AN LLM-AIDED MEASURE OF MEDIA SLANT

Mustafa Alam*

Last Updated: July 2025

Abstract

I present a method for measuring media slant using social media text data. Posts from political figures and media outlets are collected and processed through a large language model to extract "frames"—standardized declarative claims that encapsulate the core message of each post. These frames are embedded into a shared semantic space and clustered using a hierarchical density-based clustering algorithm, representing broadly held positions among politicians and media firms. For each account, vectors are constructed to represent their post frequency across each cluster. I separate the vectors for politicians and use party labels to train a classifier model that predicts political affiliation. The trained model is then applied to the vectors of media outlets, generating probabilistic scores that position each outlet on a partisan spectrum. This method contributes to the literature on measuring media slant using modern machine-learning tools to better discern semantic patterns in language and allows for granular separation of partisan positions and a nuanced measure of partisan slant without sacrificing interpretability.

*Department of Economics, 312J Wilbur O. and Ann Powers College of Business, Clemson University, Clemson, South Carolina 29634-1309, USA; email: mustafa@g.clemson.edu.

1 Introduction

Political slant in news media can shape public perception (Benedictis-Kessner et al., 2019), contribute to political polarization (Alam, 2025), and influence political and economic decision-making (Iyengar et al., 2019; McConnell et al., 2018). As a result, developing an intuitive, replicable, and relatively objective measure of media slant is essential for any empirical analyses aimed at understanding its drivers and societal consequences. In this paper, I contribute to this effort by presenting a methodology for measuring media slant from text data, leveraging the semantic representation capabilities of modern large language models (LLMs) to capture ideological positioning with high precision while preserving the interpretability and replicability offered by prominent automated approaches in economics.

Given the significant implications, efforts to quantify media slant span multiple disciplines, including computer science, political science, and economics. In computer science, researchers often employ natural language processing (NLP) and machine learning techniques to detect bias in news content. While these approaches are fast, automated, and scalable, they often suffer from limited generalizability because models are fine-tuned to specific training datasets. As a result, their predictions can reflect biases in the labeled data, and their effectiveness may degrade when applied to different news contexts or time periods[1]. In contrast, scholars in political science and communication studies have focused on carefully defining ways in which media bias may manifest in political communication to facilitate more general detection strategies. However, these methods often depend on expert judgments or manual annotations, which can introduce interpretive bias and limit their applicability to large-scale news content[2].

Within economics, media slant detection approaches can be broadly classified into explicit and implicit measures of slant. Studies of explicit media slant typically examine overt ideological/partisan endorsements, such as newspaper endorsements in political campaigns or editorial stances on policy issues (e.g., Ansolabehere et al. (2006); Ho and Quinn (2008); Puglisi and Snyder (2015)). These approaches offer direct signals of media slant but are inherently limited in scope—they capture only the most deliberate forms of slant while overlooking subtler ideological positioning embedded in news coverage. Implicit slant measures address these limitations by analyzing how media outlets report the news rather than what they explicitly endorse. This can be done by looking at whether media firms cite think tanks that are disproportionately cited by politicians belonging to a particular party (Groseclose and Milyo, 2005), the amount of coverage they give to different politically relevant topics (Puglisi (2011); Puglisi and Snyder (2011); Larcinese et al. (2011), etc), and the "tone" of their coverage (Gentzkow and Shapiro (2006); Soroka (2012); Lott and Hassett

[1]For a comprehensive discussion, see Hamborg et al. (2019).
[2]For an overview, see Groeling (2013).

(2013), etc). However, like other social science approaches, these methods often rely on substantial human input or specific datasets, which can limit generalizability or introduce subjectivity.

One prominent automated approach to measuring implicit media slant relies on phrase frequency differences between Democratic and Republican representatives in Congress (Gentzkow and Shapiro, 2010). This method captures partisan distinctions by identifying cases where Democrats and Republicans systematically use different terminology for the same issue or emphasize different topics altogether. For example, if Republican politicians predominantly refer to the inheritance tax as the "death tax" while Democrats call it the "estate tax," media outlets that disproportionately use "death tax" may be classified as Republican-leaning. Similarly, certain policy areas are discussed far more frequently by one party than the other. If Democratic politicians mention "climate change" at much higher rates than their Republican counterparts, media sources that frequently use the term may be classified as Democrat-leaning. This method offers a clear and scalable approach to detecting partisan alignment in text, making it particularly effective for large-scale media analyses.

Yet, its reliance on phrase frequency alone overlooks how ideological positions are expressed within a broader linguistic structure. If a media outlet mentions "climate change" frequently but primarily with a skeptical or critical narrative, the phrase-frequency method may misclassify the outlet as Democrat-leaning. Likewise, phrase-based methods may mistake ideological conflict for neutrality when both parties use the same politically charged phrase (e.g., "Roe v. Wade") but emphasize different aspects or draw opposing conclusions. Additionally, this approach assumes that media outlets explicitly 'talk like' politicians in Congress, potentially underestimating media slant when outlets adopt more informal, colloquial, or stylistically distinct language that signals partisan alignment but does not directly mirror Congressional speech patterns. Furthermore, the lack of broader context in this approach may lead to misidentification of phrase choices by local media as ideological positioning. For instance, Gentzkow et al. (2019) find that the bigram 'san antonio' is indicative of partisanship in the 114th Congress, possibly due to a political event or policy debate at the time. However, for a news channel based near San Antonio, TX, its frequent usage may reflect geographic relevance rather than an ideological slant.

The method presented in this paper follows a comparison approach to measuring implicit media slant, similar to phrase-frequency methods, but improves upon their limitations by incorporating broader linguistic context. Instead of relying on exact phrasing, this approach positions text in a shared semantic space using a transformer-based embedding model, allowing for a more flexible and nuanced measure of ideological alignment. By capturing semantic similarity rather than word co-occurrences, this method can differentiate between statements that contain the same phrase but convey distinct claims. Conversely, it can also identify similarities in posts that express the same claim but use distinct phrasing. Overall, the method identifies ideological positioning based on the content of the claim rather than its phrasing, allowing for a more granular separation of partisan positions in text data and reducing misclassification concerns due to lack of context.

I first collect social media (Twitter/X) posts from politicians and media outlets and extract standardized *frames* from their text content. Frames are a set of declarative statements that capture the core messages of each post, allowing for a structured comparison of claims while abstracting away from specific wording. I use an LLM (GPT-4o) to generate these frames, which reformulates posts into standardized claims while preserving their substantive meaning. This ensures that the analysis captures the ideological positions expressed by political figures and media outlets, independent of their linguistic style and phrasing choice.

Once frames are extracted, they are embedded in a high-dimensional semantic space using a transformer-based text embedding model (OpenAI's Ada) and clustered based on semantic similarity. The embedding step converts frames into continuous vector representations, allowing for precise comparisons between them. These embeddings are then clustered using a density-based unsupervised clustering algorithm that identifies broadly held ideological positions by grouping frames with similar meanings. Unlike traditional clustering methods (e.g., k-means), density-based methods do not require a predefined number of clusters or specification of cluster size and adapt to the density of the data, making them ideal for capturing the organic composition of claim categories in political discourse[3].

To assign an ideological property to the clusters, I construct engagement frequency vectors for politicians, capturing the fraction of their posts that appear in each cluster. These vectors serve as structured representations of how politicians engage with different narratives, allowing for a data-driven approach to identifying ideological positioning. To ensure that only politically meaningful clusters are used, I exclude clusters dominated by a single politician and those that contain only media firms. Clusters dominated by a single politician tend to include localized campaign messaging, event promotions, and personal endorsements, which are not useful for identifying broader ideological patterns because they reflect individual strategic communication rather than ideological alignment. Similarly, clusters consisting only of media firms often focus on topics such as general news, sports, or entertainment, which by definition are not part of the political discourse politicians engage in.

Using these frequency vectors, I train an XGBoost classifier to predict party affiliation. XGBoost is well-suited for this task because it effectively handles structured numerical data and can model nonlinear relationships in engagement patterns. Regularization is applied to prevent overfitting, ensuring the model captures consistent ideological patterns rather than transient variations in engagement. To validate the classifier, I use a random label shuffling test, where I evaluate the

[3]The frame extraction and clustering process was introduced by Ehrett (2024) as part of a broader pipeline for detecting co-ordinated information operations from social media content.

trained model on data with randomly reassigned party labels. If the classifier performs well even when tested on shuffled labels, it suggests that it is overfitting to spurious patterns rather than learning real ideological distinctions. By adjusting hyperparameters so that the model performs poorly on shuffled data but retains predictive accuracy on real labels, I ensure that the classification reflects genuine partisan engagement.

Finally, I apply the trained classifier to media outlets, assigning probabilistic ideological scores based on their participation in the partisan clusters. Media firms that predominantly engage in clusters associated with Republican/Democrat politicians receive corresponding ideological scores, while those with more balanced engagement patterns are positioned in a more central ideological position. This approach provides a continuous measure of ideological alignment, distinguishing between strongly partisan and cross-partisan media engagement. I validate the scores against an established media slant database (AllSides) and show that they align with expected ideological classifications.

This paper demonstrates how large language models enable automated identification of ideological positioning in media content by incorporating broader linguistic structure than phrase-frequency methods while preserving their comparison-based framework. Like NLP-based classification approaches, it extracts deeper semantic information from text, but unlike them, it does not rely on fine-tuned models based on specific training datasets, ensuring adaptability across political contexts and time periods. By grounding ideological classification in direct comparisons to contemporary political dialogues, this method remains robust as political narratives evolve. Moreover, as clustering algorithms and text representations by LLMs improve, the approach will identify ideological positioning with greater precision. Beyond media outlets, this framework can be extended to analyze ideological alignment of political influencers, commentators, and other digital actors. Future work could enhance this approach by incorporating multimodal content such as images, memes, and videos to capture a more comprehensive view of political messaging.

2 Data

2.1 Twitter/X Posts from Politicians and Media

I manually selected twenty-five Democratic politicians and twenty-five Republican politicians on the social media platform X, focusing on the time period January 1st, 2024 to December 31st, 2024. The selection was based on their activity levels on the platform and the diversity of their viewpoints, ensuring that major Democratic and Republican talking points of the period were represented. I collected all their social media posts from the social media listening platform Sprinklr.

For 2024, I collected a total of 80,883 posts and sampled 25,000 posts from each party for analysis.

2.2 Social Media Posts from Media Firms

In addition to collecting posts from politicians, I gathered data from a diverse range of media organizations on the social media platform X for the same time period. I collected posts from *Breitbart News Fox News Politics*, *HuffPost*, *MSNBC*, *The Daily Wire*, *New York Post Opinion*, *The New York Times*, *The New York Times Opinion* and *The Wall Street Journal*.

All posts from these media organizations were collected using the social media listening platform Sprinklr. I sampled 5,000 posts from each media outlet.

2.3 Verification Data

To validate the ideological scores produced by this method, I compare them against media bias ratings from AllSides, which assigns ideological classifications to media outlets based on multiple assessment methods. These include expert editorial reviews, blind bias surveys where participants evaluate content without knowing the source, and aggregated input from individuals with varying political perspectives. Media outlets are categorized into five groups—Left, Lean Left, Center, Lean Right, and Right—and assigned a numerical bias score ranging from -6 (Left) to +6 (Right), with 0 representing Center. This numerical scale provides a finer-grained measure of ideological positioning, particularly for outlets that fall near classification boundaries, similar to the output of the automated approach presented in this paper.

3 Measuring Slant

For a fixed time period of analysis, there is a set of N_p politician accounts with one of two party labels in $\{R, D\}$. Let the set of all politician accounts be denoted $\mathscr{P} = \{p_1, p_2, p_3, ..., p_{N_p}\}$. Similarly, let the set of N_m media outlet accounts be denoted $\mathscr{M} = \{m_1, m_2, m_3, ..., m_{N_m}\}$. The set of all accounts is

$$\mathscr{A} = \mathscr{P} \bigcup \mathscr{M}.$$

Each account $a \in \mathscr{A}$ produces N_a social media posts. Let the set of N_a posts by a be denoted $D_a = \{d_1, d_2, d_3, ..., d_{N_a}\}$. The set of all posts by all accounts is

$$\mathscr{D} = \bigcup_{a \in \mathscr{A}} D_a.$$

3.1 Frame Extraction

Each post $d \in \mathscr{D}$ is processed using an LLM (GPT 4-o) to extract one or more frames that represent the core declarative claim(s) of the post. Let F_d denote the set of frames extracted from post d. The set of all frames extracted from all posts is

$$\mathscr{F} = \bigcup_{d \in \mathscr{D}} F_d.$$

The *frame extraction* process is a mapping,

$$G : \mathscr{D} \to 2^{\mathscr{F}}.$$

The prompt used for the frame extraction process is as follows.

```
## CONTEXT
A "frame" is a factual or moral claim that has broad social significance.
This may include a political or social stance. Some social media posts
express a frame, and some do not. A post can express a frame without explicitly
stating it. A frame is always a declarative statement, with a subject,
verb, and object. Hashtags are not themselves frames, but often are used to
express a frame.

## INSTRUCTIONS
The following social media post is from the year 2024. This post is either
from a politican or a media outlet, You must list the primary frame(s) found
within it. Your response must present the frame as a python list of strings.
If no frame is expressed, return '["None"]'.

You must output only the python list containing the extracted frame,
with no other text preceding or following the list.
When referring to individuals, groups, or organizations, use full names
wherever possible.

## THE SOCIAL MEDIA POST:
```

Table 1 shows some examples from the frame extraction process.

Original Social Media Post	Extracted Frame
There is no bottom for Donald Trump. MAGA extremists at the MSG rally have called VP Harris “the Devil” and “the Antichrist” while likening her to a prostitute. Republicans must abandon the base and irresponsible rhetoric that endangers both American lives and institutions.	Republicans must reject extremist rhetoric that endangers American lives and institutions.
Schools across Israel, most of them divided along lines of religion and language, are struggling with how to help students cope since Oct 7. One Jerusalem school, where classes are taught together by Hebrew and Arabic speakers, is seeking unity.	Schools in Israel are struggling to support students in the aftermath of the October 7 events, highlighting the need for unity across religious and linguistic divides.
The Harris-Biden regime began repealing Trump-era border policies on day one. The Left is to blame for the criminals and terrorists roaming American communities. The Senate’s border insecurity bill would have codified mass parole while H.R. 2 collects dust on Schumer’s desk.	The Left is responsible for crime and terrorism in American communities due to the repeal of Trump-era border policies by the Harris-Biden administration.
Harris would turn nation into Kamalafornia, where illegal immigrants get free money trib.al/sodF9JS	Kamala Harris’s policies would lead to a nation that rewards illegal immigrants with financial benefits.
Trump’s Project 2025 would end Medicare as we know it. The more we learn about the far right extremist manifesto. The worse it gets.	Trump’s Project 2025 will destroy Medicare, leading to severe consequences for the public.

Table 1: Examples of social media posts and their extracted frames.

3.2 Frame Clustering

Each frame $f \in \mathscr{F}$ is converted to a high-dimensional vector representation using a pre-trained text embedding model (text-embedding-ada-002)[4]. The embedding process is a mapping,

$$\phi : \mathscr{F} \to \mathbb{R}^{k_e},$$

where k_e is the dimension of the embedding space and $v_f \in \mathbb{R}^{k_e}$ is the vector representation of f.

High-dimensional embedding spaces often contain noise and redundant information that can make clustering difficult. To improve clustering performance, we first apply Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018), a nonlinear dimensionality reduction technique designed to preserve both local and global structure in high-dimensional data. Unlike PCA or t-SNE, UMAP is based on manifold learning and constructs a weighted graph representation of the data, optimizing a low-dimensional embedding that retains high-density neighborhood structures (Becht et al., 2019). This step enhances the clustering process by increasing the density of semantically related frames while reducing noise from irrelevant dimensions. The mapping from vector embeddings to the dimension-reduced space is denoted

$$\rho : \mathbb{R}^{k_e} \to \mathbb{R}^{k_r},$$

where $k_r << k_e$ and v_{fr} is the reduced vector representation of f.

After dimensionality reduction, we apply Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) (Campello et al., 2013, 2015) to identify clusters of semantically related frames. HDBSCAN is well-suited for this task because it does not require a pre-specified number of clusters and can detect clusters of varying sizes and densities, which is particularly important for textual data. Unlike centroid-based methods such as k-means, HDBSCAN identifies clusters based on density variations rather than assuming uniform cluster structure. Additionally, it classifies low-density points as noise, ensuring that only semantically coherent frames are grouped together while filtering out highly idiosyncratic or ambiguous claims. The clustering process κ assigns each reduced embedding of a frame to a cluster label:

$$\kappa : \mathbb{R}^{k_r} \to \{1, 2, 3, ..., K, \emptyset\}.$$

where $\kappa(v_{fr}) = j$ if frame f belongs to cluster C_j, and $\kappa(v_{fr}) = \emptyset$ if it is classified as noise.

The resulting clusters in Table 2 illustrate how this method organizes political discourse into coherent ideological themes. Each cluster corresponds to a distinct political narrative, such as gun control advocacy, gun rights protection, contrasting views on Joe Biden's economic performance,

[4]For discussions on text clustering with LLM embeddings, see Petukhova et al., 2024, Keraghel et al., 2024.

or concerns about Donald Trump's impact on democracy. In contrast to phrase-frequency methods, these clusters emerge from broader semantic alignment in how political and media actors frame these issues.

For instance, Cluster 291, which focuses on gun violence and legislative action, consists mostly of Democratic politicians advocating for policy interventions. While a phrase-frequency approach might capture some of these posts by identifying the bigram "gun violence", it would likely miss posts that discuss gun-related tragedies without that exact phrase (e.g., "a homophobic gunman targeted Club Q... weapons of war"). This method, by contrast, captures the shared framing—a call for stronger gun laws—rather than relying on exact word matches.

Cluster 1000, which frames Donald Trump as a threat to democracy, presents a similar challenge for phrase-frequency methods. Some posts contain phrases like "threat to democracy", which might be identified as a relevant bigram (e.g., "threat democracy"), but others frame the argument differently, using words like "dictator" or "fascist", or sentences such as "we're about to see how resilient our democracy is when a full stress test of another Trump term comes". While phrase-based methods might struggle to link these variations, the clustering approach groups them based on shared meaning rather than repeated terminology.

Overall, the results indicate that this method effectively identifies partisan narratives by capturing how political actors frame key issues, offering a more flexible and scalable approach.

3.3 Training the Classifier

To systematically quantify ideological alignment, we train a classifier that learns from the engagement patterns of politicians and applies these learned patterns to media firms. The classification is based on user frequency vectors, which encode how often each account engages with different ideological clusters.

3.3.1 Constructing User Frequency Vectors

For each account $a \in \mathscr{A}$, a user frequency vector is constructed to represent the fraction of their posts assigned to each cluster. The frequency vector for account a is given by

$$u_a = \left(\frac{|F^a \cap C_1|}{|F^a|}, \frac{|F^a \cap C_2|}{|F^a|}, \ldots, \frac{|F^a \cap C_K|}{|F^a|} \right),$$

where $F^a = \{F_d\}_{d \in D_a}$ is the set of all frames generated from the social media posts of a. Each entry in u_a corresponds to the fraction of frames from a that fall into each cluster.

To ensure that only politically meaningful clusters are included in the analysis clusters dominated by a single politician are excluded as they tend to capture messages relevant to a single politician such as event promotions, campaign messages etc that are not part of a broader partisan theme. Clusters containing only media frames are also excluded, as they typically represent apolitical topics such as general news, sports, and entertainment rather than partisan discourse.

3.3.2 Training the XGBoost Classifier

The frequency vectors of politicians (u_p) are used to train an XGBoost classifier (Chen and Guestrin, 2016), which learns to predict party affiliation based on engagement patterns. XGBoost is well-suited for this task because it efficiently handles structured numerical data, captures nonlinear relationships between features, and incorporates built-in regularization to prevent overfitting. The classifier takes as input the frequency vector u_p for each politician and learns the mapping

$$f : \{u_p\}_{p \in \mathscr{P}} \to \{R, D\}.$$

Given the structure of the feature space, the model learns how different clusters are associated with partisan engagement patterns.

3.3.3 Hyperparameter Tuning and Label Randomization Test

To select hyperparameters that yield a model that can generate plausible ideology predictions from frequency vectors, I combine standard cross-validation with a label randomization test.

Cross-validation is used to estimate model performance on unseen data. The labeled politician data is split into training and validation folds, and candidate hyperparameter configurations are evaluated using mean accuracy on held-out subsets. The grid search explores combinations of tree depth, number of trees, learning rate, and regularization strength. Configurations that generalize well across folds are considered viable candidates, but additional filtering is applied to assess overfitting risk.

To identify models that learn spurious correlations or are overly sensitive to noise, a label randomization test is introduced. I form two equally sized groups of politicians by randomly assigning politicians to each group. In one group politician labels are reassigned independently with equal probability, removing any meaningful association between input features and ideology. Given a hyperparameter configuration, the model is trained on this randomized group and used to generate predicted probabilities for the other group. Since the training data carries no ideological information from frequency vectors, the model should produce outputs close uncertainty for each politician. This is measured by computing the mean absolute deviation from 0.5 across the predicted probabilities. A large mean deviation indicates that the model has learned patterns from noise that it then projects onto new users, suggesting it may also overfit when applied to media frequency vectors.

Hyperparameter configurations are selected based on both validation performance on real labels and behavior under randomized training. A suitable model achieves high accuracy on true labels while assigning low-certainty predictions when trained on random labels. This ensures that the final classifier captures meaningful ideological structure in engagement patterns rather than artifacts of the dataset.

3.3.4 Applying the Model to Media Outlets

Once trained, the classifier is applied to the frequency vectors of media outlets ($\{u_m\}_{m\in\mathscr{M}}$). The classifier produces a probabilistic ideological score for each media firm, indicating how its engagement with ideological clusters aligns with Republican or Democratic politicians. Specifically, for each media outlet m, the classifier outputs a probability

$$\hat{y}_m \in [0,1],$$

where values close to 1 indicate strong Republican alignment, values near 0 indicate strong Democratic alignment, and values around 0.5 suggest a more balanced ideological positioning.

This classification approach offers a fully automated measurement of media slant that is adaptable to different political contexts and time periods.

4 Discussion of Results

To assess the validity of the media slant scores produced by this method, I compare them with bias ratings from AllSides. AllSides assigns ideological classifications to media outlets through multi-partisan Editorial Reviews, in which trained experts analyze a media outlet's homepage, headlines, recent articles, photos, and other content to assign a bias score between -6.0 (Left) and +6.0 (Right). Additionally, AllSides incorporates Blind Bias Surveys, where individuals across the political spectrum rate content without knowing its source, ensuring that the ratings reflect a broad ideological consensus.

In contrast, the approach presented in this paper relies solely on social media text from media outlets and politicians, using an automated method based on linguistic patterns rather than direct editorial analysis. Unlike AllSides, this method does not consider multi-modal content such as images, website layout, or stylistic presentation.

Figure 1 presents a scatter plot comparing the ideological scores assigned by this method with the corresponding AllSides numerical ratings for a set of media outlets. The results show a strong positive correlation (0.903), indicating substantial agreement between the two measures. While deviations exist due to differences in data sources and methodology, the high correlation suggests that this approach effectively captures ideological positioning in media content.

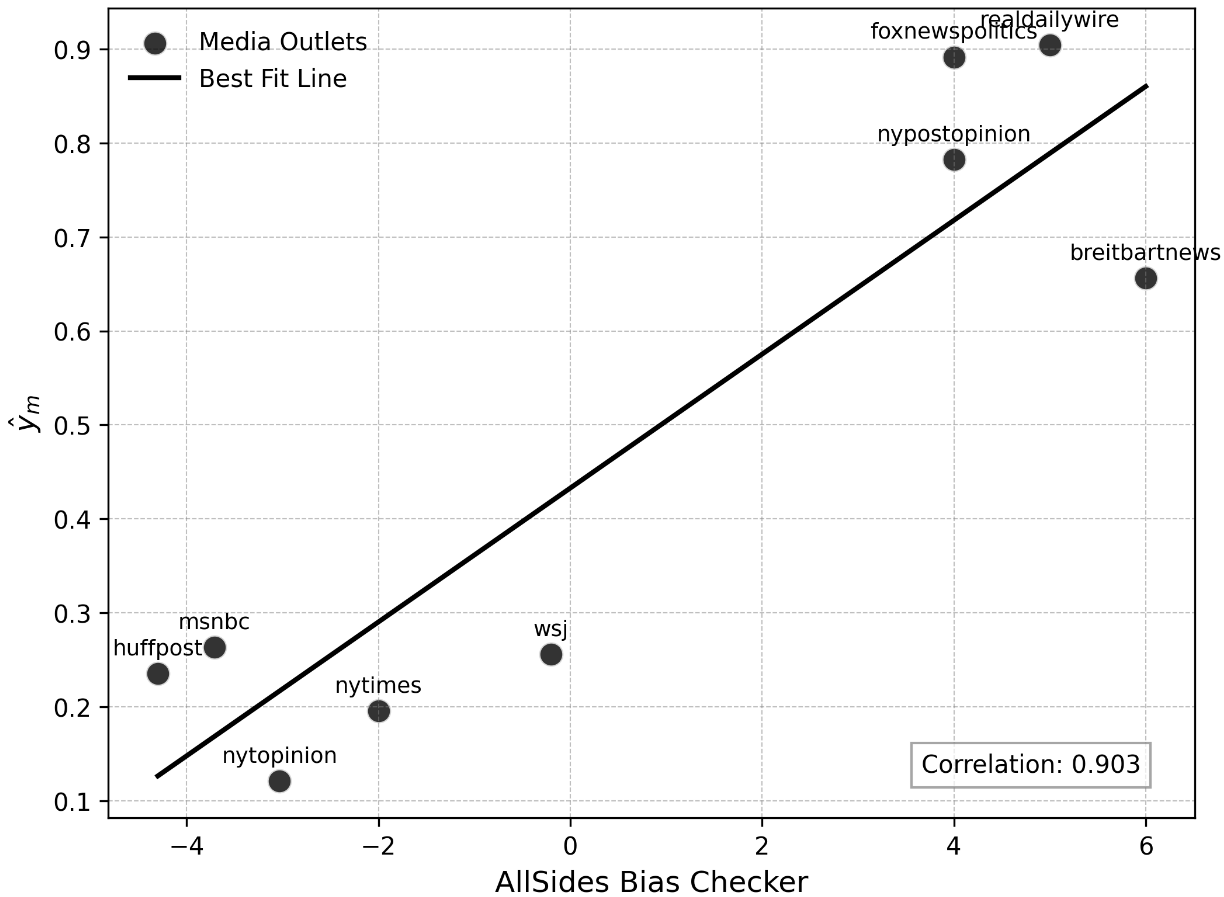

Figure 1: $\hat{y}_m$ vs. AllSides Bias Checker Score

5 Conclusion

This paper presents a fully automated approach to measuring media slant using large language models. By leveraging advances in semantic text representation, the method identifies partisan framing in social media content while preserving the comparison-based structure of existing economic approaches. This enables the incorporation of broader linguistic context, similar to NLP-based methods, without being constrained by a specific labeled dataset, political context, or time period. The results demonstrate strong alignment with AllSides ratings, suggesting that ideological signals embedded in media discourse can be effectively extracted from social media text.

The method's reliance on UMAP for dimensionality reduction and HDBSCAN for clustering introduces the inherent limitations of these algorithms. If dimensionality reduction fails to preserve essential structure when mapping high-dimensional embeddings into a lower-dimensional space, important nuances may be lost. Likewise, suboptimal clustering could lead to mis-grouping of ideologically distinct frames. However, as these techniques continue to advance, the accuracy and robustness of this approach will improve accordingly.

More broadly, this framework provides an automated tool for tracking ideological narratives in media and political discourse, with potential applications beyond traditional media outlets, including social media influencers and political commentators. Its accessibility allows researchers with social media data to integrate the method and explore critical questions regarding the causes and consequences of media slant and/or partisan political discourse. Future work could refine this approach further by incorporating multimodal content—such as images, memes, and video transcripts—to enhance ideological classification.

Cluster Label: 291		
Original Post	**Frame**	**Label**
June marks Gun Violence Awareness Month, but many kids, families, and communities are reminded of the gun violence epidemic every day. We have ample awareness and not nearly enough action. RT to tell @SpeakerJohnson that Congress must act NOW to end the scourge of gun violence. https://t.co/NtGC81OKzW"	Congress must act now to end the scourge of gun violence.	D
I wish Chairman Comer would join me in legislation to ban AR-15s and other assault weapons to keep all Americans—not just presidents but schoolchildren and firefighters—safe from more mass shootings.	Legislation to ban AR-15s and other assault weapons is necessary to ensure the safety of all Americans from mass shootings.	D
My team and I are in contact with the NYPD about the shooting today in Tompkins Square Park. This is the second act of gun violence in my district alone in the last 3 days. Our communities can't bear this. We must double down on this issue from every angle.	Gun violence is a pressing issue that our communities cannot tolerate and requires urgent, comprehensive action.	D
Two years ago a homophobic gunman targeted Club Q in Colorado Springs, CO, killing 5 people and injuring 25 others with a semiautomatic killing machine. May their memory be a blessing and a reminder to us all of the deadly urgency in removing weapons of war from our communities.	We must urgently remove weapons of war from our communities to prevent further tragedies.	D

Continued on next page

I'm celebrating the Biden-Harris Administration's one-year anniversary of the @whitehouse Office of Gun Violence Prevention & its meaningful work to address our nation's epidemic of gun violence. We refuse to accept the ongoing carnage in our communities. whitehouse.gov/briefing-room/...	The Biden-Harris Administration is positively impacting gun violence prevention in the United States, and society must actively combat the ongoing gun violence crisis.	D

Cluster Label: 290

Original Post	Frame	Label
You're not taking our guns, Joe.	The government under Joe Biden is trying to take away our guns.	R
"Everybody wants to be first. And it's great to be first. But let me tell you this, it's terrible to be wrong." Visit Breitbart.com/Downrange to equip yourself with the latest statistics and news articles to defend your right to bear arms. https://t.co/5Oa3bsaIE4	Defending the right to bear arms is crucial, and being wrong about it is a serious concern.	M
Thank you @BillHagertyTN for joining our fight to protect legal gun owners from unconstitutional government surveillance. The Protecting Privacy in Purchases Act, which I introduced in the House, would prohibit radical gun grabbing politicians from tracking lawful gun purchases....	Protecting legal gun owners from unconstitutional government surveillance is essential	R

Continued on next page

“Colorado, do NOT let Democrats in the state legislature take away your Second Amendment rights! The @CO-HouseGOP is fighting hard but they need your help! Click the link in the tweet below to find out how you can get involved!”	Democrats are attempting to take away Second Amendment rights in Colorado.	R
”Thank you to @RidersUSA, @AZCDLFreedom, and @Gun-FreedomRadio for hosting Arizona’s 2A Rally today. Anti-gun zealots are unleashing a full-frontal assault on our God-given, constitutionally protected Second Amendment rights. I don’t accept these attacks and neither should you. pic.twitter.com/gFWUkdEb0N”	Anti-gun activists threaten Second Amendment rights and should be opposed.	R

Cluster Label: 195

Original Post	**Frame**	**Label**
The Biden economy has created more than 15 million jobs in under four years. That’s more than the last three Republican Presidents. Combined.	The Biden economy is more effective at job creation than the economies of the last three Republican Presidents combined.	D
The Biden economy is strong. Why should he downplay it? Read: nyti.ms/45bI0Jq	The Biden economy is strong and should not be down-played.	M

Continued on next page

RT @RLCLeaders Ever wonder how @POTUS and House Democrats are working to create a more equitable economy? Take a look at how new initiatives, like @USDOT's Reconnecting Communities Pilot program, are making life easier and more affordable for Americans in every corner of the country." We've now added over 15 MILLION jobs and seen 40 consecutive months of job growth under Biden. A persistently strong labor market is a win for working families.	The initiatives by President Joe Biden and House Democrats are creating a more equitable economy for all Americans.	D
President Biden's economy is booming. Lowest unemployment in decades. Stock market at record highs. Wages outpacing inflation. That doesn't mean things are perfect. Price gouging and shrinkflation are still a problem. But Democrats are taking it on! axios.com/2024/05/23/us-...	President Joe Biden's economy is thriving despite ongoing challenges like price gouging and shrinkflation, thanks to the efforts of the Democratic Party.	D

Cluster Label: 898

Original Post	Frame	Label
Not really cancelled, just shifted to taxpayers and making inflation worse. Bidenomics The Total Cost of Student Debt Cancellation-2024-04-29 crfb.org/blogs/total-co	Biden's economic policies are shifting financial burdens onto taxpayers and worsening inflation.	R
Nobody can afford anything in Biden's economy. Prices up. Again. cnbc.com/2024/05/14/ppi...	Biden's economy is causing financial hardship for the public due to rising prices	R

Continued on next page

You've had almost four years to strengthen the middle class. Instead, you've driven inflation and made it more expensive to buy food, fuel, and a place to live. Your policies have decimated the middle class. So thanks, but no thanks.	The current administration's policies have harmed the middle class by driving inflation and increasing living costs.	R
Bidenomics will no longer let us 'Live Mas.' We pay more to get less.	Bidenomics causes economic hardship by increasing costs and reducing benefits.	R
Here's something else: A strong majority of Americans—both Republicans and Democrats—believe the economy is getting worse. Just look at the latest ABC News/Ipsos poll. Biden-Harris' inflationary spending policies have wrecked Americans' purchasing power. I've already shown you… pic.twitter.com/vJZ824tZyB	Biden-Harris' inflationary spending policies have wrecked Americans' purchasing power	R

Cluster Label: 1000

Original Post	**Frame**	**Label**
The stakes couldn't be higher. Donald Trump's presidency in 2025 could mean our democracy, freedoms, and future will be at risk. I am ready to stand up to protect everything we care about, but I can't do it on my own. Chip in today to help us prepare for the challenges ahead.	Donald Trump's presidency in 2025 threatens democracy, freedoms, and the future.	D

Continued on next page

John Kelly, Donald Trump's longest-serving chief of staff, said Trump met the definition of a fascist and would try to govern like a dictator if elected. nyti.ms/40fAZGH https://t.co/52NBcQFpEu	Donald Trump embodies fascism and would govern as a dictator if elected.	M
President Biden opened the 2024 election year with a dire warning about the threats he says Donald Trump poses to American democracy, declaring his predecessor is "trying to steal history the same way he tried to steal the election" on.wsj.com/48CG91b on.wsj.com/48CG91b	Donald Trump threatens American democracy by attempting to alter historical narratives.	M
We can't Trump-proof the country from the consequences of a Trump administration. Elections have consequences, and we're about to see how resilient our democracy is when a full stress test of another Trump term comes about. NYTLetters Read: nyti.ms/4gsraKv"	The consequences of a Trump administration threaten the resilience of our democracy.	M
RT @ChrisMurphyCT As Trump's messaging gets darker and darker, the stakes get higher and higher. America won't be America anymore if he wins. politico.com/news/2024/10/1…	If Donald Trump wins, America will lose its identity and core values.	D

Table 2: Selected Clusters and Some of their Posts, Frames and Party Labels

References

Stephen Ansolabehere, Erik C. Snowberg, and James M. Snyder. Television and the Incumbency Advantage in U.S. Elections. *Legislative Studies Quarterly*, 31(4):469–490, 2006. Publisher: [Wiley, Comparative Legislative Research Center].

Etienne Becht, Leland McInnes, John Healy, Charles-Antoine Dutertre, Immanuel W. H. Kwok, Lai Guan Ng, Florent Ginhoux, and Evan W. Newell. Dimensionality reduction for visualizing single-cell data using UMAP. *Nature Biotechnology*, 37(1):38–44, January 2019. Publisher: Nature Publishing Group.

Justin De Benedictis-Kessner, Matthew A. Baum, Adam J. Berinsky, and Teppei Yamamoto. Persuading the Enemy: Estimating the Persuasive Effects of Partisan Media with the Preference-Incorporating Choice and Assignment Design. *American Political Science Review*, 113(4):902–916, November 2019.

Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. Density-Based Clustering Based on Hierarchical Density Estimates. In Jian Pei, Vincent S. Tseng, Longbing Cao, Hiroshi Motoda, and Guandong Xu, editors, *Advances in Knowledge Discovery and Data Mining*, pages 160–172, Berlin, Heidelberg, 2013. Springer.

Ricardo J. G. B. Campello, Davoud Moulavi, Arthur Zimek, and Jörg Sander. Hierarchical Density Estimates for Data Clustering, Visualization, and Outlier Detection. *ACM Trans. Knowl. Discov. Data*, 10(1):5:1–5:51, July 2015.

Carl Ehrett. cehrett/social_media_frame_analysis, October 2024. URL `https://github.com/cehrett/social_media_frame_analysis`. original-date: 2024-01-18T13:32:47Z.

Matthew Gentzkow and Jesse M. Shapiro. What Drives Media Slant? Evidence From U.S. Daily Newspapers. *Econometrica*, 78(1):35–71, 2010.

Matthew Gentzkow and Jesse M. Shapiro. Media Bias and Reputation. *Journal of Political Economy*, 114(2):280–316, 2006. Publisher: The University of Chicago Press.

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 Groeling. Media Bias by the Numbers: Challenges and Opportunities in the Empirical Study of Partisan News. *Annual Review of Political Science*, 16(Volume 16, 2013):129–151, May 2013. Publisher: Annual Reviews.

Tim Groseclose and Jeffrey Milyo. A Measure of Media Bias*. *The Quarterly Journal of Economics*, 120(4):1191–1237, November 2005.

Felix Hamborg, Karsten Donnay, and Bela Gipp. Automated identification of media bias in news articles: an interdisciplinary literature review. *International Journal on Digital Libraries*, 20(4): 391–415, December 2019.

Daniel E. Ho and Kevin M. Quinn. Measuring Explicit Political Positions of Media. *Quarterly Journal of Political Science*, 3(4):353–377, December 2008. Publisher: Now Publishers, Inc.

Shanto Iyengar, Yphtach Lelkes, Matthew Levendusky, Neil Malhotra, and Sean J. Westwood. The Origins and Consequences of Affective Polarization in the United States. *Annual Review of Political Science*, 22(Volume 22, 2019):129–146, May 2019.

Valentino Larcinese, Riccardo Puglisi, and James M. Snyder. Partisan bias in economic news: Evidence on the agenda-setting behavior of U.S. newspapers. *Journal of Public Economics*, 95 (9):1178–1189, October 2011.

John R. Lott and Kevin A. Hassett. Is Newspaper Coverage of Economic Events Politically Biased?, September 2013.

Christopher McConnell, Yotam Margalit, Neil Malhotra, and Matthew Levendusky. The Economic Consequences of Partisanship in a Polarized Era. *American Journal of Political Science*, 62(1): 5–18, 2018.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*, 3(29):861, September 2018.

Riccardo Puglisi. Being The New York Times: the Political Behaviour of a Newspaper. *The B.E. Journal of Economic Analysis & Policy*, 11(1), 2011. Publisher: De Gruyter.

Riccardo Puglisi and James M. Snyder. Newspaper Coverage of Political Scandals. *The Journal of Politics*, 73(3):931–950, 2011. Publisher: [The University of Chicago Press, Southern Political Science Association].

Riccardo Puglisi and James M. Snyder. The Balanced Us Press. *Journal of the European Economic Association*, 13(2):240–264, 2015. Publisher: Oxford University Press.

Stuart N. Soroka. The gatekeeping function: Distributions of information in media and the real world. *The Journal of Politics*, 74(2):514–528, 2012. Place: United Kingdom Publisher: Cambridge University Press.