Visual Polarization Measurement Using Counterfactual Image Generation

Mohammad Mosaffa* Cornell University
Omid Rafieian Cornell University
Hema Yoganarasimhan University of Washington

February 3, 2025

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 (PMCIG) 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.

Keywords: Polarization, News Media, Politics, Generative Models, Computer Vision, Counterfactual Reasoning

*We would like to thank the participants of the 2024 PhD student workshop at Cornell University and the 2024 MarkTech Conference for their feedback. We also thank Sachin Gupta and Simha Mummalaneni for his detailed comments, which have significantly improved the paper. Please address all correspondence to: mm3322@cornell.edu, or83@cornell.edu, and hemay@uw.edu.

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 (Gentzkow and Shapiro, 2006, 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). Visuals not only shape engagement but also play a key role in spreading misinformation (Matatov et al., 2022) and influencing voter perceptions of competence and trustworthiness in politics (Hoegg and Lewis, 2011). Moreover, younger generations of consumers are showing a growing preference for news content that relies less on verbal and more on visual elements. 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 measure polarization and slant 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 slant and polarization in media?

3. How does the polarization in visual content vary across politicians and news outlets?

There are three key challenges that we need to address to answer these questions satisfactorily. First, we need a formal definition of a metric or parameter of interest that captures polarization in visual content. To address this challenge, we turn to the very structure of the problem: the editor's choice of visual content. An

ideologically slanted choice of image intuitively means that a news outlet prefers a more positive (negative) portrayal of a politician from the same (opposite) ideological side, compared to a neutral news outlet. We construct a flexible utility framework that models the editor's choice of images from a set of available options and focus on the smile as a focal feature we use to measure visual polarization given the consistent finding that a smile leads to a more positive portrayal of a politician (Sülflow and Maurer, 2019). Using a utility framework allows us to compare the utility derived from two images that are on all aspects except the focal feature (e.g., smile) on which the visual polarization is measured. This comparison enables us to build a polarization measure centered around the focal feature of interest. 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. To reflect this notion of variability for each pair of news outlets, we define the difference in this utility difference as the visual polarization measure. The farther the two outlets are in their utility difference, the higher the visual polarization measure.

Our second challenge stems from the high dimensionality of the visual content. The common approach in the literature is to use an off-the-shelf machine learning model 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 both *omitted variable bias* and *extraction bias* in the analysis (Wei and Malik, 2022). 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 lies in identifying our polarization parameter from the observed data, which is defined based on the utility functions of news outlets. This challenge arises because these utility functions are not directly identified from the data as we only observe the image selected by each outlet, not the full choice set available to them. To address this issue, we leverage the variation in image choices across outlets for similar events (e.g., a specific press hearing), under the reasonable assumption that the choice set is nearly identical for all outlets covering the same event. For instance, both Fox News and CNN likely source images from common providers such as Getty Images and the Associated Press (AP), giving them access to similar visual content. This allows us to assess whether one outlet derives greater utility from a positive portrayal of a politician compared to the other. We then theoretically link the identification of the polarization parameter to a news outlet prediction problem and develop a multi-modal deep learning model for this task. The model is designed to capture both the clustering structure in similar events and the subtle facial features that may reflect outlets' ideological preferences (if any). Finally, the estimates derived from this prediction task allow us to measure polarization at any desired level of granularity.

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 lies in its ability to overcome the potential confounding and misestimation of polarization due to information loss

from feature extraction. 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 60,000 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 multi-modal deep learning model 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, using Reuters as the base neutral news outlet, we find that compared to a neutral image of a Republican politician, a smiling image increases utility for a Republican-leaning news outlet and decreases utility for a Democratic-leaning outlet. 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 and direction of visual slanting exhibits substantial variation among news outlets, suggesting a high degree of visual polarization. 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 compare our outlet-level measures of visual slant with existing measures of ideological slanting established in the literature (Flaxman et al., 2016; Faris et al., 2017) and find strong correlation between our visual slant measure and the existing ones, providing validation for our measurement.

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, Bernie Sanders and Kamala Harris to be among the most polarizing figures who are very differently portrayed in Democratic-leaning and Republican-leaning outlets. Interestingly, Joe Manchin (D) and Susan Collins (R) rank among the least polarizing politicians, reflecting their reputations as moderates within their respective parties. Moreover, we observe low polarization levels for Liz Cheney, whose fallout with Donald Trump resulted in increased favorability among liberal outlets and decreased favorability among conservative outlets, ultimately contributing to reduced levels of visual polarization.

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 all settings where researchers seek to quantify polarization on a given feature across a given set of images and outlets. Substantively, our work demonstrates the existence of ideological slanting in the visual content used by media outlets, with substantial heterogeneity observed across news outlets and politicians.

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) quantifies 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 long term 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. Xu et al. (2024) propose a debiased embedding framework that integrates representation learning with causal inference, addressing biases inherent in traditional embedding-then-inference frameworks. 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 news outlets used in our study consist of the list of popular outlets that have been used in earlier studies on media polarization; see Flaxman et al. (2016) for details. We refer readers to Web Appendix A.1 for a full list of politicians and outlets.

The data collection is done using the SerpAPI application (SerpAPI, 2023). 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.[1] This approach was chosen because it provides a structured and reliable interface for large-scale data collection, automating the process while ensuring compliance with Google's data access policies and improving the accuracy and efficiency of the data collection pipeline.

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.[2] This information for each image in the query is compiled into a structured data frame with the following fields:

[1] For each politician-outlet combination, we use a two-year rolling window that advances by one year at a time, resulting in 10 queries spanning a total of 10 years.

[2] Some queries retrieve fewer articles/images for certain politicians-outlet combinations because the outlet may not have published 80 images for that specific politician.

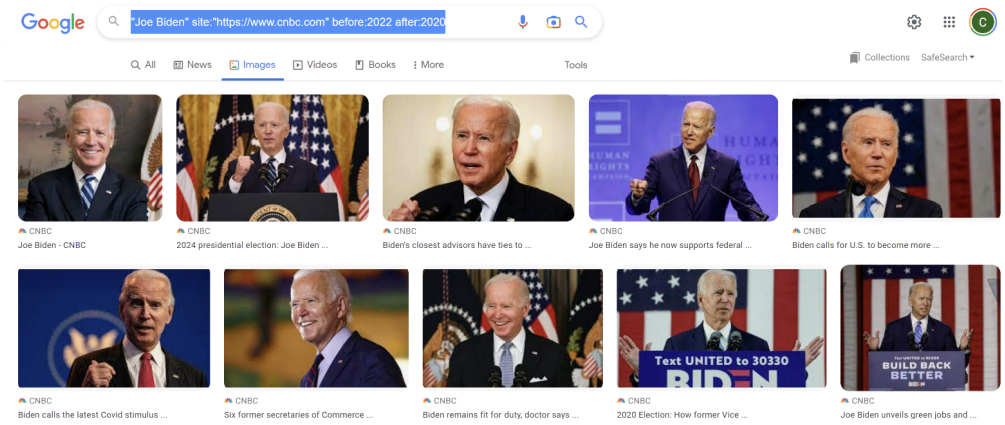


Figure 1: Search Results for "Joe Biden" site: "https://www.cnbc.com" before:2022 after:2020

- 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 article can be found.
- Title: The title of the news article or caption associated with the image.
- Query Parameter: The 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 images for 30 politicians from 20 news outlets.

3.2 Data Cleaning and Identifying Politicians

A crucial step is cleaning the data to ensure that each image contains the face of the politician referred to in the query. We face the following challenges in the data-cleaning step:

- Some 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 analysis of each politician's records.

To address these challenges, we design a two-phase computer vision framework. We provide a brief overview of this two-step framework here and refer readers to Web Appendix A.2 for the technical details. In the first phase, we use a series of computer vision models to keep only images with one face presented in them. In the second phase, we build and deploy a face-verification tool to ensure that the face in an image belongs to the intended politician. First, we first manually select 20 high-quality, single-face images for each of the 30 politicians. These selected images serve as true labels identifying the correct politician, ensuring a reliable foundation for training the face verification model. Next, we use the trained verification 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. After applying the above data cleaning procedure, we are left with a set of 63,188 images, where each image shows a single face that belongs to the intended

politician. Web Appendix A.2 provides further details on our two-step model and includes a comprehensive table (Figure A.3) listing the number of images for each politician-outlet combination after the cleaning.

4 Problem Definition

Recall that our objective is to see if and how images of politicians in media outlets can be used to measure political polarization. As such, our goal here is to develop measures of *visual slant* and *visual polarization* that can be estimated from data on news articles (with images of politicians).

Consider a data set $\mathscr{D}$ 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 text, topic, 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. 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.

Next, we define the image choice problem for a given article i from the perspective of a news editor. Prior research has shown that 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, which 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 (Jakesch et al., 2022). Figure 2 shows an example news article, highlighting the decision stage where an editor picks an accompanying image for the given article.

Figure 2: Example of an editor (CNN) selecting an image of Joe Biden from a set of two images, for a given news article.

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

$$U_i = u(Z_i, X_i, P_i, Y_i) + \xi_i, \tag{1}$$

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 ξ_i denotes the idiosyncratic error term, which follows an i.i.d. Type 1 Extreme Value distribution.[3] This utility function captures all the main considerations of the editor when making image

[3]We model the news outlet's decision-making using a utility function rather than a conventional profit function, as ideological preferences may drive them to deviate from profit-maximizing choices. For other theoretical models that micro-found agents'

choices, 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 compared to a neutral outlet. The key challenge in developing a measure of ideological slant in visual content stems 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 (Sülflow 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, we start with a utility difference measure that compares the utility from two articles that are identical in all respects, except for the smile feature. Similar to the causal inference literature, we therefore consider two versions of an image Z: $Z(T=1)$ 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 (with a neutral expression). For ease of exposition, let $Z^{(-T)}$ denote all the information in the 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 versions of the image are identical in all features other than smile. Using this set of two 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:

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

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, this is still not a measure of ideological slant in visual content because we cannot fully attribute this utility difference to ideological preferences. For instance, there can be vertical preferences for the smile feature where all news outlets prefer a smiling image of a politician. As such, the difference in the utility difference $\Delta^T u(p,y)$ across two news outlets helps cancel out the vertical preference for the feature and the remaining difference can be linked to ideological preferences. For example, we expect a news outlet with a 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) even if they both have a vertical preference for smiling images. Since this difference is defined for any two news outlets, it measures *visual polarization* of the two outlets with respect to the feature of interest. However, it is important to note that this is not a measure of *visual slant*, because both outlets may have slanted preferences. In what follows, we present formal definitions for both *visual polarization* and *visual slant*. We first present the formal definition for *visual polarization* as follows:

decision-making to examine polarization in equilibrium, see Gentzkow and Shapiro (2010), Iyer and Yoganarasimhan (2021), and Bondi et al. (2023).

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

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

This definition of *visual polarization*, $\rho^T(p, y_1, y_2)$, is flexible and allows for detailed analysis at the level of individual politicians and news outlets. For example, by examining the above measure for a specific politician like Donald Trump across specific outlets such as CNN and Fox News, we can measure the extent to which these outlets are polarized or differentiated in their visual portrayal of Trump. Naturally, for the measure of *visual polarization* to capture *visual slant*, we need the news outlet y_2 to be neutral. Let y_n denote the neutral news outlet. We can formally define *visual slant* as follows:

Definition 2. *For a given politician p and any news outlet y_1,* **visual slant** *with respect to feature T is denoted by ρ_s^T and defined as follows:*

$$\rho_s^T(p, y_1) = \Delta^T u(p, y_1) - \Delta^T u(p, y_n), \tag{4}$$

where y_n is the neutral news outlet.

To better understand the differences between *visual polarization* and *visual slant*, we return to the example with the portrayal of Donald Trump in CNN and Fox News, but add a neutral news outlet like Reuters. Suppose that the *visual polarization* between Fox News and CNN for Donald Trump is equal to one, i.e., $\rho^T(\text{Donald Trump}, \text{Fox News}, \text{CNN}) = 1$. This measure is a composite of both Fox News' preference for a positive portrayal of Trump and CNN's preference for a negative portrayal of him. Our *visual slant* measure helps decompose the *visual polarization* measure. For example, we may find that there is a positive *visual slant* for Donald Trump at Fox News such that $\rho_s^T(\text{Donald Trump}, \text{Fox News}) = 0.4$, but a negative *visual slant* for him at CNN such that $\rho_s^T(\text{Donald Trump}, \text{CNN}) = -0.6$. It is easy to verify the following relationship between the two definitions:

$$\rho^T(p, y_1, y_2) = \rho_s^T(p, y_1) - \rho_s^T(p, y_2). \tag{5}$$

Lastly, a notable feature of our measures is that they are defined at a high degree of granularity, which allows us to capture the varying degrees of ideological slant in visual content specific to each outlet or each politician. This specification allows us to consider different types of aggregation:

- First, by aggregating the *visual slant* measure over Democratic (Republican) politicians for a given outlet, we can obtain insights into how a given outlet portrays liberal (conservative) politicians. This analysis can tell us how conservative outlets like Fox News and liberal outlets like CNN portray the two groups of politicians differently.
- Second, by aggregating a given politician's *visual slant* measure across all the outlets and then comparing these measures across politicians, we can derive insights into how different politicians are portrayed in the media. For instance, this can help us understand questions such as whether the portrayal of Donald Trump in the media is more (or less) polarizing than that of Joe Biden.

In the next few sections, we discuss how we can measure *visual slant* and *visual polarization* using observational data on news articles and accompanying images, and in §7.3, we present the substantive results on *visual slant*.

5 Standard Reduced-Form Approach

In this section, we discuss 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, in §5.2, we provide a theoretical and conceptual discussion of the drawbacks of this approach. Finally, in §5.3, we present empirical evidence from our setting demonstrating the challenges with the two-step approach.

5.1 Two-step Model

Recall that our goal is to measure *visual polarization*, $\rho^T(p, y_1, y_2)$, with respect to a focal feature T, such as whether the politician in the image is smiling or not. However, this is inherently challenging when dealing with unstructured image data, because unlike the standard causal inference literature, where treatment is observed (Imbens and Rubin, 2015), here we do not directly observe T_i, the presence or absence of the treatment (smile in this case) in a given article 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.[4] 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 $f_1 : Z \to T$ is used to derive the feature(s) T:

$$\hat{T} = f_1(Z) \tag{6}$$

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:

$$Y = f_2^{\text{T-independent}}(\hat{T}, \boldsymbol{X}) + \varepsilon, \quad \text{or,} \tag{7a}$$

$$\hat{T} = f_2^{\text{T-dependent}}(Y, \boldsymbol{X}) + \varepsilon, \tag{7b}$$

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

[4] Beyond 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).

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).

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++ (Face++, 2023), Google Vision (Vision, 2023)[5], 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. For notational 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:

$$\log\left(\frac{\Pr(\hat{T}=1 \mid Y, \boldsymbol{X})}{\Pr(\hat{T}=0 \mid Y, \boldsymbol{X})}\right) = \alpha_0 + \alpha_1 \boldsymbol{X} + \beta Y \tag{8}$$

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:

$$\begin{aligned} \rho^T(p, Y=1, Y=0) &= \Delta^T u(p, Y=1) - \Delta^T u(p, Y=0) \\ &= (\alpha_0 + \alpha_1 \boldsymbol{X} + \beta) - (\alpha_0 + \alpha_1 \boldsymbol{X}) \\ &= \beta \end{aligned} \tag{9}$$

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.

[5] Azure (2023)'s Face API, a popular model previously used in studies such as Boxell (2021); Caprini (2023), has been unavailable for use since June 2022 due to updated responsible AI policies (Microsoft, 2022).

5.2 Drawbacks of the Two-Step Approach

While the two-step model described in §5.1 is easy to interpret and apply, it relies on the assumption that the regression model is well-specified. However, this assumption can fail due to two natural reasons and lead to incorrect inference: (1) *extraction bias*, and (2) *omitted variable bias*. In this section, we discuss these biases and explain how they can manifest in our setting.

Extraction bias occurs due to imperfections in the feature extraction process, or the first step of the two-step process (Wei and Malik, 2022). That is, the true label T may differ from the extracted label $\hat{T} = f_1(Z)$ as follows:

$$T = f_1(Z) + \varepsilon_1 = \hat{T} + \varepsilon_1 \tag{10}$$

The error in Equation (10) can be viewed as a measurement error. It is well-known that systematic dependence of this error on other relevant factors in the model can introduce biases or inconsistencies in the estimates; see Chapter 4 of Wooldridge (2010). 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 expressions (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 can rewrite the relationship between T and $\hat{T}$ as:

$$T = \hat{T} + \varepsilon_r + \varepsilon_e. \tag{11}$$

Then, the predicted feature $\hat{T}$ is a combination of the true feature of interest T (e.g., smile) and a correlated nuisance feature ε_e (e.g., brightness). Thus, when we perform the second stage estimation, the estimated effect (β) captures the effect of the outlet on the biased measurement $\hat{T}$ rather than the true feature T.

Next, *omitted variable bias* arises from relevant factors omitted from the second step. This issue can arise even if we have the true labels T. To illustrate this point, consider the second-step relationship between the true label and covariates:

$$T = f_2(Y, X) + \varepsilon_2, \tag{12}$$

where ε_2 consists of unobserved variables that are not captured 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 correctly. 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 likely that the US flag is present more often in images from 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 (12) will lead to function f_2 picking up the association between the presence of the flag and the smile in the image because the presence of the 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 (12) as follows:

$$T = f_2(Y,X) + \varepsilon_0 + \varepsilon_{ov}, \tag{13}$$

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. It is worth noting that image features that lead to extraction bias can also lead to omitted variable bias. For example, omitting image features like brightness can lead to omitted variable bias if brightness is correlated with both the actual presence of a smile in the image (not the extracted one) and the news outlet.

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 (11) also appear in the second step equation as follows:

$$\hat{T} = f_2(Y,X) + \varepsilon_0 + \varepsilon_{ov} + \varepsilon_r + \varepsilon_e, \tag{14}$$

where ε_{ov} leads to omitted variable bias, ε_e leads to extraction bias due to the use of a machine learning model in the first step, and ε_r contributes to higher uncertainty in model estimates. In summary, the information loss resulting from extracting a single feature from an image can introduce both extraction bias and omitted variable bias. Readers interested in a more formal characterization of these biases and their derivations are encouraged to consult the proofs provided in the Web Appendix B for further details.

5.3 Empirical Evidence

In the previous section, we theoretically characterized how the two-step approach based on feature extraction results in information loss that can appear in forms of extraction bias and omitted variable bias. In this section, we empirically examine if the sources of bias 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 the focal politician 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 (8). For illustrative purposes, we focus on the subset of our data containing images of Hillary Clinton (comprising 1,021 articles). We refer readers to Web Appendix C for a comprehensive analysis 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 developed by Meta as the machine learning algorithm f_1 (Taigman et al., 2014). We extract the presence ($Z(\hat{T}=1)$) or lack of a smile ($Z(\hat{T}=0)$) in the images of Hilary Clinton using DeepFace.[6] 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 as:

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

[6]DeepFace 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 in a variety of social science contexts (Dvorkin et al., 2021; Luca et al., 2022).

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

However, the above specification ignores all the non-smile-related information in the image (represented by $Z^{(-T)}$). As discussed earlier, doing so can lead to biased estimates of β. To that end, we now consider a modified specification that accounts for a few other image characteristics. For illustrative purposes, we select three arbitrary image features: edge density, brightness, and contrast, which capture a few relevant aspects of the visual context. Thus, the extended specification of the logistic regression model becomes:

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

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 estimates for the model without other image characteristics. Here, 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 other image characteristics in Column (2), the coefficient for *Democratic Outlets* changes to 0.278 and loses statistical significance ($p < 0.10$). This reduction in both magnitude and significance suggests that part of the association initially attributed to *Democratic Outlets* in Column (1) was actually capturing differences in how images were presented across outlets. In Column (3), we show 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 important to emphasize that the correlation structure between the Democratic Outlet dummy and three arbitrary images only highlight the potential for the bias in the estimates from the reduced form model, and the correct estimate that can be obtained with including all relevant controls can have any directional relationship with the estimates in Column (1).

In sum, 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. Even in this simple illustrative example, we see that failing to account for correlated image characteristics can lead to biased estimates and incorrect conclusions about media bias and political polarization. Finally, 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.

6 Our Approach: Polarization Measurement Using Counterfactual Image Generation

As discussed in §5, the feature extraction approach is subject to information loss since it ignores the high-dimensional information contained in the image, which in turn can bias the estimated measure of polariza-

Variable	*Dependent Variable: $\hat{T}$ (Happiness)*		*Dependent Variable: Y*
	(1)	(2)	(3)
Intercept	-1.170***	-1.809***	-1.604***
	(0.104)	(0.257)	(0.235)
Democratic Outlets (Y)	0.369**	0.278†	
	(0.141)	(0.145)	
Brightness Level		0.356	0.325
		(0.561)	(0.512)
Edge Density		0.878†	2.312***
		(0.477)	(0.456)
Contrast		0.537	1.450**
		(0.561)	(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***
Note:		†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.

tion. Therefore, we now develop a novel algorithm – Polarization Measurement Using Counterfactual Image Generation (PMCIG) – that recovers the polarization measure defined in §4 without incurring information loss.

Recall that the *visual polarization* measure is defined as (see Definition 1):

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

where $\Delta^T u(p,y) = \mathbb{E}_X\left[u\left(Z(T=1), X, P, Y\right) - u\left(Z(T=0), X, P, Y\right) \mid P = p, Y = y\right]$. 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 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 2. However, we do not see what would be the probability of this image being chosen by y_1 if everything else about it was held the same, with only the feature/treatment of interest turned off (i.e., a non-smiling version of the same image of Biden with $T = 0$). 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, 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= CNN	✓	X
Outlet y_2 = FoxNews	X	X

Table 2: Factual and Counterfactual outcomes for the example from Figure 2, 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 versions of an image that only differ in one feature (T). Using two versions of an image where the only difference is the presence of a smile ($T = 1$ vs. $T = 0$) 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, for any given image, we only have one version, where the image either has a smile or does not. Thus, for any focal image Z, the first challenge consists of obtaining two versions of the image that are exactly the same in all aspects except the smile.
- **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 the pair of news outlets (y_1, y_2), which is defined based on the news outlets' utility. However, identifying the utility function is not feasible because we do not observe the news outlets' choice set. Note that, in standard discrete choice models, identification of the utility function comes from observing which option an agent chooses from a set of alternatives (Train, 2009). However, in our setting, we only observe the chosen image. Therefore, we need a systematic approach to map the observed data to the polarization parameter without directly observing/identifying the utility functions.

The rest of this section is organized as follows. First, in §6.1, we present the overview of our solution to these two challenges. Next, in §6.2, we present our unified algorithm for quantifying visual polarization.

6.1 Overview of the Solution

In this section, we present an overview of our solution to the two challenges described earlier. We first present the generative component in §6.1.1. We then discuss our identification strategy in §6.1.2.

6.1.1 Counterfactual Image Generation

To address the first 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 been employed in several recent studies on 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) utilize these models to manipulate facial features and explore how these alterations affect judicial decisions. Finally, in Luo and Toubia (2024), these models create facial images to assess the impact of perceived gender traits on discrimination within online marketplaces.

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 that only differ in the feature of interest (e.g., the facial expression or smile).[7]

- 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 in the original image. Formally, applying π^1 to an image Z results in:

$$\pi^1(Z) = (T=1, Z^{(-T)}) = Z^1. \tag{15}$$

- 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:

$$\pi^0(Z) = (T=0, Z^{(-T)}) = Z^0. \tag{16}$$

For a set of original images, together, these two operators give us pairs of treated and control images.

Implementation Details: We now discuss how we generate counterfactual images in our study. For each politician p, we first select three neutral images $\tilde{Z}^0_{p1}, \tilde{Z}^0_{p2}, \tilde{Z}^0_{p3}$ (where $T=0$) from our dataset to ensure a balanced and unbiased representation of the politician.[8] As such, the function $\pi^0(\cdot)$ to generate the neutral image without the smile feature is an identity function. To generate the counterfactual version of each image with a smile, we employ AILabTools as our π^1 operator; this tool utilizes conditional Generative Adversarial Networks (cGANs) (Mirza and Osindero, 2014) for advanced image manipulation and generates

[7] It is worth mentioning that one could easily extend this operation to a continuous case, where we manipulate the extent to which the feature is activated, e.g., the extent to which the politician smiles.

[8] Using three instead of one image ensures that the findings are not driven by the peculiarities of a specific image. In principle, it is possible to use a larger/fewer number of images for this task. We choose three because it balances noise reduction with the cost of generating counterfactual images.

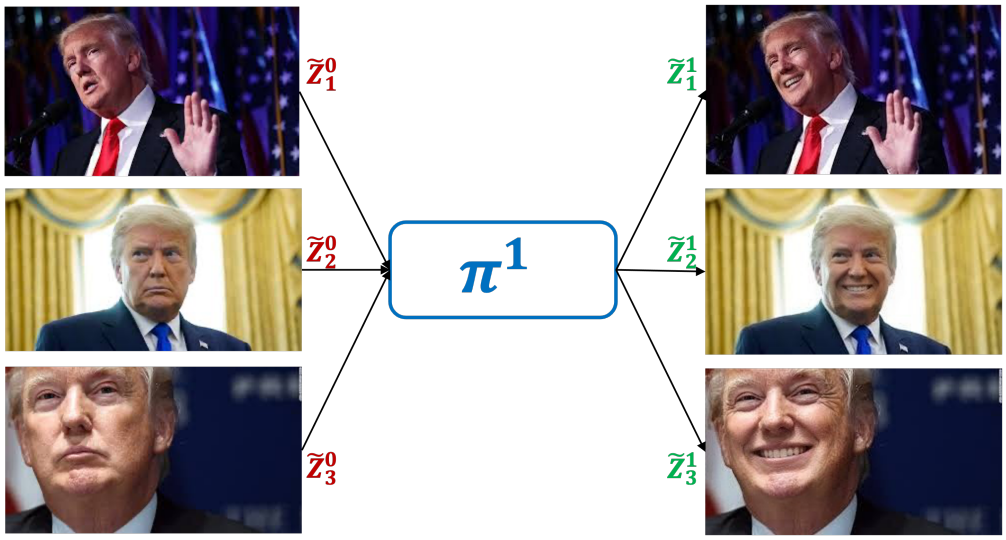


Figure 3: Counterfactual image generation for three neutral images of Trump

high quality and realistic outputs.[9] Thus, for each neutral image, we generate corresponding smiling versions $\tilde{Z}^1_{p1}, \tilde{Z}^1_{p2}, \tilde{Z}^1_{p3}$, where the only difference is the added attribute (smile). Figure 3 illustrates how we apply π^1 to three neutral images of Donald Trump, which in turn gives us three corresponding neutral images of Donald Trump with a smile. We repeat this process for all the 30 politicians in our study to construct the set $\{(\tilde{Z}^0_{p1}, \tilde{Z}^0_{p2}, \tilde{Z}^0_{p3}, \tilde{Z}^1_{p1}, \tilde{Z}^1_{p2}, \tilde{Z}^1_{p3})\}_p$.

6.1.2 Identification: Linking Polarization to the News Outlet Prediction Problem

We now discuss how we overcome the second challenge of identifying the *visual polarization* parameter defined in Equation (3) using observed data. Recall that our measure is defined for a politician p and any pair of news outlets y_1 and y_2. Consider a pair of images $\{Z^0_p, Z^1_p\}$ for politician p, where Z^0_p is the version of the image Z with the feature T turned off and Z^1_p is the version with the feature T turned on. Then, we can write the sample analogue estimator for the polarization parameter for this pair of images as follows:

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

where $\mathscr{X}_p$ is the set of all articles featuring politician p, so the sample analog estimator takes the average over all these articles featuring politician p. Clearly, if the utility function is known, we can directly estimate $\hat{\rho}(p, y_1, y_2)$ by integrating the above equation over all the articles. However, as highlighted in Challenge 2, the underlying utility function is not identified with our data because we do not observe the editors' choice sets of images.

We propose a novel solution to this identification challenge that directly links the *visual polarization* parameter to an outlet-prediction model that can be identified from the data given without observing/identifying utility functions. We now introduce some additional notation to help with this task. In a setting with two

[9] While many such tools are available, we use AILabTools because it allows us to precisely modify the focal images (by adding a smile) while maintaining the authenticity of the other features in the original image.

outlets y_1 and y_2, we define $\tilde{Y}$ as a pseudo-variable that denotes the news outlet with he highest utility for producing an article with the content x, politician p, and image z. Formally, we can define $\tilde{Y}$ as follows:

$$\tilde{Y}(z,x,p) = \underset{y_1,y_2}{\operatorname{argmax}}\{u(z,x,p,y_1)+\xi_1,\ \ u(z,x,p,y_2)+\xi_2\}, \tag{18}$$

where ξ_1 and ξ_2 are independent and identically distributed terms that come from the Type 1 Extreme Value distribution. We can link the elements of the right-hand side (RHS) of Equation (17) to pieces identifiable from the data at hand by re-writing the sample analog estimator for *visual polarization* as:

$$\begin{aligned}
\hat{\rho}(p,y_1,y_2) &= \frac{1}{|\mathscr{X}_p|}\sum_{x\in\mathscr{X}_p}\Big(\big[u(Z_p^1,x,p,y_1)-u(Z_p^0,x,p,y_1)\big]-\big[u(Z_p^1,x,p,y_2)-u(Z_p^0,x,p,y_2)\big]\Big)\\
&= \frac{1}{|\mathscr{X}_p|}\sum_{x\in\mathscr{X}_p}\Big(\big[u(Z_p^1,x,p,y_1)-u(Z_p^1,x,p,y_2)\big]-\big[u(Z_p^0,x,p,y_1)-u(Z_p^0,x,p,y_2)\big]\Big)\\
&= \frac{1}{|\mathscr{X}_p|}\sum_{x\in\mathscr{X}_p}\left[\log\left(\frac{e^{u(Z_p^1,x,p,y_1)}}{e^{u(Z_p^1,x,p,y_2)}}\right)-\log\left(\frac{e^{u(Z_p^0,x,p,y_1)}}{e^{u(Z_p^0,x,p,y_2)}}\right)\right]\\
&= \frac{1}{|\mathscr{X}_p|}\sum_{x\in\mathscr{X}_p}\left[\log\left(\frac{\Pr\left(\tilde{Y}=y_1\mid Z_p^1,x,p\right)}{\Pr\left(\tilde{Y}=y_2\mid Z_p^1,x,p\right)}\right)-\log\left(\frac{\Pr\left(\tilde{Y}=y_1\mid Z_p^0,x,p\right)}{\Pr\left(\tilde{Y}=y_2\mid Z_p^0,x,p\right)}\right)\right],
\end{aligned} \tag{19}$$

In deriving Equation (19), 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 (19) allows us to define the polarization parameter as the difference between two log-odds ratios related to the pseudo-variable $\tilde{Y}$.

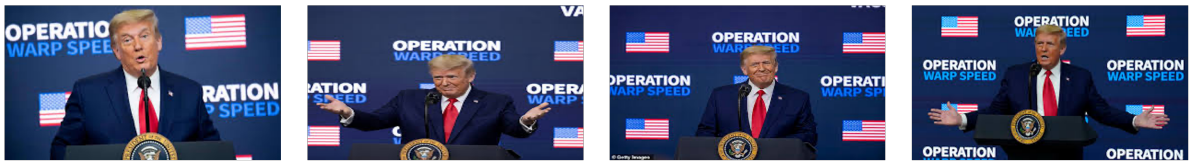
We now link the identification of the polarization parameter to an outlet prediction problem (given article and image features). This link depends on the link between the pseudo-variable $\tilde{Y}$ and the actual Y variable that determines the news outlet for an article. The following proposition characterizes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

Proposition 1. *Suppose that we have data $\mathscr{D}=\{(X_i,P_i,Y_i,Z_i)\}_i$, where if an article i is produced by news outlet Y_i, then Y_i has higher utility from this article than other outlets. Then, 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 outlet producing an article i has the highest utility from it (compared to other outlets), we have $\tilde{Y}=Y$. Therefore, the identification of log-odds in Equation (19) becomes equivalent to log-odds for variable Y. Let $\hat{g}(y\mid z,x,p)$ denote the conditional distribution that the probability of an article with image z, characteristics x, and politician p is produced by news outlet y. If the condition in Proposition 1 is satisfied, we can write *visual polarization* as follows:

$$\hat{\rho}(p,y_1,y_2) = \frac{1}{|\mathscr{X}_p|}\sum_{x\in\mathscr{X}_p}\left[\log\left(\frac{\hat{g}\left(y_1\mid Z_p^1,x,p\right)}{\hat{g}\left(y_2\mid Z_p^1,x,p\right)}\right)-\log\left(\frac{\hat{g}\left(y_1\mid Z_p^0,x,p\right)}{\hat{g}\left(y_2\mid Z_p^0,x,p\right)}\right)\right] \tag{20}$$

As such, the critical question is where in our data the condition in Proposition 1 is satisfied, that is, the production of an article by an outlet implies having the highest utility from producing it compared to other



ABC News **News Max** **Daily Mail** **Washington Post**

Figure 4: Visual coverage of Donald Trump's speech at "Operation Warp Speed" by different news outlets.

outlets. Such areas in the data satisfy $\tilde{Y} = Y$ and characterize our identifying variation. For that purpose, we focus on 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 4. 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. Therefore, the news outlet prediction from the set $\mathscr{Y}$ can potentially mimic the true choice model over the set of images $\mathscr{Z}$. Later in §7.2.2, we present results highlighting how our prediction model utilizes this identifying variation in the data.

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.

Model Architecture for the Multi-Modal News Outlet Prediction Problem: We now discuss the estimation of our news outlet prediction model, $g(y_i \mid Z_i, \boldsymbol{X}_i, P_i; \theta)$. 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, 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 condition in Proposition 1. 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 a 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, as illustrated in Figure 5.

To address the first challenge of leveraging variations in similar events, we use the contextual information in news articles. This includes textual data (X^{text}), publication dates (X^{date}), politicians' names (P), political affiliations (P^{aff}), and image data (Z). Textual data is processed using Latent Dirichlet Allocation (LDA), which encodes each news title as a 40-dimensional topic vector. Then, categorical data is embedded into dense vectors, enabling the model to capture patterns in metadata such as dates, names, and affiliations. Then, a *Attention Mechanism* processes the structured data (X^{text}, X^{date}, P, P^{aff}), learning to prioritize the most relevant metadata features for the classification task (Vaswani, 2017). Next, *ResNet-101* processes the entire image Z, leveraging its capabilities from general image recognition tasks to extract hierarchical and

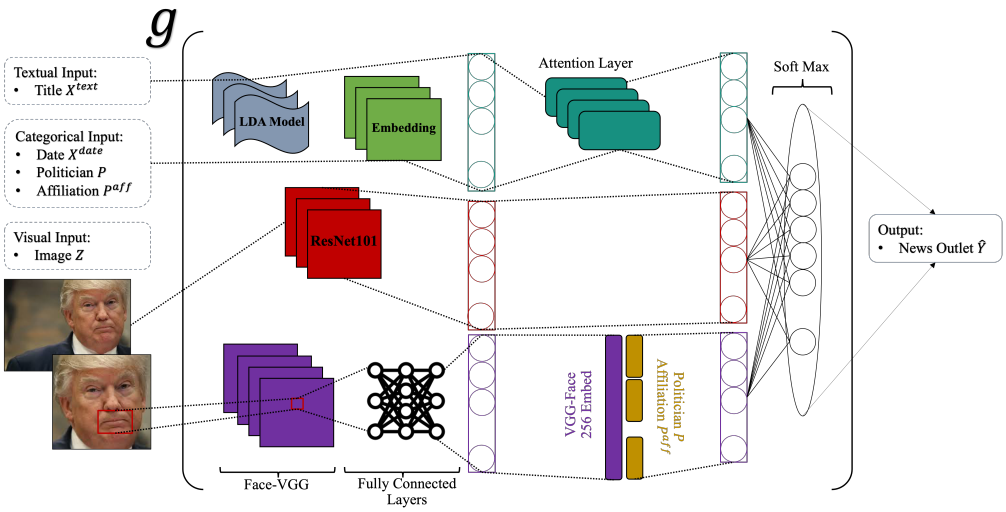


Figure 5: The multi-modal deep learning model for the news outlet prediction task

context-rich features from the broader visual content and scene information (He et al., 2016). Intuitively, given similarities in the event/news covered, these two parts of the model architecture effectively capture a given outlet's stylistic preference for certain styles of text and image features.

To address the second challenge of accurately capturing smiles and linking them to news outlets, we use *MTCNN* for precise face detection and cropping the face (Zhang et al., 2016). Then, the detected face is passed through the *VGG-Face* network, which is well-suited for our 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). Finally, *Chunk attention* is applied to the VGG-Face embeddings, combining them with categorical data (politicians' names P and affiliations P^{aff}) to capture correlations between facial features and structured metadata related to the image. This approach ensures that the model accurately captures the relationship between smiles and news outlets, mitigating the risk of misattribution to correlated features.

The final classification layer combines all modalities. The model is trained using the AdamW optimizer (Loshchilov and Hutter, 2017), which ensures efficient optimization and robust regularization. To train the model, we maximize the following entropy function over the training data:

$$\textbf{Entropy:} \quad \max_{\theta} \mathscr{H}(\theta) = \max_{\theta} \left(-\sum_{i=1}^{\mathscr{N}} \sum_{y \in \mathscr{Y}} g(Y_i = y \mid Z_i, X_i, P_i; \theta) \log g(Y_i = y \mid Z_i, X_i, P_i; \theta) \right) \tag{21}$$

The resulting model $\hat{g}$ is then used to estimate the polarization parameter using Equation (19). Further details on implementation and model training are provided in Web Appendix E.2.

6.2 Algorithm

We now present our algorithm, *Polarization Measurement Using Counterfactual Image Generation (PM-CIG)* in Algorithm 1. We split our data $\mathscr{D}$ into training and test data, denoted by $\mathscr{D}_{\text{train}}$ and $\mathscr{D}_{\text{test}}$ respectively. We build model estimates using the train data and apply these estimates to measure polarization using the

test data. Since our main goal is polarization measurement, this form of cross-fitting helps avoid overfitting bias. The algorithm consists of three steps as described below:

Algorithm 1 Polarization Measurement Using Counterfactual Image Generation (PMCIG)

Input:
Dataset $\mathscr{D} = \{(Z_i, \boldsymbol{X}_i, P_i, y_i)\}_{i \in \mathscr{N}}$
Split $\mathscr{D}$ into training set $\mathscr{D}_{\text{train}}$ and test set $\mathscr{D}_{\text{test}}$
Selected neutral images $\tilde{Z}_p \notin \mathscr{D}_{\text{train}}$ for each $p \in P$
Number of images for each politician used for counterfactual image generation S
News outlets $\mathscr{Y} = \{y^k\}_{k=1}^K$, with a baseline y^0 determined as the neutral outlet

Output:
Polarization Measurement $\hat{\rho}^T(p, y^k, y^0)$ for each observation in $\mathscr{D}_{\text{test}}$

Step 1: Apply Transformation π in §6.1.1

For each $p \in \mathscr{P}$ and image index s ($1 \leq s \leq S$) for the neutral image, apply π^1 to obtain:

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

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

Step 2: Train the Multi-Modal Deep Learning Model g in §6.1.2 on $\mathscr{D}_{\text{train}}$

Given the training set $\mathscr{D}_{\text{train}} = \{(Z_i, \boldsymbol{X}_i, P_i, y_i)\}_{i \in \mathscr{N}_{\text{train}}}$, learn parameters of model $g(\cdot; \theta) \in \mathscr{G}$:

$$\hat{\theta} \leftarrow \underset{\theta}{\text{argmax}}\, \mathscr{H}(\theta) = \underset{\theta}{\text{argmax}} \left(- \sum_{i=1}^{\mathscr{N}_{train}} \sum_{y \in \mathscr{Y}} g(Y_i = y \mid Z_i, X_i, P_i; \theta) \log g(Y_i = y \mid Z_i, X_i, P_i; \theta) \right)$$

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

Step 3: Estimate Probabilities and Calculate $\hat{\rho}^T(p, y^k, y^0)$ on $\mathscr{D}_{\text{test}}$

For each politician $p \in \mathscr{P}$ and any news outlet y^k, calculate the article-level estimated Polarization Measurement for the each article $j \in \mathscr{D}_{\text{test}}$ as $\hat{\rho}^T_j(p, y^k, y^0)$ following:

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

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

$$\hat{\rho}^T(p, y^k, y^0) \leftarrow \frac{1}{N^{\text{test}}_p} \sum_{j \in \mathscr{D}_{\text{test}}, P_j = p} \hat{\rho}^T_j(p, y^k, y^0)$$

- In the first step, we generate the counterfactual image versions for each politician. To do so, we take three neutral images from each politician p from the test data and apply the function π^1 using the procedure outlined in §6.1.1. 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 §6.1.2 to estimate the model $\hat{g}$ using training data $\mathscr{D}_{train}$, which is a random subset of 85% of all articles in data $\mathscr{D}$.
- Finally, in the third step, we measure polarization on the held-out test data set $\mathscr{D}_{test}$, 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, and ensures that our estimates do not exhibit

overfitting bias (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. To the extent that Reuters is the neutral point zero on the spectrum, one could interpret our polarization estimates as measures of visual slant.

7 Empirical Evaluation and Findings

In this section, we present the results from our empirical analysis. We divide our results into three main parts. First, §7.1, we provide evidence demonstrating that the generated counterfactual images are consistent with the original images, preserving all features except the added smile. Next, §7.2, we present results on the performance of our news outlet prediction model and how it captures the feature of interest. Finally, in §7.3, we share our results on political polarization in visual content among news outlets, and present findings at both the news outlet and politician levels.

7.1 Validating Counterfactual Image Consistency

We briefly discuss the validation of the counterfactual images using our GAN toolkit. Specifically, it is important to ensure that adding a smile does not systematically change other contextual features of the image. Ensuring this consistency is essential for isolating the effect of the smile without introducing biases from other image characteristics, such as brightness or colorfulness (see §5.2). To that end, we measure a few other image characteristics (such as brightness and colorfulness) before and after applying the smile operator π^1 and use the Kolmogorov-Smirnov (K-S) test to compare the distributions of these characteristics for the original and smiley images. Our tests suggest that there are no statistically significant differences between the original and smiley images for other contextual features. This confirms that the operator π^1 does not introduce significant changes in other characteristics, ensuring that the counterfactual images remain consistent with the original ones. Detailed histograms and a complete analysis of the test results are shown in the Web Appendix §E.1.1.

7.2 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 an 85%-15% split between training and test data. We consider four measures to evaluate the model's predictive performance – (1) Accuracy, (2) Precision, (3) Recall, and (4) Weighted Cross-Entropy (WCE). Detailed explanations of these metrics, including WCE, are provided in Appendix F.1. We use these performance measures to evaluate three models on the test data and present the results in Table 3.

Each row shows a more complex model that progressively adds more explanatory variables/features. The row considers a model that only uses categorical inputs such as politician identity, party affiliation, and date. We first add textual information in the second row and add image features in the final row. Two key points emerge from Table 3. First, our final model achieves a remarkably good performance of approximately 44%

Features Used	Accuracy (%)	Precision (%)	Recall (%)	WCE Loss
Politician, Party, Date	11.39	10.99	15.18	2.76
Politician, Party, Date, Text	16.72	15.01	17.08	2.66
Politician, Party, Date, Text, Image	43.81	43.65	41.40	2.29

Table 3: Performance metrics for 20-class classification task using 5% smoothing factor with different inputs on the test data.

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 achieves accuracy of 43.81%.

We present additional details on the model's performance measures across outlets in Web Appendix F.1. In the rest of this section, we add provide some interpretation and intuition to the results from the black-box predictive model. In §7.2.1, we discuss how our model captures the information in smile and in §7.2.2, we examine how outlet predictions shift when we add a smile to an image.

7.2.1 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. However, this can be challenging since visual features are high-dimensional. In particular, it is important to examine whether the model is really utilizing information from facial features and emotions or whether it is simply utilizing the broader contextual information in the image for prediction (such as background color, text, etc.).

We employ a recently developed, popular tool – Grad-CAM – to visualize the contribution of specific visual features to the model's predictions by helping us understand the attention patterns of different components in the model (Selvaraju et al., 2017). Grad-CAM highlights the regions of an image that drive the decision-making process by examining the gradient of the model's prediction function $g(\cdot)$ with respect to its inputs. This technique is instrumental in our architecture, where two different CNN branches—Face-VGG and ResNet101—process facial features and global image context, respectively. We can therefore use Grad-CAM to analyze attention patterns in both CNN branches separately.

As an illustrative example, consider the last image of Donald Trump in Figure 4, from *Washington Post*. Figure 6 shows visualizations for two versions of this image: 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. Second, 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

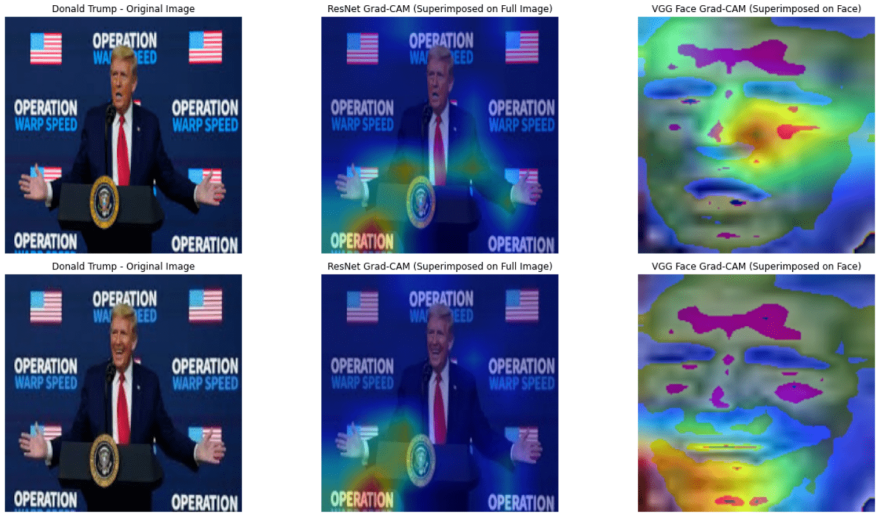


Figure 6: Grad-CAM Visualizations for Original and Smile-Added Versions of a *Washington Post* Image. The ResNet101 heatmaps (right column) capture information on the global image context, while the Face-VGG heatmaps (middle column) highlight the mouth region and facial expression changes.

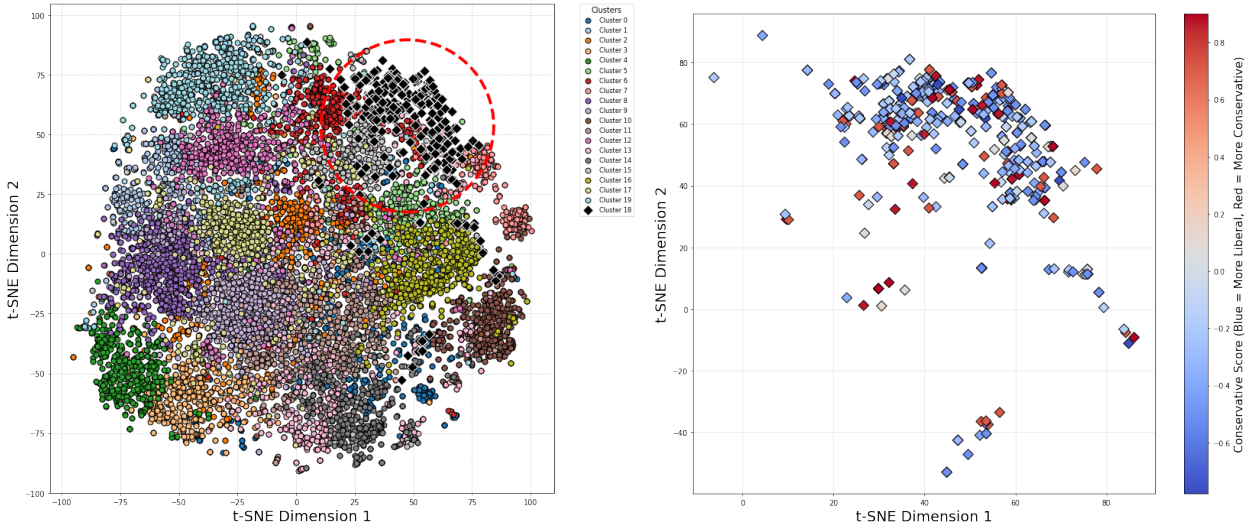
other facial areas associated with smile dynamics. This pattern suggests that Face-VGG is able to successfully detect smile-related features and subtle changes in facial expressions; which provides strong evidence in support of including Face-VCG in our network architecture.

7.2.2 How Does Adding a Smile Change Predictions?

In the previous section, we saw that our outlet prediction model is able to capture the information in the smile and/or the facial features of the focal politician. We now examine how adding a smile to a focal image changes news outlet predictions. As discussed in §6.1.2, our identifying assumption requires that the dataset contains similar images (e.g., from the same/similar events, as in the case of Operation Warp Speed) from a diverse set of news outlets, where the main difference between the images is the facial expression of the focal politician. As such, 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 background features, 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 7a 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". For instance, consider Cluster 18, which is highlighted in Figure 7b.



(a) Clusters for Donald Trump: Images grouped into 20 clusters based on t-SNE projections.

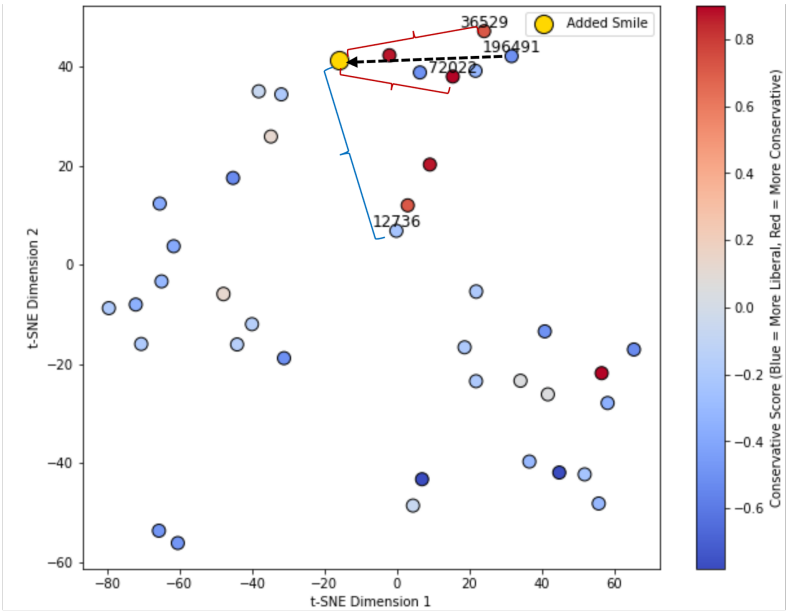
(b) t-SNE Plot for Cluster 18: Highlighting conservative scores for images in this cluster.

Figure 7: t-SNE-based clustering of images. (a) Displays clusters for Donald Trump; (b) Focuses on Cluster 18 with conservative scores.

Here, each point represents an image, and the color indicates the conservative score of the news outlet that selected the image, ranging from liberal (blue) to conservative (red). The diversity in news outlets within the cluster suggests that multiple news outlets show images that are similar in structure/background (likely covering similar events). This variation is critical since it allows the outlet prediction model to capture how variation in facial features influences the likelihood of an image being chosen by a given outlet, conditional on the other image features (e.g., background, text, texture) being similar.

Figure 8a illustrates examples from Cluster 18, showcasing images used by different news outlets, while Figure 8b displays the corresponding images for the highlighted IDs. 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 8a, 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 to the left. Now, the updated position results in a shorter distance to images such as ID 36529 from the *Daily Mail* and ID 72022 from *Newsmax*, compared to its distance to the original image ID 196491 by the *Washington Post* and ID 12736 by *ABC 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 do not capture the full complexity of the image features that our model $\hat{g}$ captures. Therefore, in Table 4, we present our model's outlet predictions for the fourth image from Figure 4 (the one that was originally shown



(a) Embedding positions of the original image (ID: 196491), similar images, and the image with a smile added (highlighted in yellow).

(b) Images corresponding to highlighted points in the t-SNE plot.

Figure 8: t-SNE plot of the 18th cluster of Donald Trump's images illustrating the effect of adding a smile.

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

Table 4: Outlet predictions for the original and smile-added versions of the Donald Trump image.

in Washington Post) for two cases – the original image and smile-added image version. We see substantial differences in the outlet predictions between the original image and the counterfactual version with the smile. 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 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%).

In summary, we see that adding a smile systematically shifts the outlet prediction probabilities. Specifically, we see an increase (decrease) in the predicted probabilities for right-leaning (left-leaning) news outlets when we add a smile to an image of Donald Trump. Nevertheless, 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 to quantify visual polarization.

7.3 Visual Slant

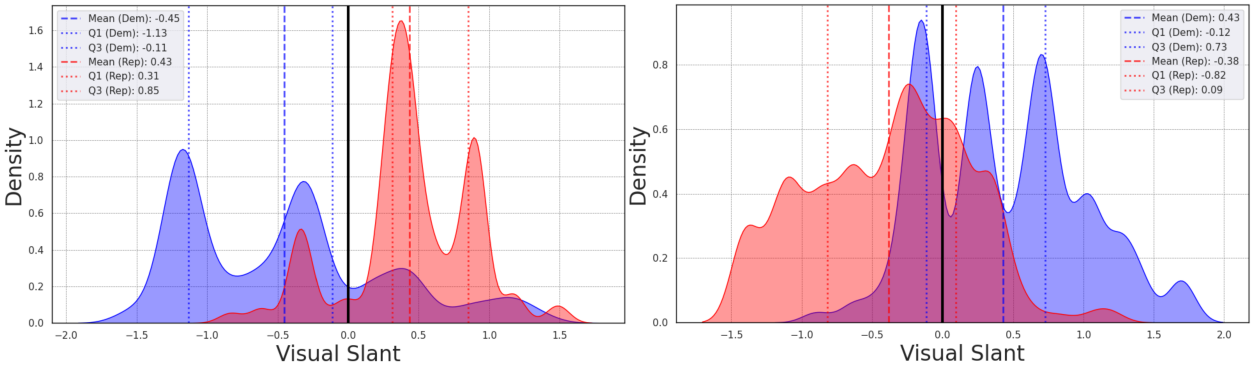
We now present our main results on visual slant. All results are direct applications of Algorithm 1 in §6.2. Before presenting the results, we review a few important considerations in applying our algorithm. First,

recall that our visual slant parameter requires a neutral outlet denoted by y_n in Definition 2. We use *Reuters* as the baseline outlet y_n to measure visual slant, as *Reuters* is a largely neutral and fact-based outlet.[10] Second, we use a train-test split, and all the polarization and slant measures are shown for the test data $\mathscr{D}_{test}$ using the estimates obtained from the training data $\mathscr{D}_{train}$. Third, for each politician, we start with three neutral images, generate counterfactual smiling versions of these three images, and use these six images in our polarization and slant measurement for each politician.

This section is organized as follows. First, in §7.3.1, we demonstrate the overall distribution of visual slant across all outlets and politicians. Next, in §7.3.4, we correlate our measure with existing measures of slant and news outlet partisanship at the news outlet level to examine the extent to which it is consistent with older measures. We then explore the extent of heterogeneity across news outlets in §7.3.2 to see which outlets exhibit higher levels of visual slant. Finally, in §7.3.3, we document the heterogeneity in visual slant at the politician level.

7.3.1 Distribution of Visual Slant

An important feature of our algorithm is its ability to produce an individual-level measure of visual slant $\hat{\rho}_i^T(p_i, y_i^k, y_i^{Reuters})$. As such, for each article i featuring politician p, we can estimate 19 visual slant measures corresponding to all the 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*. Figure 9 shows the distributions of visual slant for Democratic and Republican politicians as featured in Democratic- and Republican-leaning news outlets. We denote the set of Republican and Democratic politicians by $\mathscr{P}_R$ and $\mathscr{P}_D$, respectively. Similarly, we denote the sets of Republican- and Democratic-leaning outlets defined above are denoted by $\mathscr{Y}_R$ and $\mathscr{Y}_D$, respectively.



(a) Republican politicians.

(b) Democratic politicians.

Figure 9: Histograms of *visual slant* for Democratic and Republican Politicians in Democratic and Republican News Outlets.

If there were no visual slant or polarization, all distributions would be tightly centered around zero.

[10] It is worth emphasizing that one could easily change the choice of base. Although the visual slant measure will change with a different choice, the overall visual polarization between two outlets remains unchanged.

However, we observe that for both Republican- and Democratic-leaning outlets, the individual-level visual slant measures deviate from zero and exhibit significant dispersion. This indicates the presence of a visual slant in these outlets. Additionally, we notice a clear divide in the distributions of Democratic- and Republican-leaning outlets for both Republican and Democratic politicians, pointing to evidence of visual polarization.

Next, we examine if there is a systematic difference between Republican- and Democratic-learning outlets by examining their averages in each panel. For Republican politicians, in Figure 9a and , the distributions of $\hat{\rho}_i^T$ show distinct differences between left- and Republican-leaning news outlets. The visual slant 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 visual slant 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:

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

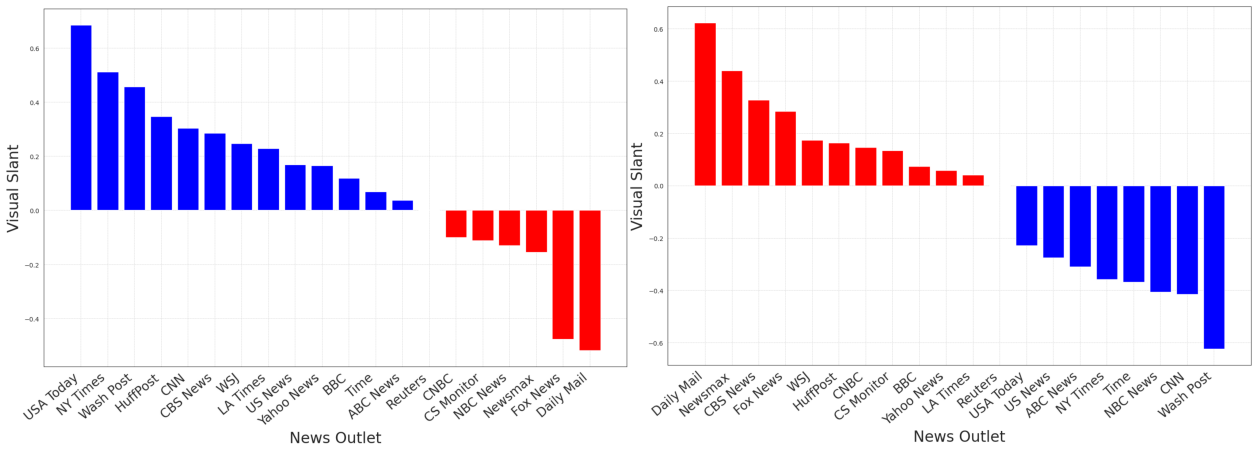
For Democratic politicians, in Figure 9b, the visual slant distribution for Democratic and Republican news outlets also shows clear differences. The average visual slant for Democratic news outlets is *0.43*, indicating that, on average, images of smiling Democratic politicians increase the utility of a news outlet classified as Democratic-leaning compared to the neutral baseline, Reuters. Conversely, the average visual slant 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:

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

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. More specifically, while Figures 9a and 9b highlight a distinct divide in the portrayal of politicians, the separation is more pronounced for Republican politicians, as evidenced by the smaller overlapping region in their distributions. In the following sections, we delve deeper into the heterogeneity across both news outlets and individual politicians.

7.3.2 Visual Slant Across News Outlets

We now document the extent to which visual slant varies across outlets. Figure 10 shows our visual slant measure for each news outlet ($y^k \in \mathscr{Y}$) for Republican politicians as $\hat{\rho}^T(p \in \mathscr{P}_R, y^k, y^{Reuters})$), and for Democratic politicians as $\hat{\rho}^T(p \in \mathscr{P}_D, y^k, y^{Reuters})$, ranked in bar charts. Specifically, Figure 10a shows that Democratic news outlets tend to have a positive visual slant measurement for Democratic politicians and a negative visual slant for Republican politicians. Conversely, Figure 10b indicates that Republican news outlets exhibit a positive visual slant measurement for Republican politicians and a negative visual slant measurement for Democratic politicians.



(a) Democratic politicians.

(b) Republican politicians.

Figure 10: Overall visual slant measures for Democratic and Republican politicians by news outlet.

Specifically, outlets such as *Fox News* and *Daily Mail* display a more positive visual slant measurement for Republican politicians, indicating a more favorable portrayal of these politicians. Simultaneously, these outlets show a more negative visual slant measure of Democratic politicians, displaying a less favorable portrayal of these politicians. On the other hand, outlets like *CNN*, *Washington Post*, and *The New York Times* exhibit a more positive visual slant measurement for Democratic politicians and a more negative visual slant for Republican politicians. Interestingly, some outlets, such as the *Wall Street Journal* and *CBS News*, demonstrate a positive visual slant for both Democratic and Republican politicians. This suggests that their editorial approach may balance portrayals of politicians from both parties, reflecting a more centrist ideological positioning rather than strong partisan alignment.

To establish a single outlet-specific visual slant measurement, we compute the difference between the visual slant measures for Republican and Democratic politicians within each outlet. A larger difference indicates a stronger conservative visual slant. Accordingly, we define our unified measure as the *Conservative Visual Slant (CVS)* as follows:

$$\text{CVS}(y_k) = \hat{\rho}^T(p \in \mathscr{P}_R, y^k, y^{\text{Reuters}}) - \hat{\rho}^T(p \in \mathscr{P}_D, y^k, y^{\text{Reuters}}), \quad \text{where } y^k \in \mathscr{Y}. \tag{22}$$

Intuitively, this metric captures the degree to which an outlet portrays Republican politicians more favorably compared to Democratic politicians. For example, if a Republican-leaning outlet scores high on this measure, it implies that it emphasizes showing Republican politicians with a smile while depicting Democratic politicians without a smile. Figure 11 illustrates the conservative visual slant scores for all outlets in our dataset, sorted in increasing order. The results reveal that outlets such as *Daily Mail* and *Fox News* demonstrate high conservative visual slant, strongly favoring Republican politicians. On the other hand, outlets like *Washington Post*, *USA Today*, and *NY Times* exhibit a pronounced liberal slant, favoring Democratic politicians. Meanwhile, outlets such as *Wall Street Journal*, *BBC News*, and *CBS News* exhibit relatively neutral or low levels of slant. Overall, there is significant heterogeneity in visual slant across outlets, providing insights into which outlets are the most and least polarized visually.

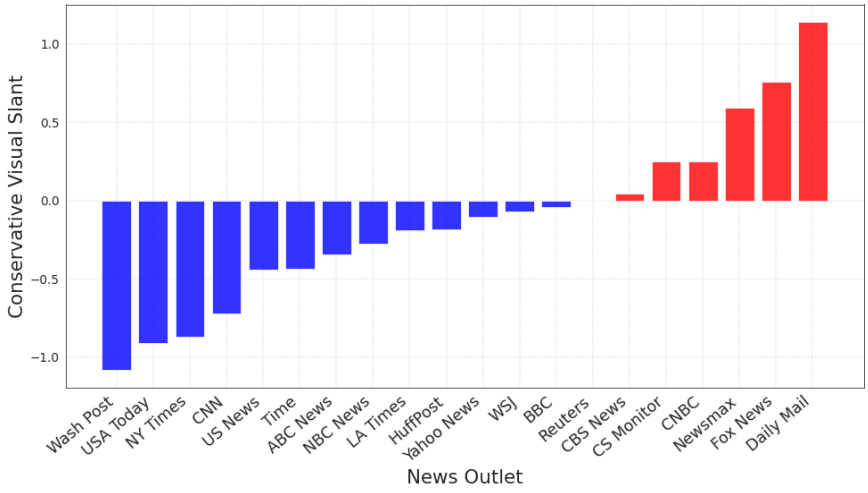


Figure 11: Overall visual slant between Republican and Democratic politicians across outlets.

Finally, in Web Appendix F.6, we provide a detailed analysis of outlet-level visual slant, including individual histograms, statistical summaries, and additional tests to quantify bias and polarization across news outlets.

7.3.3 Visual Polarization Across Politicians

We now examine how different politicians are portrayed across news outlets and identify the politicians with the most polarizing depictions in media. Figure 12 presents the visual slant measures for two prominent figures: Barack Obama and Donald Trump. Consistent with our earlier findings, we observe a clear divide in visual slant scores-Obama receives more favorable portrayals in Democratic-leaning outlets, while Trump is depicted more positively in Republican-leaning outlets. We extend this analysis to other politicians, with detailed results provided in Web Appendix F.7.

Next, we develop a measure for the overall extent of visual polarization for each politician. Intuitively, we expect to observe a greater extent of variability in visual slant measures across outlets for a more polarizing politician. As such, we quantify the Overall Visual Polarization (OVP) for each politician, which is measured by the standard deviation of the polarization across all news outlets as:

$$\text{OVP}(p) = \sqrt{\frac{1}{|\mathscr{Y}|}\sum_{y^k \in \mathscr{Y}} \left(\hat{\boldsymbol{\rho}}^T(p, y^k, y^{\text{Reuters}}) - \mu_p\right)^2}, \quad \text{where} \quad \mu_p = \frac{1}{|\mathscr{Y}|}\sum_{y^k \in \mathscr{Y}} \hat{\boldsymbol{\rho}}^T(p, y^k, y^{\text{Reuters}}) \tag{23}$$

This metric provides a measure of the overall polarization each politician faces in the media. Figure 13 ranks the politicians from most polarized to least polarized based on this criterion.

We see that Donald Trump and Barack Obama are the top two politicians with the most visually polarizing portrayal across all outlets. Given that both were president/presidential candidate for significant periods of time and at the forefront of multiple polarizing discussions and events, this is understandable. Further, Rand Paul and Ted Cruz, both prominent Republican politicians, also show high levels of polariza-

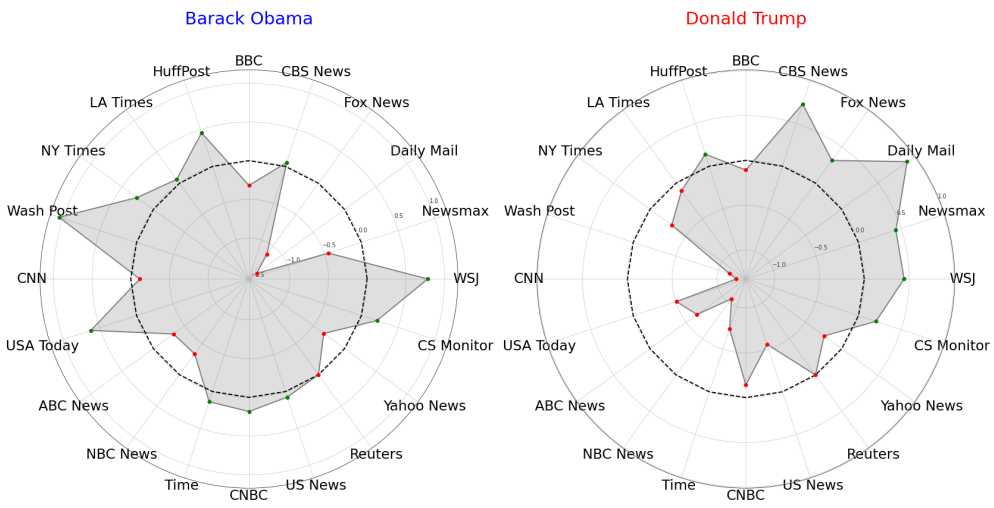


Figure 12: Radar plots of visual slant in each news outlet for Barack Obama and Donald Trump

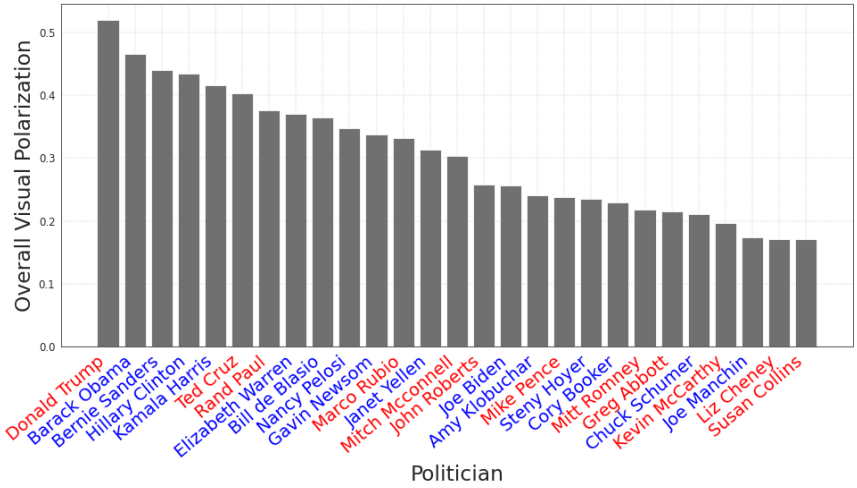


Figure 13: Overall visual polarization of politicians across all news outlets, ranked from most to least polarized

tion, likely attributed to their roles in policy debates and media prominence during key political events. On the Democratic side, figures such as Bernie Sanders, Hillary Clinton, and Kamala Harris display significant polarization, reflecting their leadership roles and ideological positions within the party.

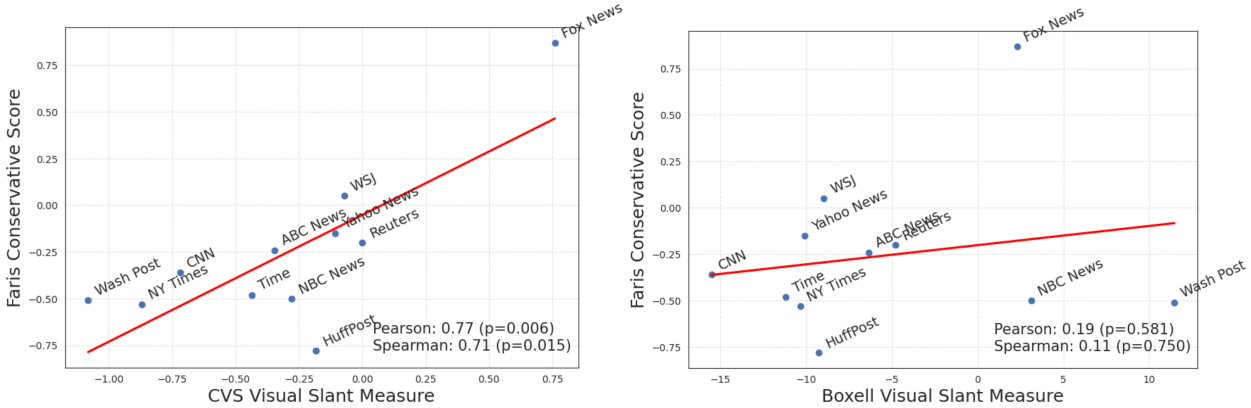
On the right end of Figure 13, we observe politicians with lower OVP scores. Notably, Susan Collins and Joe Manchin rank among the least polarized, aligning with their reputations for bipartisanship and moderation (Killough, 2018). We also observe low overall slant scores for Liz Cheney, who is not typically considered a moderate Republican. However, her open criticism likely led to a shift in her portrayal within each partisan outlet, resulting in a low outlet-level visual slant score. These findings have important implications for political strategy, particularly in how politicians shape their election campaigns when deciding

whether to mobilize their base or appeal to a broader, centrist electorate.

7.3.4 Measurement Validation

In this section, we validate the measures extracted from our PMCIG algorithm, performing the following tests:

- *Consistency with external measures of media slant:* We begin by assessing the correlation between our CVS measure and existing media slant measures derived from independent sources not used in our algorithm. Specifically, we consider three measures from Faris et al. (2017), Flaxman et al. (2016), and AllSides (2024). For instance, Faris et al. (2017) quantifies media slant based on the proportion of a media outlet's stories shared on Twitter by users who predominantly retweet conservative-leaning sources. Since our algorithm does not incorporate the data used to construct these external slant measures, a positive correlation between our CVS measure and these benchmarks would provide validation for our approach. For the 20 outlets in our dataset, we find significant and positive correlations of 0.79, 0.55, and 0.81 with the measures from Faris et al. (2017), Flaxman et al. (2016), and AllSides (2024), respectively. We visualize these correlations and conduct formal statistical tests in Web Appendix §F.4 to further support this validation.
- *Comparison to the existing measures of visual slant:* Second, we compare our Conservative Visual Slant (CVS) measure with existing outlet-specific visual slant measures from prior literature. Specifically, we examine the visual slant measure proposed by Boxell (2021), which employs a reduced-form approach similar to that discussed in §5. Our objective is to determine which measure better aligns with polarization benchmarks used in our earlier validation. In the main text, we focus on the conservative share score by Faris et al. (2017), a widely recognized benchmark in media slant research, including in Boxell (2021). To quantify the alignment between these visual slant measures and the conservative share score from Faris et al. (2017), we conduct both Pearson and Spearman correlation analyses, assessing how well each measure captures the overall slant of news outlets. Given that our dataset and Boxell (2021) share 11 common outlets, we focus on this subset for direct comparison.



(a) Conservative Visual Slant (CVS) from PMCIG

(b) Visual Slant Measure from Boxell (2021).

Figure 14: Comparison of CVS and Boxell (2021) Visual Slant against Conservative Share Score from Faris et al. (2017).

Figure 14 presents a comparative analysis of visual slant measures across news outlets. In both plots, the x-axis represents each outlet's *conservative share score* from Faris et al. (2017), while the y-axis represents either our CVS metric (Figure 14a) or the visual slant measure from Boxell (2021) (Figure 14b). Each point corresponds to a news outlet, with a red trendline illustrating the relationship between visual slant and conservative share scores. As shown in these figures, our CVS measure exhibits a statistically significant and strong Pearson and Spearman correlation with the conservative share score from Faris et al. (2017), whereas the visual slant measure from Boxell (2021) shows only a weak, statistically insignificant correlation. This finding further validates our method, demonstrating that our CVS measure more accurately captures media slant compared to prior visual slant metrics. Additionally, in Web Appendix §F.5, we replicate the analysis presented in this part using two other widely recognized conservative share scores to further assess the robustness of our findings.

- *Consistency with external measures of politician's polarization:* Lastly, we validate our politician-specific Overall Visual Polarization (OVP) measure by comparing it with external indicators of a politician's level of polarization. A key challenge in this validation is the absence of a widely accepted, politician-specific polarization metric. We propose that a politician's ideological alignment with their primary constituency serves as a meaningful proxy, as more polarizing politicians are likely to perform better in ideologically aligned constituencies and struggle in misaligned ones. To quantify this alignment, we use the politician's party success in the 2016 election within their state, by measuring their party's percentage point advantage in that election. We then analyze the correlation between this measure and our OVP metric. Figure 15 presents the results, showing a strong and significant correlation between the two, further supporting the validity of our measure.

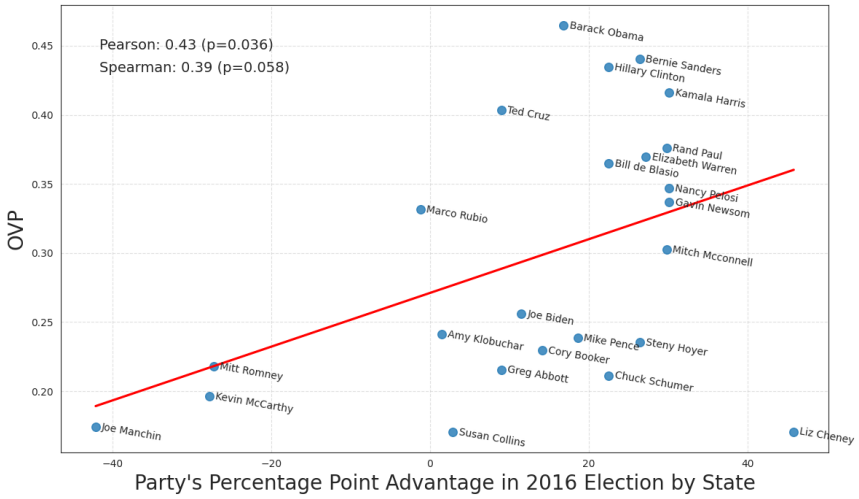


Figure 15: The Relationship Between a Politician's Overall Visual Polarization and Their Ideological Alignment with Constituency

This finding supports the idea that politicians from states with strong partisan leanings have greater ideological freedom, allowing them to adopt more pronounced partisan positions without needing to appeal to a broad electorate. For example, Bernie Sanders (Vermont) and Ted Cruz (Texas) come from

states with clear ideological identities and exhibit high OVP values, likely reflecting their strong partisan stances and the resulting polarized media portrayals. In contrast, Joe Manchin (West Virginia) and Susan Collins (Maine), who represent states where their party is in the minority, have lower OVP values, which reflects their moderate positions.

In summary, we validate our CVS and OVP measures by comparing them to external benchmarks based on independent sources of information. Notably, our algorithm's ability to capture variations in these external measures using only outlet images and articles highlights the importance of visual information and demonstrates the strength of our algorithm. Additionally, our comparison with prior work on visual slant shows that our measure captures variations in external benchmarks more accurately, further demonstrating the strength of our approach.

8 Conclusion

In this paper, we present a framework for analyzing political polarization in the visual content of news media, combining theoretical insights with empirical analysis. At the center of this study is the Polarization Measurement Using Counterfactual Image Generation (PMCIG) algorithm, which investigates news outlets' preference slanted imagery—such as smiling images—to convey positive (vs. negative) representation of politicians in their articles. Our framework systematically examines ideological slanting and political polarization in visual content used by news outlets by integrating a news outlet utility model with advanced machine learning methods, including multi-modal deep learning and GANs. Notably, our algorithm overcomes the key limitations in traditional descriptive methods due to information loss in the feature extraction phase by using the rich information contained in images.

Applying the PMCIG framework to a decade-long dataset that covers 20 major news outlets and 30 prominent politicians, we identify clear patterns of ideological slanting and political polarization in the visual representation of political figures. We validate our measure of visual slant by demonstrating a high correlation between our measure and the existing measures used for media slant and partisanship. Our framework measures visual slant and polarization with detailed granularity, highlighting differences both at the outlet level and for individual politicians. Among outlets, we find that Daily Mail and Fox News display the strongest Republican-leaning visual slant, while Washington Post and New York Times exhibit the strongest Democratic-leaning slant. In contrast, CBS News and Wall Street Journal are among the outlets with the lowest overall visual slant. At the individual level, Donald Trump and Barack Obama stand out as the most polarizing figures, whereas Joe Manchin and Susan Collins are among the least polarizing in their visual portrayal across news outlets.

In summary, the PMCIG framework addresses key challenges in studying visual bias, offering a systematic approach to analyzing how ideological preferences shape visual content in news media and contribute to polarization. Nevertheless, our paper has limitations that serve as excellent avenues for future research. For instance, our analysis is based on data from the United States, and extending the framework to other regions, such as Europe, could uncover cross-cultural differences in visual polarization. Additionally, while the framework measures visual slant and polarization, it does not examine the downstream impact on individuals' beliefs or behavior, which could be a fruitful area for future study. Expanding this work, future research could also apply the PMCIG framework to other forms of media, such as social media or adver-

tising, and investigate additional visual content features to provide a more comprehensive understanding of visual polarization. These extensions would deepen insights into how visual media influences public discourse and societal outcomes.

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.

Microsoft Azure. Face api - cognitive services, 2023. URL https://azure.microsoft.com/en-us/services/cognitive-services/face/. Accessed on September, 2023.

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.

Tommaso Bondi, Omid Rafieian, and Yunfei Jesse Yao. Privacy and polarization: An inference-based framework. *Available at SSRN 4641822*, 2023.

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.

Face++. Emotion recognition, 2023. URL https://www.faceplusplus.com/emotion-recognition/. Accessed on September, 2023.

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. Media bias and reputation. *Journal of political Economy*, 114(2): 280–316, 2006.

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.

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.

Joandrea Hoegg and Michael V Lewis. The impact of candidate appearance and advertising strategies on election results. *Journal of Marketing Research*, 48(5):895–909, 2011.

Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press, 2015.

Ganesh Iyer and Hema Yoganarasimhan. Strategic polarization in group interactions. *Journal of Marketing Research*, 58(4):782–800, 2021.

Maurice Jakesch, Mor Naaman, and MACY Michael. Belief in partisan news depends on favorable content more than on a trusted source. 2022.

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.

Ashley Killough. Moderate senators feel boost as shutdown ends. *CNN*, 2018. URL https://www.cnn.com/2018/01/22/politics/moderate-senators-shutdown-end-bipartisan-group-senate/index.html.

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.

Hana Matatov, Mor Naaman, and Ofra Amir. Stop the [image] steal: The role and dynamics of visual content in the 2020 us election misinformation campaign. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–24, 2022.

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.

Microsoft. Responsible ai investments and safeguards for facial recognition, 2022. URL https://azure.microsoft.com/en-us/blog/responsible-ai-investments-and-safeguards-for-facial-recognition.

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.

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

SerpAPI. Serpapi - real-time search engine results api, 2023. URL https://serpapi.com/. Accessed on January, 2023.

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 Sülflow 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.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

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.

Kenneth E Train. *Discrete choice methods with simulation*. Cambridge university press, 2009.

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.

Google Vision. Ml kit: Face detection api, 2023. URL https://developers.google.com/ml-kit/vision/face-detection. Accessed on September, 2023.

Yanhao Wei and Nikhil Malik. Unstructured data, econometric models, and estimation bias. SSRN, 2022.

Jeffrey M. Wooldridge. *Econometric Analysis of Cross Section and Panel Data*. MIT Press, Cambridge, MA, 2nd edition, 2010.

Sikun Xu, Dennis J. Zhang, Zhenling Jiang, and Zhengling Qi. Causal inference when controlling for unstructured data. *Working Paper*, 2024. Olin Business School, Washington University in St. Louis. Accessed: 2025-01-21.

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 A.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 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

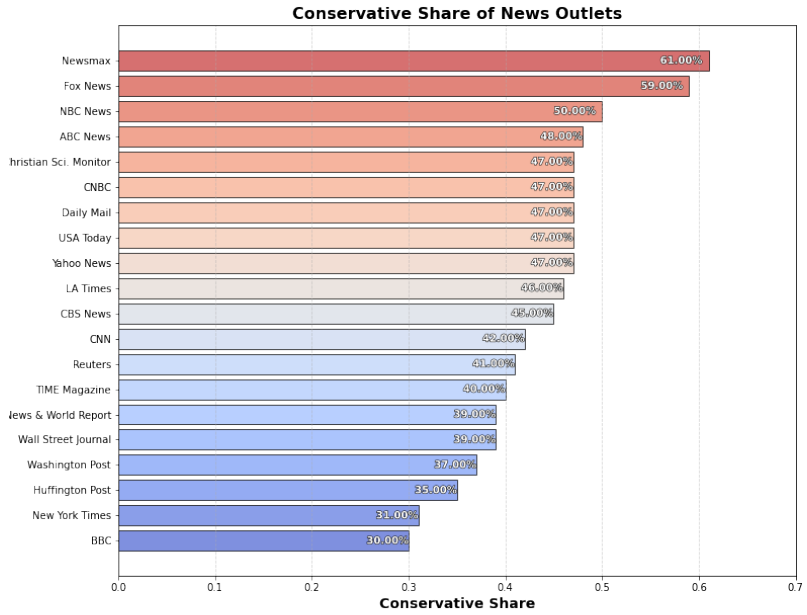


Figure A.1: Conservative share of news outlets. (Flaxman et al., 2016)

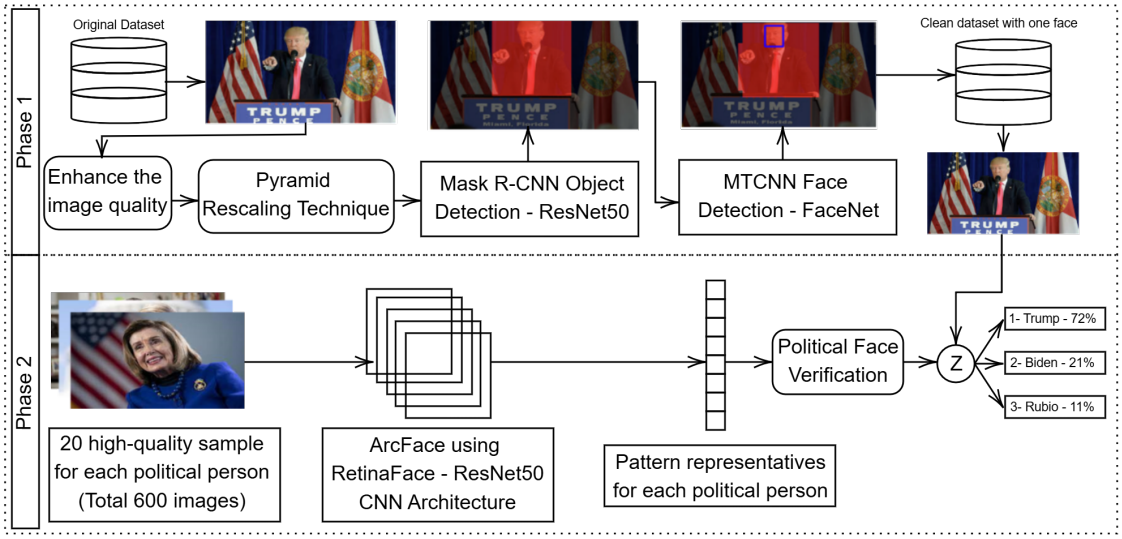


Figure A.2: Two-phase computer vision framework for political person verification

detected face with our database of political figures and ranks the top three matches, providing a similarity percentage for each.

In total, we retain 63,188 images with exactly one face, that have been 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.3. This dataset allows us to accurately analyze media representation of political figures, ensuring that each image is correctly attributed to the intended individual.

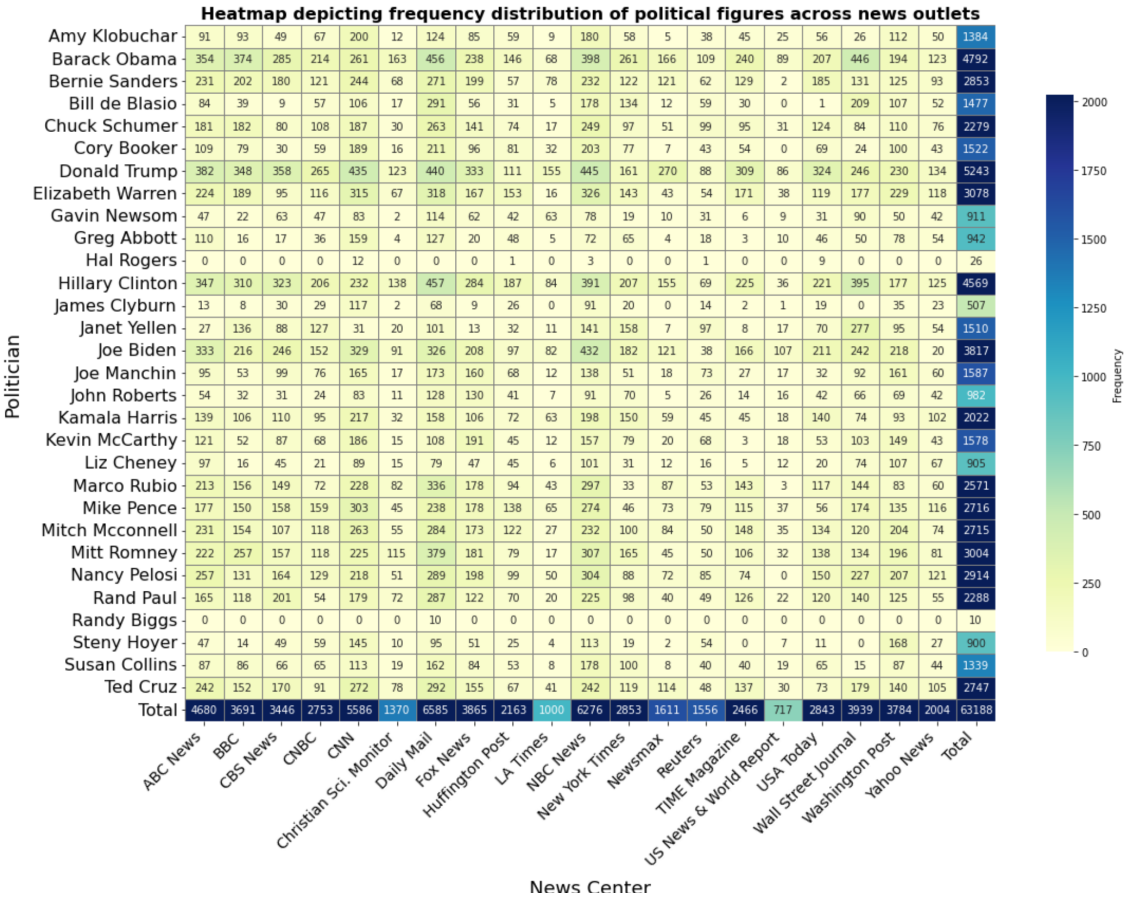


Figure A.3: Frequency Distribution of Political Figures Across News Outlets

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_r) and Extraction Error (ε_e).

Remark 1. *Measurement Error (ε_r) 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:*

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

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:*

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

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:*** *Let (Y,X) denote observed regressors, where Y is the variable of interest and X captures other explanatory variables, ensuring no omitted variable bias. Consider the model:*

$$\begin{aligned} &\hat{T}_{oracle} = T + \varepsilon_r, \quad where \quad \hat{T}_{oracle} = Human, \\ &T = \alpha + \beta Y + \gamma X + e \end{aligned} \tag{A.24}$$

where e and ε_r are i.i.d. errors, and $\mathrm{Cov}(T,\varepsilon_r) = 0$ (Remark 1). Suppose $\hat{\beta}$ is the OLS estimator of the parameter of interest (coefficient of Y). Then $\hat{\beta}$ is an unbiased estimator of β. As $n \to \infty$:

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0, \Sigma_\beta = \mathbf{e}_2^\top (A_0^{-1} \Sigma A_0^{-1}) \mathbf{e}_2),$$

where $\mathbf{Z} = (1,Y,X)^\top$, $A_0 = \mathbb{E}[\mathbf{Z}\mathbf{Z}^\top]$, $\Sigma = \mathbb{E}[\mathbf{Z}\theta\theta\mathbf{Z}^\top]$, $\theta = e + \varepsilon_r$, and $\mathbf{e}_2 = (0,1,0)^\top$.

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 linear model where Y is a binary variable representing the news outlet (1 for Fox News, 0 for CNN).

Let $\hat{T}_{oracle} = \alpha + \beta Y + \gamma X + \theta$, where $\theta = e + \varepsilon_r$. Define $\mathbf{Z} = (1,Y,X)^\top$ and stack the observations:

$$\hat{\mathbf{T}}_{oracle} = \mathbf{Z}\boldsymbol{\beta}_{\text{true}} + \boldsymbol{\theta}, \quad \boldsymbol{\beta}_{\text{true}} = (\alpha,\beta,\gamma)^\top.$$

Here, the OLS estimator is:

$$\hat{\boldsymbol{\beta}} = (\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top\hat{\mathbf{T}}_{oracle} = \boldsymbol{\beta}_{\text{true}} + (\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top\boldsymbol{\theta}.$$

Since $\mathrm{Cov}(T,\varepsilon_r) = 0$ (Remark 1), ε_r is uncorrelated with (Y,X). Therefore:

$$\mathbb{E}[\theta\mathbf{Z}] = \mathbb{E}[(e+\varepsilon_r)\mathbf{Z}] = \mathbf{0}.$$

Taking the expectation of $\hat{\boldsymbol{\beta}}$:

$$\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}_{\text{true}} + (\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbb{E}[\mathbf{Z}^\top\boldsymbol{\theta}] = \boldsymbol{\beta}_{\text{true}},$$

so $\hat{\boldsymbol{\beta}}$ is unbiased. In particular, $\hat{\beta}$ is an unbiased estimator of β. By the Law of Large Numbers (LLN) and Central Limit Theorem (CLT):

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_{\text{true}}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, A_0^{-1}\Sigma A_0^{-1}),$$

where:

$$A_0 = \mathbb{E}[\mathbf{Z}\mathbf{Z}^\top], \quad \Sigma = \mathbb{E}[\mathbf{Z}\boldsymbol{\theta}\boldsymbol{\theta}\mathbf{Z}^\top].$$

For $\hat{\beta}$, focusing on the second component:

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}\left(0, \mathbf{e}_2^\top A_0^{-1}\Sigma A_0^{-1}\mathbf{e}_2\right),$$

where $\mathbf{e}_2 = (0,1,0)^\top$. 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. □

The Proposition above confirms the reliability of using such models in well-specified scenarios.

Proposition 3. ***Partial-specified:*** *Let (Y,X) again denote observed regressors, where Y is the variable of interest and X captures other explanatory variables, ensuring no omitted variable bias. Consider the model:*

$$\begin{aligned} &\hat{T}_{ml} = T + \varepsilon_r + \varepsilon_e, \quad \textit{where} \quad \hat{T}_{ml} = f_1(Z), \\ &T = \alpha + \beta Y + \gamma X + e \end{aligned} \tag{A.25}$$

where e and ε_r are i.i.d. errors uncorrelated with (Y,X), and ε_e is an extraction error that is endogenous $\text{Cov}(T,\varepsilon_e) \neq 0$ (Remark 2). Suppose $\hat{\beta}$ is the OLS estimator of the parameter of interest (coefficient of Y). Then, $\hat{\beta}$ is a biased estimator due to the endogeneity introduced by ε_e. As $n \to \infty$:

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(\mu = \mathbf{e}_2^\top A_0^{-1}\mathbb{E}[\mathbf{Z}\varepsilon_e], \Sigma_\beta = \mathbf{e}_2^\top (A_0^{-1}\Sigma A_0^{-1})\mathbf{e}_2),$$

where $\mathbf{Z} = (1,Y,X)^\top$, $A_0 = \mathbb{E}[\mathbf{Z}\mathbf{Z}^\top]$, $\Sigma = \mathbb{E}[\mathbf{Z}\boldsymbol{\theta}\boldsymbol{\theta}\mathbf{Z}^\top]$, $\theta = e + \varepsilon_r + \varepsilon_e$, and $\mathbf{e}_2 = (0,1,0)^\top$.

Proof. In real-world scenarios, $\hat{T}$ often includes extraction errors, leading to biases.

Let $\hat{T}_{ml} = \alpha + \beta Y + \gamma X + \theta$, where $\theta = e + \varepsilon_r + \varepsilon_e$. Define $\mathbf{Z} = (1,Y,X)^\top$ and stack the observations:

$$\hat{\mathbf{T}}_{ml} = \mathbf{Z}\boldsymbol{\beta}_{\text{true}} + \boldsymbol{\theta}, \quad \boldsymbol{\beta}_{\text{true}} = (\alpha,\beta,\gamma)^\top.$$

Here, the OLS estimator is:

$$\hat{\boldsymbol{\beta}} = (\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top\hat{\mathbf{T}}_{ml} = \boldsymbol{\beta}_{\text{true}} + (\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top\boldsymbol{\theta}.$$

Since $\mathbb{E}[\boldsymbol{\theta}] \neq 0$ due to $\text{Cov}(T,\varepsilon_e) \neq 0$ (Remark 2), we have:

$$\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}_{\text{true}} + (\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbb{E}[\mathbf{Z}^\top\boldsymbol{\theta}] \neq \boldsymbol{\beta}_{\text{true}}.$$

By the LLN and CLT:

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_{\text{true}}) \xrightarrow{d} \mathcal{N}(A_0^{-1}b_0, A_0^{-1}\Sigma A_0^{-1}),$$

where $A_0 = \mathbb{E}[\mathbf{Z}\mathbf{Z}^\top]$, $b_0 = \mathbb{E}[\mathbf{Z}\boldsymbol{\theta}]$ and $\Sigma = \mathbb{E}[\mathbf{Z}\boldsymbol{\theta}\boldsymbol{\theta}\mathbf{Z}^\top]$.

To isolate the second component (β) in $\hat{\boldsymbol{\beta}}$, let $\mathbf{e}_2 = (0,1,0)^\top$. Then:

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(\mu = \mathbf{e}_2^\top A_0^{-1}b_0, \Sigma_\beta = \mathbf{e}_2^\top A_0^{-1}\Sigma A_0^{-1}\mathbf{e}_2),$$

where:

$$A_0^{-1} b_0 \;=\; \begin{pmatrix} a_{11}\,\mathbb{E}[\varepsilon_e] \;+\; a_{12}\,\mathbb{E}[Y\,\varepsilon_e] \;+\; a_{13}\,\mathbb{E}[X\,\varepsilon_e] \\ a_{21}\,\mathbb{E}[\varepsilon_e] \;+\; a_{22}\,\mathbb{E}[Y\,\varepsilon_e] \;+\; a_{23}\,\mathbb{E}[X\,\varepsilon_e] \\ a_{31}\,\mathbb{E}[\varepsilon_e] \;+\; a_{32}\,\mathbb{E}[Y\,\varepsilon_e] \;+\; a_{33}\,\mathbb{E}[X\,\varepsilon_e] \end{pmatrix}.$$

Since $\mathbf{e}_2^\top = (0,\ 1,\ 0)$, it follows that

$$\mu \;=\; \mathbf{e}_2^\top \left(A_0^{-1} b_0\right) \;=\; a_{21}\,\mathbb{E}[\varepsilon_e] \;+\; a_{22}\,\mathbb{E}[Y\,\varepsilon_e] \;+\; a_{23}\,\mathbb{E}[X\,\varepsilon_e].$$

From Remark 2, we know:

$$\mathrm{Cov}(T,\varepsilon_e) = \beta \mathrm{Cov}(Y,\varepsilon_e) + \gamma \mathrm{Cov}(X,\varepsilon_e) \neq 0.$$

This implies that at least one of $\mathrm{Cov}(Y,\varepsilon_e)$ or $\mathrm{Cov}(X,\varepsilon_e)$ must be nonzero. Consequently, at least one of $\mathbb{E}[Y\varepsilon_e]$ or $\mathbb{E}[X\varepsilon_e]$ must also be nonzero. Therefore, $\mu = a_{21}\mathbb{E}[\varepsilon_e] + a_{22}\mathbb{E}[Y\varepsilon_e] + a_{23}\mathbb{E}[X\varepsilon_e] \neq 0$. Thus, $\hat{\beta}$ is asymptotically biased.

□

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.

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:*

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

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

$$\begin{aligned} T &= \beta Y + \delta k(Z^{(-T)}) + \varepsilon_0 \quad \textit{(True Model)}, \\ T &= \tilde{\beta} Y + \varepsilon \quad \textit{(Observed Model)}, \end{aligned} \tag{A.26}$$

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:

$$Bias(\hat{\tilde{\beta}}) = \tilde{\beta} - \beta = \lambda\,\delta, \tag{A.27}$$

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:

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

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

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

This simplifies to:

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

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 (Conservative 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	*Dependent Variable: Residual*	
	(1)	(2)
Intercept	0.1097***	0.0991**
	(0.0081)	(0.035)
Label	-0.4758***	-0.4817***
	(0.0115)	(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***
Note:	†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 Counterfactual Image Generation

E.1.1 Validating Generated Counterfactual Images

In this part, we explain generating counterfactual images. First, for each politician p, we select three images $\tilde{Z}^0_{p1}, \tilde{Z}^0_{p2}, \tilde{Z}^0_{p3}$ 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 proposed by Mirza and Osindero (2014), we generate a smiley version of each image while keeping other factors constant. In other words, we apply the transformation π^1 on images $\tilde{Z}^0_{p1}, \tilde{Z}^0_{p2}, \tilde{Z}^0_{p3}$ to obtain $\pi^1(\tilde{Z}^0_{p1}) = \tilde{Z}^1_{p1}, \pi^1(\tilde{Z}^0_{p2}) = \tilde{Z}^1_{p2}, \pi^1(\tilde{Z}^0_{p3}) = \tilde{Z}^1_{p3}$. Each $\tilde{Z}^1$ now embodies the slant notion, specifically a smile, as shown in Figure 3.

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.4:

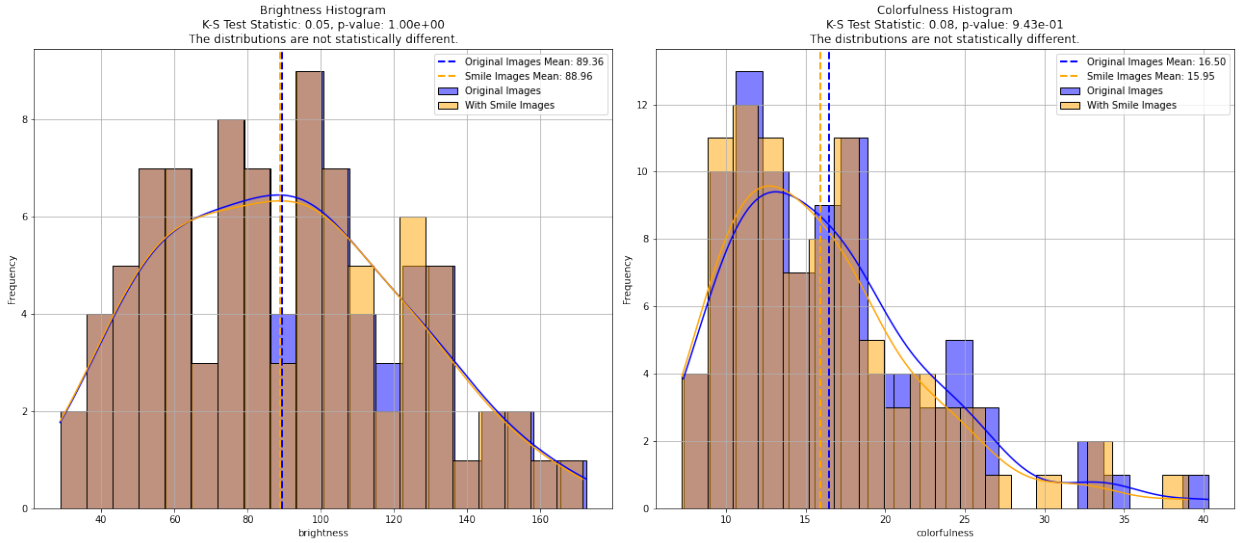


Figure A.4: Histograms of brightness and colorfulness for original and smiley images.

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

First, to fully exploit the presence of similar events, we use the contextual information in articles. The contextual information comprises textual data X^{text} from article titles, publication date X^{date}, politicians' names P, and political affiliations P^{aff}, and image data Z. We want our model architecture to fully capture the clustering structure of the contextual information for two reasons. First, identifying clusters with diverse news outlets allows for identification of 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 5 offers a visual overview of the model. To extract meaningful patterns in textual data X^{text}, 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 40-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 X^{date}, politicians' names P, and political affiliations P^{aff}, 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 X^{date}, 8 for the politicians' names P, and 2 for political affiliations P^{aff}. 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 P^{aff}) 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:

$$\textbf{Entropy:} \quad \max_{\theta} \mathscr{H}(\theta) = \max_{\theta} \left(-\sum_{i=1}^{\mathscr{N}} \sum_{y \in \mathscr{Y}} g(Y_i = y \mid Z_i, X_i, P_i; \theta) \log g(Y_i = y \mid Z_i, X_i, P_i; \theta) \right) \tag{A.28}$$

The resulting model is what we use to estimate the polarization parameter using Equation (19).

E.2.1 Textual Data Processing using LDA

Textual data X^{text} 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 arti-

cles.

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)

for each topic k in $\{1,\ldots,K\}$ **do**
    Draw a distribution over words $\phi_k \sim \text{Dir}(\beta)$
end for
for each document d in $\{1,\ldots,M\}$ **do**
    Draw a distribution over topics $\theta_d \sim \text{Dir}(\alpha)$
    for each word n in $\{1,\ldots,N_d\}$ **do**
        Draw a topic $z_{dn} \sim \text{Multinomial}(\theta_d)$
        Draw a word $w_{dn} \sim \text{Multinomial}(\phi_{z_{dn}})$
    end for
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 40, and the model is trained with 40 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 (Řehůřek 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 40-dimensional vector representation for each news article title.

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:

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

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

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

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

$$(\mathbf{R}_{\text{final}}, \mathbf{L}) = f_3(\mathbf{R}_2; \boldsymbol{\theta}_3),$$

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 $g_{\text{VGG}}(\cdot; \boldsymbol{\Theta}_{\text{VGG}})$ denote the Face-VGG model, where $\boldsymbol{\Theta}_{\text{VGG}}$ represents its parameters. The extracted feature vector $\mathbf{v} \in \mathbb{R}^{512}$ is computed as:

$$\mathbf{v} = g_{\text{VGG}}(\mathbf{F}; \boldsymbol{\Theta}_{\text{VGG}}).$$

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 $\mathbf{v}_{\text{red}}$. The normalized and regularized feature vector $\mathbf{v}_{\text{norm}}$ 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:

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

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 $\mathbf{m}$:

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

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 $\mathbf{v}_{\text{att}}$ is concatenated with features from the ResNet branch $\mathbf{f}_{\text{ResNet}}$ and the metadata branch $\mathbf{f}_{\text{meta}}$ 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 $P \in \mathbb{R}^{d_{\text{person}}}$, $P^{\text{aff}} \in \mathbb{R}^{d_{\text{party}}}$, and $X^{\text{date}} \in \mathbb{R}^{d_{\text{date}}}$, respectively, while the topic vector is $X^{\text{text}} \in \mathbb{R}^{d_{\text{topics}}}$, where $d_{\text{text}} = 40$. The structured input vector $\mathbf{x}$ is thus formed as:

$$\mathbf{x} = [P; P^{\text{aff}}; X^{\text{date}}; X^{\text{text}}],$$

where $[\cdot]$ 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:

$$\mathbf{O}_{\text{single}} = \mathbf{A}\mathbf{V}.$$

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 $i = 1, \ldots, h$ 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:

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

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 $g_{\text{ResNet}}(\cdot; \mathbf{\Theta}_{\text{ResNet}})$, where $\mathbf{\Theta}_{\text{ResNet}}$ 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:

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

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

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

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:

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

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.

E.2.6 Model Training and Fine-Tuning

To ensure robust model evaluation, the dataset is split into training (85%) and test (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 test data set is used for assessing model performance and polarization measurement to avoid overfitting bias.

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 $\mathscr{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N}$ 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:

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

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 (Szegedy et al., 2016). The optimizer updates the model parameters $\boldsymbol{\theta}$ at each iteration by minimizing the total loss $\mathscr{L}$. The parameter updates follow the AdamW rule:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \frac{\hat{\mathbf{m}}_t}{\sqrt{\hat{\mathbf{v}}_t} + \varepsilon} - \lambda \boldsymbol{\theta}_t,$$

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

Input: Learning rate η, decay rates β_1, β_2, weight decay coefficient λ, small constant ε
Initialize: $\mathbf{m}_0 = 0$, $\mathbf{v}_0 = 0$, $t = 0$
for each iteration **do**
    $t = t + 1$
    Compute gradients of the loss: $\mathbf{g}_t = \nabla_{\boldsymbol{\theta}} \mathscr{L}(\boldsymbol{\theta}_{t-1})$
    Update biased first moment estimate:

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

    Update biased second moment estimate:

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

    Compute bias-corrected first moment estimate:

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

    Compute bias-corrected second moment estimate:

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

    Update parameters:

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} - \eta \frac{\hat{\mathbf{m}}_t}{\sqrt{\hat{\mathbf{v}}_t} + \varepsilon} - \lambda \boldsymbol{\theta}_{t-1}$$

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 $\lambda \boldsymbol{\theta}_{t-1}$ 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 Detailed Multi-Modal ML Performance

In this section, we detail the performance of our multi-modal multi-class classification problem for news outlet prediction. As mentioned earlier, we use an 85%-15% split between training and test data. We consider four measures to evaluate the model's predictive performance – (1) Accuracy, (2) Precision, (3) Recall, and (4) Weighted Cross-Entropy (WCE). Accuracy is simply measured as:

$$\text{Accuracy} = \frac{\sum_{i=1}^{N} \mathbf{1}(Y_i = \text{argmax}_c \hat{Y}_{ic})}{N},$$

where $\hat{Y}_{ic}$ is our predicted probability for news outlet c producing article i. Precision evaluates the reliability of the model's predictions by measuring the proportion of correctly predicted articles for a given outlet

among all articles predicted for that outlet. Formally, it is defined as:

$$\text{Precision} = \frac{\sum_{i=1}^{N} \mathbf{1}(Y_i = c \wedge \text{argmax}_c \hat{Y}_{ic} = c)}{\sum_{i=1}^{N} \mathbf{1}(\text{argmax}_c \hat{Y}_{ic} = c)},$$

where $\mathbf{1}(Y_i = c \wedge \text{argmax}_c \hat{Y}_{ic} = c)$ counts the true positives (correct predictions for outlet c), and $\mathbf{1}(\text{argmax}_c \hat{Y}_{ic} = c)$ counts all predictions made for outlet c. Recall, on the other hand, measures the model's ability to identify all true articles for a given outlet. It is defined as:

$$\text{Recall} = \frac{\sum_{i=1}^{N} \mathbf{1}(Y_i = c \wedge \text{argmax}_c \hat{Y}_{ic} = c)}{\sum_{i=1}^{N} \mathbf{1}(Y_i = c)},$$

where $\mathbf{1}(Y_i = c)$ counts all true articles for outlet c. While Precision emphasizes minimizing false positives, Recall ensures that false negatives are minimized, making them complementary metrics to assess the model's performance comprehensively. Finally, WCE Loss is defined as:

$$\text{WCE Loss} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} w_c \left(Y_{ic}(1-\varepsilon) + \varepsilon/C\right) \log\left(\hat{Y}_{ic}\right),$$

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.

Table A4 provides information on the multi-modal model performance at each class level (news outlet) and presents the test 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.

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 $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{d'} \in \mathbb{R}^d$, where each component maximizes the variance of the projected data. The reduced embedding for an image is given by:

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

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

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: Predictive accuracy results on the test set for different news outlets (sorted by accuracy)

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 $\mathscr{E}^p = \{\mathbf{e}_i^{\text{PCA}} \,|\, i \in \mathscr{I}_p\}$, where $\mathscr{I}_p$ 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:

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

where $\mathscr{C}_j^p$ 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
	Republican Outlets Mean Test	-82.654572	0.0000	15864
Republican Politicians	K-S Test	0.672592	0.0000	24978
	Democratic Outlets Mean Test	-73.504981	0.0000	12489
	Republican 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.

F.4 Correlation Analysis for Existing Partisanship Scores

We extend our analysis to all 19 outlets, $y^k \in \mathscr{Y}$, to evaluate whether polarization patterns observed in §7.3.1 hold more broadly. Specifically, we calculate outlet-level CVS measurement for all outlets following Equation 22. To validate our measure, we compare CVS with existing benchmarks, focusing on the *conservative share score* by Faris et al. (2017), AllSides (2024), Flaxman et al. (2016). Figures A.5, A.6, and A.7 visualize these relationships alongside correlation analysis using Pearson and Spearman.

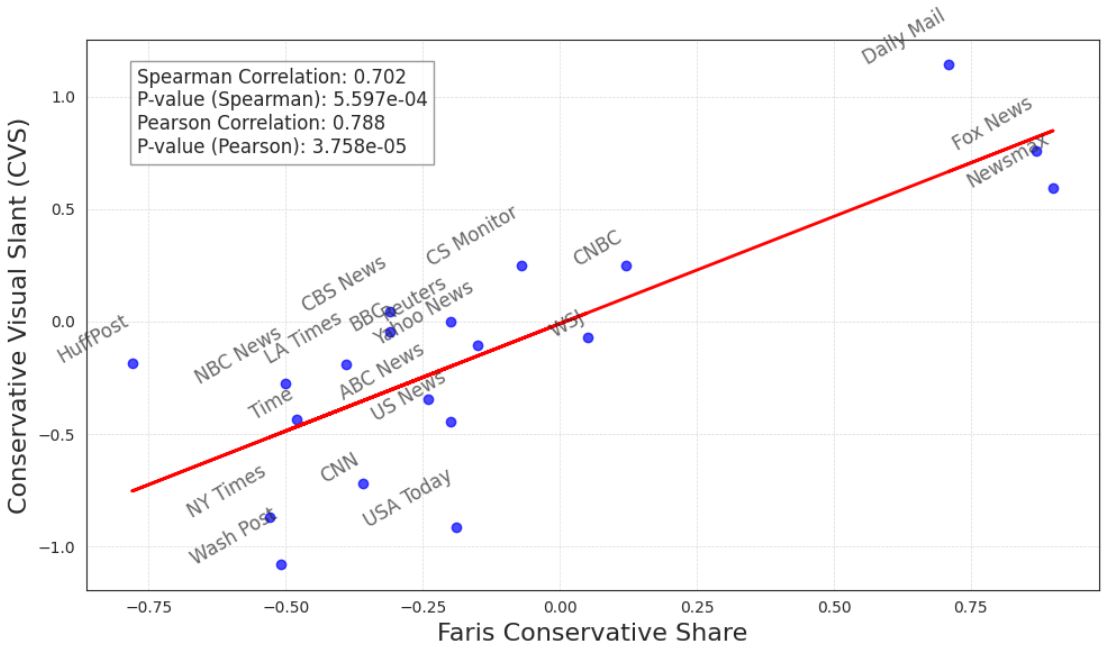


Figure A.5: Correlation between *CVS* and the conservative share score of news outlets. (Based on Faris et al. (2017) Conservative Score)

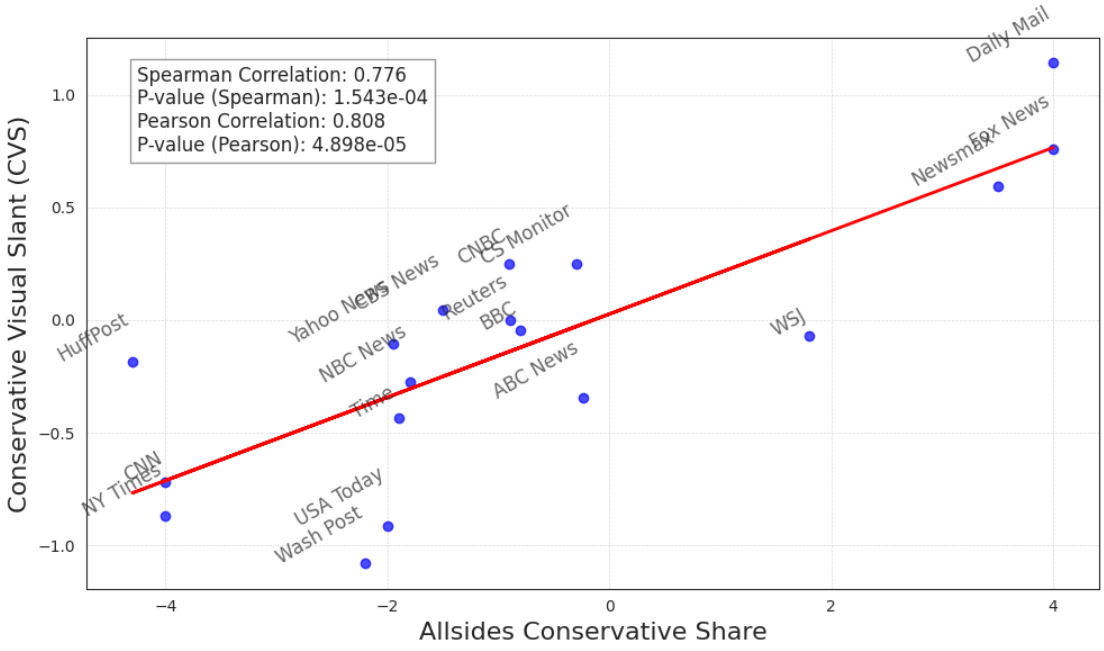


Figure A.6: Correlation between *CVS* and the conservative share score of news outlets. (Based on AllSides (2024) Conservative Score)

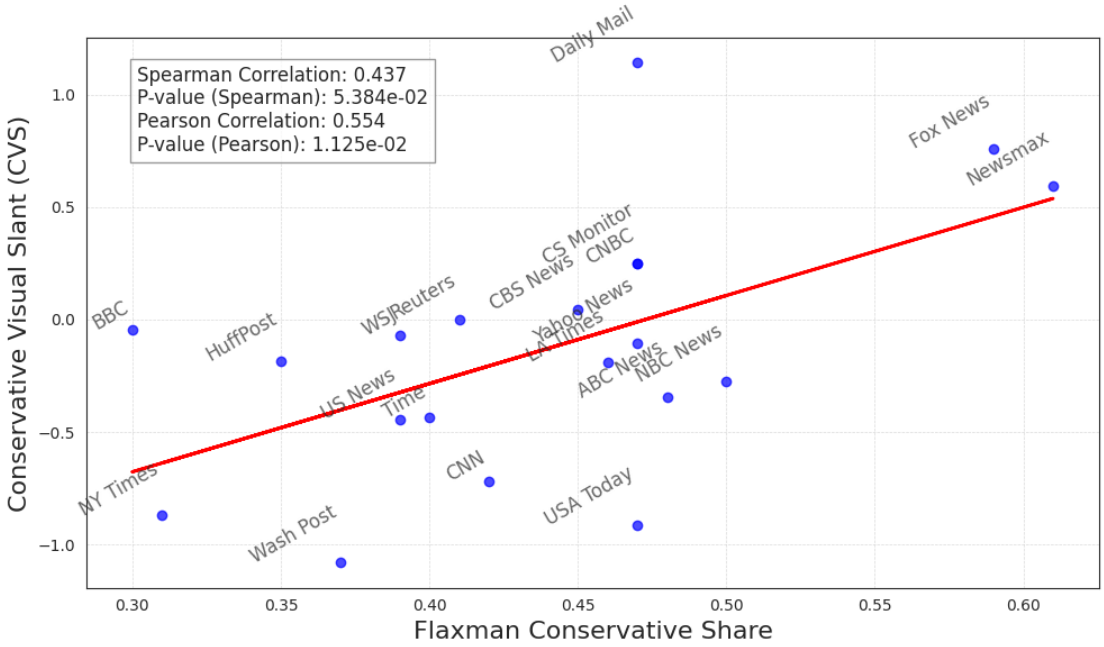


Figure A.7: Correlation between *CVS* and the conservative share score of news outlets. (Based on Flaxman et al. (2016) Conservative Score)

Across all three benchmarks, we observe a significant positive correlation between the conservative share scores and our CVS measure. The strength of correlation varies across sources, with AllSides (2024) showing the highest Spearman ($\rho = 0.776$, $p < 0.001$) and Pearson ($r = 0.808$, $p < 0.001$) correlations, followed by Flaxman et al. (2016) ($\rho = 0.702$, $r = 0.788$), and Faris et al. (2017) showing the weakest but still statistically significant association ($\rho = 0.437$, $r = 0.554$). These results support the validity of CVS as an independent measure of conservative visual slant, aligning well with established textual and audience-based conservative share scores.

F.5 Comparison of Slant Visual from PMCIG and Two-stage Approach

In this section, we present a comparison of our proposed method, PMCIG, with the established two-stage approach for analyzing slant in visual content. For this purpose, we collect the overall visual slant scores from Boxell (2021), which are calculated as the difference between the relative favorability towards Republicans minus Democrats (similar to CSV we present in 22, but using the two-stage approach).

Our dataset and Boxell (2021) share 11 common news outlets; therefore, the comparison focuses on these shared outlets. To evaluate the performance, we use both Pearson correlation analysis and Spearman rank correlation analysis. Each of these analyses is conducted by comparing the respective methods (PMCIG and Boxell) against three existing conservative share measures, providing an assessment of the alignment and ranking accuracy across approaches. Figures A.8 A.9 A.10 summarize our findings, providing statistical insights into the relative effectiveness of each method.

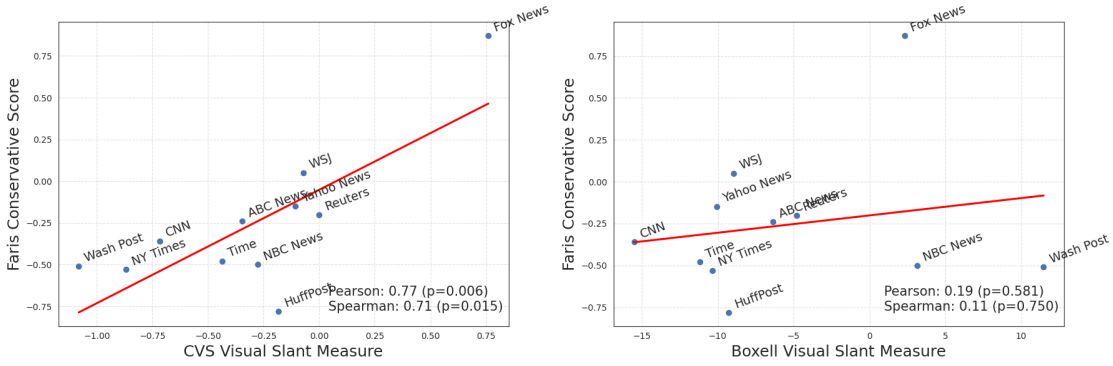


Figure A.8: Comparison of visual slant from our PMCIG method (left) and Boxell (2021) approach (right) against the conservative score from Faris et al. (2017).

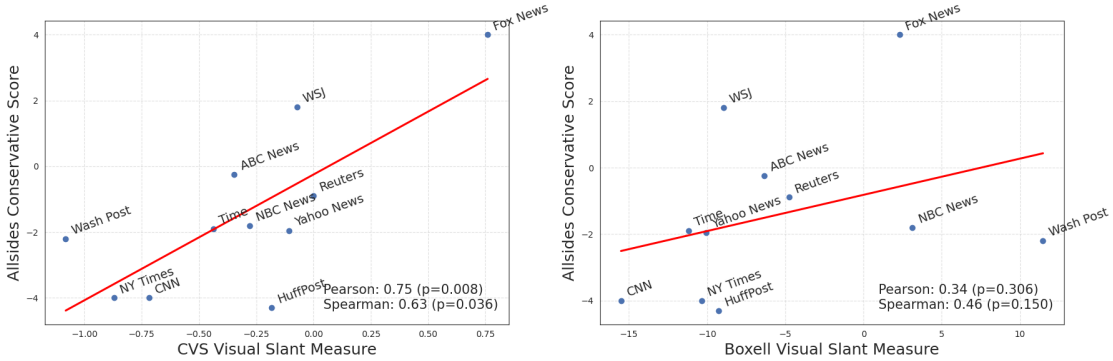


Figure A.9: Comparison of visual slant from our PMCIG method (left) and Boxell (2021) approach (right) against the conservative score from AllSides (2024).

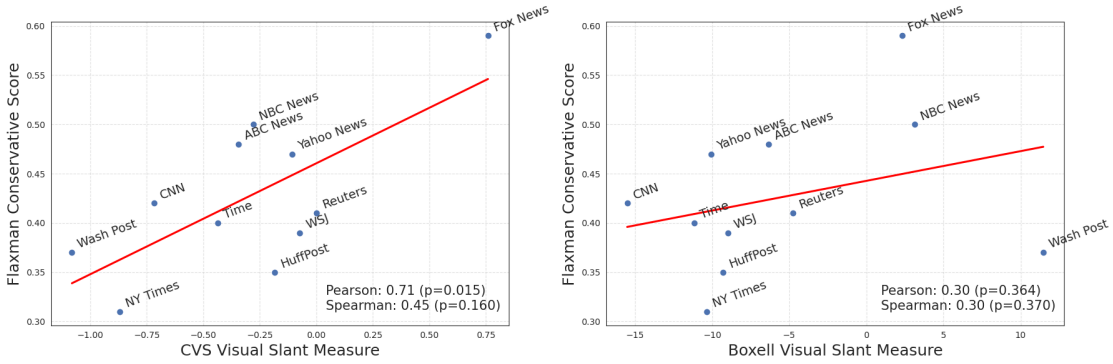


Figure A.10: Comparison of visual slant from our PMCIG method (left) and Boxell (2021) approach (right) against the conservative score from Flaxman et al. (2016).

The results clearly show that PMCIG outperforms Boxell (2021) two-stage approach in measuring visual slant. Across all three benchmarks, PMCIG achieves higher Pearson and Spearman correlations, demonstrating stronger alignment with established conservative share measures. Additionally, PMCIG's correlations are statistically significant, with lower p-values, whereas Boxell (2021) approach often produces weaker

and less reliable correlations. PMCIG also excels in ranking accuracy, as reflected in its consistently higher Spearman rank correlations, confirming that it orders news outlets' visual slant more accurately.

F.6 Individual News Outlet Results

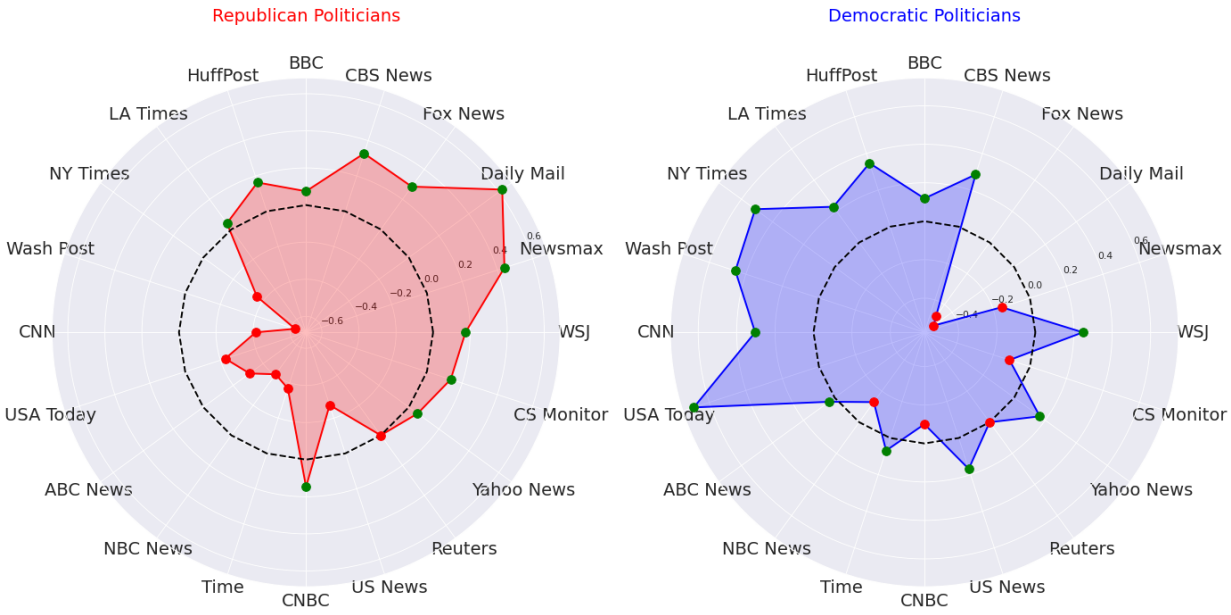


Figure A.11: Radar plots showing the mean polarization measurement for Republican and Democratic politicians across various news outlets. *Reuters* polarization is zero since it serves as the baseline.

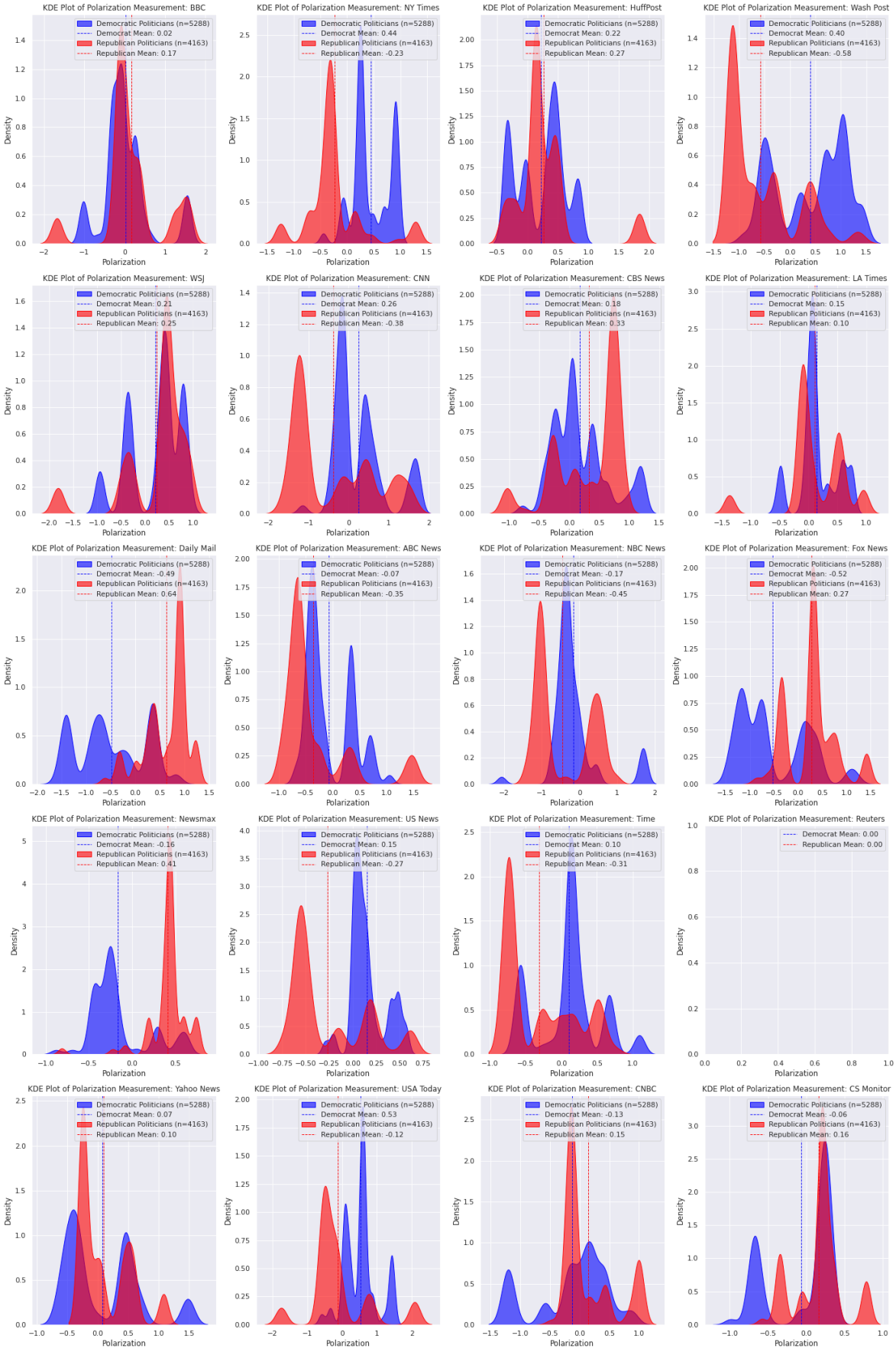


Figure A.12: Histograms of polarization for Democratic and Republican Politicians in each News Outlet

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 A6: Statistical Analysis of Polarization by News Center and Politician Side

F.7 Individual Politicians Results

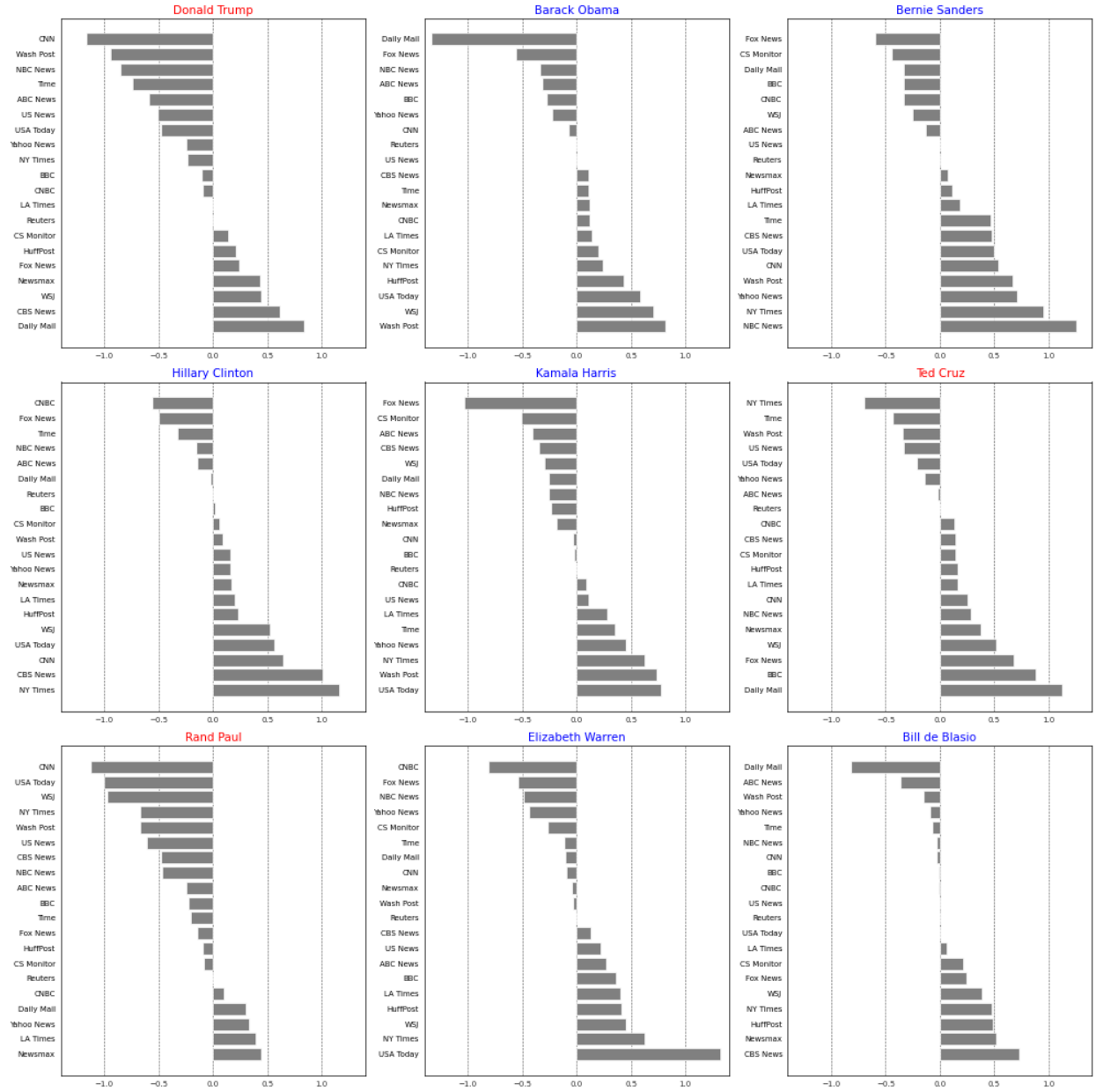


Figure A.13: Bar plot showcasing the polarization across various Politicians, segmented by different News Outlets, Part I

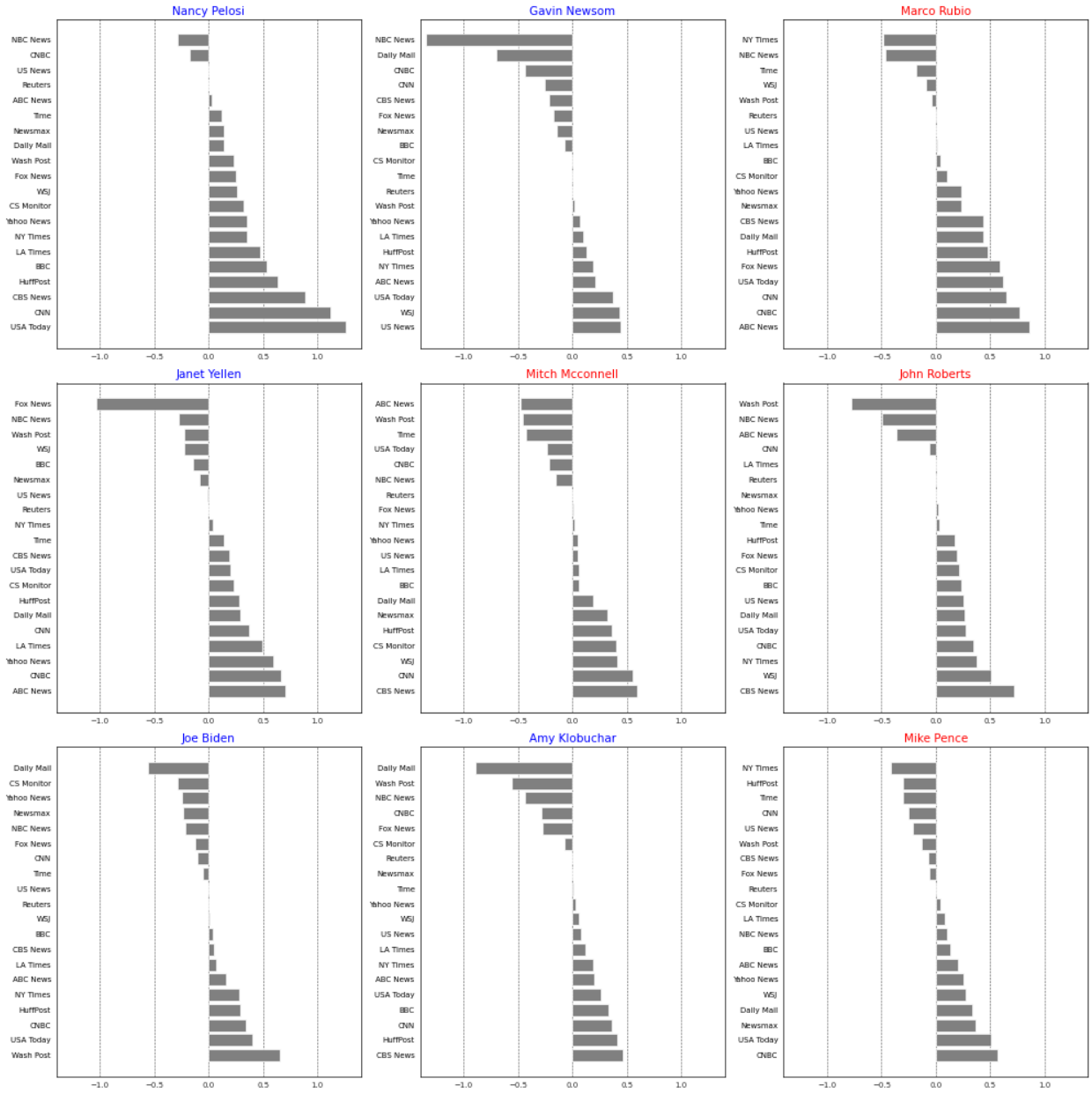


Figure A.14: Bar plot showcasing the polarization across various Politicians, segmented by different News Outlets, Part II

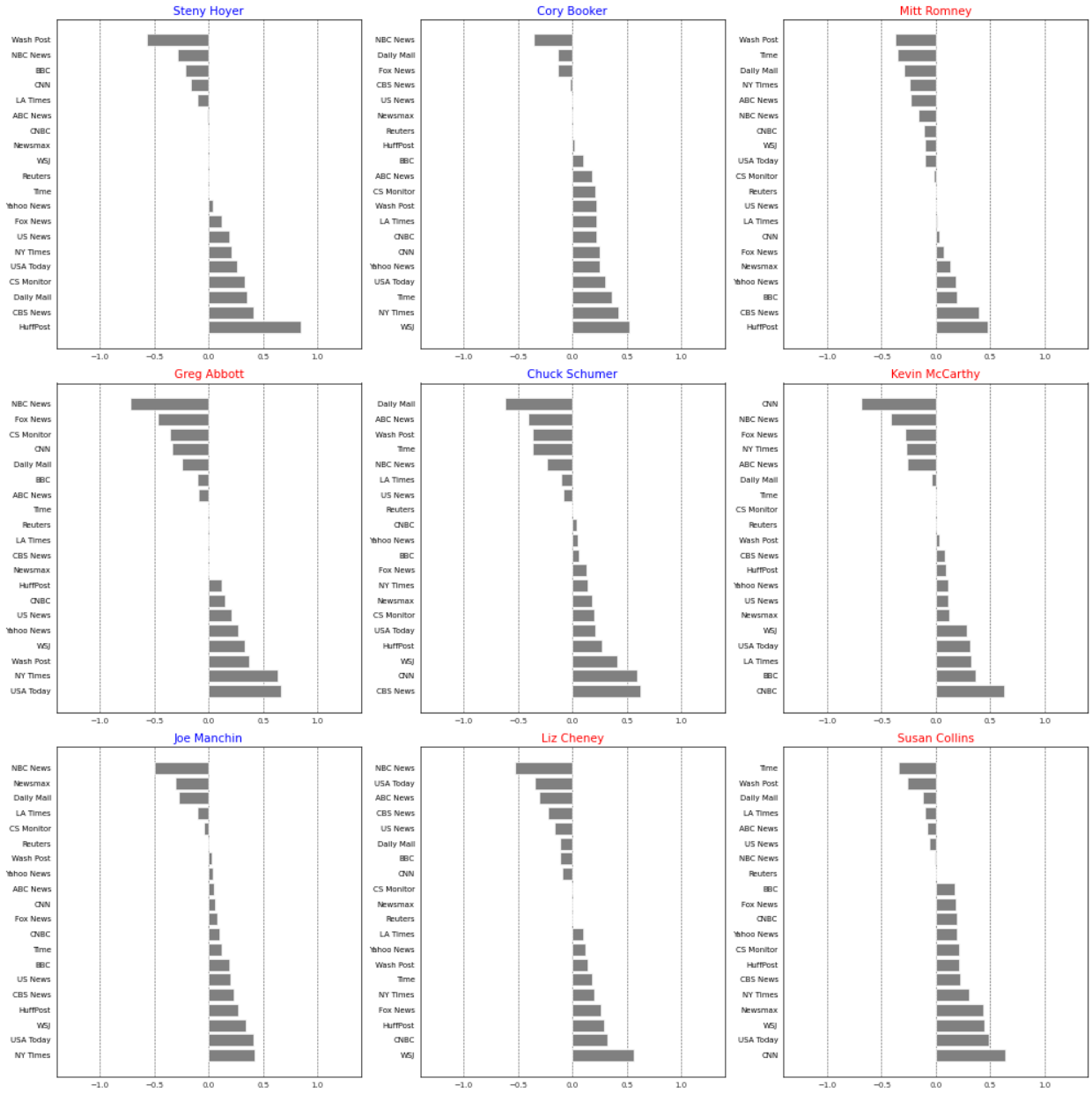


Figure A.15: Bar plot showcasing the polarization across various Politicians, segmented by different News Outlets, Part III

F.8 Ideological Partisanship in States and the Polarization of Politicians

In this section, we examine the relationship between state-level ideological partisanship and the visual polarization of politicians. Our goal is to understand whether politicians from states with strong partisan leanings experience greater or lesser media polarization compared to those from more ideologically mixed states. To measure ideological partisanship at the state level, we rely on the presidential vote margin in the 2016 election—specifically, the difference between the percentage of votes received by the politician's party and the opposing party. This vote difference serves as a proxy for how politically secure or competitive a politician's home state is. We then regress each politician's OVP from Equation 23 on this vote difference to assess whether electoral environment influences how divisive a politician's visual representation is across

media outlets. Figure 15 presents the results of this analysis, revealing a statistically significant positive relationship.

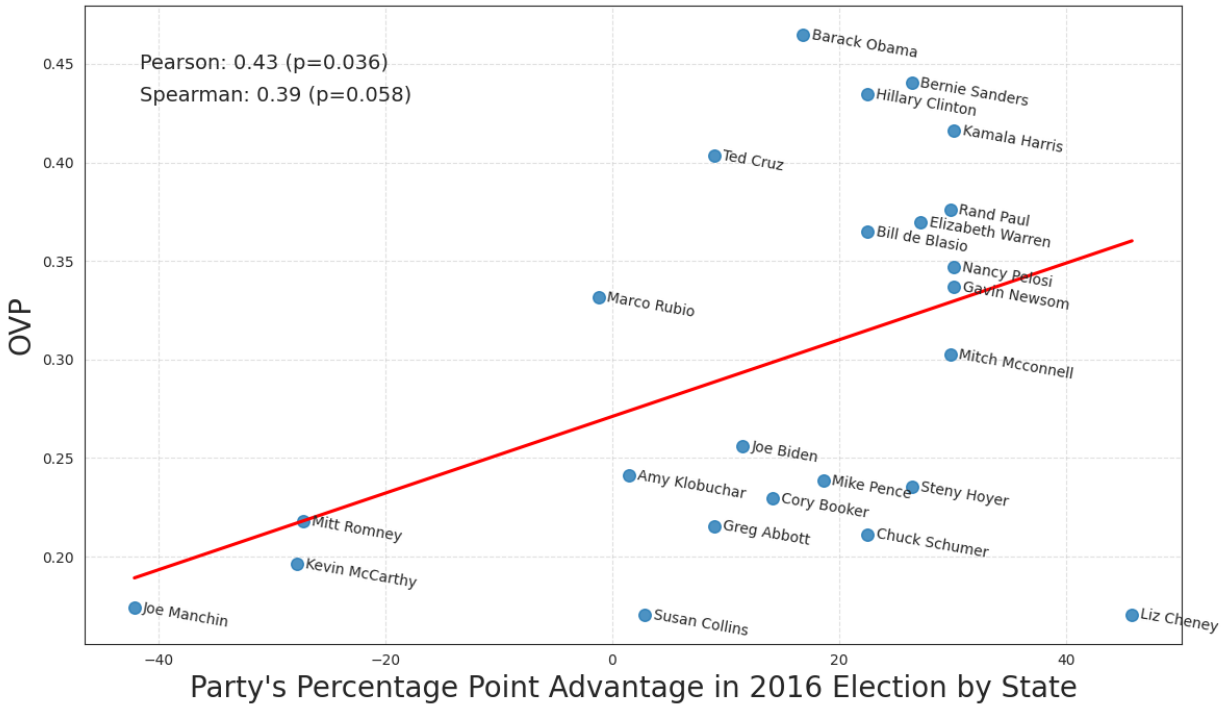


Figure A.16: Relationship Between State Partisanship and Politician's Visual Polarization

This finding aligns with the idea that politicians from safe states enjoy greater ideological freedom, enabling them to take stronger partisan stances without the need to appeal to a broad electorate. Consequently, they may attract more partisan media portrayals, reinforcing their image as highly divisive figures. For instance, Bernie Sanders (Vermont) and Ted Cruz (Texas), both from states with strong partisan identities, exhibit high OVP values, likely reflecting their strong ideological positions and the way they are framed by different media outlets. In contrast, Joe Manchin (West Virginia) and Susan Collins (Maine), who represent states where their party is in the minority, display lower OVP values, suggesting that politicians from ideologically mixed states receive more moderate media portrayals.

Overall, these results suggest that state-level ideological partisanship influences how politicians are visually portrayed in the media, with those from politically homogeneous states being more polarizing figures.