**Term Project: Decoding Political Issue Framing in News Media Using Word Embeddings**

Zhiqiang Ji[[1]](#footnote-1)♦

Georgetown University

This study employs word embedding methods to quantify how news reports from different media outlets associate sentiment with political terms, aiming to uncover latent biases in their ideologically-oriented discourse. Building upon the work of (Rozado & al-Gharbi, 2022), the project analyzes a dataset of articles from 9 news outlets (3 left-leaning, 3 centrist, and 3 right-leaning) spanning from 2000 to 2019. The research explores four key aspects: issue ownership, political orientation sentiment associations, long-term temporal dynamics, and detailed sentiment associations. The findings reveal differences in the coverage of political issues by media outlets with varying ideological leanings, the tendency of left-leaning and right-leaning outlets to associate positive sentiment with their own ideological camp and negative sentiment with the opposing side, and the increasing polarization of media sentiment over time, particularly among left-leaning outlets. The detailed sentiment associations analysis provides a nuanced understanding of the specific types of negative associations linked to different political orientations. Despite challenges in constructing political issue axes, the project demonstrates the potential of using word embeddings and cultural axes to quantify and describe political biases in text data. The insights gained contribute to our understanding of the complex dynamics of media bias and polarization, laying the groundwork for further research in this area.

Keywords: media bias, political polarization, word embeddings, sentiment analysis

In the age where the term “fake news” has entered common parlance and “alternative truths” obscures the line between fact and fiction, public trust in news media stands on precarious ground. While prestigious news media platforms have been putting great effort to avoid distortion and mistake about facts, they may craft narratives with subtle political biases, conveying different sentiments out of the same facts. The ideological discrepancy among news media has become a widely recognized fact. Disbelieving in the objectivity of news media and indulged by the ease to get free news headlines, the audience adapt to care more about how the news stories make they feel rather than the authenticity of the facts behind the reporting to consume more content from outlets that align with their political leanings. Consequentially, news media may be more driven to align their ideological stance with their separate audience bases.  
Assuming most prestigious news media platforms cover similar major news stories at a given time, how do they consistently attach their political leanings to fact reporting, maintain their ideological images, and frame their narratives? Can such framing be quantitatively portrayed? More importantly, has this politically loaded narrative framing intensified in sync with the gradual polarization of the political climate in recent decades?  
Recent studies have demonstrated the potential of word embeddings in analyzing ideological placement, sentiment, and cultural meanings in political texts. (Rheault & Cochrane, 2020) and (Kozlowski et al., 2019) used word embeddings to capture latent concepts such as ideology and social class, while (Gennaro & Ash, 2022) and (Rodman, 2020) explored the use of emotion and the changing meanings of political concepts over time. Building upon these advancements, the present study applies word embedding methods to quantify how news reports from different media outlets associate sentiment with political terms, aiming to uncover latent biases in their ideologically-oriented discourse.

Inspired by the work of (Rozado & al-Gharbi, 2022), this study employs word embedding methods to quantify how news reports from different media outlets associate positive or negative sentiment words with political terms related to various parties, such as prominent party figures and key issues, in order to construct their ideologically-oriented discourse. By calculating the projections of sentiment words onto semantic axes representing political orientations in the embedding space, this study aims to reveal the latent political biases embedded in the language used by various media organizations.

Building upon the foundation laid by Rozado & al-Gharbi, this project partially replicates and expands their research. Utilizing the preprocessed text data provided by the authors, which includes articles from 9 news outlets (3 left-leaning, 3 centrist, and 3 right-leaning) spanning from 2000 to 2019, word2vec models were trained for each outlet in 5-year intervals. Within these models and their corresponding embedding spaces, several key aspects were explored: (1) issue ownership, (2) positive or negative sentiments attached to different parties and personal ideologies, (3) the dynamics of sentiment associations in left-leaning and right-leaning news media outlets from 2000 to 2019, and (4) the creation of political issue axes to examine how sentimental words align with these axes across different news outlets, with the goal of identifying the closest-aligned positive and negative words.

While the analysis of issue ownership and sentiment associations with political parties yielded clear results, the creation of political issue axes proved to be more challenging and did not produce conclusive findings. The subsequent sections of this report will delve into the data, methods, and results of this study in greater detail, providing a comprehensive overview of the research process and its outcomes. The report will conclude by discussing key lessons learned and exploring potential avenues for future research in this field.

Data and Methods

The present study utilizes the preprocessed text data provided by Rozado & al-Gharbi (2022), which includes news and opinion articles from various media outlets. The articles were collected from the outlets' online domains and public cache repositories such as Google Cache, The Internet Wayback Machine, and Common Crawl. The temporal coverage of articles varies across outlets, with some dating back to the 1970s, while others have more limited availability. The preprocessed data is limited to the articles' headlines and main text, excluding elements such as figure captions.

To ensure consistency and relevance, a subset of 9 news outlets (3 left-leaning, 3 centrist, and 3 right-leaning) was selected based on the volume of articles available between 2000 and 2019 and the bias ratings provided by AllSides.com. The selected articles were limited to online accessible content not behind paywalls.

The preprocessing steps performed by (Rozado & al-Gharbi 2022) included lowercasing tokens, removing markup language tags, URLs, non-alphanumeric characters, punctuation, digits, and 330 common stop words. Additionally, the remaining words were randomly scrambled within a sliding window of size 10 to render the articles incomprehensible to human readers while preserving the local context for training word2vec models. These preprocessing steps have been shown to have minimal impact on the Continuous Bag of Words (CBOW) word2vec model, and the results reported in their work are similar when using the scrambled articles text to train outlet-specific embedding models.

Methods Following the methodology outlined by (Rozado & al-Gharbi 2022), temporal and outlet-specific word2vec embedding models were trained using the preprocessed text data. The models were trained for each outlet in 5-year intervals from 2000 to 2019, allowing for the analysis of semantic shifts over time.

To validate the effectiveness of the analysis methods, several cultural axes were created, including those related to personal ideology, political party affiliation, and presidents from different parties. Additionally, a health-disease axis was constructed to serve as a reference for evaluating the validity of the approach. Furthermore, cultural axes representing political issues such as immigration, abortion, and gun control were created to examine the alignment of sentimental words with these topics across different news outlets.

The word2vec models were trained using the Gensim library with the following hyperparameters: vector dimensions = 300, window size = 10, negative sampling = 10, down sampling frequent words = 0.0001, minimum frequency count = 5 (only terms that appear more than 5 times in the corpus were included in the word embedding model vocabulary), and number of training iterations (epochs) through the corpus = 5. The exponent used to shape the negative sampling distribution was the default 0.75.

**Analysis**

This project explores four main aspects of media bias and political ideology using word embeddings: issue ownership, political orientation sentiment associations, long-term temporal dynamics, and detailed sentiment associations.

***Issue Ownership***

Issue ownership refers to how political parties or candidates are perceived as being more competent in managing certain issues than their competitors. The news media may play a significant role in reflecting and shaping public perceptions of issue ownership. One question in analyzing how the media frames political issues is whether media outlets with different political leanings invest different levels of attention in covering the handling of policy topics by the parties and ideological camps they favor, thereby emphasizing the competency of those parties or political camps.

(Fagan, 2021) conducted an interesting study based on the records of issue setting by think tanks and Congress with different political leanings. Taking its findings as a reference, this project uses a straightforward approach to investigate whether media outlets with different political leanings exhibit partisan or ideological biases when reporting on political issues. By projecting the topics mentioned in Fagan's article onto a two-dimensional plane formed by party and ideology axes, it becomes possible to visually observe differences in issue coverage among various media outlets. Although simple, this method effectively reveals the potential role media plays in shaping public perceptions of issue ownership by political parties.

In the future, this research could be further extended by incorporating the approach of (Gyllensten & Sahlgren, 2018), which utilizes word embeddings to measure the correlation between different media outlets on specific issues, thus enabling a deeper exploration of this question.

***Political Orientation Sentiment Associations***

Replicating the work of (Rozado & al-Garbi 2022), this project next examines the political orientation sentiment associations displayed by influential media outlets across two different political axes. For the 9 outlets, positive (in cyan) and negative (in salmon) words from 19 external sentiment lexicons (N = 15,704) are projected onto two political orientation axes: the Republican-Democratic party axis and the Conservative-Liberal ideological axis.

***Long-term Temporal Dynamics***

This project also replicates the analysis of temporal changes in the political orientation of news media from 2000 to 2019 using sentiment associations' dynamics measured using outlets' content at 5-year intervals. The study employs the Sentiment Political Bias (SPB) measure, which quantifies the emotional bias of media outlets toward different political parties or ideologies by associating positive and negative sentiment words with specific political poles.

***Detailed Sentiment Associations***

Leveraging antonym pairs from WordNet to construct mini semantic axes and measuring their similarity to political orientation axes, the project explores which specific positive and negative words are associated with media outlets of different stances.

A common limitation of most sentiment lexicons is that the label "negativity" conflates a variety of different phenomena, such as shame, anxiety, selfishness, suffering, sadness, fear, anger, aggressiveness, or dislike, under the umbrella category of negativity. Thus, a news outlet that emphasizes the marginalization or suffering of a particular political group could display measurable negative associations with that group, as words that describe victimization and suffering are often labeled as 'negative' in sentiment lexicons. Similarly, a news outlet that emphasizes the negative character traits of a political group by highlighting traits such as dishonesty, bigotry, violent behavior, intolerance, or selfishness, will also display negative associations with that group, because such terms are also labeled as 'negative.' Thus, it is important to disentangle what specific types of negative associations with a given political group are more salient in news outlet content.

This project uses the 3,872 antonym pairs contained in the WordNet lexicon and creates cultural axes with each antonym pair. That is, we can use a WordNet antonym pair, such as unselfish-selfish, to create a cultural axis that traces the spectrum going from altruism to selfishness in embedding space. Then, we calculate the cosine similarity between these mini axes and the three political cultural axes (ideology, party, and presidents). A high degree of similarity indicates that the word embedding model tends to associate each word in the antonym pair with a specific pole of the political axis. To circumscribe the measurement to sentiment associations, antonym pairs that do not belong to a large and popular sentiment lexicon, the Harvard General Inquirer positive/negative lexicon (or HGI for short, N = 3623), are filtered out. This leaves 357 antonym pairs with a clear sentiment valence.

To validate the effectiveness of this method, a "disease-health" cultural axis was constructed as a reference. Next, the analysis of the ideological cultural axis was replicated. Furthermore, several different strategies were attempted to construct political issues axes, including using common words (e.g., "pro\_choice" for abortion rights), names of relevant organizations (e.g., "Planned Parenthood"), well-known activists and political figures, names or slogans of movements, or a combination of these elements. However, the results were not satisfactory, and this will be discussed further in the subsequent sections.

**Results**

***Issue Ownership***

The analysis of Issue Ownership reveals interesting patterns in how left, center, and right-leaning media outlets associate policy topics with political parties and ideologies. By comparing the bottom-left (Republican-Conservative) and top-right (Democratic-Liberal) quadrants in Figure 1., we can see that the media's associations largely align with the traditional understanding of issue ownership by the two major parties. For example, immigration is more strongly associated with the Republican party, while education and labor are more closely linked to the Democratic party.

However, some outlets, such as CBS News, The New York Times, and The Washington Times, appear to assign most policy topics to the party and ideology that align with their own leanings. This suggests that these media outlets may be devoting more coverage to the party they favor in general.

Interestingly, Slate and The Daily Mail exhibit the opposite trend. Using the center row as a reference point, Slate seems to associate more topics with the Republican party and conservative ideology than expected based on its left-leaning orientation. This connection could be driven by negative coverage and criticism of the Republican party and conservative policies.

Overall, the Issue Ownership analysis provides valuable insights into how media outlets across the political spectrum establish connections between policy topics and political parties. While some associations confirm traditional issue ownership patterns, others reveal potential biases in media coverage, with outlets either over-associating topics with their preferred party or linking topics to the opposing party, possibly through negative coverage.

***Political Orientation Sentiment Associations***

The analysis of Political Orientation Sentiment Associations yields expected results, confirming the presence of media bias in the emotional framing of political parties and ideologies. The graph (Figure 2.) clearly shows that left-leaning media outlets tend to associate liberals and Democrats with positive words, as evidenced by the concentration of data points in the upper-right quadrant of the plane. Conversely, these same outlets are more likely to associate conservatives and Republicans with negative words, as indicated by the clustering of data points in the lower-left quadrant.

A group of graphs with different colored dots

Description automatically generated with medium confidenceInterestingly, the three center-leaning media outlets in the middle row appear to exhibit a sentiment association pattern more similar to that of the left-leaning outlets. This may suggest that even media sources that are considered to be ideologically centrist may have a slight bias towards the left in terms of their emotional framing of political parties and ideologies.

Figure 1. Issue ownership shows similar results across news outlets

***A screenshot of a computer screen

Description automatically generated***

Figure 2. Projections of sentiment-labeled words show clear patterns

***Long-term Temporal Dynamics***

Although this study only analyzed data from 2000 to 2019, as opposed to (Rozado & al-Gharbi's 2022) analysis starting from 1975, similar conclusions were reached. Across the three ideological orientation axes analyzed, a notable increasing trend in polarization was observed among left-leaning media outlets (Figure 3.). Left-leaning outlets appear to have become more partisan over time, increasingly associating positive terms with words that denote their own ideological tribe and negative terms with words that denote their ideological outgroup. However, the results for right-leaning media outlets are less robust due to data limitations.

Interestingly, the plot depicting the correlation of sentiment with the two poles of the party axis reveals intriguing patterns. The changes in the party affiliation of the president (from Bush to Obama to Trump) seem to have different effects on the blue and red lines, representing left-leaning and right-leaning media outlets, respectively. During the terms of Bush and Obama, it appears that media sentiment tends to converge towards the center line when the president belongs to the same party. The red line moves closer to the center line during Bush's presidency, while the blue line exhibits a similar trend during Obama's term. However, after Trump's election, both the red and blue lines appear to diverge from the center line, suggesting an increase in polarization.

This observation potentially leads to a conclusion that differs from Rozado & al-Gharbi's findings, indicating that Trump's election may have indeed exacerbated media polarization. The sentiment associations of both left-leaning and right-leaning media outlets seem to have become more extreme following Trump's victory, with each side displaying stronger positive associations with their own ideological camp and more pronounced negative associations with the opposing side.

A graph of political parties

Description automatically generated with medium confidence

Figure 3. Time series analysis

***Detailed Sentiment Associations***

Using the "disease-health" cultural axis (see Figure 4.) as a reference and the results obtained from the ideology axis, the findings align with expectations and the previous observations made by Rozado & al-Gharbi. The results indicate that negative words associated with conservatives often describe negative character traits, such as dishonesty and hostility, while negative words linked to liberals include incompetence and insanity. Although the causal roots of these negative associations are not immediately clear, and factors beyond simple animosity against the ideological outgroup may be at play, this method provides more information for analysis compared to a simple positive/negative dichotomy.

However, difficulties were encountered when attempting to construct political issue axes. First, the process of projecting acronym pairs onto cultural axes was not always successful (the reasons for failure require further investigation by the author). Moreover, even when the projection and visualization of results were successful, they did not reflect patterns that could differentiate between media with different political leanings. For example, in the results for the immigration axis across the nine news outlets, not only was it impossible to discern distinct distributions of positive/negative words, but it was also challenging to differentiate between outlets at the word level. At the anti-immigration end, all media outlets used terms like "illegal" and "unlawful."[[2]](#footnote-2)

Several factors may contribute to the difficulty in constructing political issue cultural axes:

1. **Insufficient corpus size**: Political articles generally constitute a small proportion of all media text corpora, and content related to specific issues may be even more limited.
2. **Limited semantic information covered by political issues**: Compared to axes like "gender," "ideology," or "health," which encompass a wide range of information, specific political issues may involve a much smaller and more closed semantic space. For instance, when gender serves as the main cultural axis, content from various domains such as business, entertainment, and sports can be projected onto it, constructing a rich subspace. However, when using a specific political issue like "gun control" as an axis, it only involves a very small and closed semantic space, with most words being equally distant or close to both poles of the axis.
3. **Inherent incommensurable constraints in the construction of political issue axes**: The reasons for supporting or opposing a particular issue may not be captured by antonyms along a single dimension. For example, support for a president may stem from perceptions of competence, while opposition may arise from a perceived lack thereof. However, in the case of immigration, support may be based on humanitarian considerations and the value of community diversity, while opposition may stem from concerns about the security risks and enforcement costs associated with illegal immigration. Opponents may not necessarily object to the benefits of diversity, just as supporters may not dismiss concerns about enforcement costs. Consequently, the method used in this study may struggle to capture these complex position choices.

**Discussion**

This project aimed to replicate and expand upon the work of Rozado & al-Gharbi (2022) in quantifying political biases in news media using word embeddings and cultural axes. The findings largely corroborate the original study, demonstrating the effectiveness of this approach in capturing sentiment associations and polarization trends across different media outlets and time periods.

The analysis of issue ownership, political orientation sentiment associations, and long-term temporal dynamics yielded clear and interpretable results. The project successfully identified differences in the coverage of political issues by media outlets with different ideological leanings, revealed the tendency of left-leaning and right-leaning outlets to associate positive sentiment with their own ideological camp and negative sentiment with the opposing side, and traced the increasing polarization of media sentiment over time, particularly among left-leaning outlets.

Furthermore, the detailed sentiment associations analysis, using antonym pairs from WordNet, provided a more nuanced understanding of the specific types of negative associations linked to different political orientations. This approach moved beyond the simple positive/negative dichotomy and shed light on the character traits and behaviors that media outlets tend to emphasize when portraying their ideological outgroups.

However, the project encountered challenges in constructing political issue axes. The attempts to create axes based on acronym pairs and project them onto the embedding space were not always successful, and even when the visualization was possible, the results did not reveal clear patterns distinguishing between media outlets with different political leanings. Several factors may have contributed to this difficulty, including the limited size of the corpus, the narrow semantic scope of specific political issues, and the inherent incommensurable constraints in capturing the complex reasons behind support or opposition to a given issue.

Despite the lack of success in constructing political issue axes, this project demonstrates the potential of using word embeddings and cultural axes to quantify and describe political biases in text data. In fact, recent research has begun to explore this avenue further, suggesting that this approach holds promise for future investigations.

The next steps for this research could involve:

1. Training models on corpora that are more explicitly political and focused on discussing specific issues, such as congressional debate records, academic journal articles, or think tank reports. This could help address the problem of sparse coverage of political issues in general media texts.
2. Incorporating Named Entity Recognition (NER) methods to identify patterns and regularities in how different media outlets use sentiment-laden language when framing political issues. By focusing on the entities and actors involved in these issues, this approach could provide a more targeted analysis of the sentiment associations and biases present in media discourse.

In conclusion, while the project encountered some limitations, particularly in the construction of political issue axes, it successfully replicated and extended the findings of (Rozado & al-Gharbi 2022), demonstrating the value of word embeddings and cultural axes in quantifying political biases and sentiment associations in news media. The insights gained from this analysis contribute to our understanding of the complex dynamics of media bias and polarization and lay the groundwork for further research in this area.

References

Fagan, E. (2021). Issue ownership and the priorities of party elites in the United States, 2004–2016. *Party Politics*, *27*(1), 149–160. https://doi.org/10.1177/1354068819839212

Gennaro, G., & Ash, E. (2022). Emotion and Reason in Political Language. *The Economic Journal*, *132*(643), 1037–1059. https://doi.org/10.1093/ej/ueab104

Gyllensten, A. C., & Sahlgren, M. (2018). *Measuring Issue Ownership using Word Embeddings* (arXiv:1811.00127). arXiv. https://doi.org/10.48550/arXiv.1811.00127

Kozlowski, A. C., Taddy, M., & Evans, J. A. (2019). The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings. *American Sociological Review*, *84*(5), 905–949. https://doi.org/10.1177/0003122419877135

Rheault, L., & Cochrane, C. (2020). Word Embeddings for the Analysis of Ideological Placement in Parliamentary Corpora. *Political Analysis*, *28*(1), 112–133. https://doi.org/10.1017/pan.2019.26

Rodman, E. (2020). A Timely Intervention: Tracking the Changing Meanings of Political Concepts with Word Vectors. *Political Analysis*, *28*(1), 87–111. https://doi.org/10.1017/pan.2019.23

Rozado, D., & al-Gharbi, M. (2022). Using word embeddings to probe sentiment associations of politically loaded terms in news and opinion articles from news media outlets. *Journal of Computational Social Science*, *5*(1), 427–448. https://doi.org/10.1007/s42001-021-00130-y

1. Zhiqiang Ji: zj117@georgetown.edu

   Date submitted: 2024-05-01 [↑](#footnote-ref-1)
2. See the resulting Figures 4-6 files for this section inside the “/figures” directory of the GitHub repository. [↑](#footnote-ref-2)