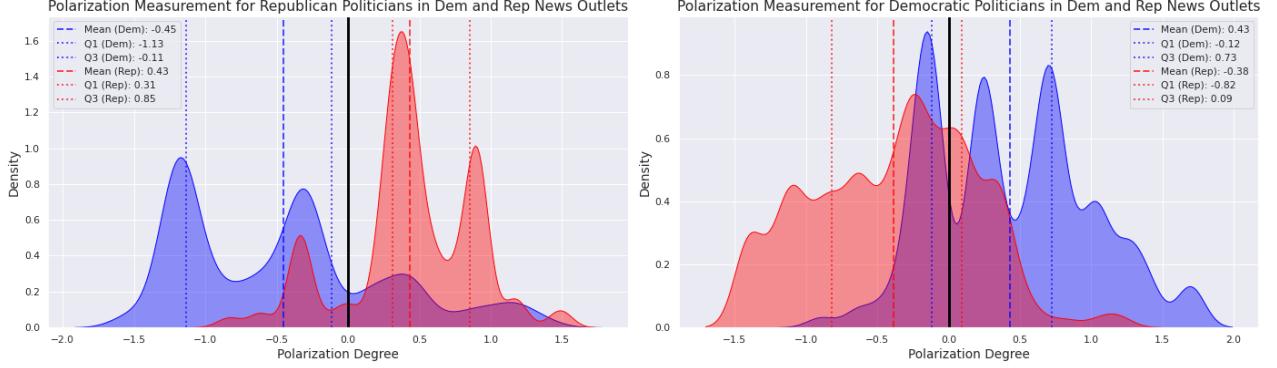


Figure 11: Histograms of Polarization for Democratic and Republican Politicians in Democratic and Republican News Outlets

of visual polarization at the politician level.

7.2.1 Distribution of Visual Polarization

An important feature of algorithm is its ability to produce an individual-level measure of polarization $\hat{\rho}_i^T(p_i, y_i^k, y_i^{Reuters})$. As such, for each article i featuring politician p , we can estimate 19 polarization measures corresponding to outlets other than Reuters. Based on the scores from [Faris et al. \(2017\)](#), [AllSides \(2024\)](#), and [Flaxman et al. \(2016\)](#), we categorize the three most evident Republican-leaning news outlets as: *Fox News*, *News Max*, and *Daily Mail*. Similarly, the three most evident Democratic-leaning news outlets are: *Washington Post*, *CNN*, and *New York Times*. The histograms in Figure 11 illustrate the distribution of polarization measurement for Democratic and Republican politicians as featured in Democratic- and Republican-leaning news outlets. We denote the set of Republican and Democratic politicians by \mathcal{P}_R and \mathcal{P}_D , respectively. Similarly, we denote the sets of Republcan- and Democratic-leaning outlets defined above by \mathcal{Y}_R and \mathcal{Y}_D , respectively.

We start with the overall dispersion of the individual $\hat{\rho}_i^T(p_i, y_i^k, y_i^{Reuters})$ parameters, irrespective of whether the numbers are from Republican- or Democratic-leaning outlets. If there is no visual polarization, we would have seen a distribution tightly centered around zero. However, we see extensive variation in individual-level polarization parameters. If we assume Reuters to be at the neutral point zero, our measure will have a slant interpretation. To see if there is a systematic difference between Republican- and Democratic-learning outlets, we examine their averages in each panel.

For Republican politicians, in the left panel of Figure 11, the distributions of $\hat{\rho}_i^T$ show distinct differences between left- and Republican-leaning news outlets. The polarization measurement for left-leaning news outlets is -0.45 , implying that, on average, smiling images of Republican politicians decrease the utility of a news outlet classified as Democratic-leaning compared to Reuters. In contrast, the polarization measurement for Republican-leaning news outlets is 0.43 , suggesting that, on average, smiling images of Republican politicians significantly increase their utility compared to Reuters. Therefore, for Republican politicians, we can conclude that:

$$\hat{\rho}^T(p \in \mathcal{P}_R, y^k \in \mathcal{Y}_D, y^{Reuters}) < 0 < \hat{\rho}^T(p \in \mathcal{P}_R, y^k \in \mathcal{Y}_R, y^{Reuters})$$

For Democratic politicians, in the right panel of Figure 11, the polarization distribution for Democratic and Republican news outlets also shows clear differences. The polarization for Democratic news outlets is 0.43, indicating that, on average, images of smiling Democratic politicians increase the utility of a news outlet classified as Democratic-leaning compared to the neutral baseline, Reuters. Conversely, the polarization for Republican-leaning new outlets is -0.38, suggesting that, on average, smiling images of democratic politicians decrease their utility compared to Reuters. Therefore, for Democratic politicians, we can conclude that:

$$\hat{\rho}^T(p \in \mathcal{P}_D, y^k \in \mathcal{Y}_R, y^{Reuters}) < 0 < \hat{\rho}^T(p \in \mathcal{P}_D, y^k \in \mathcal{Y}_D, y^{Reuters})$$

In Appendix F.3, we employ two statistical tests to analyze these differences statistically: the Kolmogorov-Smirnov (K-S) test and one-sample t-tests and confirm the significant differences between the distributions and means of Republican- and Democratic-leaning outlets for both Democratic and Republican politicians.

7.2.2 Correlation with Existing Measures of Polarization

To investigate whether the observed polarization pattern holds for a broader range of news outlets, we extend our analysis beyond the initial three well-known Democratic- and Republican-leaning outlets. Specifically, we calculate the polarization for each outlet $y^k \in \mathcal{Y}$ when featuring Republican politicians, $\hat{\rho}^T(p \in \mathcal{P}_R, y^k, y^{Reuters})$, and Democratic politicians, $\hat{\rho}^T(p \in \mathcal{P}_D, y^k, y^{Reuters})$. This measurement allows us to compare the relative utility of smiling images for politicians in each group against the baseline outlet, Reuters. We can then examine these polarization values in relation to existing measures of polarization to see if there is any consistency. This can also serve as a validation test for our visual polarization measure.

We start with one of the more established measures of polarization based on the conservative share scores (Faris et al., 2017). Figure 12 shows two separate panels for Democratic and Republican politicians, with each point representing both polarization measures for each news outlet: the x-axis corresponds to each outlet's conservative share score, while the y-axis represents our visual polarization measure relative to Reuters.

In this regard, Figure 12 visualizes the regression analysis using the Faris et al. (2017) conservative share score, with each point representing our visual polarization measurement $\hat{\rho}^T(p \in D, y^k, y^{Reuters})$ for Democratic politicians (left panel) or $\hat{\rho}^T(p \in R, y^k, y^{Reuters})$ for Republican politicians (right panel). The x-axis corresponds to each outlet's conservative share score, while the y-axis represents the polarization relative to Reuters. In the left panel, the strong negative slope illustrates that outlets with higher conservative share scores, such as “Fox News” and “Newsmax,” perceive smiling images of Democratic politicians less favorably, with negative polarization values. Conversely, Democratic-leaning outlets like “HuffPost” and “Washington Post” exhibit positive polarization values, reflecting a favorable interpretation of smiling Democratic politicians. In the right panel, the positive slope indicates that outlets with higher conservative share scores strongly favor smiling Republican politicians, as seen for “Daily Mail” and “Newsmax,” while Democratic-leaning outlets like “NBC News” and “New York Times” show negative polarization values.

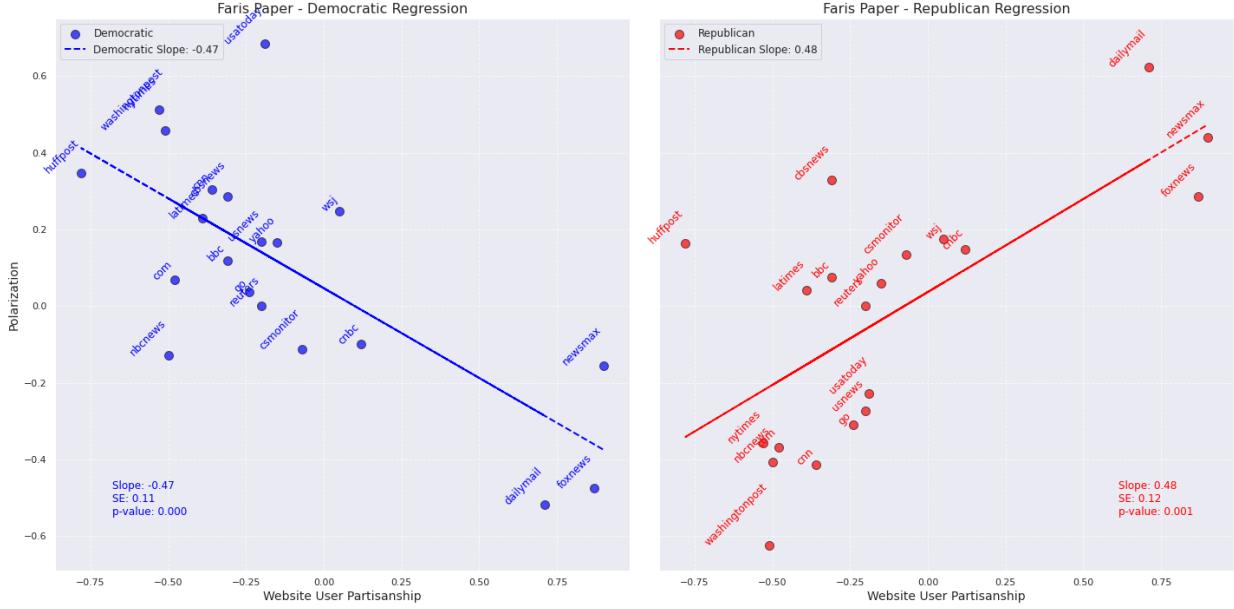


Figure 12: Regression analysis comparing mean smile effect and website user partisanship for Democratic and Republican politicians, adapted from Faris et al. (2017).

In summary, the results demonstrate great consistency between our visual polarization measure and the existing partisanship scores. In Appendix F.4, we go beyond the conservative share measure and conduct regression analysis to provide further validation using other existing measures.

7.2.3 Heterogeneity Across Outlets

In the previous section, we show how our visual polarization measure co-vary with the existing polarization measures. In this section, we focus more on the extent of this variation and examine heterogeneity across outlets. Figure 13 provides descriptive insights into how different news outlets portray politicians from both the Republican and Democratic parties.

The radar plot in Figure 13 illustrates our visualization polarization measurement for a given news outlet $y^k \in \mathcal{Y}$ across Republican politicians as $\hat{\rho}^T(p \in \mathcal{P}_R, y^k, y^{Reuters})$, and Democratic politicians as $\hat{\rho}^T(p \in \mathcal{P}_D, y^k, y^{Reuters})$. Each axis represents a different news outlet, allowing a comparison of how these outlets portray politicians based on their political affiliation. Reuters polarization measurement is zero since it serves as the baseline. The plot reveals clear distinct patterns among the news outlets in their portrayal of politicians from different political parties.

Notably, outlets such as *Fox News* and *Daily Mail* display a positive visual polarization measurement for Republican politicians, indicating a more favorable portrayal consistent with their conservative editorial stance. In contrast, these outlets show a negative or less favorable effect for Democratic politicians. On the other hand, outlets like *CNN*, *Washington Post*, and *The New York Times* exhibit a positive visual polarization measurement for Democratic politicians, suggesting a favorable portrayal. These outlets, however, show a negative or less favorable smile effect for Republican politicians.

Interestingly, some outlets, such as the *Wall Street Journal*, demonstrate positive polarization for both

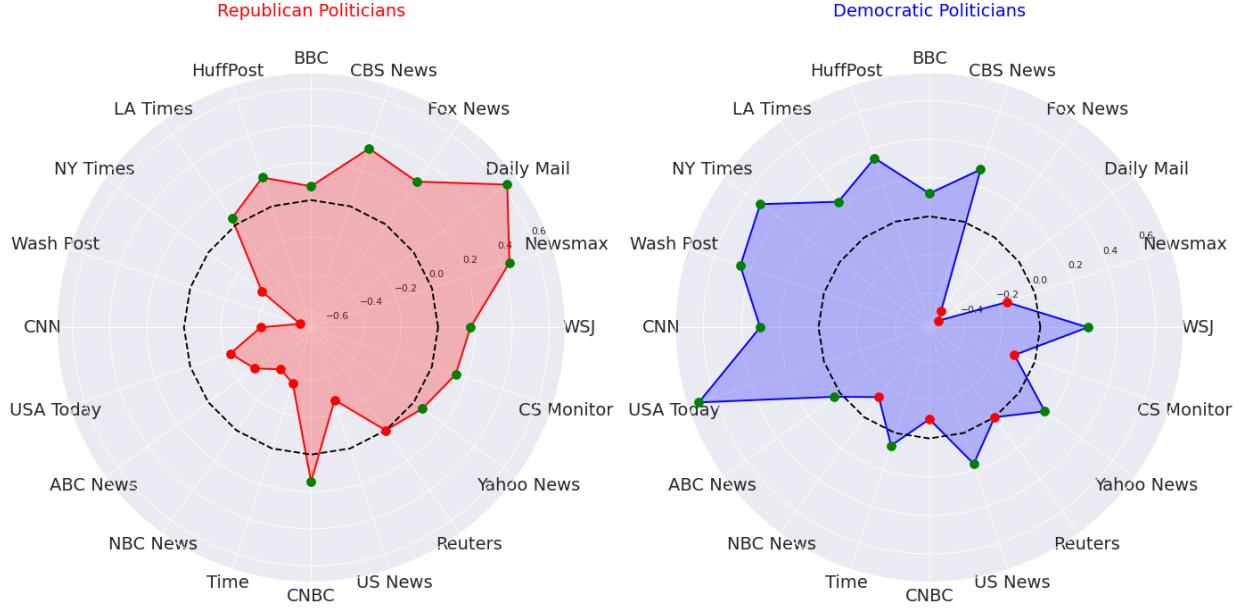


Figure 13: Radar plots showing the mean polarization measurement for Republican and Democratic politicians across various news outlets

Democratic and Republican politicians, which could be due to the presence of underlying mixed policies within this outlet. This suggests that its editorial approach may balance portrayals of politicians from both parties, reflecting a more centrist or business-oriented perspective rather than strong partisan alignment. This variation in polarization measurement across different news outlets underscores the presence of partisan biases in visual representation. To further explore these findings, Figure 14 provides additional insights by presenting the visual polarization measurements for Democratic and Republican politicians ranked in bar charts.

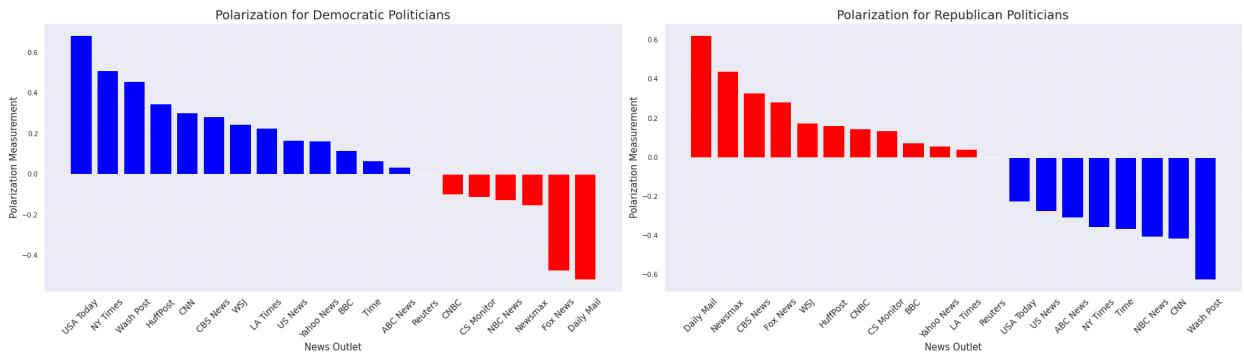


Figure 14: Side-by-side bar charts showing the polarization measurement for Democratic (left) and Republican (right) politicians across various news outlets

In Figure 14, the left chart shows that Democratic news outlets tend to have positive polarization measurement for Democratic politicians and negative effects for Republican politicians. Conversely, the right chart indicates that Republican news outlets exhibit positive polarization measurement for Republican politi-

cians and negative effects for Democratic politicians. Democratic-leaning outlets such as *Washington Post* and *New York Times* display a strong positive bias towards Democratic politicians and a negative bias towards Republican politicians. Similarly, Republican-leaning outlets like *Fox News* and *Daily Mail* show a noticeable positive bias towards Republican politicians and a negative bias towards Democratic politicians.

To provide a comprehensive view of the overall polarization in the news media landscape, Figure 15 illustrates the absolute values of the polarization measurement for both Democratic and Republican politicians in each news outlet. This plot effectively captures the extent of polarization by summing the absolute values of these measurements such that:

$$|\hat{\rho}^T(p \in \mathcal{P}_R, y^k, y^{Reuters})| + |\hat{\rho}^T(p \in \mathcal{P}_D, y^k, y^{Reuters})| \quad \text{where } y^k \in \mathcal{Y}$$

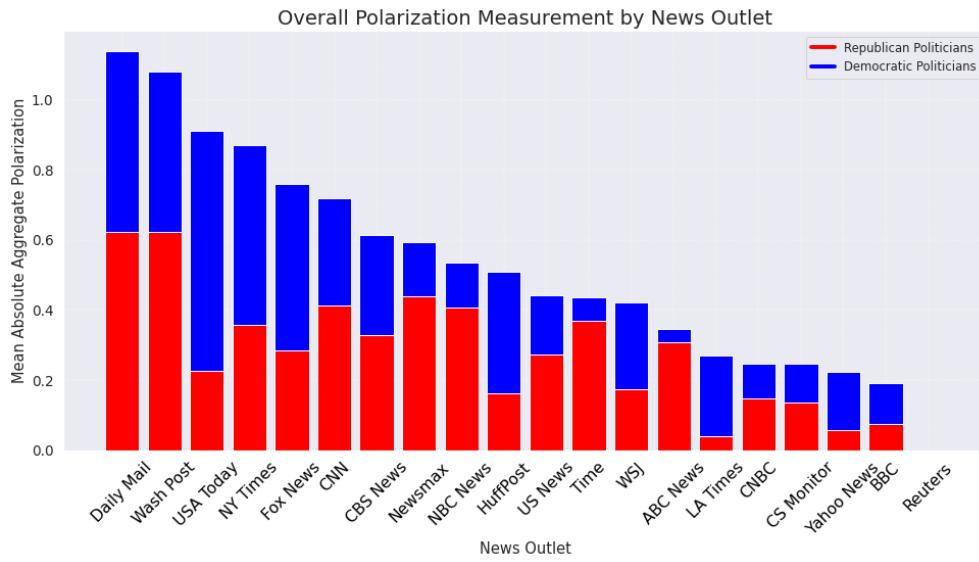


Figure 15: Bar plot showing the absolute value of polarization measurement for Republican and Democratic politicians across various news outlets

From the bar plot in Figure 15, it is evident that news outlets such as *Daily Mail* and *Washington Post* exhibit high levels of collective polarization, with *Daily Mail* showing strong favorability towards Republican politicians and *Washington Post* strongly favoring Democratic politicians. Conversely, outlets like *Yahoo News* and *BBC News* display comparatively lower polarization. Additionally, we conduct a thorough analysis of the polarization measurement distributions for each news outlet to provide a deeper understanding of their visual slant. Appendix F.5 contains individual histograms for each news outlet, visually depicting the distribution of polarization measurement for both Democratic and Republican politicians. Moreover, detailed statistical reports that accompany these histograms, including key metrics and test results that quantify the extent of bias and polarization are presented.

7.2.4 Heterogeneity Across Politicians

In this section, we examine polarization in images at the level of individual politicians such that for a given $p \in \mathcal{P}$, polarization measurement is $\hat{\rho}^T(p, y^k, y^{Reuters})$ where $y^k \in \mathcal{Y}$. Specifically, we aim to understand

how each politician is portrayed across various news outlets and identify which politicians are most polarized in the media. Figure 16 illustrates the polarization measurement for six prominent politicians: Democratic politicians Joe Biden, Hillary Clinton, and Barack Obama, and Republican politicians Donald Trump, Ted Cruz, and Mitt Romney. The rest of the results for all politicians can be found in Appendix F.6.

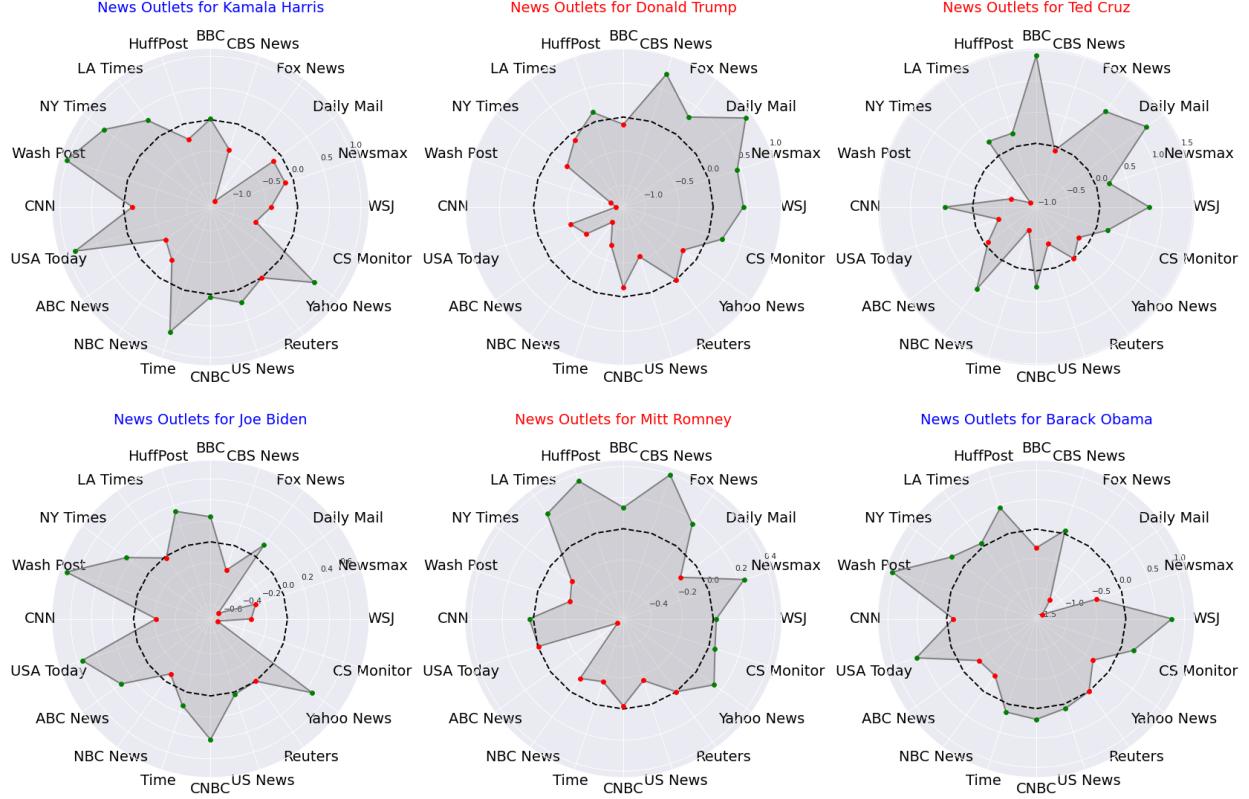


Figure 16: Radar plots of polarization measurement in each news outlet for six prominent politicians

As shown in Figure 16, adding a smile to Democratic politicians tends to increase the utility of Democratic-leaning news outlets such as *The New York Times* and *Washington Post*, while it decreases the utility of Republican-leaning news outlets such as *Fox News*, *Daily Mail*, and *News Max*. Similarly, adding a smile to Republican politicians decreases the utility of Democratic-leaning news outlets but increases the utility of Republican-leaning news outlets.

It is also important to note that the degree of polarization and bias varies among politicians. For example, Joe Biden and Mitt Romney exhibit less heterogeneity across news outlets, whereas Donald Trump and Barack Obama are more controversial, showing greater variability in how they are portrayed. To quantify the overall polarization for each politician, we sum the absolute polarization measurement for each politician across all news outlets such that:

$$\sum_{y^k \in \mathcal{Y}} |\hat{\rho}^T(p, y^k, y^{Reuters})| \quad \text{where } p \in \mathcal{P}$$

This provides a measure of the total polarization each politician faces in the media. Figure 17 ranks the politicians from most polarized to least polarized based on this criterion.

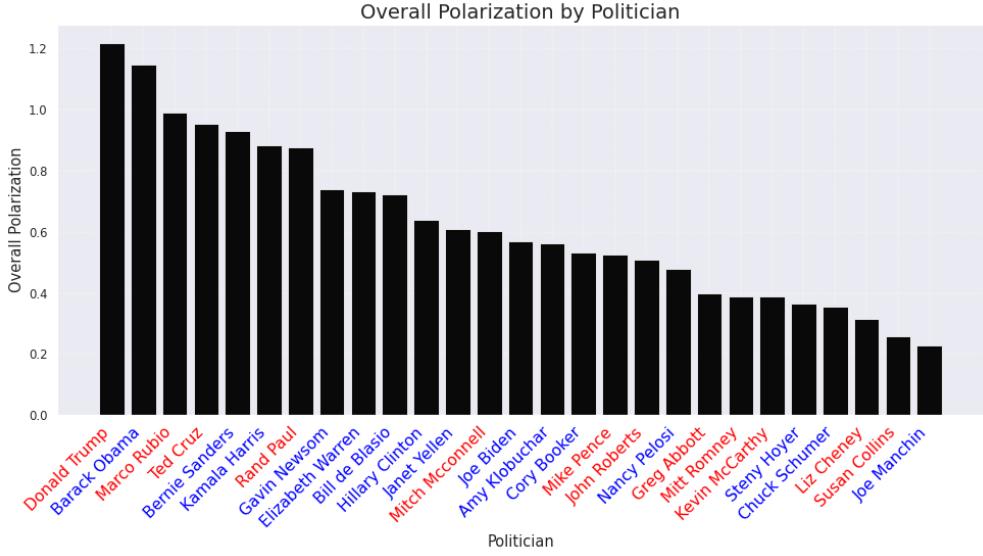


Figure 17: Overall absolute polarization measurement of politicians across all news outlets, ranked from most to least polarized

In Figure 17, the bars represent the total absolute polarization for each politician across all news outlets. The plot reveals that Donald Trump, Barack Obama, and Marco Rubio are among the most polarized politicians. Donald Trump, as a highly divisive figure and central to numerous political controversies, exhibits the highest overall polarization, reflecting the intense media focus and contrasting portrayals across partisan outlets. Similarly, Barack Obama, despite being from the opposing party, demonstrates a comparable level of polarization due to his presidency's significant political impact and media coverage. Notably, Marco Rubio and Ted Cruz, both prominent Republican politicians, also show high levels of polarization, likely attributed to their roles in policy debates and media prominence during key political events. On the Democratic side, figures such as Bernie Sanders and Kamala Harris display significant polarization, reflecting their leadership roles and ideological positions within the party.

Interestingly, lower levels of polarization are observed for politicians such as Joe Manchin, Susan Collins, and Liz Cheney. These are all figures with some bipartisan support, leading to less extreme portrayals across media outlets. For instance, Joe Manchin's centrist stance within the Democratic Party might contribute to his relatively low polarization measurement. The analysis highlights that polarization levels are influenced not just by party affiliation but also by the individual politician's prominence, ideological stance, and role in significant political events. This nuanced variation underscores how media coverage contributes to shaping public perception through varying degrees of polarization for different politicians.

8 Conclusion

This paper introduces a unified framework for analyzing political polarization in the visual content of news media, blending theoretical perspectives with empirical evidence. Central to this study is the Polarization Measurement Using Counterfactual Image Generation (PMCIG) framework, which critically examines whether news outlets intentionally choose slanted imagery, particularly smiling images, for positive representation in their articles. By integrating economic demand choice modeling with cutting-edge machine learning techniques, including multi-modal deep learning and GANs, the PMCIG framework offers a robust foundation for understanding political polarization in visual content among news media.

The study specifically examines the impact of smiling images on the utility that news outlets derive from their news articles. By addressing challenges in traditional descriptive approaches widely used in social science, such as extraction bias and omitted variable bias, PMCIG offers a more precise measurement of polarization in visual content. The framework's approach uses GANs to generate counterfactual (smiling) images and a non-parametric multi-modal deep learning model for learning the complex relationships between image features alongside other informative elements such as article titles and politician names to predict each news outlet's choices.

Applying the PMCIG framework to a decade-long dataset spanning 20 major news outlets and 30 prominent politicians, the study discovers an evident political polarization in the representation of political figures across news outlets. The framework's strength lies in its ability to capture polarization with a high degree of granularity, both at the level of individual news outlets and specific politicians. The findings indicate that the use of smiling images in news articles has a notable impact on the perceived utility of the content, and this effect is closely aligned with the political leanings of the outlet. For example, Republican-leaning outlets such as Fox News and the Daily Mail tend to derive greater utility from articles that feature smiling images of Republican politicians, as these images reinforce positive associations with their audience. Conversely, Democratic-leaning outlets like Washington Post and New York Times show a preference for smiling images of Democratic politicians, suggesting a strategic use of imagery to enhance the appeal of their content to their respective audiences. Similarly, these outlets often diminish the use of smiling images of politicians from the opposing party, thereby reinforcing their audience's ideological preferences.

At the individual politician level, the PMCIG framework provides important insights into how different politicians are visually portrayed, highlighting those who are most subject to polarized representation. The study notably reveals that figures such as Donald Trump are consistently depicted in a highly polarized manner, with smiling images enhancing his portrayal in Republican-leaning outlets while diminishing it in Democratic-leaning ones. Conversely, Democratic politicians experience the opposite trend, with their images being positively represented in Democratic-leaning outlets and negatively in Republican-leaning ones.

In conclusion, the PMCIG framework addresses critical methodological challenges in the study of visual bias, offering a unified approach to understanding the impact of political polarization on visual content in news media. As visual media continues to play an increasingly influential role in shaping public discourse, this framework provides insights that extend beyond the scope of this study. Future research could expand on these findings by applying the PMCIG framework to other media formats, such as social media or adver-

tising, and by exploring additional visual content features that may contribute to polarization or any other parameter of interest.

References

- AllSides. Media bias rating methods, 2024. URL <https://www.allsides.com/media-bias/media-bias-rating-methods>.
- Piyush Anand and Vrinda Kadiyali. Frontiers: Smoke and mirrors: Impact of e-cigarette taxes on underage social media posting. *Marketing Science*, 2024.
- Joshua D. Angrist and Jörn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, 2009.
- Susan Athey, Dean Karlan, Emil Palikot, and Yuan Yuan. Smiles in profiles: Improving fairness and efficiency using estimates of user preferences in online marketplaces. Technical report, National Bureau of Economic Research, 2022.
- Anna C Belkina, Christopher O Ciccolella, Rina Anno, Richard Halpert, Josef Spidlen, and Jennifer E Snyder-Cappione. Automated optimized parameters for t-distributed stochastic neighbor embedding improve visualization and analysis of large datasets. *Nature communications*, 10(1):5415, 2019.
- Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. ” O'Reilly Media, Inc.”, 2009.
- David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022, 2003.
- Kenneth A Bollen and Walter R Davis. Causal indicator models: Identification, estimation, and testing. *Structural Equation Modeling: A Multidisciplinary Journal*, 16(3):498–522, 2009.
- Levi Boxell. Slanted images: Measuring nonverbal media bias during the 2016 election. Available at SSRN 3837521, 2021.
- Giulia Caprini. Visual bias. 2023.
- Keyu Chen, Xu Yang, Changjie Fan, Wei Zhang, and Yu Ding. Semantic-rich facial emotional expression recognition. *IEEE Transactions on Affective Computing*, 13(4):1906–1916, 2022.
- Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters, 2018.
- Thomas H Davenport. How analytics has changed in the last 10 years (and how it's stayed the same). *Harvard Business Review*, 28(08):2017, 2017.
- Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Carroll Doherty, Jocelyn Kiley, Nida Asheer, and Talie Price. Americans’ feelings about politics, polarization and the tone of political discourse, 2023. URL https://www.pewresearch.org/wp-content/uploads/sites/20/2023/09/PP_2023.09.19_views-of-politics_REPORT.pdf.

Maximiliano Dvorkin, Juan M Sánchez, Horacio Sapriza, and Emircan Yurdagul. Sovereign debt restructurings. *American Economic Journal: Macroeconomics*, 13(2):26–77, 2021.

Robert Faris, Hal Roberts, Bruce Etling, Nikki Bourassa, Ethan Zuckerman, and Yochai Benkler. Partisanship, propaganda, and disinformation: Online media and the 2016 us presidential election. *Berkman Klein Center Research Publication*, 6, 2017.

Seth Flaxman, Sharad Goel, and Justin M Rao. Filter bubbles, echo chambers, and online news consumption. *Public opinion quarterly*, 80(S1):298–320, 2016.

Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3146–3154, 2019.

Matthew Gentzkow and Jesse M Shapiro. What drives media slant? evidence from us daily newspapers. *Econometrica*, 78(1):35–71, 2010.

Matthew Gentzkow, Jesse M Shapiro, and Matt Taddy. Measuring group differences in high-dimensional choices: method and application to congressional speech. *Econometrica*, 87(4):1307–1340, 2019.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

Tim Groseclose and Jeffrey Milyo. A measure of media bias. *The quarterly journal of economics*, 120(4):1191–1237, 2005.

Bruce Hansen. *Econometrics*. Princeton University Press, 2022.

Tam Harbert. Tapping the power of unstructured data. *MIT Sloan. Feb*, 1:3, 2021.

Jochen Hartmann, Mark Heitmann, Christina Schamp, and Oded Netzer. The power of brand selfies. *Journal of Marketing Research*, 58(6):1159–1177, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

Jacob Jensen, Suresh Naidu, Ethan Kaplan, Laurence Wilse-Samson, David Gergen, Michael Zuckerman, and Arthur Spirling. Political polarization and the dynamics of political language: Evidence from 130 years of partisan speech [with comments and discussion]. *Brookings Papers on Economic Activity*, pages 1–81, 2012.

Jeffrey K Lee. Emotional expressions and brand status. *Journal of Marketing Research*, 58(6):1178–1196, 2021.

Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations & trends in multimodal machine learning: Principles, challenges, and open questions. *ACM Computing Surveys*, 56(10):1–42, 2024.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

Liu Liu, Daria Dzyabura, and Natalie Mizik. Visual listening in: Extracting brand image portrayed on social media. *Marketing Science*, 39(4):669–686, 2020.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Shijie Lu, Dai Yao, Xingyu Chen, and Rajdeep Grewal. Do larger audiences generate greater revenues under pay what you want? evidence from a live streaming platform. *Marketing Science*, 40(5):964–984, 2021.

Michael Luca, Elizaveta Pronkina, and Michelangelo Rossi. Scapegoating and discrimination in times of crisis: Evidence from airbnb. Technical report, National Bureau of Economic Research, 2022.

Jens Ludwig and Sendhil Mullainathan. Machine learning as a tool for hypothesis generation. *The Quarterly Journal of Economics*, 139(2):751–827, 2024.

Lan E Luo and Olivier Toubia. Using ai for controllable stimuli generation: An application to gender discrimination with faces. *Available at SSRN 4865798*, 2024.

J. Nathan Matias, Kevin Munger, Marianne Aubin Le Quere, and Charles Ebersole. The upworthy research archive, a time series of 32,487 experiments in u.s. media. *Scientific Data*, 8(195), 2021. doi: 10.1038/s41597-021-00934-7.

Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

Omkar Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *BMVC 2015-Proceedings of the British Machine Vision Conference 2015*. British Machine Vision Association, 2015.

Ling Peng, Geng Cui, Yuho Chung, and Wanyi Zheng. The faces of success: Beauty and ugliness premiums in e-commerce platforms. *Journal of Marketing*, 84(4):67–85, 2020.

Yilang Peng. Same candidates, different faces: Uncovering media bias in visual portrayals of presidential candidates with computer vision. *Journal of Communication*, 68(5):920–941, 2018.

Radim Řehůřek and Petr Sojka. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. <http://is.muni.cz/publication/884893/en>.

Amandeep Singh and Bolong Zheng. Causal regressions for unstructured data. In *Causal Representation Learning Workshop at NeurIPS 2023*, 2023.

Geoffrey Skelley and Holly Fuong. 3 in 10 americans named political polarization as a top issue facing the country, 2022. URL <https://fivethirtyeight.com/features/3-in-10-americans-named-political-polarization-as-a-top-issue-facing-the-country/>.

Michael Sülfow and Marcus Maurer. The power of smiling. how politicians' displays of happiness affect viewers' gaze behavior and political judgments. *Visual political communication*, pages 207–224, 2019.

Denis G Sullivan and Roger D Masters. "happy warriors": Leaders' facial displays, viewers' emotions, and political support. *American Journal of Political Science*, pages 345–368, 1988.

Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708, 2014.

Claudia Townsend and Barbara E. Kahn. The visual preference heuristic: Effects of visual vs. verbal depiction on assessment of others, choice of hedonic products, and voting. *Journal of Consumer Research*, 41(2):392–411, 2014.

Damian Trilling, Petro Tolochko, and Björn Burscher. From newsworthiness to shareworthiness: How to predict news sharing based on article characteristics. *Journalism & Mass Communication Quarterly*, 94(1):38–60, 2016.

Isamar Troncoso and Lan Luo. Look the part? the role of profile pictures in online labor markets. *Marketing Science*, 2022.

A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

Yanhao'Max' Wei and Nikhil Malik. Unstructured data, econometric models, and estimation bias. In *Unstructured Data, Econometric Models, and Estimation Bias: Wei, Yanhao'Max'—uMalik, Nikhil*. [SI]: SSRN, 2022.

Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10):1499–1503, 2016.

Web Appendix

A Details of Data Collection and Cleaning

A.1 Details of Politicians and News Outlets

Two important details about our data collection relate to the set of politicians and news outlets. Below we present these two lists:

- **Politicians:** To select the sample of 30 politicians, we first created a large set of individuals who have run for public offices from either Democratic or Republican party. We then sampled a set of top 30 politicians based on their search volume on Google Trends. As such, the higher the search volume for a politician, the more likely it is to select that politician for our main sample. The resulting sets of Democratic and Republican politicians is presented below:
 - *Democratic Politicians:* This list contains (1) Barack Obama, (2) Joe Biden, (3) Hilary Clinton, (4) Bernie Sanders, (5) Nancy Pelosi, (6) Kamala Harris, (7) Chuck Schumer, (8) Corey Booker, (9) Amy Klobuchar, (10) Elizabeth Warren, (11) Gavin Newsom, (12) Bill de Blasio, (13) James Clyburn, (14) Janet Yellen, (15) Joe Manchin, and (16) Steny Hoyer.
 - *Republican Politicians:* This list contains (1) Donald Trump, (2) Ted Cruz, (3) Marco Rubio, (4) Mitt Romney, (5) Mitch McConnell, (6) Greg Abbott, (7) Rand Paul, (8) Mike Pence, (9) Kevin McCarthy, (10) Susan Collins, (11) Liz Chenney, (12) John Roberts, (13) Hal Rogers, (14) Andy Biggs.
- **News Outlets:** We selected top 20 news outlets following the work by [Flaxman et al. \(2016\)](#). Unlike politicians, news outlets do not have a clear political affiliation and party. However, there are several indices for media slant based on the readership and coverage ([Groseclose and Milyo, 2005](#); [Flaxman et al., 2016](#)). For example, Figure A.1 shows the news outlets in our data using the *conservative share* measure used by [Flaxman et al. \(2016\)](#). In our main analysis, we do not use the media slant indices for model building and measuring political polarization in visual content. However, we use these measures for validation of our main measures.

A.2 Details of Data Cleaning

In this section, we present details about data cleaning. We first present the manual verification procedure we used to ensure that SerpAPI does not miss important data. To address the limitations of the automated method, we employ an additional manual process for quality control. This involves saving HTML files of search results locally and using Beautiful Soup to extract images and metadata. This method ensures the accuracy and completeness of our dataset.

We now present the two-phase data cleaning procedure in Figure 2 in greater detail. The first phase, designed for accurate face detection, consists of two main steps. Initially, we perform pyramid rescaling, adjusting each image to three scales (1.1, 0.8, and 0.6) to enhance focus, similar to adjusting a camera lens ([Lin et al., 2017](#)). Following this, Mask R-CNN, which uses the ResNet50 architecture as its backbone, identifies and outlines humans within these images ([He et al., 2017](#)), referring to an object detection and

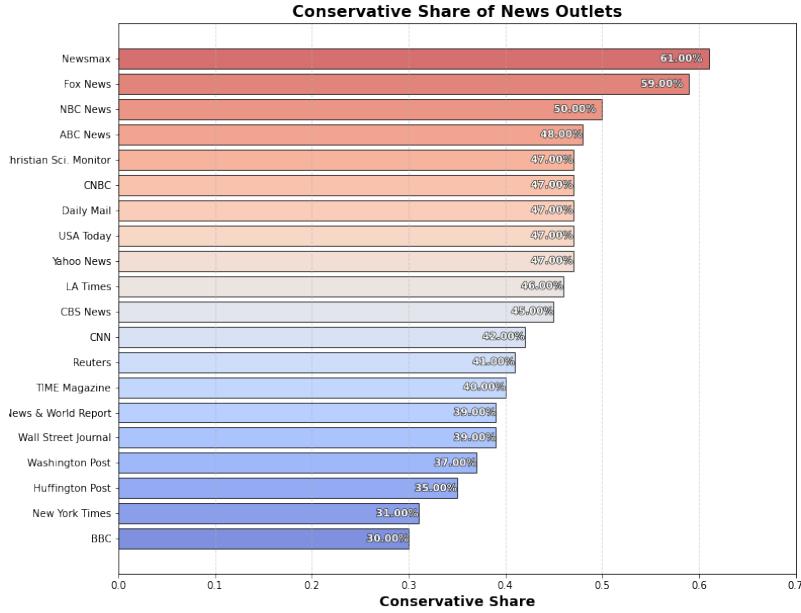


Figure A.1: Conservative share of news outlets. ([Flaxman et al., 2016](#))

instance segmentation task. Next, we use multitask cascaded Convolutional Neural Networks (MTCNN) for face detection within the regions identified by Mask R-CNN. The detected faces are subsequently processed by FaceNet for face recognition ([Zhang et al., 2016](#)). This dual approach not only increases the efficiency of our face detection process but also significantly improves its accuracy.

The second phase is the face verification stage of our computer vision framework. We initiate the process by feeding ArcFace with twenty samples for each of the thirty politicians to create a unique facial pattern representative for each political person ([Deng et al., 2019](#)). Then, we take the images from the first phase, which are confirmed to have just one face, and run them through ArcFace. This system compares the detected face with our database of political figures and ranks the top three matches, providing a similarity percentage for each.

Ultimately, we obtained 53,563 images with exactly one face, verified by our dual framework. These images represent a clean and reliable dataset for our study. The distribution of the final data is shown in Figure A.2. This dataset allows us to accurately analyze media representation of political figures, ensuring that each image is correctly attributed to the intended individual.

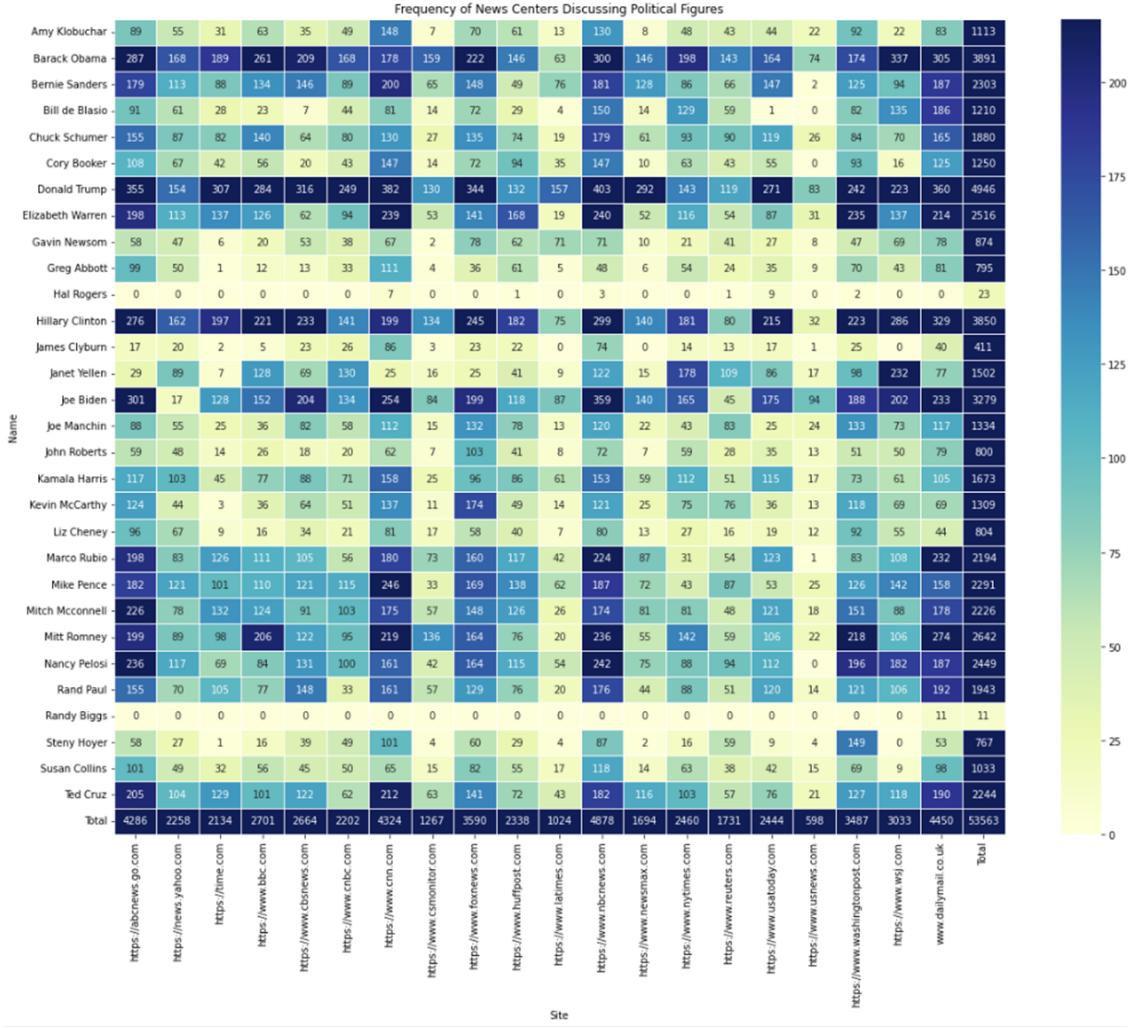


Figure A.2: Heatmap depicting the association between political figures and news outlets

B Theoretical Analysis of Bias for the Two-step Model

B.1 Extraction Bias

As discussed in §5.2, *extraction bias* arises from the use of the predicted variable \hat{T} in the analysis because the true variable T is not observable. This issue is primarily related to the machine learning model f_1 , where transfer learning is employed to predict the feature of interest. We provide the econometric theoretical framework demonstrating how extraction bias can lead to biased estimation of β .

Recall that T is not directly observed; rather, it is extracted from the high-dimensional unstructured image data Z using a first-stage machine learning model f_1 . This leads to two types of errors in the estimate \hat{T} : Measurement Error (ε_m) and Extraction Error (ε_e).

Remark 1. *Measurement Error (ε_m) arises from inaccuracies in predictions or measurements that are not*

systematically correlated with the true value of T (Hansen, 2022). This can be formally defined as:

$$T = \hat{T} + \varepsilon_m \quad \text{where} \quad \text{Cov}(T, \varepsilon_m) = 0$$

Remark 2. Extraction Error (ε_e) occurs when the estimation method includes additional, indirectly related features (Wei and Malik, 2022). This error implies that the error term is correlated with the true value T :

$$T = \hat{T} + \varepsilon_e \quad \text{where} \quad \text{Cov}(T, \varepsilon_e) \neq 0$$

Measurement error can be viewed as random noise added to the estimate \hat{T} , resulting in a variable that is not systematically biased. In contrast, extraction errors occur when the machine learning model, trained to predict the specific feature of interest T , inadvertently captures additional non-focal features that are correlated with T . For instance, a machine learning model designed to detect the facial expression of faces whether a person is happy may also inadvertently capture brightness levels in the image, as brighter images are often associated with happier expressions. This can introduce an extraction error ε_e into the model's predictions, potentially biasing the results.

In our setting, suppose that images used by CNN are generally brighter. In this case, the model may also capture brightness rather than true happiness. As a result, the predicted variable \hat{T} reflects not only the intended feature (happiness) but also the unintended influence of brightness. This issue complicates the analysis by conflating the natural effect of emotional expression with confounding factors, making it difficult to separate their impacts. To assess this challenge theoretically, we can divide it into *Well-specified* and *Partially-specified* models.

Consider scenarios where T represents an oracle estimate of the population. Achieving this oracle estimate can be done by using human judgment to classify photos as Happy ($T = 1$) or Unhappy ($T = 0$), though this method is expensive. Now, assume a simple linear model where Y is a binary variable indicating the news outlet (1 for CNN, 0 for Fox news). In this case, the parameter of interest, β , can be estimated without bias because of the following Proposition.

Proposition 2. Well-specified: Consider the following model for \hat{T}_{oracle} :

$$\begin{aligned} \hat{T}_{\text{oracle}} &= T + \varepsilon_m, \quad \text{where} \quad \hat{T}_{\text{oracle}} = \text{Human}, \\ \hat{T}_{\text{oracle}} &= \beta Y + \varepsilon + \varepsilon_m \end{aligned} \tag{A.20}$$

Given that ε and ε_m are i.i.d. and uncorrelated with other variables, $\hat{\beta}$ is an unbiased estimator of β . Using the Central Limit Theorem (CLT), we get:

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N\left(0, \frac{\sigma_\varepsilon^2 + \sigma_{\varepsilon_m}^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}\right)$$

Proof. In scenarios where T represents an oracle estimation of the population, it is possible to leverage human judgment to categorize photos into Happy ($T = 1$) or Unhappy ($T = 0$). Assume a simple linear model where Y is a binary variable representing the news outlet (1 for Fox News, 0 for CNN). In this case,

the predicted variable \hat{T}_{oracle} can be formulated as:

$$\hat{T}_{oracle} = T + \varepsilon_m, \quad \text{where } \hat{T}_{oracle} = \text{Human Estimate of } T$$

$$\hat{T}_{oracle} = \beta Y + \varepsilon + \varepsilon_m \quad \text{such that } \varepsilon_m \sim N(0, \sigma_{\varepsilon_m}^2) \quad \& \quad \varepsilon \sim N(0, \sigma_{\varepsilon}^2)$$

We define $\theta = \varepsilon + \varepsilon_m$. Thus, we have:

$$\hat{T}_{oracle} = \beta Y + \theta$$

Given that both error terms ε and ε_m are assumed to be Independent and Identically Distributed (iid) and uncorrelated with Y , we can show:

$$E[\theta | Y] = E[\varepsilon + \varepsilon_m | Y] = E[\varepsilon | Y] + E[\varepsilon_m | Y] = 0$$

The least squares estimator for β is given by:

$$\hat{\beta} = \frac{\sum_{i=1}^n (Y_i - \bar{Y}) \hat{T}_{oracle,i}}{\sum_{i=1}^n (Y_i - \bar{Y}) Y_i}$$

Substituting $\hat{T}_{oracle,i}$ with its expression:

$$\hat{T}_{oracle,i} = \beta Y_i + \theta_i$$

we get:

$$\hat{\beta} = \beta + \frac{\sum_{i=1}^n (Y_i - \bar{Y}) \theta_i}{\sum_{i=1}^n (Y_i - \bar{Y}) Y_i}$$

Dividing the numerator and denominator by n :

$$\hat{\beta} = \beta + \frac{\frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y}) \theta_i}{\frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y}) Y_i}$$

Using the Weak Law of Large Numbers (WLLN), as $n \rightarrow \infty$:

$$\frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y}) \theta_i \rightarrow E[(Y_i - \bar{Y}) \theta_i] = E[(Y_i - \bar{Y}) \varepsilon] + E[(Y_i - \bar{Y}) \varepsilon_m] = 0$$

Therefore, in probability:

$$\hat{\beta} \xrightarrow{P} \beta$$

The variance of $\hat{\beta}$, given $\theta_i \sim \text{i.i.d.}(0, \sigma_{\theta}^2)$, is:

$$\text{Var}(\hat{\beta}) = \text{Var}\left(\frac{\sum_{i=1}^n (Y_i - \bar{Y}) \theta_i}{\sum_{i=1}^n (Y_i - \bar{Y})^2}\right) = \frac{\sigma_{\varepsilon}^2 + \sigma_{\varepsilon_m}^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$

Using the Central Limit Theorem (CLT), we can show:

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N\left(0, \frac{\sigma_\varepsilon^2 + \sigma_{\varepsilon_m}^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}\right)$$

This confirms that in a well-specified model, the estimator $\hat{\beta}$ remains unbiased, and only the variance is affected by the inclusion of the error terms. This approach validates the reliability of using such models in well-specified scenarios, such as categorizing images of Trump as either happy or unhappy when used by Fox News or CNN.

□

The Proposition above confirms that in a well-specified model, the estimator $\hat{\beta}$ remains unbiased, and only the variance is affected by the inclusion of the error terms. This approach validates the reliability of using such models in well-specified scenarios.

Proposition 3. Partial-specified: Consider the following model for \hat{T}_{ml} :

$$\begin{aligned} \hat{T}_{ml} &= T + \varepsilon_m + \varepsilon_e, \quad \text{where } \hat{T}_{ml} = f_1(Z), \\ T &= \beta Y + \varepsilon, \\ \hat{T}_{ml} &= \beta Y + (\varepsilon + \varepsilon_m + \varepsilon_e) \end{aligned} \tag{A.21}$$

Given that ε and ε_m are i.i.d., but ε_e is correlated with T , $\hat{\beta}$ would be a biased estimator of β as:

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N\left(\frac{\text{Cov}(Y, \varepsilon_e)}{\sum_{i=1}^n (Y_i - \bar{Y})^2}, \frac{\sigma_\varepsilon^2 + \sigma_{\varepsilon_m}^2 + \sigma_{\varepsilon_e}^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}\right)$$

Proof. In real-world scenarios, \hat{T} often includes extraction errors, leading to biases. Consider the following model for \hat{T}_{ml} :

$$\begin{aligned} \hat{T}_{ml} &= T + \varepsilon_m + \varepsilon_e, \quad \text{where } \hat{T}_{ml} = f_1(Z), \\ T &= \beta Y + \varepsilon, \\ \hat{T}_{ml} &= \beta Y + (\varepsilon + \varepsilon_m + \varepsilon_e) \quad \text{s.t. } \varepsilon_m \sim N(0, \sigma_{\varepsilon_m}^2) \quad \& \quad \varepsilon \sim N(0, \sigma_\varepsilon^2) \quad \& \quad \text{Cov}(T, \varepsilon_e) \neq 0 \end{aligned}$$

Start with expanding $\text{Cov}(T, \varepsilon_e)$:

$$\text{Cov}(T, \varepsilon_e) = \beta \cdot \text{Cov}(Y, \varepsilon_e) + \text{Cov}(\varepsilon, \varepsilon_e)$$

Given that ε is i.i.d., $\text{Cov}(\varepsilon, \varepsilon_e) = 0$. Therefore, $\text{Cov}(T, \varepsilon_e) \neq 0$ directly implies:

$$\text{Cov}(Y, \varepsilon_e) \neq 0$$

The Ordinary Least Squares (OLS) estimator for β is:

$$\hat{\beta} = \frac{\sum_{i=1}^n (Y_i - \bar{Y}) \hat{T}_{ml,i}}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$

Substituting $\hat{T}_{ml,i}$ with its expression:

$$\hat{T}_{ml,i} = \beta Y_i + \varepsilon_i + \varepsilon_{m,i} + \varepsilon_{e,i}$$

we get:

$$\hat{\beta} = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(\beta Y_i + \varepsilon_i + \varepsilon_{m,i} + \varepsilon_{e,i})}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$

Simplifying further:

$$\hat{\beta} = \beta + \frac{\sum_{i=1}^n (Y_i - \bar{Y})(\varepsilon_i + \varepsilon_{m,i} + \varepsilon_{e,i})}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$

Dividing the numerator and denominator by n :

$$\hat{\beta} = \beta + \frac{\frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y})(\varepsilon_i + \varepsilon_{m,i} + \varepsilon_{e,i})}{\frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y})^2}$$

Using the Weak Law of Large Numbers (WLLN), as $n \rightarrow \infty$:

$$\frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y})(\varepsilon_i + \varepsilon_{m,i} + \varepsilon_{e,i}) \xrightarrow{p} \text{Cov}(Y, \varepsilon_e) \neq 0$$

Therefore, in probability:

$$\hat{\beta} \xrightarrow{p} \beta + \frac{\text{Cov}(Y, \varepsilon_e)}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$

This indicates that $\hat{\beta}$ is a biased estimator of β due to the non-zero covariance between Y and ε_e . The variance of $\hat{\beta}$, given $\theta_i = \varepsilon_i + \varepsilon_{m,i} + \varepsilon_{e,i} \sim \text{i.i.d.}(0, \sigma_\theta^2)$, is:

$$\text{Var}(\hat{\beta}) = \text{Var} \left(\frac{\sum_{i=1}^n (Y_i - \bar{Y}) \theta_i}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \right) = \frac{\sigma_\varepsilon^2 + \sigma_{\varepsilon_m}^2 + \sigma_{\varepsilon_e}^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$

Using the Central Limit Theorem (CLT), we can show:

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N \left(\frac{\text{Cov}(Y, \varepsilon_e)}{\sum_{i=1}^n (Y_i - \bar{Y})^2}, \frac{\sigma_\varepsilon^2 + \sigma_{\varepsilon_m}^2 + \sigma_{\varepsilon_e}^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \right)$$

This confirms that in a partially-specified model, the estimator $\hat{\beta}$ is biased due to the correlation between Y and ε_e . Additionally, the variance of $\hat{\beta}$ is inflated by the variances of the error terms. This bias arises specifically because $\text{Cov}(Y, \varepsilon_e) \neq 0$, which is now directly implied without the presence of X . \square

In summary, in a partially-specified model where \hat{T}_{ml} includes extraction errors correlated with Y , the OLS estimator $\hat{\beta}$ becomes biased. This bias stems from the inherent correlation between the predictor Y

and the error term ε_e . Consequently, while the estimator remains consistent in its variance under the CLT, its expectation deviates from the true parameter β , undermining its unbiasedness.

B.2 Omitted Variable Bias

This bias can occur when the variable of interest Y is correlated with the information not captured in T . Recall that $Z^{(-T)}$ represents the components of Z that are not captured by the extracted feature T . To quantify this bias, we first start by characterizing the true relationship between the variable of interest Y , the extracted feature T , and the omitted components $Z^{(-T)}$ as $T = f_2(Y, k(Z^{(-T)}); \beta) + \varepsilon_0$, where $k(Z^{(-T)})$ is a non-parametric function that captures the potentially complex effects of the omitted variables $Z^{(-T)}$. In contrast, the econometric model that researchers estimate (as described earlier) is given by $T = f_2(Y; \beta) + \varepsilon$. The key challenge is that the error term ε is composed of both the original noise ε_0 and the bias due to the exclusion of $k(Z^{(-T)})$.

Remark 3. *Omitted Variable Bias (ε_{ov}) occurs when the error term includes the effect of omitted variables that are correlated with the independent variable Y . Formally, the error term can be expressed as:*

$$\varepsilon = \varepsilon_0 + \varepsilon_{ov} \quad \text{where} \quad \text{Cov}(Y, \varepsilon_{ov}) \neq 0$$

Proposition 4. Omitted Variable Bias: Consider the linear model for T :

$$\begin{aligned} T &= \beta Y + \delta k(Z^{(-T)}) + \varepsilon_0 \quad (\text{True Model}), \\ T &= \tilde{\beta} Y + \varepsilon \quad (\text{Observed Model}), \end{aligned} \tag{A.22}$$

where $\varepsilon = \varepsilon_0 + \varepsilon_{ov}$ and $\varepsilon_{ov} = \delta k(Z^{(-T)})$ represents the impact of the omitted variable. Then, the bias in the estimator $\tilde{\beta}$ is given by:

$$\text{Bias}(\hat{\tilde{\beta}}) = \tilde{\beta} - \beta = \lambda \delta, \tag{A.23}$$

where λ is the coefficient from the projection of Y on $k(Z^{(-T)})$ based on the Frisch-Waugh-Lovell Theorem (Hansen, 2022).

Proof. We start by considering the true model:

$$T = \beta Y + \delta k(Z^{(-T)}) + \varepsilon_0$$

and the estimated model, which omits the variable $k(Z^{(-T)})$:

$$T = \tilde{\beta} Y + \varepsilon$$

where $\varepsilon = \varepsilon_0 + \varepsilon_{ov}$ and $\varepsilon_{ov} = \delta k(Z^{(-T)})$. The OLS estimator for $\tilde{\beta}$ is given by:

$$\hat{\tilde{\beta}} = \frac{\text{Cov}(Y, T)}{\text{Var}(Y)}$$

Substituting the true model into this equation, we have:

$$\text{Cov}(Y, T) = \text{Cov}\left(Y, \beta Y + \delta k(Z^{(-T)}) + \varepsilon_0\right)$$

Expanding the covariance:

$$\text{Cov}(Y, T) = \beta \text{Cov}(Y, Y) + \delta \text{Cov}(Y, k(Z^{(-T)})) + \text{Cov}(Y, \varepsilon_0)$$

Assuming $\text{Cov}(Y, \varepsilon_0) = 0$, we get:

$$\text{Cov}(Y, T) = \beta \text{Var}(Y) + \delta \text{Cov}(Y, k(Z^{(-T)}))$$

Thus, the OLS estimator becomes:

$$\hat{\beta} = \frac{\beta \text{Var}(Y) + \delta \text{Cov}(Y, k(Z^{(-T)}))}{\text{Var}(Y)}$$

The expected value of $\hat{\beta}$ is therefore:

$$\mathbb{E}[\hat{\beta}] = \frac{\beta \text{Var}(Y) + \delta \text{Cov}(Y, k(Z^{(-T)}))}{\text{Var}(Y)} = \beta + \frac{\delta \text{Cov}(Y, k(Z^{(-T)}))}{\text{Var}(Y)}$$

This simplifies to:

$$\text{Bias}(\hat{\beta}) = \frac{\delta \text{Cov}(Y, k(Z^{(-T)}))}{\text{Var}(Y)}$$

where $\frac{\text{Cov}(Y, k(Z^{(-T)}))}{\text{Var}(Y)} = \lambda$ is the coefficient from the projection of Y on $k(Z^{(-T)})$.

□

C Complete Regression Analysis Results

We now expand the simple empirical context from §5.3 into a more comprehensive analysis. To that end, the emotion score (happiness) for a given image is modeled as a function of both politician-specific and outlet-specific fixed effects, alongside the interaction between the partisanship of the news outlet and the political affiliation of the politician. This approach aligns with the framework used by Boxell (2021). The formal expression of the regression model is:

$$y_{ijt} = \alpha + p_i + o_j + \beta \cdot (c_j \cdot \mathbb{I}(i \in R)) + e_{ijt}$$

In this model, y_{ijt} represents the binary dependent variable that indicates whether happiness is recognized in instance t of politician i on website j (with $y_{ijt} = 1$ if happiness is recognized, and $y_{ijt} = 0$ otherwise). The term α is the intercept, capturing the baseline level of the emotion score across all observations. The fixed effect p_i accounts for characteristics unique to each politician, while o_j represents fixed

effects specific to each news outlet. The variable c_j denotes the conservative share score for news outlet j , reflecting its political leaning.

The interaction term $\beta \cdot (c_j \cdot \mathbb{I}(i \in R))$ captures the effect of the news outlet's conservative-leaning when the politician is a member of the Republican party. Here, $\mathbb{I}(i \in R)$ is an indicator function that equals 1 if the politician i is Republican, and 0 if the politician is Democratic. This interaction term thus models how the conservative share score of a news outlet influences the portrayal of Republican politicians differently compared to Democratic politicians. Finally, e_{ijt} represents the error term, which captures unobserved factors that might affect the emotion score. This formulation allows the model to assess the differential impact of news outlet partisanship on the emotional portrayal of politicians from different parties, with a specific focus on how conservative-leaning outlets treat Republican politicians compared to their Democratic politicians.

Table A1: Regression Results for Emotion (Happiness)

Variable	Coef.	Std. Err.	t	P> t 	[0.025	0.975]
Intercept	0.1021	0.003	29.392	0.000	0.095	0.109
ABC News	0.0131	0.002	5.293	0.000	0.008	0.018
Time	0.0069	0.003	2.050	0.040	0.000	0.014
BBC	-0.0050	0.003	-1.589	0.112	-0.011	0.001
CBS News	0.0044	0.003	1.461	0.144	-0.001	0.010
CNBC	0.0041	0.003	1.233	0.218	-0.002	0.011
CNN	-0.0033	0.002	-1.389	0.165	-0.008	0.001
CS Monitor	0.0176	0.004	4.089	0.000	0.009	0.026
Daily Mail	-0.0075	0.002	-3.125	0.002	-0.012	-0.003
Fox News	-0.0122	0.003	-4.091	0.000	-0.018	-0.006
HuffPost	0.0094	0.003	2.882	0.004	0.003	0.016
LA Times	0.0098	0.005	2.084	0.037	0.001	0.019
NBC News	0.0025	0.002	1.046	0.296	-0.002	0.007
Newsmax	-0.0073	0.004	-1.804	0.071	-0.015	0.001
NY Times	0.0262	0.003	8.221	0.000	0.020	0.032
USA Today	0.0121	0.003	3.839	0.000	0.006	0.018
US News	0.0109	0.006	1.778	0.075	-0.001	0.023
Washington Post	0.0042	0.003	1.573	0.116	-0.001	0.009
Wall Street Journal	0.0162	0.003	5.741	0.000	0.011	0.022
Amy Klobuchar	0.0561	0.006	9.492	0.000	0.045	0.068
...
Ted Cruz	-0.0277	0.006	-4.644	0.000	-0.039	-0.016
Conservative Share	0.0515	0.018	2.801	0.005	0.015	0.088

Based on the results of the regression analysis, the significance of the partisanship-related variable (Con-

servative Share, $\beta = 0.0515$) with $p - value = 0.005$ indicates that visual bias exists in how politicians are portrayed. Specifically, news outlets with higher conservative shares are more likely to depict politicians in a happier light who are categorized as Republican politicians.

D Empirical Evidence - Extraction Bias

In this section, we demonstrate that when using off-the-shelf machine learning models for tasks such as facial expression analysis, formulated as $f_1(Z) = \hat{T}$, the residual $\varepsilon_e = \hat{T} - T$ is correlated with the true label T . This correlation represents Extraction Bias, as outlined in Remark 2. Since the true variable T (happiness) is not directly annotated in our dataset, we utilize a dataset created by [Chen et al. \(2022\)](#), which includes 135 distinct facial expressions with accurate true labels. For our analysis, we randomly select 2,500 images labeled as Happy ($T = 1$) and 2,500 images labeled as Sad ($T = 0$).

Variable	<i>Dependent Variable: Residual</i>	
	(1)	(2)
Intercept	0.1097*** (0.0081)	0.0991** (0.035)
Label	-0.4758*** (0.0115)	-0.4817*** (0.012)
Brightness		0.0000877 (0.000)
Edge Density		0.0030*** (0.001)
Contrast		-0.0007 [†] (0.0004)
No. of Obs.	4997	4997
R ²	0.256	0.258
Log-Likelihood	-2586.8	-2580.0
F-statistic	1715.0***	433.0***
<i>Note:</i>		[†] p<0.10; *p<0.05; **p<0.01; ***p<0.001

Table A2: OLS regression results for residuals. Model 1 shows the relationship between the true label (happy/sad) and the residuals. Model 2 includes image characteristics such as brightness, edge density, and contrast as additional predictors.

Each image is passed through the DeepFace model, which generates predicted scores for seven emotions. The predicted label is determined by identifying the emotion with the highest score: if Happy has the highest score, the predicted label is set to 1 (happy); otherwise, it is set to 0 (not happy). The residual $\varepsilon_e = \hat{T} - T$ measures the prediction error, where a positive residual indicates over-prediction of happiness, and a negative residual indicates under-prediction. In addition to the predicted labels, we also extract three image characteristics: brightness, edge density, and contrast. These characteristics are included to assess whether

the visual properties of the images influence the model’s prediction errors. To analyze the relationship between the residuals, true labels, and image characteristics, we employ two OLS regression models.

The first model regresses the residuals on the true labels to determine if the model’s prediction errors are systematically related to T . In this case, we evaluate whether the residual ε_e is a function of T , i.e., $\varepsilon_e = \alpha + \beta T + \eta$, where η is the error term and α, β are coefficients to be estimated. The second model extends this analysis by adding image characteristics (brightness, edge density, and contrast) to investigate whether these features contribute to the systematic prediction errors, leading to the model $\varepsilon_e = \alpha + \beta T + \gamma_1 X_{\text{Brightness}} + \gamma_2 X_{\text{EdgeDensity}} + \gamma_3 X_{\text{Contrast}} + \eta$, where X represents the image characteristics.

The results are summarized in Table A2. In Model 1, the true label explains approximately 25.6% of the variance in the residuals ($R^2 = 0.256$), indicating a significant relationship between T and ε_e . The coefficient for the true label is -0.4758 ($p < 0.001$), suggesting that the DeepFace model systematically under-predicts happiness when the true label is happy and over-predicts happiness when the true label is sad. Additionally, the intercept of 0.1097 shows that, on average, the model tends to over-predict happiness for sad images. The high significance level ($p < 0.001$) confirms that T plays a key role in influencing the residuals, thus highlighting the presence of extraction bias in the model.

Model 2, which controls for the image characteristics, suggests that while image characteristics such as brightness, edge density, and contrast provide some additional information, the extraction bias persists. Notably, edge density is the only statistically significant characteristic ($p < 0.001$), with a positive coefficient of 0.0030. This suggests that images with more pronounced edges are more likely to lead to larger residuals, meaning the model tends to over-predict happiness in these cases. Despite controlling for these image characteristics, the residual remains strongly correlated with the true label, supporting the conclusion that extraction bias is inherent in the model’s feature extraction process and is not fully explained by the visual properties of the images. The extraction bias originates from the model’s tendency to systematically over- or under-predict certain emotions, independently of the visual features included in the analysis.

E Appendix: Implementation Details

E.1 Generative Adversarial Networks (GANs)

We utilize Generative Adversarial Networks (GANs) to generate slanted images by altering specific attributes while retaining other features. GANs, introduced by [Goodfellow et al. \(2014\)](#), consists of a generator G and a discriminator D that train together through an adversarial process. The generator G creates synthetic images from random noise z' or an input image z , while the discriminator D differentiates between real and generated images. The training objective can be framed as a minimax game:

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{\text{data}}(x)} [\log D(x)] + \mathbb{E}_{z' \sim p_{z'}(z')} [\log (1 - D(G(z')))]$$

Here, x denotes a real image drawn from the true data distribution $p_{\text{data}}(x)$, and z' is a random noise vector sampled from a prior distribution $p_{z'}(z')$, typically Gaussian. The generator G takes z' and generates an image $G(z')$. The discriminator D aims to distinguish between real images x and generated images $G(z')$. In this adversarial setup, the generator G strives to produce images that appear real, while the discriminator D seeks to accurately identify real versus generated images. Conditional GANs (cGANs) are used to generate

slanted images conditioned on specific attributes, such as a smiling expression ([Mirza and Osindero, 2014](#)). The objective function is:

$$\min_G \max_D V(D, G) = \mathbb{E}_{x, T \sim p_{\text{data}}(x, T)} [\log D(x, T)] + \mathbb{E}_{z' \sim p_{z'}(z'), T \sim p_T(T)} [\log (1 - D(G(z', T), T))]$$

In this context, x is a real image with the attribute T , sampled from the true data distribution $p_{\text{data}}(x, T)$. The generator G takes both the random noise vector z' and the attribute vector T as inputs to produce an image reflecting the specified attribute. The discriminator D is tasked with distinguishing real images with the attribute T from generated ones conditioned on the same attribute. This setup motivates the generator to create images that not only look real but also exhibit the given attribute. During training, the generator G and discriminator D are optimized iteratively. The generator's goal is to produce realistic images with the specified attribute, while the discriminator aims to classify real and generated images correctly. The loss functions for the discriminator and generator are defined as:

$$\mathcal{L}_D = -\mathbb{E}_{x, T \sim p_{\text{data}}(x, T)} [\log D(x, T)] - \mathbb{E}_{z' \sim p_{z'}(z'), T \sim p_T(T)} [\log (1 - D(G(z', T), T))]$$

The discriminator's loss function aims to maximize the probability of accurately classifying real images as real and generated images as fake, thereby maximizing L_D .

$$\mathcal{L}_G = -\mathbb{E}_{z' \sim p_{z'}(z'), T \sim p_T(T)} [\log D(G(z', T), T)]$$

The generator's loss function seeks to minimize the probability that the discriminator correctly classifies generated images as fake, thus minimizing L_G . The overall objective is to find the optimal G and D that minimize and maximize these losses, respectively, until G produces images indistinguishable from real ones with the desired attribute T .

Once trained, the generator G evolves into G' , and it is used for generating new images. Given an original image z and a target attribute T , the trained generator G' creates a slanted image z^1 from random noise z' that incorporates the target attribute while preserving other features:

$$z^1 = G'(z, T) \approx \pi^1(z)$$

For example, to generate a slanted image of Donald Trump with a smile, follow these steps: 1. Input the original image z , 2. Set the target attribute T to "smiling", and 3. The generator G' modifies the random vector z' to produce z^1 with a smile:

$$z^1 = G'(z, T_{\text{smile}})$$

A notable instance in this domain is the work by [Luo and Toubia \(2024\)](#), who designed a GAN to adjust facial femininity in images while keeping other features unchanged. Similarly, our task involves altering facial expressions, such as adding a smile to an image.

Developing a high-quality cGAN that generates images closely resembling the original while altering

only a specific target attribute is challenging. It demands extensive training data and meticulous tuning of the model parameters to ensure that the generated images are realistic and only the specified attribute is changed. This is crucial to maintain the authenticity of the images and the validity of the study.

To achieve this, we employ advanced image editing tools as our G' , such as those available on [AILabTools](#). AILabTools offers a suite of powerful image editing and enhancement tools designed to improve the quality of images generated by cGANs. By utilizing these tools, we ensure our manipulated images maintain their original authenticity while incorporating the desired attributes. This platform aids in generating the slanted image z^1 and the neutral image z^0 from an original image z , ensuring high-quality results with precise attribute manipulation.

E.1.1 Generating Counterfactual Images

In this part, we explain generating counterfactual images. First, for each politician p , we select three images s_{p1}, s_{p2}, s_{p3} that are not used to train our multi-modal ML model, sum up to 84 images. To ensure unbiasedness, these images are chosen to represent the politician in a neutral manner, avoiding any inherent slant. Consequently, this selected image set inherently represents the neutral version z^0 . Additionally, selecting three images helps mitigate any specific biases that might arise from using a single image and ensures that our analysis is robust across different visual contexts.

Using the approach explained in Appendix E.1, we generate a smiley version of each image while keeping other factors constant. In other words, we apply the transformation π^1 on images s_{p1}, s_{p2}, s_{p3} to obtain $\pi^1(s_{p1}) = s_{p1}^1, \pi^1(s_{p2}) = s_{p2}^1, \pi^1(s_{p3}) = s_{p3}^1$. Each s^1 now embodies the slant notion, specifically a smile, as shown in Figure 4.

Since the isolated change in smile and facial expression is crucial, we further assess the generated images to see if other image characteristics have changed. The generated photos should have no other significant differences. As discussed, one of the challenges with the two-stage approach is that image characteristics such as brightness or colorfulness can cause bias in the model (see §5.2). Therefore, we measure the brightness and colorfulness of images before and after applying the smile operator π^1 . The histograms of these characteristics are shown in Figure A.3:

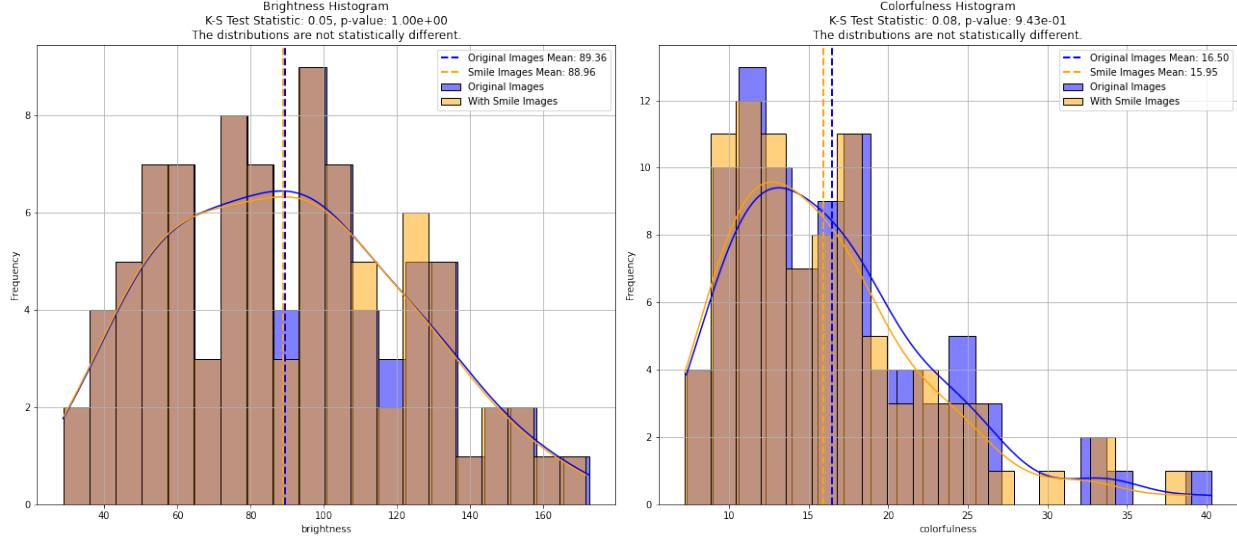


Figure A.3: Histograms of brightness and colorfulness for original and smiley images.

For brightness, the mean value for the original images is 89.36, while for the smiley images, it is 88.96. The Kolmogorov-Smirnov (K-S) test statistic is 0.05 with a p-value of 1.00, indicating that the distributions are not statistically different. This suggests that the introduction of a smile does not significantly alter the brightness of the images. Similarly, for colorfulness, the mean value for the original images is 16.50, compared to 15.95 for the smiley images. The K-S test statistic here is 0.08 with a p-value of 0.943, again indicating no statistical difference between the distributions.

This analysis confirms that the generated smiley images do not exhibit significant changes in brightness and colorfulness compared to the original images. This supports the validity of using these counterfactual images to isolate the effect of the smile on the parameter of interest, without introducing additional biases from other image characteristics.

E.2 Multi-Modal ML Implementation

E.2.1 Textual Data Processing using LDA

Textual data T_i is pivotal as it provides contextual and content information from the articles. We evaluated several approaches for modeling this data, including BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018) and LDA (Latent Dirichlet Allocation) (Blei et al., 2003). Despite BERT’s advanced capabilities in understanding context and semantics, LDA solely demonstrated superior performance in our prediction task, particularly in identifying topics and their distributions across the articles.

LDA is a generative statistical model that explains sets of observations by unobserved groups, revealing why certain parts of the data are similar. The model assumes that each document is a mixture of a small number of topics and that each word in the document is attributable to one of the document’s topics. The hyperparameters α and β are set to ‘auto’ to allow the model to learn these parameters during training. Mathematically, LDA posits the following generative process for a corpus D consisting of M documents,

each containing N words:

Algorithm 2 LDA Generative Process ([Blei et al., 2003](#))

```

1: for each topic  $k$  in  $\{1, \dots, K\}$  do
2:   Draw a distribution over words  $\phi_k \sim \text{Dir}(\beta)$ 
3: end for
4: for each document  $d$  in  $\{1, \dots, M\}$  do
5:   Draw a distribution over topics  $\theta_d \sim \text{Dir}(\alpha)$ 
6:   for each word  $n$  in  $\{1, \dots, N_d\}$  do
7:     Draw a topic  $z_{dn} \sim \text{Multinomial}(\theta_d)$ 
8:     Draw a word  $w_{dn} \sim \text{Multinomial}(\phi_{z_{dn}})$ 
9:   end for
10: end for

```

Here, α and β are hyperparameters of the Dirichlet distributions, θ_d is the topic distribution for document d , ϕ_k is the word distribution for topic k , z_{dn} is the topic assignment for the n -th word in document d , and w_{dn} is the n -th word in document d . In our model, we set the number of topics K to 20, and the model is trained with 20 passes over the corpus to ensure topic extraction.

The preprocessing steps for the textual data include cleaning and tokenizing the news titles, followed by lemmatization and stopword removal using NLTK’s WordNetLemmatizer and stopwords list ([Bird et al., 2009](#)). The processed tokens are then used to create bigrams and trigrams using Gensim’s Phrases model ([Řehůrek and Sojka, 2010](#)). These n-grams help capture contextual relationships between words, making the textual data more informative for the LDA model.

The trained LDA model generates topic distributions for each document. For each document d , LDA provides the distribution θ_d over topics, which can be interpreted as the document’s composition in terms of latent topics. These distributions are then scaled and normalized for input into the neural network. This transformation converts textual data into a structured format encapsulating underlying topics and their relevance to each document. As a result, we have a 20-dimensional vector representation for each news article title which is represented by F_{LDA} .

E.2.2 Face Information Branch: Framework and Operations

The Face Analysis branch of the proposed architecture is designed to extract, refine, and integrate facial information using a multi-step process that combines advanced detection, feature extraction, dimensionality reduction, and attention mechanisms. This section details each operation, starting from face detection via the MTCNN, progressing to feature extraction using Face-VGG, and culminating with chunked attention for integrating structured metadata. Given an input image $\mathbf{I} \in \mathbb{R}^{H \times W \times C}$, the MTCNN algorithm detects faces and aligns them for further processing. The detection process is divided into three stages. In the first stage, a proposal network is used to generate candidate bounding boxes:

$$\mathbf{R}_1 = f_1(\mathbf{I}; \theta_1),$$

where \mathbf{R}_1 denotes the set of candidate bounding boxes, and θ_1 are the parameters of the proposal network. These candidate regions are refined in the second stage using a refine network:

$$\mathbf{R}_2 = f_2(\mathbf{R}_1; \theta_2),$$

where \mathbf{R}_2 represents the refined bounding boxes. Finally, the output network predicts the bounding boxes and facial landmarks:

$$(\mathbf{R}_{\text{final}}, \mathbf{L}) = f_3(\mathbf{R}_2; \theta_3),$$

where $\mathbf{L} \in \mathbb{R}^{5 \times 2}$ are the coordinates of the facial landmarks. Using $\mathbf{R}_{\text{final}}$ and \mathbf{L} , the aligned face crop $\mathbf{F} \in \mathbb{R}^{160 \times 160 \times 3}$ is extracted from the original image. Each aligned face crop \mathbf{F} is passed through a pre-trained Face-VGG network to extract a 512-dimensional feature vector. Let $g_{\text{VGG}}(\cdot; \Theta_{\text{VGG}})$ denote the Face-VGG model, where Θ_{VGG} represents its parameters. The extracted feature vector $\mathbf{v} \in \mathbb{R}^{512}$ is computed as:

$$\mathbf{v} = g_{\text{VGG}}(\mathbf{F}; \Theta_{\text{VGG}}).$$

The last five layers of the Face-VGG model are fine-tuned to adapt to the specific task. The extracted feature vector \mathbf{v} is projected into a lower-dimensional space using a fully connected layer. The reduced feature vector $\mathbf{v}_{\text{red}} \in \mathbb{R}^{256}$ is given by:

$$\mathbf{v}_{\text{red}} = \mathbf{W}_{\text{red}}\mathbf{v} + \mathbf{b}_{\text{red}},$$

where $\mathbf{W}_{\text{red}} \in \mathbb{R}^{256 \times 512}$ and $\mathbf{b}_{\text{red}} \in \mathbb{R}^{256}$ are learnable parameters of the layer. To improve generalization, batch normalization and dropout are applied to \mathbf{v}_{red} . The normalized and regularized feature vector \mathbf{v}_{norm} is computed as:

$$\mathbf{v}_{\text{norm}} = \text{BatchNorm}(\mathbf{v}_{\text{red}}),$$

and the final output after dropout is:

$$\mathbf{v}_{\text{drop}} = \text{Dropout}(\mathbf{v}_{\text{norm}}, p = 0.3).$$

The reduced feature vector $\mathbf{v}_{\text{drop}} \in \mathbb{R}^{256}$ is divided into $k = 8$ equally sized chunks, each of dimension 32:

$$\mathbf{v}_{\text{chunk}} = \{\mathbf{v}_i \in \mathbb{R}^{32} \mid i = 1, \dots, k\}, \quad \mathbf{v}_i = \mathbf{v}_{\text{drop}}[32(i-1) : 32i].$$

This chunking operation allows the model to analyze localized regions of the image feature vector (Fu et al., 2019). To incorporate structured metadata, such as person and party embeddings, the chunked features $\mathbf{v}_{\text{chunk}}$ interact with a metadata embedding $\mathbf{m} \in \mathbb{R}^{32}$. The metadata embedding is computed as:

$$\mathbf{m} = \mathbf{W}_{\text{meta}}\mathbf{x} + \mathbf{b}_{\text{meta}},$$

where \mathbf{x} represents the concatenated embeddings of person and party, $\mathbf{W}_{\text{meta}} \in \mathbb{R}^{32 \times d_{\text{meta}}}$, and $\mathbf{b}_{\text{meta}} \in \mathbb{R}^{32}$ are learnable parameters. An attention mechanism assigns weights α_i to each chunk \mathbf{v}_i based on its similarity to

the metadata embedding \mathbf{m} :

$$\alpha_i = \frac{\exp(\mathbf{v}_i^\top \mathbf{m})}{\sum_{j=1}^k \exp(\mathbf{v}_j^\top \mathbf{m})}.$$

The attention-weighted feature representation $\mathbf{v}_{\text{att}} \in \mathbb{R}^{256}$ is computed as:

$$\mathbf{v}_{\text{att}} = \sum_{i=1}^k \alpha_i \mathbf{v}_i.$$

The output of the Face Analysis branch \mathbf{v}_{att} is concatenated with features from the ResNet branch $\mathbf{f}_{\text{ResNet}}$ and the metadata branch \mathbf{f}_{meta} to form the final aggregated feature vector:

$$\mathbf{z} = [\mathbf{v}_{\text{att}}; \mathbf{f}_{\text{ResNet}}; \mathbf{f}_{\text{meta}}].$$

This feature vector $\mathbf{z} \in \mathbb{R}^{d_{\text{final}}}$ is passed through subsequent layers for classification.

E.2.3 Structured Metadata Branch: Framework and Operations

The Structured Data Branch in the proposed architecture processes tabular metadata, including embeddings for categorical features (e.g., person, party, and date) and topic distributions derived from LDA. The key component of this branch is the multi-head attention mechanism, inspired by the Transformer model ([Vaswani, 2017](#)), which dynamically highlights relevant interactions among structured features. This section provides the mathematical foundations and operations in a coherent manner.

Let the structured input be represented as $\mathbf{x} \in \mathbb{R}^{d_{\text{struct}}}$, where $d_{\text{struct}} = 54$, obtained by concatenating the embeddings of categorical features and topic distributions. The embeddings for person, party, and date are denoted as $\mathbf{e}_{\text{person}} \in \mathbb{R}^{d_{\text{person}}}$, $\mathbf{e}_{\text{party}} \in \mathbb{R}^{d_{\text{party}}}$, and $\mathbf{e}_{\text{date}} \in \mathbb{R}^{d_{\text{date}}}$, respectively, while the topic vector is $\mathbf{e}_{\text{topics}} \in \mathbb{R}^{d_{\text{topics}}}$, where $d_{\text{topics}} = 40$. The structured input vector \mathbf{x} is thus formed as:

$$\mathbf{x} = [\mathbf{e}_{\text{person}}; \mathbf{e}_{\text{party}}; \mathbf{e}_{\text{date}}; \mathbf{e}_{\text{topics}}],$$

where $[.]$ denotes vector concatenation. This input vector is then projected into a latent representation space of dimension $d_{\text{latent}} = 64$ via a linear transformation:

$$\mathbf{h} = \mathbf{W}_{\text{linear}} \mathbf{x} + \mathbf{b}_{\text{linear}},$$

where $\mathbf{W}_{\text{linear}} \in \mathbb{R}^{d_{\text{latent}} \times d_{\text{struct}}}$ and $\mathbf{b}_{\text{linear}} \in \mathbb{R}^{d_{\text{latent}}}$ are learnable weights and biases. The resulting vector $\mathbf{h} \in \mathbb{R}^{d_{\text{latent}}}$ serves as the input to the multi-head attention mechanism. The multi-head attention mechanism begins by projecting \mathbf{h} into three distinct subspaces corresponding to the query (\mathbf{Q}), key (\mathbf{K}), and value (\mathbf{V}) representations ([Vaswani, 2017](#)). These projections are computed as:

$$\mathbf{Q} = \mathbf{W}_Q \mathbf{h}, \quad \mathbf{K} = \mathbf{W}_K \mathbf{h}, \quad \mathbf{V} = \mathbf{W}_V \mathbf{h},$$

where $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{d_{\text{head}} \times d_{\text{latent}}}$ are learnable weight matrices, and d_{head} represents the dimensionality of each attention head. The scaled dot-product attention mechanism computes the compatibility between

queries and keys, resulting in an attention score matrix:

$$\mathbf{A} = \text{softmax} \left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_{\text{head}}}} \right),$$

where $\text{softmax}(\cdot)$ normalizes the scores across all keys for each query, and the scaling factor $\sqrt{d_{\text{head}}}$ stabilizes gradients during training by mitigating the effect of large dot-product magnitudes. The attention scores $\mathbf{A} \in \mathbb{R}^{d_{\text{head}} \times d_{\text{head}}}$ modulate the value vectors \mathbf{V} , yielding the attention output:

$$\mathbf{O}_{\text{single}} = \mathbf{AV}.$$

To capture diverse interactions among the structured features, multiple attention heads are employed. For each attention head i , the process is repeated independently, producing:

$$\mathbf{O}_i = \mathbf{A}_i \mathbf{V}_i = \text{softmax} \left(\frac{\mathbf{Q}_i \mathbf{K}_i^\top}{\sqrt{d_{\text{head}}}} \right) \mathbf{V}_i,$$

where $i = 1, \dots, h$ and h is the number of heads. The outputs from all attention heads are concatenated and projected back into the latent representation space using a linear transformation:

$$\mathbf{O}_{\text{multi}} = \mathbf{W}_{\text{out}} [\mathbf{O}_1; \mathbf{O}_2; \dots; \mathbf{O}_h],$$

where $\mathbf{W}_{\text{out}} \in \mathbb{R}^{d_{\text{latent}} \times (h \cdot d_{\text{head}})}$ is a learnable projection matrix. The multi-head attention output $\mathbf{O}_{\text{multi}} \in \mathbb{R}^{d_{\text{latent}}}$ encodes a rich and dynamic representation of the structured data. To ensure stability and prevent overfitting, the output $\mathbf{O}_{\text{multi}}$ is normalized using batch normalization and regularized with dropout. These operations are defined as:

$$\mathbf{h}_{\text{norm}} = \text{BatchNorm}(\mathbf{O}_{\text{multi}}), \quad \mathbf{h}_{\text{drop}} = \text{Dropout}(\mathbf{h}_{\text{norm}}, p),$$

where p is the dropout probability. The final representation of the structured data branch is $\mathbf{h}_{\text{struct}} = \mathbf{h}_{\text{drop}}$, a 64-dimensional vector ready to be fused with the outputs from other branches in the model. The multi-head attention mechanism allows the model to dynamically prioritize different aspects of the structured data by assigning relevance scores based on the compatibility of queries and keys. The queries (\mathbf{Q}) represent the model's current focus, the keys (\mathbf{K}) encode contextual information about all features, and the values (\mathbf{V}) provide the corresponding information content. By using multiple attention heads, the model captures diverse patterns and relationships, resulting in a robust and task-specific representation of the structured metadata.

E.2.4 Image Contextual Information Branch: Framework and Operations

The ResNet101 branch is responsible for extracting high-dimensional feature representations from the full raw image data. Let $\mathbf{I} \in \mathbb{R}^{H \times W \times C}$ represent an input image, where $H = 244$ and $W = 244$ are the height and width, and C is the number of color channels. The input image passes through a pre-trained ResNet101 network, with the last 10 layers trainable. Denote the ResNet101 transformation as $g_{\text{ResNet}}(\cdot; \Theta_{\text{ResNet}})$, where Θ_{ResNet} are the trainable weights of the final layers. The output feature map $\mathbf{F}_{\text{ResNet}} \in \mathbb{R}^{d_{\text{ResNet}}}$ is given by:

$$\mathbf{F}_{\text{ResNet}} = g_{\text{ResNet}}(\mathbf{I}; \Theta_{\text{ResNet}}),$$

where $d_{\text{ResNet}} = 2048$ is the dimensionality of the extracted feature vector. This high-dimensional feature vector is flattened and passed through a linear layer for dimensionality reduction:

$$\mathbf{f}_{\text{ResNet}} = \mathbf{W}_{\text{reduce}} \mathbf{F}_{\text{ResNet}} + \mathbf{b}_{\text{reduce}},$$

where $\mathbf{W}_{\text{reduce}} \in \mathbb{R}^{d_{\text{reduce}} \times d_{\text{ResNet}}}$, $\mathbf{b}_{\text{reduce}} \in \mathbb{R}^{d_{\text{reduce}}}$, and $d_{\text{reduce}} = 512$. Batch normalization and dropout are applied to enhance generalization:

$$\mathbf{f}_{\text{norm}} = \text{BatchNorm}(\mathbf{f}_{\text{ResNet}}), \quad \mathbf{f}_{\text{drop}} = \text{Dropout}(\mathbf{f}_{\text{norm}}, p),$$

where $p = 0.6$ is the dropout probability. The final output of the ResNet101 branch is $\mathbf{f}_{\text{drop}} \in \mathbb{R}^{512}$, which is fused with outputs from other branches for classification.

E.2.5 Final Model Architecture

The final architecture of the multi-modal model is summarized in Table A3. This table provides an overview of each layer in the model.

The outputs from the three branches—Face Analysis ($\mathbf{f}_{\text{face}} \in \mathbb{R}^{d_{\text{face}}}$), Structured Data ($\mathbf{h}_{\text{struct}} \in \mathbb{R}^{d_{\text{struct}}}$), and ResNet101 ($\mathbf{f}_{\text{drop}} \in \mathbb{R}^{d_{\text{reduce}}}$)—are concatenated to form a unified feature vector:

$$\mathbf{z} = [\mathbf{f}_{\text{face}}; \mathbf{h}_{\text{struct}}; \mathbf{f}_{\text{drop}}] \in \mathbb{R}^{d_{\text{final}}},$$

where $d_{\text{final}} = d_{\text{face}} + d_{\text{struct}} + d_{\text{reduce}}$. In this case, $d_{\text{face}} = 256$, $d_{\text{struct}} = 64$, and $d_{\text{reduce}} = 512$, resulting in $d_{\text{final}} = 832$. To enhance generalization and mitigate overfitting, dropout is applied to the concatenated representation:

$$\mathbf{z}_{\text{drop}} = \text{Dropout}(\mathbf{z}, p),$$

where $p = 0.5$ is the dropout rate. The resulting vector $\mathbf{z}_{\text{drop}} \in \mathbb{R}^{d_{\text{final}}}$ is passed through a fully connected classification layer to compute the logits for the n_{class} target classes:

$$\mathbf{y} = \mathbf{W}_{\text{class}} \mathbf{z}_{\text{drop}} + \mathbf{b}_{\text{class}},$$

where $\mathbf{W}_{\text{class}} \in \mathbb{R}^{n_{\text{class}} \times d_{\text{final}}}$ and $\mathbf{b}_{\text{class}} \in \mathbb{R}^{n_{\text{class}}}$ are the weights and biases of the classification layer. The predicted class probabilities are obtained by applying the softmax function to the logits:

$$\hat{\mathbf{y}}_i = \frac{\exp(y_i)}{\sum_{j=1}^{n_{\text{class}}} \exp(y_j)}, \quad i = 1, \dots, n_{\text{class}},$$

where y_i is the i -th component of \mathbf{y} , and $\hat{\mathbf{y}}_i$ represents the probability of the i -th class. The final output $\hat{\mathbf{y}} \in \mathbb{R}^{n_{\text{class}}}$ is a normalized probability distribution over the target classes, enabling robust multi-class predictions.

Layer (type)	Output Shape	Param #
Branch 1: Face Information		
MTCNN (Face Detection)	[-1, 160, 160, 3]	0
Face-VGG (Last 5 layers trainable)	[-1, 512]	23,560,896
Linear (Face-VGG to 256)	[-1, 256]	131,328
BatchNorm1d (Face-VGG normalization)	[-1, 256]	512
Dropout (Face-VGG, 0.3)	[-1, 256]	0
Chunked VGG Features (8 chunks of 32)	[-1, 8, 32]	0
Linear (person + party to 32)	[-1, 1, 32]	384
Chunked Attention (Chunked VGG attention with person + party)	[-1, 256]	65,792
Branch 2: Image Contextual Information		
ResNet101 (Last 10 layers trainable)	[-1, 2048]	44,549,160
Flatten (ResNet feature vector)	[-1, 2048]	0
Linear (ResNet to 512)	[-1, 512]	1,049,088
BatchNorm1d (ResNet normalization)	[-1, 512]	1,024
Dropout (ResNet, 0.6)	[-1, 512]	0
Branch 3: Structured Metadata		
Embedding (person: 8, party: 2, date: 4)	[-1, 14]	0
Latent Dirichlet Allocation (LDA topics: 40)	[-1, 40]	0
Concatenation (person, party, date, topics)	[-1, 54]	0
Linear (Structured data to 64)	[-1, 64]	3,520
MultiHeadAttention (Structured data attention)	[-1, 64]	16,512
BatchNorm1d (Structured data normalization)	[-1, 64]	128
Dropout (Structured data, 0.4)	[-1, 64]	0
Final: Concatenation		
Concatenation (VGG + Structured + ResNet)	[-1, 832]	0
Dropout (Final dropout, 0.5)	[-1, 832]	0
Linear (Final classification layer)	[-1, num_classes]	8,320
Total params		69,398,560
Trainable params		10,580,820
Non-trainable params		58,817,740

Table A3: Architecture of the Multi-Modal Machine Learning Model

E.2.6 Model Training and Fine-Tuning:

To ensure robust model evaluation, the dataset is split into training (85%) and validation (15%) sets. The splitting process preserves the distribution of key features by stratifying based on both news center and date, ensuring balanced records across all classes. The training set is used to optimize the model parameters, while the validation set assesses model performance and tunes hyperparameters to prevent overfitting ([Goodfellow et al., 2016](#)).

The training process employs the AdamW optimizer ([Loshchilov and Hutter, 2017](#)), configured with a learning rate $\eta = 0.0001$ and a weight decay coefficient $\lambda = 0.01$. This optimizer is particularly well-suited for large-scale data due to its adaptive learning rates and decoupled weight decay, effectively balancing convergence speed and regularization. The loss function used for training is a weighted cross-entropy loss, adjusted to account for class imbalance and enhanced with label smoothing. Let $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$ represent

a batch of N training samples, where $\mathbf{y}_i \in \{0, 1\}^C$ is the one-hot encoded true label for the i -th sample, and $C = 20$ denotes the number of classes. The predicted class probabilities are given by $\hat{\mathbf{y}}_i = \text{softmax}(\mathbf{z}_i)$, where $\mathbf{z}_i \in \mathbb{R}^C$ is the logit vector for the i -th sample. The weighted cross-entropy loss with label smoothing is defined as:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^C w_c \tilde{y}_{i,c} \log \hat{y}_{i,c},$$

where w_c is the weight for class c , computed as the inverse of its relative frequency in the training set to mitigate the impact of class imbalance:

$$w_c = \frac{1}{\text{freq}(c)} \cdot \frac{\sum_{j=1}^C \text{freq}(j)}{C}.$$

The smoothed label $\tilde{y}_{i,c}$ for class c is defined as:

$$\tilde{y}_{i,c} = (1 - \alpha)y_{i,c} + \frac{\alpha}{C},$$

where $\alpha = 0.05$ is the label smoothing coefficient. Label smoothing redistributes a small fraction of the ground-truth probability mass to all classes, reducing overconfidence in predictions and improving generalization (?). The optimizer updates the model parameters θ at each iteration by minimizing the total loss \mathcal{L} . The parameter updates follow the AdamW rule:

$$\theta_{t+1} = \theta_t - \eta \frac{\hat{\mathbf{m}}_t}{\sqrt{\hat{\mathbf{v}}_t} + \epsilon} - \lambda \theta_t,$$

where $\hat{\mathbf{m}}_t$ and $\hat{\mathbf{v}}_t$ are the bias-corrected first and second moments of the gradient, respectively, and ϵ is a small constant to ensure numerical stability. The algorithm for parameter updates is formally presented below:

Algorithm 3 AdamW Optimizer

1: **Input:** Learning rate η , decay rates β_1, β_2 , weight decay coefficient λ , small constant ϵ 2: **Initialize:** $\mathbf{m}_0 = 0, \mathbf{v}_0 = 0, t = 0$ 3: **for** each iteration **do**4: $t = t + 1$ 5: Compute gradients of the loss: $\mathbf{g}_t = \nabla_{\theta} \mathcal{L}(\theta_{t-1})$

6: Update biased first moment estimate:

$$\mathbf{m}_t = \beta_1 \mathbf{m}_{t-1} + (1 - \beta_1) \mathbf{g}_t$$

7: Update biased second moment estimate:

$$\mathbf{v}_t = \beta_2 \mathbf{v}_{t-1} + (1 - \beta_2) \mathbf{g}_t^{\odot 2}$$

8: Compute bias-corrected first moment estimate:

$$\hat{\mathbf{m}}_t = \frac{\mathbf{m}_t}{1 - \beta_1^t}$$

9: Compute bias-corrected second moment estimate:

$$\hat{\mathbf{v}}_t = \frac{\mathbf{v}_t}{1 - \beta_2^t}$$

10: Update parameters:

$$\theta_t = \theta_{t-1} - \eta \frac{\hat{\mathbf{m}}_t}{\sqrt{\hat{\mathbf{v}}_t + \epsilon}} - \lambda \theta_{t-1}$$

11: **end for**

The optimization process continues over 30 epochs, where each epoch consists of processing mini-batches of data of size 64. At each iteration, the model updates its parameters using the computed gradients and the AdamW update rule. The weight decay term $\lambda \theta_{t-1}$ ensures regularization by penalizing large weights, helping to improve generalization. Metrics such as validation loss and classification accuracy are monitored at the end of each epoch to assess the model’s performance and ensure generalization to unseen data.

F Appendix: Supplementary Results

F.1

Table A4 provides information on the multi-modal model performance at each class level (news outlet) and presents the validation performance of the model across various news sources, sorted by accuracy. The model performs best with *Daily Mail* and shows strong, consistent results across other sources, such as *Newsmax* and *CNN*. The model effectively addresses class imbalances by employing a weighted cross-entropy loss, ensuring fair representation for both frequent and rare classes. This is reflected in the balanced

precision, recall, and F1 scores across sources with different sample sizes. The weighted loss function helps prevent overfitting to dominant classes, resulting in reliable performance across various news outlets.

Source	Accuracy	Precision	Recall	F1-Score	No. of observations
Daily Mail	0.654	0.613	0.654	0.633	987
Newsmax	0.609	0.511	0.609	0.556	241
LA Times	0.576	0.629	0.576	0.601	151
Fox News	0.493	0.509	0.493	0.501	577
ABC News	0.489	0.314	0.489	0.381	703
CNN	0.470	0.417	0.470	0.441	832
CNBC	0.444	0.373	0.444	0.405	410
New York Times	0.440	0.380	0.440	0.408	426
Wall Street Journal	0.433	0.454	0.433	0.444	593
CS Monitor	0.429	0.668	0.429	0.522	203
Time	0.416	0.418	0.416	0.417	369
BBC	0.403	0.636	0.403	0.492	554
NBC News	0.394	0.448	0.394	0.419	937
Washington Post	0.373	0.316	0.373	0.342	567
CBS News	0.297	0.463	0.297	0.360	517
Reuters	0.295	0.382	0.295	0.332	233
HuffPost	0.285	0.316	0.285	0.300	321
US News	0.249	0.411	0.249	0.309	109
USA Today	0.234	0.303	0.234	0.264	425
Yahoo News	0.196	0.202	0.196	0.199	296

Table A4: Validation Results for Different News Sources (Sorted by Accuracy)

F.2 Details of PCA and t-SNE for Reducing Image Dimensionality to Two

PCA is used to reduce the dimensionality of the embeddings from d to $d' \ll d$ by projecting them onto a lower-dimensional subspace that retains most of the variance in the data. Mathematically, PCA identifies a set of orthogonal components $\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_{d'} \in \mathbb{R}^d$, where each component maximizes the variance of the projected data. The reduced embedding for an image is given by:

$$\mathbf{e}_i^{\text{PCA}} = \mathbf{W}^\top \mathbf{e}_i, \quad \mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_{d'}],$$

where \mathbf{W} is the matrix of the top d' eigenvectors of the covariance matrix of the embeddings, sorted by their corresponding eigenvalues. This transformation ensures that the majority of the information in the original embeddings is preserved in the reduced representation, making subsequent computations more efficient.

After reducing the embeddings to d' dimensions using PCA, we project them into a two-dimensional space, $\mathbf{e}_i^{\text{t-SNE}} \in \mathbb{R}^2$, using t-SNE for visualization. t-SNE, or t-Distributed Stochastic Neighbor Embedding, is a dimensionality reduction technique designed specifically for high-dimensional data visualization (Belkina et al., 2019). It aims to preserve the local structure of the data by modeling the similarity between points i and j in the original high-dimensional space as probabilities $P = \{p_{ij}\}$, and then finding a 2D embedding where the similarities, represented by $Q = \{q_{ij}\}$, approximate P . In this 2D space, the similarity between

points i and j is modeled using a Student-t distribution, where the pairwise similarity is defined as:

$$q_{ij} = \frac{\left(1 + \|\mathbf{e}_i^{\text{t-SNE}} - \mathbf{e}_j^{\text{t-SNE}}\|^2\right)^{-1}}{\sum_{k \neq l} \left(1 + \|\mathbf{e}_k^{\text{t-SNE}} - \mathbf{e}_l^{\text{t-SNE}}\|^2\right)^{-1}},$$

where $\mathbf{e}_i^{\text{t-SNE}}$ and $\mathbf{e}_j^{\text{t-SNE}}$ represent the 2D coordinates of points i and j , $\|\mathbf{e}_i^{\text{t-SNE}} - \mathbf{e}_j^{\text{t-SNE}}\|$ is the Euclidean distance between points i and j in the 2D space, and k and l iterate over all points in the dataset to compute the denominator of the similarity measure for proper normalization. t-SNE attempts to arrange the 2D points such that their similarities, denoted as $Q = \{q_{ij}\}$, approximate the similarities in the high-dimensional space, denoted as $P = \{p_{ij}\}$. Here, $P = \{p_{ij}\}$ represents the pairwise similarities computed from the PCA-reduced high-dimensional embeddings, and $Q = \{q_{ij}\}$ represents the pairwise similarities in the 2D space computed using the formula above. The arrangement of points in 2D is achieved by minimizing the Kullback-Leibler (KL) divergence:

$$\text{KL}(P \| Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}},$$

where p_{ij} and q_{ij} denote the similarity between points i and j in the high-dimensional and 2D spaces, respectively. By minimizing this divergence, t-SNE ensures that points with high similarity in the high-dimensional space (large p_{ij}) remain close in the 2D space (large q_{ij}), while dissimilar points (small p_{ij}) are positioned farther apart. The resulting 2D embeddings, $\mathbf{e}_i^{\text{t-SNE}}$, enable visualization of images' space,

To further analyze the patterns in the embedding space for a specific politician p , we perform clustering on the reduced embeddings associated with their images. For each politician p , let the set of embeddings corresponding to their images be denoted as $\mathcal{E}^p = \{\mathbf{e}_i^{\text{PCA}} \mid i \in \mathcal{I}_p\}$, where \mathcal{I}_p is the set of image indices for politician p . These embeddings are grouped into $k = 20$ clusters using K-Means clustering. The clustering objective for politician p is to minimize the within-cluster sum of squared distances:

$$\mathcal{C}_i^p = \arg \min_{\mathcal{C}} \sum_{j=1}^k \sum_{\mathbf{e}_i^{\text{PCA}} \in \mathcal{C}_j^p} \|\mathbf{e}_i^{\text{PCA}} - \mu_j^p\|^2,$$

where \mathcal{C}_j^p is the set of embeddings assigned to cluster j for politician p , and μ_j^p is the centroid of cluster j for p . Intuitively, each cluster represents a group of visually similar images, which may correspond to specific events or contextual features, such as similar camera angles and settings. For instance, images of politician p during a particular speech or event are likely to form a distinct cluster, capturing the shared visual context of those images. Now, for a politician p and a given cluster k , we can investigate how different news outlets select images.

F.3 Hypothesis Tests for the Distribution of Visual Polarization

We employ two statistical tests to analyze these differences statistically: the Kolmogorov-Smirnov (K-S) test and one-sample t-tests. The K-S test determines if two samples come from the same distribution. The null hypothesis H_0 states that the two samples are drawn from the same distribution. On the other hand, the one-sample t-test determines if the mean of a sample is different from a known value (zero in this case). The null hypothesis H_0 states that the sample mean is equal to the known value. A low p-value indicates that we

can reject the null hypothesis. The results of these statistical tests are summarized in Table A5.

Politicians	Test	Statistic	P-Value	n
Democratic Politicians	K-S Test	0.526790	0.0000	31728
	Democratic Outlets Mean Test	96.016984	0.0000	15864
	Opposing Outlets Mean Test	-82.654572	0.0000	15864
Republican Politicians	K-S Test	0.672592	0.0000	24978
	Republican Outlets Mean Test	-73.504981	0.0000	12489
	Opposing Outlets Mean Test	114.545837	0.0000	12489

Table A5: Summary of Statistical Test Results

The results presented in Table A5 provide strong statistical evidence of media polarization in the portrayal of smiling images for both Democratic and Republican politicians. For Democratic politicians, the statistical tests confirm that smiling images significantly enhance alignment with Democratic-leaning outlets, while leading to a marked decrease in perceived utility in Republican-leaning outlets. These findings reflect a clear partisan divide, as shown by the significant test results.

For Republican politicians, the table demonstrates an opposing trend. Smiling images are positively associated with Republican-leaning outlets, reinforcing alignment with their narratives, whereas they are perceived negatively in Democratic-leaning outlets. The statistical significance across all tests underscores the robustness of these patterns. Table A5 highlights how visual elements, such as smiles, are interpreted differently depending on the political context and outlet alignment, illustrating the media's role in amplifying partisan biases.

F4 Regression Analysis for Existing Partisanship Scores

To further validate these findings, we conduct additional regression analyses using two other sources of conservative share scores, summarized in Tables A6 and A7. For Democratic politicians, the results consistently show negative slopes across all sources, highlighting that outlets with stronger Republican leanings exhibit lower—and even negative—utility for smiling images of Democratic politicians. For Republican politicians, the slopes remain positive, confirming that Republican-leaning outlets strongly favor smiling images of Republican politicians. Figures A.4 and A.5 visualize these relationships.

Table A6: Regression Analysis Results for Democratic Politicians

Source	Democratic Slope	Democratic SE	p-value	Support (n)
Faris et al. (2017)	-0.468304	0.108549	0.000418	20
AllSides (2024)	-0.093878	0.020359	0.000289	18
Flaxman et al. (2016)	-2.241561	0.734285	0.006852	20

Table A7: Regression Analysis Results for Republican Politicians

Source	Republican Slope	Republican SE	p-value	Support (n)
Faris et al. (2017)	0.484955	0.123893	0.001017	20
AllSides (2024)	0.090209	0.024491	0.002012	18
Flaxman et al. (2016)	1.677110	0.902191	0.079466	20

This analysis demonstrates that smiling images amplify partisan biases across media outlets. By incorporating multiple sources of conservative share scores, we validate the robustness of our findings and align them closely with existing literature on media polarization.

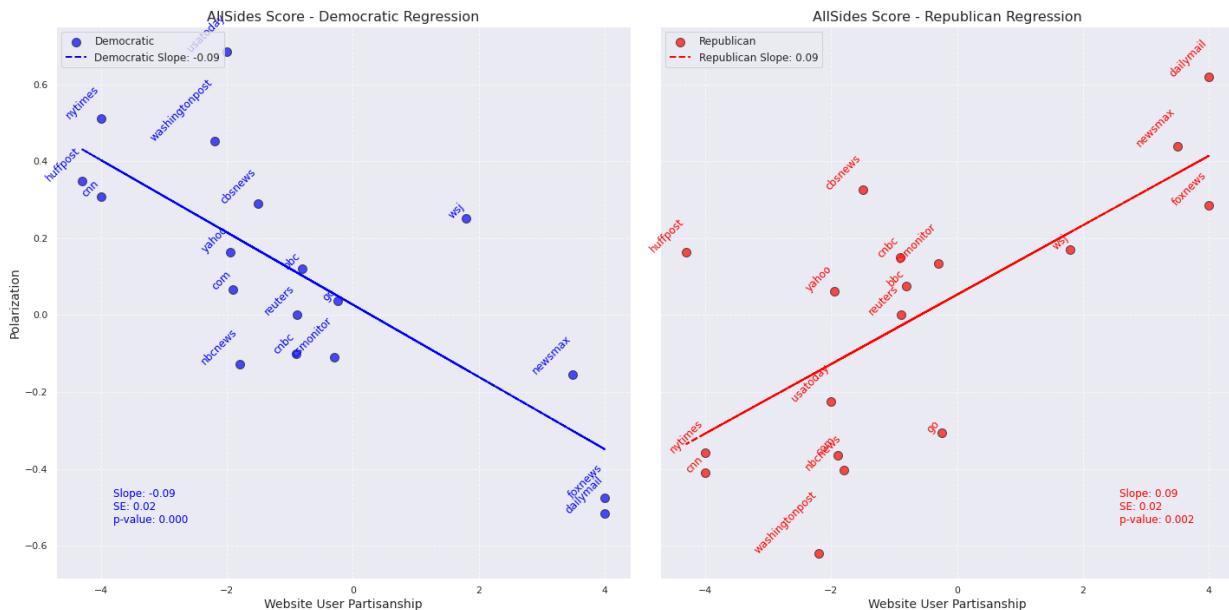


Figure A.4: Regression Plot for Democratic and Republican Politicians Based on [AllSides \(2024\)](#) Conservative Score

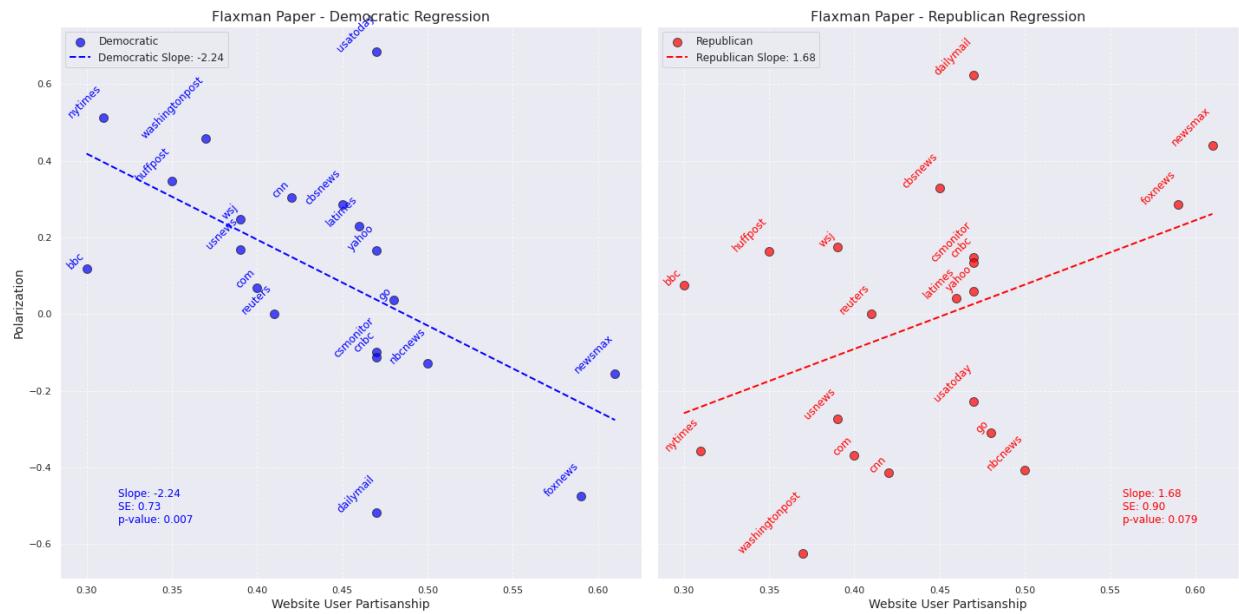
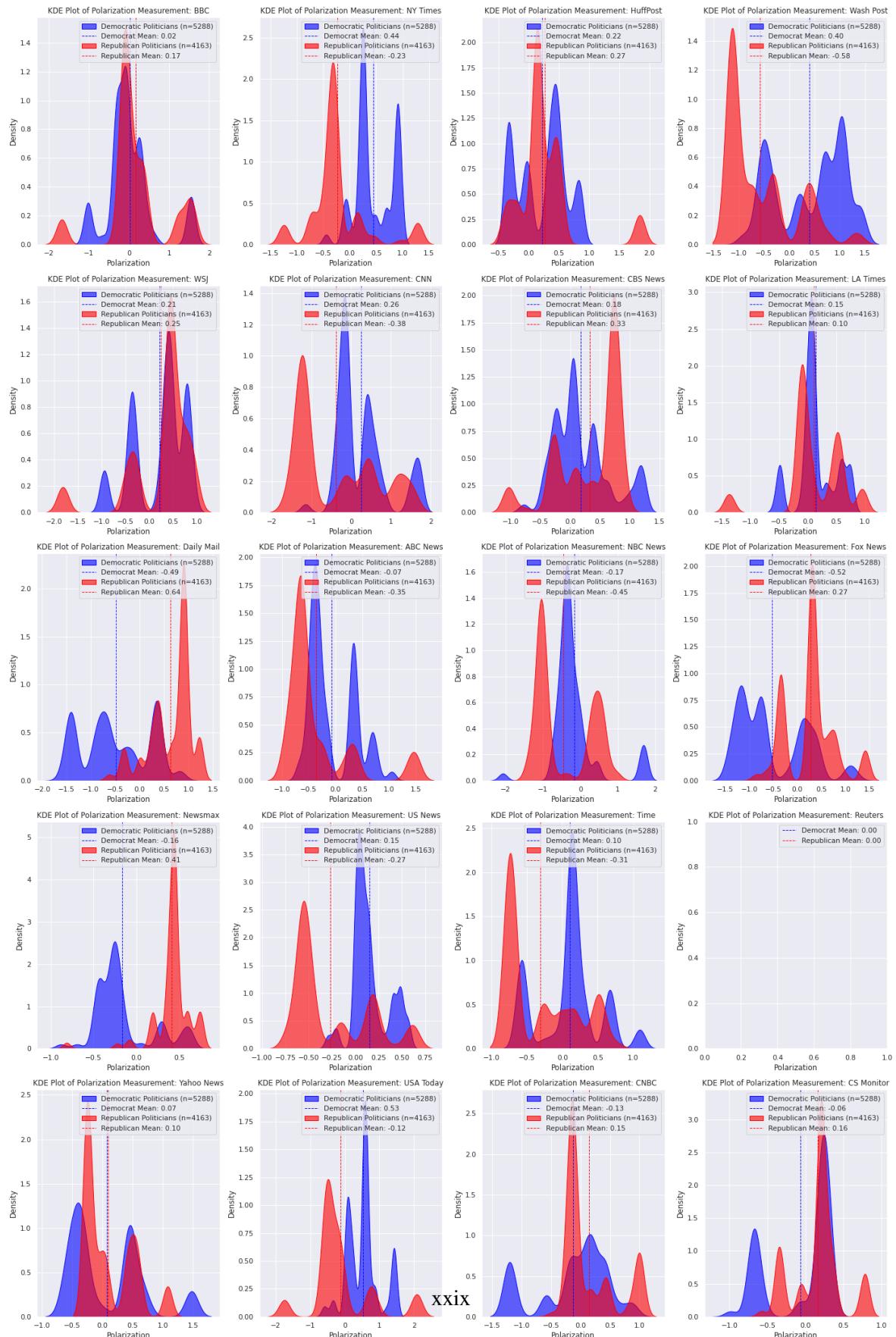


Figure A.5: Regression Plot for Democratic and Republican Politicians Based on Flaxman et al. (2016) Conservative Score

F.5 Individual News Outlet Results



News Center	Pol. Side	Mean	Q1	Q3	T-Test Stat	T-Test P-Val	Significant?	Support (n)
BBC	Dem	0.02	-0.32	0.63	2.89	3.90×10^{-3}	True	5288
BBC	Rep	0.17	-0.11	1.53	14.87	8.47×10^{-49}	True	4163
NY Times	Dem	0.44	-0.01	0.93	90.19	0.00	True	5288
NY Times	Rep	-0.23	-0.71	0.44	-27.63	1.97×10^{-154}	True	4163
HuffPost	Dem	0.22	-0.34	0.77	41.33	0.00	True	5288
HuffPost	Rep	0.27	-0.19	0.49	36.33	2.87×10^{-251}	True	4163
Wash Post	Dem	0.40	-0.52	1.05	42.34	0.00	True	5288
Wash Post	Rep	-0.58	-1.13	0.49	-55.35	0.00	True	4163
WSJ	Dem	0.21	-0.42	0.78	28.15	1.49×10^{-162}	True	5288
WSJ	Rep	0.25	-0.48	0.77	24.34	1.93×10^{-122}	True	4163
CNN	Dem	0.26	-0.36	1.51	29.14	2.81×10^{-173}	True	5288
CNN	Rep	-0.38	-1.22	1.15	-24.19	4.70×10^{-121}	True	4163
CBS News	Dem	0.18	-0.25	1.01	28.14	1.80×10^{-162}	True	5288
CBS News	Rep	0.33	-0.29	0.73	39.12	2.11×10^{-285}	True	4163
LA Times	Dem	0.15	-0.48	0.63	31.82	1.77×10^{-203}	True	5288
LA Times	Rep	0.10	-0.10	0.56	13.27	2.21×10^{-39}	True	4163
Daily Mail	Dem	-0.49	-1.41	0.36	-52.10	0.00	True	5288
Daily Mail	Rep	0.64	0.00	0.90	94.11	0.00	True	4163
ABC News	Dem	-0.07	-0.46	0.66	-11.31	2.40×10^{-29}	True	5288
ABC News	Rep	-0.35	-0.81	0.34	-36.04	8.04×10^{-248}	True	4163
NBC News	Dem	-0.17	-0.67	0.44	-19.09	1.37×10^{-78}	True	5288
NBC News	Rep	-0.45	-1.05	0.61	-38.75	1.06×10^{-280}	True	4163
Fox News	Dem	-0.52	-1.26	0.37	-53.96	0.00	True	5288
Fox News	Rep	0.27	-0.35	0.78	35.00	1.74×10^{-235}	True	4163
Newsmax	Dem	-0.16	-0.45	0.51	-35.51	5.97×10^{-248}	True	5288
Newsmax	Rep	0.41	0.19	0.69	115.30	0.00	True	4163
US News	Dem	0.15	0.02	0.48	58.18	0.00	True	5288
US News	Rep	-0.27	-0.56	0.20	-42.57	0.00	True	4163
Time	Dem	0.10	-0.57	0.66	17.50	1.00×10^{-66}	True	5288
Time	Rep	-0.31	-0.74	0.49	-40.21	5.77×10^{-299}	True	4163
Yahoo News	Dem	0.07	-0.48	0.73	8.95	4.75×10^{-19}	True	5288
Yahoo News	Rep	0.10	-0.24	0.58	15.71	4.78×10^{-54}	True	4163
USA Today	Dem	0.53	0.07	1.42	82.28	0.00	True	5288
USA Today	Rep	-0.12	-0.51	0.93	-9.03	2.64×10^{-19}	True	4163
CNBC	Dem	-0.13	-1.21	0.42	-15.40	2.13×10^{-52}	True	5288
CNBC	Rep	0.15	-0.23	0.99	21.14	2.73×10^{-94}	True	4163
CS Monitor	Dem	-0.06	-0.67	0.27	-10.63	4.03×10^{-26}	True	5288
CS Monitor	Rep	0.16	-0.34	0.73	33.06	4.96×10^{-213}	True	4163

Table A8: Statistical Analysis of Polarization by News Center and Politician Side
 XXX

F.6 Individual Politicians Results



Figure A.7: Bar plot showcasing the polarization across various Politicians, segmented by different News Outlets, Part I

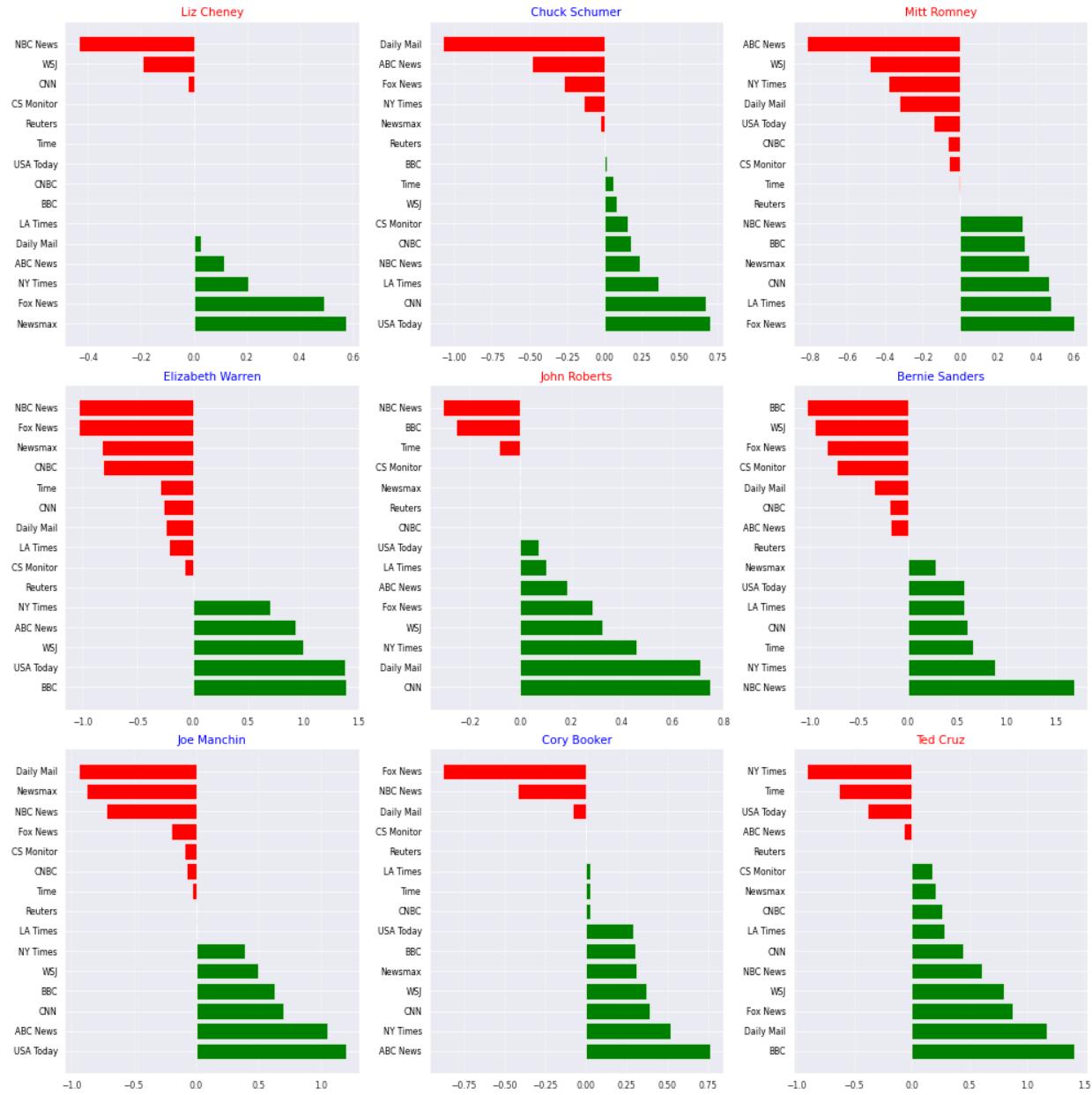


Figure A.8: Bar plot showcasing the polarization across various Politicians, segmented by different News Outlets, Part II

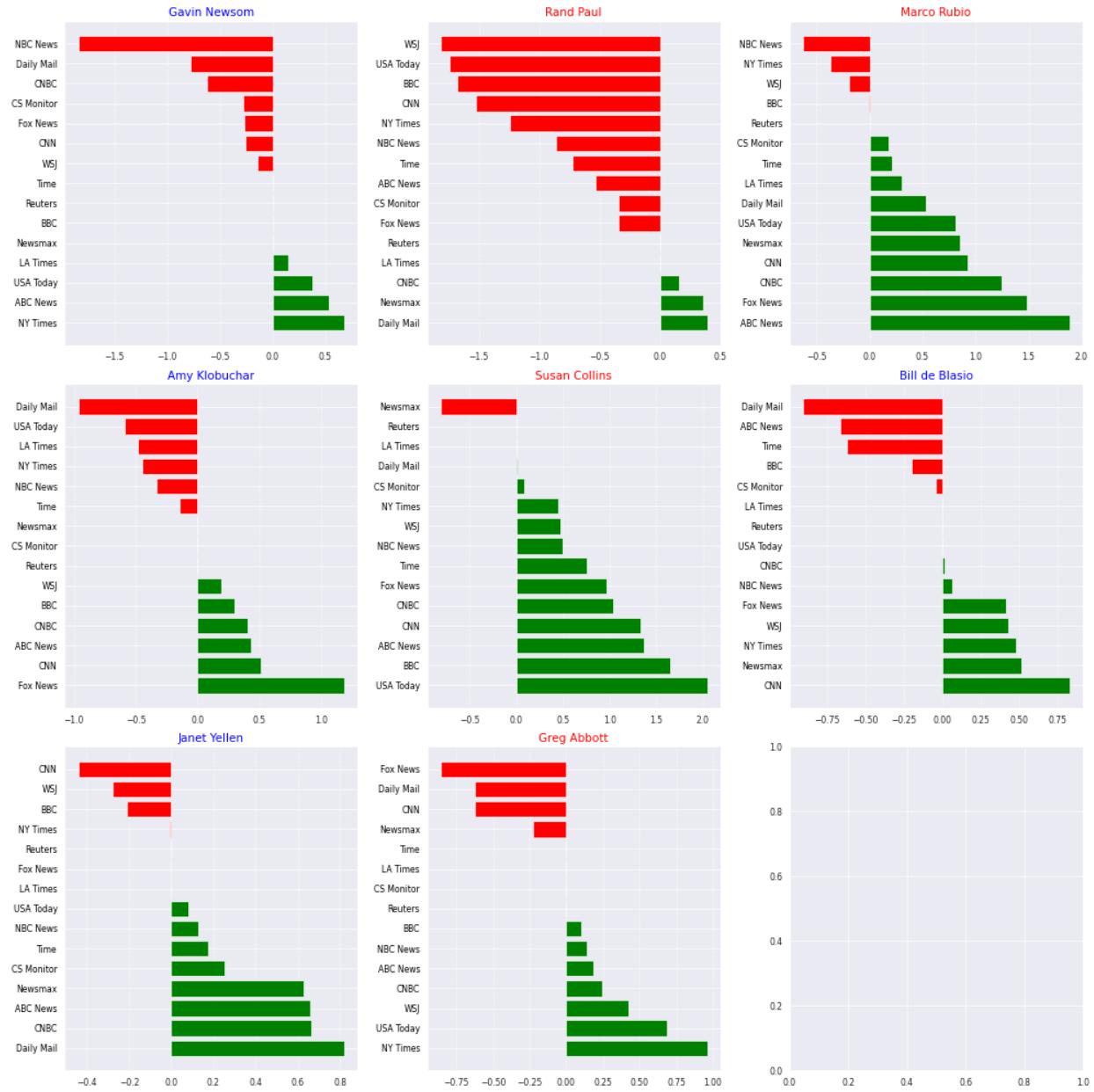


Figure A.9: Bar plot showcasing the polarization across various Politicians, segmented by different News Outlets, Part III