Analyzing Media Narratives with Machine Learning: Topic Alignment and Sentiment in Trump's Rally Speeches and News Media Coverage

Tashvi Bansal,[1] Akhila Pasupunuri,[2] and Manushri Pitale[3]

[1]Monta Vista High School, tashvi.bansal@gmail.com
[2]Henry M. Jackson High School, akhila.pnuri07@gmail.com
[3]Evergreen Valley High School, manushripitale@gmail.com

ABSTRACT

Considering pivotal democratic elections like those of 2020 and 2024, understanding the role of news media in shaping political polarization is critical: media influence profoundly impacts democratic processes. Using agenda-setting theory and ingroup/outgroup dynamics from social identity theory as frameworks, this study examines the framing of Trump's rally speeches and their coverage by liberal and conservative news sources. Study 1 uses a BERT-based LLM and the probabilistic LDA model to analyze the relationship between topics in liberal versus conservative media coverage, examining how media outlets' political alignments influence the alignment and selective emphasis of topics in relation to the topics present in Trump's rally speeches. This led to further inductive analysis of topics using a neural-network based topic modeling pipeline. In Study 2, the political alignment of media outlets in connection to the sentiment of news articles was examined using another BERT-based LLM, while also considering the sentiment expressed in Trump's rally speeches. Our findings revealed that despite significant differences in topic alignment throughout sources, both liberal and conservative sources exhibited predominantly negative sentiments. Further analysis revealed that each outlet selectively criticizes the other to express ideological bias, rather than merely advocating its own views. This conclusion exemplifies a pervasive issue in contemporary media: the emphasis on adversarial coverage over constructive discourse. Such coverage inundates the public with divisive narratives, heightening political polarization and weakening trust in media institutions, thus impairing informed decision-making. Our research demonstrates the need for increased media literacy to safeguard against a misinformed electorate, reduce polarization, and strengthen democratic processes.

Keywords: topic modeling, sentiment analysis, agenda-setting theory, ingroup-favoritism, outgroup-discrimination, social identity theory, media literacy, polarization

I. INTRODUCTION

The rally speeches of Trump, the 45th President of the United States, have profoundly impacted American politics. Known for their polarizing nature, his speeches address controversial political topics such as immigration, national security, and economic policy.

These speeches are highly polarizing, as they not only mobilize strong support among his base but also generate significant opposition from others, thereby deepening the political divide in the country. Studies have shown that Trump's rhetoric often employs a divisive "us versus them" approach, and amplifies partisan differences (Evers et al., 2019). In today's media-dependent society, the portrayal of political events and actors by the news media plays a crucial role in shaping public perception.

The manner in which different news outlets, particularly liberal versus conservative sources, cover Trump's rally speeches provides insight into how media bias manifests through agenda-setting and social identity theory. Agenda-setting refers to how news media influences public perception by determining what issues are covered and shaping how those issues are presented and interpreted (McCombs & Valenzuela, 2014). Within social identity theory, the concept of in-group favoritism and out-group hostility describes the tendency to favor one's own group while disfavoring other groups (Tajfel & Turner, 1979). In fact, research indicates that conservative media often frame Trump's speeches in a positive light, emphasizing themes of patriotism and economic success, while liberal media highlight divisive and controversial aspects, such as his stance on immigration and alleged misinformation (Tyson, 2018). This study seeks to investigate media bias through agenda-setting theory and social identity theory with the main research question:

How do liberal and conservative news sources differ in their portrayal of Trump's rally speeches?

This main question is divided into two sub-research questions, each with distinct sets of concepts, with the exception of one that recurs across both sub-research questions: the political alignment of the news sources.

In the Western world, political alignment has divided the news media, manifesting as either liberal (left-leaning) or conservative (right-leaning) (Groseclose & Milyo, 2005). News media outlets express their political alignments through different contexts. For instance, through McCombs' definition of agenda-setting theory (McCombs et al., 2014), it can be suggested that different media outlets may prioritize certain issues over others to align with their specific political ideologies and shape the public agenda or public attitude toward political matters. Additionally, prominently featuring certain political actors or certain topics, media outlets set their own agenda. Political alignment is not only expressed through directly advocating for a preferred political party but also through the criticism of the opposing political party (Budak et al., 2016). Understanding the political alignment of news media sources is crucial for recognizing their potential ideological biases (Shultziner, 2019), which helps individuals assess the information presented, ultimately influencing the public agenda.

RQ1: How closely do the topics discussed in liberal versus conservative news articles align with the content of Donald Trump's rally speeches?

As previously discussed, the political alignment of a media outlet represents its ideological bias, which can manifest in various ways. This study also examines the topics mentioned in Trump’s speeches, analyzing the political alignment of news sources (whether conservative or liberal) and the specific topics discussed in Trump’s speeches in relation to the topics emphasized in the news media’s coverage of those speeches.

In a broader context, the topics discussed in news media matter because they have the power to set the public agenda, and by extension, the policy agenda. Using agenda-setting theory (McCombs et al., 2014) as a framework, when media outlets increase salience or attention towards certain topics they can shape the public’s perception of which issues to focus on. The emphasis on certain topics by news media outlets can also contribute to creating ingroups and outgroups. That is, by emphasizing certain topics that align with the news outlet’s audience’s interests they can enforce ingroup beliefs and exacerbate outgroup discrimination.

Trump’s speeches often lead to disagreements and conflict due to his tendency to employ outgroup hostility. In fact, studies have shown that during his 2016 presidential campaign, outgroup hostility was a key predictor of support for his candidacy and was one of the primary rhetorical strategies used throughout both the primary and general elections (Matos et al., 2021). In order to create this environment, the topics Trump emphasizes in his speeches play a crucial role. These topics have the power to polarize audiences by addressing controversial issues, thereby reinforcing group identities. Consequently, the topics emphasized in Trump's speeches can influence the topics emphasized by differently politically aligned news media outlets, resulting in either alignment or misalignment of topics. In this study, political alignment in the context of topic emphasis specifically refers to the emphasis on topics by liberal versus conservative media outlets. Finally, the relevance of these topics lies in the idea that prior research (Gunther, 1998) has found that news media and the topics they address can influence public agenda.

RQ2: How does the sentiment of Donald Trump's rally speeches compare to the sentiment portrayed in liberal and conservative news articles covering these speeches?

As stated before, the political alignment of news sources refers to the tendency of media outlets to consistently frame and prioritize issues in ways that reflect their ideological bias, in this case, a liberal and conservative alignment.

Sentiment in news content matters because it sets the tone for how readers interpret the information presented. In fact, news content has been shown to predict public mood, shape public opinion, and reflect media biases (Rozado et al., 2022). For example, second-level agenda setting suggests that if a topic is framed positively with more optimistic tones, it can lead the public to view that topic more favorably. Furthermore, sentiment in news content matters because it can increase the salience of the issue being discussed (Coleman & Wu 2010; Sheafer, 2007); if certain media outlets cover an issue with a more emotional charge, the public may place higher importance or perceive greater salience of that issue.

In the context of our study, the sentiment used to describe Trump's rally speeches in liberal vs conservative news media explains how different media outlets can frame the same events in different ways regarding the sentiment they use towards that event and thus influence public perception of these events. Thus, the differing sentiments in the coverage of Trump's speeches can lead to increased political polarization among the public. For instance, positive sentiment from conservative outlets can encourage support among Trump's base, while negative sentiment from liberal outlets can reinforce opposition among his critics.

Sentiment in this study is defined in two separate contexts. In liberal and conservative coverage of Trump's rally speeches, sentiment is defined as the tone or attitude toward the issues discussed in the articles within the broader context of the rally speech. Whereas, within Trump's rally speeches themselves, sentiment reflects how Trump personally feels about these issues and constitutes the overall emotional tone Trump conveys regarding the topics he addresses in his speeches.

Additionally, understanding the sentiment difference between these three media communication forms (Trump's speeches himself, liberal media coverage, and conservative media coverage) can help the public audience become more well-informed about the potential ideological biases news media has. A higher public awareness of media biases also ensures a more balanced news intake (Spinde et al., 2022) and reflective news consumption (Spinde et al., 2022).

While previous studies employ audience-based approaches (Mullainathan and Shleifer, 2005), content-parsing approaches (Gentzkow and Shapiro, 2010), and crowdsourced content analysis (Budak, 2016), this study introduces two new perspectives. First, it examines Trump's rally speeches, focusing on a recent and significant political figure, especially in light of his 2024 reelection campaign. Second, it utilizes a machine learning approach to systematically analyze and compare the topics and sentiments in Trump's rally speeches and their coverage by liberal and conservative sources.

This study explores the nuances of media bias in the portrayal of Trump's rally speeches by contrasting liberal and conservative news coverage. It analyzes both the topics and sentiments within the speeches and their respective media coverage, shedding light on how ideological biases influence political polarization and public opinion. Understanding these dynamics is essential for cultivating more informed media consumption among the public.

II. METHODOLOGY

A. Dataset

In this study, data was collected from three sources. The first dataset is a collection of 35 transcriptions of Trump's rally speeches from Kaggle during the 2020 presidential election cycle (Lillelund, 2020). For each transcription, manual data collection was conducted to find one news article from a liberally aligned news source (N = 35) and one from a conservatively aligned news source (N = 35), with the political alignment of sources verified using Media Bias / Fact Check or MBFC (Zandt, 2024). Additionally, to ensure an equal distribution of speeches news sources were collected from across national news outlets (e.g., CNN, ABC, CBS) and local news outlets (e.g., Minnesota Reformer, Denver Post) for liberally-aligned news sources. A similar distribution was attempted to be achieved regarding conservative news sources.

A set of criteria was established during the data collection process. First, articles needed to be of a sufficient length(>1500 words) to provide a comprehensive analysis of the rally speech. Short articles were excluded to ensure substantial content for sentiment analysis. Second, articles were required to focus specifically on the rally speech being examined and general articles about Trump that may have been published at the time of the rally speech were excluded. As stated previously, it was attempted to achieve an even distribution of national and local news outlets to capture a range of perspectives within each political alignment. Lastly, articles published within a week of the rally were preferred to ensure timeliness and relevance.

B. Methodologies for Research Question 1

To check the alignment of topics across the three datasets, the topics first needed to be identified in all the article and rally speech data. While political alignment was measured implicitly through our dataset collection (i.e., using Media Bias / Fact Check) other concepts of interest such as the topics in Trump's rally speeches and the topics in the corresponding media coverage were explored through topic modeling.

Topic modeling typically utilizes unsupervised machine learning to generate topic clusters in the form of words based on text documents (Vayansky & Sathish, 2020). Additionally, topic modeling does not return the names of the topics themselves but rather the words that constitute that topic. Topic modeling also requires a corpus of documents to perform the topic modeling.

In this study, both BERTopic and Latent Dirichlet Allocation (LDA) topic modeling were considered. BERTopic leverages the capabilities of BERT to perform topic modeling. BERT, which stands for Bidirectional Encoder Representations from Transformers, is a language representation model developed by Google AI Language (Devlin et al., 2019). BERT utilizes a bidirectional approach, allowing it to understand the context from both directions (Devlin et al., 2019). This differs from traditional language models used in studies that process text left-to-right or right-to-left (Peters et al., 2018), thus enabling topic modeling that takes contextual factors into account. Furthermore, this bidirectional approach differs BERT from LDA modeling (Jelodar et al., 2018), which is a probabilistic model that generates topics based on word co-occurrence patterns. Building off of BERT's capabilities, BERTopic can thus handle the complexity and nuance of language better than LDA (Egger, 2022). By using BERT embeddings (Devlin et al., 2019), it captures the meaning of words in context, which is especially useful in this context due to the large length of Trump's articles (>20000 words). Thus, BERTopic was chosen as the topic modeling model to use for this study.

Before topic modeling can be carried out though, several preprocessing steps were first applied to the text data to reduce noise. These steps included tokenization (Schmidt, 2024), stemming, and the removal of stopwords. Tokenization into sentences (rather than feeding the entire text in) and then feeding the sentences as documents into BERTopic proved beneficial because, as mentioned in previous studies (Grootendorst, 2022), it enables the sentences' individual semantic contributions to generate topics to provide more granular performance. Stop word removal is similarly beneficial as it reduces the influence and smoothing effect of commonly occurring, non-informative terms on the model (Schofield et al., 2017). Stemming has also been used in previous studies with topic modeling (Liu et al., 2010; Lo et al., 2015).

Due to the unsupervised nature of topic modeling with BERT, the model requires the user to specify the number of topics to identify. Consequently, BERTopic outputs clusters of words that represent the discovered topics, without predefined labels or categories. Thus, in this study, an optimal number of 5 topics was chosen for BERTopic, as increasing the number of topics led to noise and reduced clarity. The noise was inferred through researcher interpretation.

While topic modeling produced good results when applied to Trump's rally speeches, identifying 5 clearly distinct topics, it was less successful in the liberal and conservative media coverage of these speeches. The main reason for this was the length of the articles (Groot et al., 2022). The articles were relatively short and lacked the variety of content found in the rally speeches. Consequently, BERTopic struggled to find enough meaningful patterns, often returning only a single outlier topic (i.e., a topic with noisy words in its clustering, noise was defined through researcher interpretation), similar to past studies (Groot et al., 2022), which indicates that the content did not fit well into any coherent topic clusters. To resolve this, LDA was found to be better suited for handling shorter texts and provided more coherent topic modeling compared to BERTopic, despite previous studies by Gan et al. (2024) suggesting otherwise. However, this presented two significant challenges.

First, it was necessary to determine how to compare the topics. Since the outputs were unsupervised and did not include predefined topic names, one option was to manually label the topics and then compare the similarities between topic categories. However, this approach was rejected due to the risk of bias (Wegge et al., 2024). Instead, calculating the cosine similarity between the three sources was chosen as a metric for comparison. This decision was based on prior research (Gan et al., 2024) that demonstrated cosine similarity as a reliable measure. Computing the cosine similarity also helped fit time constraint limitations as it allowed this study to compare the words in topic clusters rather than having to manually label them.

Cosine similarity calculates the cosine of the angle between two vectors to determine how similar two vectors are(Xia et al., 2015), regardless of their magnitude. This metric is often used in text analysis to compare the similarity between vectorized words or documents (Park et al., 2020; Lahitini, 2016). The cosine similarity ranges from -1 to 1, where 1 indicates identical vectors, 0 indicates orthogonal (no similarity) vectors, and -1 indicates opposite vectors.

Second, while preprocessing steps were kept the same, LDA outputs topics as collections of N-grams (multi-word phrases), whereas BERTopic generates clusters of single words. This difference created a challenge in comparing topics, as the formats were not directly compatible for cosine similarity analysis. This challenge was overcome by applying a preprocessing function to transform N-grams into single words to align the formats of LDA and BERTopic outputs.

The text was also further flattened by removing weights and colons from outputs returned by BERTopic and joining words returned by LDA into a single string. Using this newly flattened data, three corpora were created: one combining liberal articles with Trump's rally speeches, another combining conservative articles with Trump's rally speeches, and the third combining liberal articles with conservative articles.

These corpora were then vectorized using TF-IDF (Term Frequency-Inverse Document Frequency). TF-IDF is the significance of a word to a document in a corpus (MonkeyLearn, 2019). The TfidfVectorizer converts a collection of raw documents into a matrix of TF-IDF features (Awan, 2022) and was chosen due to its validity in previous studies (Park et al., 2020). This step is necessary because it transforms text into numerical vectors, allowing for the quantification and comparison of textual data with cosine similarity. Without this transformation, it would be impossible to compute cosine similarity, as raw text cannot be directly compared in a meaningful way. Additionally, another advantage of TFIDFVectorizer versus other common vectorizers (e.g., CountVectorizer (Suryaningrum, 2023)) is that it reduces the impact of commonly occurring words (Simha, 2021).

Then, for each rally speech, cosine similarities are calculated for three comparisons: liberal vs. conservative, liberal vs. Trump’s rally speeches, and conservative vs. Trump’s rally speeches. This involves computing the cosine similarity between each topic from one source (e.g., Trump’s speeches) and each topic from another source (e.g., liberal articles). The result is a two-dimensional similarity matrix (Figure 1), where each row represents a topic from one source and each column represents a topic from the other source. Each element in the matrix is the cosine similarity score between a topic from each source. These similarity scores are then averaged to produce a single cosine similarity score for each comparison. This process is repeated for all 35 rally speeches and is visualized through Figure 2.

To provide additional context to the cosine similarity scores and topic alignment, the top ten topics for the liberal and conservative news media coverage corpus were identified. To achieve this, Top2Vec, a neural network-based topic modeling method (Angelov, 2020), was utilized to extract the top ten topics from both liberal and conservative news sources. These vectors were reduced to a lower-dimensional space with UMAP and clustered using HDBSCAN to group similar document vectors into topics, represented by the centroid of each cluster. Top2Vec identified the top ten topics through their keywords and frequencies (Angelov, 2020). A GPT-4 architecture-based model, ChatGPT, was then used to classify these topics by inputting the keywords for each one. Additionally, while BERT and LDA were tested to perform overall corpus-based topic modeling, they resulted in topics that were too context-specific, whereas Top2Vec provided more general topics. Due to Top2Vec's requirement for several thousand documents, data was replicated: 35 articles (from both liberal and conservative sources) were duplicated 58 times, resulting in 2,030 documents for the analysis. The topics identified were then visualized using Matplotlib (Figure 3 and Figure 4).

C. Methodologies for Research Question 2

Sentiment constitutes the emotional tone of a piece of text. Sentiment analysis acts as a computational tool used to evaluate the emotional tone by classifying the text as positive, negative, or neutral or by evaluating the degree of these three sentiment categories (Wankhade, 2022).

Once again, political alignment was inherently measured through our dataset collection (i.e., using Media Bias / Fact Check) whereas sentiment was measured through sentiment analysis on a scale of -1 to 1. Based on previous quantifications of sentiment (Qi & Shabrina, 2023; Hutto & Gilbert, 2014), in this study -1 represents extremely negative sentiment, 0 represents neutral sentiment, and 1 represents extremely positive sentiment.

Initially, a selection of the appropriate type of sentiment analysis model was determined. A comparison was made between BERT-based models and VADER. VADER (Valence Aware Dictionary and sentiment Reasoner) is a lexicon and rule-based sentiment analysis model (Hutto & Gilbert, 2014). However, it relies on a predefined dictionary and simple rules, which often results in a lack of nuance in understanding complex language (Barik et al., 2024). That is, it may not perform well in certain domains or with complex sentences because VADER relies heavily on the lexicon, so if a word is not present in the lexicon, it may not be accurately classified. On the other hand, BERT-based models utilize deep learning and transformer architecture, which allows them to handle more nuanced language by taking into account bidirectional context (Egger, 2022). Additionally, Trump's rally speeches are often very long and contain extensive segments of discursive content. Though there is a token limitation to BERT-based models (Vijayan et al., 2024), it can process these long texts effectively with proper modifications (e.g., CogLTX (Ding et al., 2020)).

After choosing a BERT-based approach, further model specification was needed. Initially, the DistilBERT model was tested as it was the default model for sentiment analysis. The DistilBERT model is a distilled version of BERT, that is it aspires for the same performance as BERT while being smaller, faster, and more efficient (Sanh et al., 2020). Regarding preprocessing, tokenization was performed by sentences. This tokenization approach was chosen for two reasons: to facilitate sentiment analysis using the BERT model on individual sentences by averaging sentiment scores for each sentence post-hoc and to meet the token limitations of the BERT model.

However, despite splitting the text up into sentences and averaging post-hoc, a significant challenge arose. The DistilBERT model, like other BERT-based models, has a maximum input sequence length of 512 tokens. This means that the model can process inputs up to 512 tokens in length. Yet, even after sentence-based tokenization, the input text length exceeds the maximum sequence length that the RoBERTa model can handle (512 tokens) causing a tensor dimension mismatch. To resolve this, the preprocessed text is split into smaller chunks of maximum size (192 tokens), ensuring that each chunk fits within the model's input limits. This prevents the tensor size mismatch error by processing each segment separately for sentiment analysis. This conservative token size, well below the RoBERTa model's 512-token maximum, was chosen as it accounts for additional tokens added during tokenization and avoids input size errors.

The DistilBERT also outputs probability scores—not sentiment scores—as a list of dictionaries, where each dictionary contains two key-value pairs. The key is a label indicating positive or negative sentiment classification and the value indicates the probability of the text being classified into the particular sentiment category of the label (i.e., a probability score). To convert these probabilities scores into an interpretable format a compound score formula (Figure 5) was used.

However, after predicting sentiment scores with the DistilBERT model and applying the compound score formula, its predicted sentiment scores did not align well with manual sentiment evaluations. This occurred because it did not adequately discount neutral sentiments. Thus, the Cardiffnlp roBERTa-based model was selected instead as it allowed for discounting neutral sentiments and thus provided sentiment scores that better matched manual evaluation. The model discounted neutral sentiments by providing a third key-value pair that indicates the probability of the text being classified as neutral (Barbieri et al., 2020), which was then factored into the compound score formula by ignoring the neutral score. This approach improved the compound score formula by focusing solely on the positive and negative sentiment, while also taking into account neutral sentiment.

Finally, the sentiment analysis process was repeated for all articles and visualized using histograms and density plots. The histograms displayed the frequency distribution of sentiment scores for liberal, conservative, and actual speech categories, while the density plots provided a smoothed representation of the sentiment score distributions across these categories (Figure 6).

III. RESULTS

The analysis of cosine similarity scores reveals no statistically significant differences in topic alignment between Trump's rally speeches and media coverage from liberal and conservative sources. This was determined through an ANOVA test, which returned a p-value of statistical significance, as it was greater than the chosen alpha value of 0.05 (F = 1.238, p = 0.298).

Despite the lack of statistically significant differences, there are spikes within the low values, defined as more than two standard deviations away from the mean cosine similarity. For liberal coverage versus Trump's speeches, cosine similarity scores ranged from 0.003 to 0.054. With a mean cosine similarity of 0.017 and a standard deviation of 0.013, the spikes are values greater than 0.04. In this range, there is one spike with a cosine similarity score of 0.055 (New Mexico rally on Sep. 16, 2019). For conservative coverage versus Trump, cosine similarity scores varied from 0.0 to 0.057. With a mean cosine similarity of 0.025 and a standard deviation of 0.016, spikes are values greater than 0.057. In this range, there is one spike at a score of 0.084 (Minden rally on Sep. 12, 2020). The similarity scores between liberal and conservative media range from 0.0 to 0.122. With a mean cosine similarity of 0.027 and a standard deviation of 0.027, spikes are values greater than 0.081. In this range, there is one spike with a cosine similarity score of 0.12 (Freeland rally on Sep. 10, 2020). However, while there are moments of higher alignment, these values are still low overall within the context of cosine similarity's range from -1 to 1.

To analyze the context of these low cosine similarity scores, topic modeling was performed on the overall corpus—rather than individual articles from each news source—of both liberal and conservative news sources. Ten topics were extracted for each corpus. The results from our Top2Vec topic modeling reveal significant differences in the topics emphasized by liberal and conservative news sources, with liberals focusing on controversial issues like immigration and deportation, while conservatives highlight topics like economics and partisan politics. All of the topics can be seen through Figure 3 and Figure 4.

Additionally, in terms of the spikes, a qualitative analysis was performed to evaluate the statistical results. For instance, in the comparison of average cosine similarities comparing topics identified in Trump's rally speeches to topics identified in conservative coverage, the Minden rally on Sep.12, 2019, stood out as a spike or high outlier (i.e., more than two standard deviations away from the average), indicating relatively higher similarity. To perform qualitative analysis, three evaluators read through the corresponding articles and determined this relatively high similarity resulted from the unique context of the Nevada rally. Specifically, this rally occurred during a critical period of debate over mail-in voting, which was a significant local issue in Nevada at the time. Trump's emphasis on alleged voter fraud aligned strongly with the conservative media's agenda. Similar trends emerged through qualitative analysis of other high outliers (e.g., Trump's New Mexico rally on Sep. 16, 2,019 within cosine similarities between liberal coverage versus Trump's rally speeches).

An overall qualitative analysis was also conducted across articles. Three evaluators came to the consensus that overall both liberal and conservative news media outlets tend to focus on different topics in the same negative context, creating a heavy negative sentiment. For instance, taking the Minden rally as an example, it was found that liberal media coverage, instead of focusing on the issues presented by Trump in his rally speech, instead focused on the consequences of holding that rally speech.

When analyzing sentiment scores with regards to political alignment for this study's second research question, as qualitatively observed through Figure 6, it was found that the sentiment scores for transcriptions of

Trump's rally speeches were generally positive. In contrast, liberal news coverage demonstrated predominantly negative sentiment, while conservative news coverage, although also negative overall, demonstrated a more varied sentiment score distribution. The liberal sentiment scores revealed a standard deviation of 0.32; conservative sentiment scores demonstrated a standard deviation of 0.43, indicating a relatively higher degree of variability; and Trump's rally speeches exhibited a standard deviation of 0.15, indicating a relatively low variability among the distribution of sentiment scores.

These findings can be more clearly visualized through Figure 7 as they aggregate these results into a three-column bar graph of the average sentiment score. Figure 7 aligns with prior qualitative observations, revealing that both Liberal (-0.32) and Conservative (-0.28) news sources demonstrate negative average sentiment scores, with Liberal sources skewing slightly more negatively. In contrast, the actual sentiment scores from Trump's speeches themselves are slightly positive (0.15). Consequently, an ANOVA was conducted to test for significant differences, followed by t-tests to determine where these significant differences occurred, if they did. Choosing the standard alpha level of 0.05, an ANOVA test confirmed significant statistical differences across all groups ($F = 23.500$, $p < 0.001$). However, three separate t-tests indicated no significant difference between liberal and conservative media sentiment ($t = -0.389$, $p = 0.704$), but significant differences were found when comparing both liberal ($t = -7.843$, $p < 0.001$) and conservative ($t = -5.704$, $p < 0.001$) coverage to the actual speeches. Therefore, the most significant difference in sentiment scores is between liberal coverage and Trump's rally speeches, followed by conservative coverage and Trump's rally speeches, while there is no significant difference between liberal and conservative media coverage.

To investigate this disparity it was found by reading and analyzing articles and comparing the evaluations of three readers that both liberal and conservative media use more negative language when criticizing the other party's views on Trump's speeches. For instance, in the coverage of Trump's Minden rally, conservative news outlets prominently highlighted Trump's negative statements about Democrats, using language such as "damn," "democrats [...] allow churches to burn," and "domestic terrorists." In contrast, liberal coverage focused heavily on Trump's comments about COVID-19 statistics, emphasizing the death tolls and other negative aspects of the pandemic, which the conservative articles did not mention at all. Similar trends were observed across other articles, showcasing a consistent pattern of bias through selective emphasis and language use.

IV. DISCUSSION

In regards to research question one, as stated previously, the analysis of cosine similarity scores between Trump's rally speeches and media coverage from both liberal and conservative sources reveals negligible statistically significant differences in topic alignment. The average similarity scores remain low across all comparisons, indicating that the topics discussed in Trump's speeches are generally not strongly mirrored in media coverage. According to second-level agenda setting, media outlets present news stories to emphasize certain attributes while downplaying others, influencing how the audience perceives issues (McCombs & Valenzuela, 2014). In alignment with this theory, both liberal and conservative media might frame Trump's speeches in ways that align with their own narrative goals, which would account for the low cosine similarity scores, as each outlet may selectively highlight different elements of the speeches to fit their narrative and political alignment.

The inductive (data-first) analysis conducted in this study validates prior findings of overall low cosine similarity scores. The analysis empirically provides context to the circumstances surrounding these low cosine similarity scores, providing evidence rather than merely hypothesizing that the differences are due to the framing of these issues. By seeing which of the topics in Trump's speeches are emphasized by the different sources it is justified that the similarity scores were quite low. This analysis supports this study's prior statements that liberal and conservative outlets selectively frame and emphasize different topics of the speeches to align with their respective narratives.

Previous studies (Mitchell et al., 2017) find that right-leaning or conservative media outlets were more likely to express positive sentiments regarding Trump's rally speeches, while liberal source coverage tended to refute Trump's claims. However, findings from this study contradict these assumptions, revealing that both conservative and liberal news coverage exhibit a predominantly negative sentiment compared to Trump's rally speeches. A qualitative analysis provides further insights into this contradiction. Specifically, it was found that while both liberal and conservative media outlets focus on different aspects of Trump's speeches, they do so within a negative context. Liberal media tends to disproportionately criticize conservative ideologies by emphasizing particular topics from Trump's speeches and portraying them negatively. On the other hand, conservative media, instead of highlighting Trump's positive rhetoric, often focuses on his negative critiques of the opposing liberal party. These findings align with previous studies indicating that news media outlets exhibit ideological bias not by explicitly endorsing a political party but by disproportionately criticizing the opposing group, thereby moderating overall differences (Budak et al., 2016).

However, while overall cosine similarity scores remained low, there were still relative peaks in the scores which were qualitatively analyzed for trends. These peaks in similarity were found to occur largely due to the unique context of the rally. The significance of these peaks lies in the view that by emphasizing certain topics, such as voter fraud during the Minden rally as discussed in the results, a certain comparison group (e.g., conservative coverage of Trump for the Minden rally) is shaping the public agenda and ensuring these issues remain at the forefront of public discourse. This correlates with ingroup and outgroup dynamics, where a certain comparison group can form an ingroup by selectively giving certain topics greater emphasis within their group. The selective emphasis on certain topics or topic attributes relates back to the agenda-setting theory, which posits the media shapes public perception by giving specific issues and attributes greater salience(McCombs & Valenzuela, 2014).

Furthermore, the implications of this study are important for political communication. The portrayal of events, such as Trump's rally speeches, by various news sources not only influences public perception of these events and political figures but also influences voter decision-making and exacerbates existing political polarization in America, which can have a significant impact on electoral outcomes (Bernhardt et al., 2008).

This exacerbation also directly relates to ingroup and outgroup dynamics, where media outlets cultivate a political ingroup by increasing negative perceptions and hostility towards their political outgroup. In this study, this is evident through the differing emphasis on topics by the outlets. This topic emphasis aligns with first-level agenda setting, where media outlets increase the salience of certain issues by featuring them more prominently.

In future studies, the timeframe of the data analyzed could be expanded, making this study a longitudinal one. Rather than examining these relationships over one presidential election cycle, they could be examined over multiple election cycles, or even across several years, outside of the campaign season to see whether similar patterns are validated or contradicted. This would also enable a larger dataset and more comprehensive machine-learning techniques. Future research could focus on other political actors, such as governors, judges, or Supreme Court justices, rather than only presidential candidates like Trump. It could also examine whether demographic factors like race, gender, and age influence media coverage. This is particularly relevant to the 2024 election cycle when considering Kamala Harris as a potential presidential candidate. As a Black and South Asian woman, she represents a significant demographic difference compared to Trump.

Building on the insights from topic alignment, the analysis of sentiment scores in response to the second research question also provides an understanding of how media coverage shapes public perception of Trump's speeches.

Although it is generally assumed that there is a significant disparity in the sentiment in which news sources with opposing political alignments portray Trump (McCombs & Shaw, 1972), findings in this study reveal that there is not a significant difference in sentiment between these sources. That is, statistical analyses revealed negligible differences between sentiment expressed between liberal and conservative news coverage. These results are further reinforced by findings from previous studies which found that, except for political scandals, major news organizations present topics in a largely nonpartisan manner. (Budak et al., 2016).

However, when comparing the sentiment scores of liberal and conservative news coverage to Trump's actual speeches using statistical analysis, significant differences were found. Both liberal and conservative outlets tend to portray a more negative sentiment than what Trump conveys in his rally speeches. Additionally, liberal articles demonstrated more negative distribution compared to conservative articles. These results were explained through a qualitative analysis: the analysis revealed that both liberal and conservative media use more negative language when criticizing the opposing party's views on Trump's speeches.

Once again, these findings align with previous studies that news media outlets show their ideological bias not by explicitly advocating for a chosen political party, but instead by disproportionately criticizing the other group, a norm that further moderates overall differences (Budak et al., 2016). This criticism then reinforces negative sentiment throughout both sources. The findings also suggest that media outlets use in-group and out-group dynamics to frame their coverage. This means they may not directly support their preferred political party but instead use more negative language to criticize the opposing side.

Recognizing these patterns can help individuals interpret news with more insight and factor that into their voting decision-making process. With better media literacy, individuals can make more informed decisions about which news sources to trust and how to interpret the information presented(Eagle, 2007).

This also aligns with second-level agenda-setting theory, which posits that the media can shape public perception by emphasizing certain attributes (McCombs & Valenzuela). This selective framing of the issue by emphasizing issue attributes reinforces the idea that media outlets can influence public perception through the strategic presentation of information. This study provides awareness of media bias, and this awareness can potentially reduce the likelihood of individuals being influenced by biased reporting (Lord & Vogt, 2021). In the political scene, this enables voters to see through selective emphasis and form more balanced opinions.

For a more complex analysis, future studies may include more speeches and news articles over a longer period (i.e., longitudinal studies). This may help determine if the patterns observed in this study are consistent over time and across different contexts. Additionally, studies could involve a more detailed analysis of various aspects of sentiment related to Trump, such as emotional tone, language complexity, and the context in which Trump is mentioned, aspects that were not examined in the sentiment evaluation. This further research would not only substantiate the findings of the current study but also potentially uncover new trends not identified in the initial analysis.

Although this study employed thorough methodologies, it is not without limitations. One limitation is the lack of variety in the news sources analyzed, which may not fully represent the media spectrum and could affect the generalizability of the findings. For instance, conservative news coverage was disproportionately represented by Fox News due to a lack of coverage of Trump's rally speeches from other conservative outlets. Another limitation was the inability to apply BERT models to the liberal and conservative news media coverage for topic modeling due to the shorter length of these articles—instead, LDA topic modeling was used for the latter news sources. BERT models typically perform better with longer texts, and the short length of the news articles limited the effectiveness of this approach. It is important to note, however, that this limitation was unique to the context of this study as previous studies (Gan et al., 2024) have suggested that BERT-based models provide more coherent topic clusters than LDA on short texts. Using BERT to analyze topics in liberal and conservative articles may have been advantageous because it leverages deep learning to understand the context, semantics, and nuances of language, leading to more accurate sentiment analysis. Another limitation is a shortage of data due to Trumponly giving 35 rally speeches in the period examined in this study. This compromised the robustness of the Top2Vec model, as Top2Vec requires a larger dataset to function effectively.

Ultimately, the combined findings from both research questions in this study present novel insights. Although they covered the same rally speeches, liberal and conservative media outlets emphasized different topics in their reporting. The manner in which the sources discussed certain topics—focusing on criticism rather than advocacy—resulted in similar negative sentiments in both liberal and conservative coverage, contradicting initial beliefs. The implications of this study rely on two key points. First, media framing can shape the public agenda, reinforcing existing biases and contributing to political polarization by presenting news in a way that aligns with the ingroup and outgroup's ideological stance. Then, this study also revealed that media practices contribute to an environment of reinforcing negative perceptions through outgroup criticism. However, by revealing and validating these findings with computational techniques, this study lays the groundwork for improved media literacy. This awareness enables the public to critically evaluate news sources, influencing the public agenda and potentially shaping the policy agenda. This, in turn, promotes informed decision-making and may contribute to reducing polarization.

REFERENCES

Angelov, D. (2020). Top2Vec: Distributed Representations of Topics. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2008.09470

Awan, A. A. (2022, September 7). *Convert text documents to a TF-IDF matrix with tfidfvectorizer*. KDnuggets. https://www.kdnuggets.com/2022/09/convert-text-documents-tfidf-matrix-tfidfvectorizer.html

Barik, K., & Misra, S. (2024c). Analysis of customer reviews with an improved Vader Lexicon classifier. *Journal of Big Data*, *11*(1). https://doi.org/10.1186/s40537-023-00861-x

Barbieri, F., Camacho-Collados, J., Espinosa Anke, L., & Neves, L. (2020). Tweeteval: Unified benchmark and comparative evaluation for Tweet Classification. *Findings of the Association for Computational Linguistics: EMNLP 2020*. https://doi.org/10.18653/v1/2020.findings-emnlp.148

Benkler, Y., Tilton, C., Etling, B., Roberts, H., Clark, J., Faris, R., Kaiser, J., & Schmitt, C. (2020). Mail-in voter fraud: Anatomy of a disinformation campaign. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3703701

Bernhardt, D., Krasa, S., & Polborn, M. K. (2006). Political polarization and the electoral effects of media bias. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.892475

Budak, C., Goel, S., & Rao, J. M. (2016). Fair and balanced? quantifying media bias through crowdsourced content analysis. *Public Opinion Quarterly*, *80*(S1), 250–271. https://doi.org/10.1093/poq/nfw007

Coleman, R., & Wu, H. D. (2010). Proposing emotion as a dimension of affective agenda setting: Separating affect into two components and comparing their second-level effects. *Journalism & Mass Communication Quarterly*, *87*(2), 315–327. https://doi.org/10.1177/107769901008700206

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018a). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1810.04805

Ding, M., Zhou, C., Yang, H., & Tang, J. (2020). CogLTX: Applying BERT to Long Texts. *Neural Information Processing Systems*, *33*, 12792–12804. https://proceedings.neurips.cc/paper/2020/file/96671501524948bc3937b4b30d0e57b9-Paper.pdf

Eagle, L. (2007). Commercial media literacy: What does it do, to whom—and does it matter? *Journal of Advertising*, *36*(2), 101–110. https://doi.org/10.2753/joa0091-3367360207

Egger, R., & Yu, J. (2022). A topic modeling comparison between LDA, NMF, Top2Vec, and Bertopic to demystify twitter posts. *Frontiers in Sociology*, *7*. https://doi.org/10.3389/fsoc.2022.886498

Evers, M. M., Fisher, A., & Schaaf, S. D. (2019). Is there a trump effect? an experiment on political polarization and audience costs. *Perspectives on Politics*, *17*(02), 433–452. https://doi.org/10.1017/s1537592718003390

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2203.05794

Groseclose, T., & Milyo, J. (2005). A measure of media bias. *The Quarterly Journal of Economics*, *120*(4), 1191–1237. https://doi.org/10.1162/003355305775097542

GUNTHER, A. C. (1998). The Persuasive Press Inference. *Communication Research*, *25*(5), 486–504. https://doi.org/10.1177/009365098025005002

Hutto, C., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, *8*(1), 216–225. https://doi.org/10.1609/icwsm.v8i1.14550

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2018). Latent dirichlet allocation (LDA) and topic modeling: Models, applications, a survey. *Multimedia Tools and Applications*, *78*(11), 15169–15211. https://doi.org/10.1007/s11042-018-6894-4

Kuypers, J. A. (2020). *President Trump and the news media: Moral Foundations, framing, and the nature of press... Bias in America*. Rowman & Littlefield.

Lahitani, A. R., Permanasari, A. E., & Setiawan, N. A. (2016). Cosine similarity to determine similarity measure: Study case in Online essay assessment. *2016 4th International Conference on Cyber and IT Service Management*. https://doi.org/10.1109/citsm.2016.7577578

Lillelund, C. (2020, September 26). *Trumprally speeches*. Kaggle. https://www.kaggle.com/datasets/christianlillelund/donald-trumps-rallies

Lo, S. L., Cornforth, D., & Chiong, R. (2015). Effects of training datasets on both the extreme learning machine and support vector machine for target audience identification on Twitter. *Proceedings of ELM-2014 Volume 1*, 417–434. https://doi.org/10.1007/978-3-319-14063-6_35

Matos, Y., & Miller, J. L. (2021). The politics of pronouns: How trump framed the ingroup in the 2016 presidential election. *Politics, Groups, and Identities*, *11*(3), 507–525. https://doi.org/10.1080/21565503.2021.2007964

McCombs, M. E., Shaw, D. L., & Weaver, D. H. (2014). New Directions in agenda-setting theory and research. *Mass Communication and Society*, *17*(6), 781–802. https://doi.org/10.1080/15205436.2014.964871

Media Bias/Fact Check News. Media Bias/Fact Check. (2024, February 19). https://mediabiasfactcheck.com/

Muriël, D. G., Aliannejadi, M., & Haas, M. R. (2022). Experiments on Generalizability of BERTopic on Multi-Domain Short Text. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2212.08459

Mullainathan, S., & Shleifer, A. (2005). The market for news. *American Economic Review*, *95*(4), 1031–1053. https://doi.org/10.1257/0002828054825619

Park, K., Hong, J. S., & Kim, W. (2020). A methodology combining cosine similarity with classifier for text classification. *Applied Artificial Intelligence*, *34*(5), 396–411. https://doi.org/10.1080/08839514.2020.1723868

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers).* https://doi.org/10.18653/v1/n18-1202

McCombs, M. E., & Shaw, D. L. (1972). The agenda-setting function of mass media. *Public Opinion Quarterly*, *36*(2), 176. https://doi.org/10.1086/267990

Qi, Y., & Shabrina, Z. (2023). Sentiment analysis using Twitter data: A Comparative Application of lexicon- and machine-learning-based approach. *Social Network Analysis and Mining*, *13*(1). https://doi.org/10.1007/s13278-023-01030-x

Ramaraj, V., Appa Swamy, M. V., Prince, E. E., & Kumar, C. (2024). Improving the bert model for long text sequences in question answering domain. *International Journal of Advances in Applied Sciences*, *13*(1), 106. https://doi.org/10.11591/ijaas.v13.i1.pp106-115

Rozado, D., Hughes, R., & Halberstadt, J. (2022). Longitudinal analysis of sentiment and emotion in news media headlines using automated labeling with Transformer language models. *PLOS ONE*, *17*(10). https://doi.org/10.1371/journal.pone.0276367

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1910.01108

Schmidt, C. W., Reddy, V., Zhang, H., Alameddine, A., Uzan, O., Pinter, Y., & Tanner, C. (2024). Tokenization Is More Than Compression. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2402.1837.

Schofield, A., Magnusson, M., & Mimno, D. (2017). Pulling out the stops: Rethinking stopword removal for topic models. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. https://doi.org/10.18653/v1/e17-2069

Shultziner, D., & Goldberg, S. (2018). The stages of mass mobilization: Separate phenomena and distinct causal mechanisms. *Journal for the Theory of Social Behaviour*, *49*(1), 2–23. https://doi.org/10.1111/jtsb.12187

Simha, A. (2021, October 6). *Understanding TF-IDF (term frequency-inverse document frequency)*. GeeksforGeeks. https://www.geeksforgeeks.org/understanding-tf-idf-term-frequency-inverse-document-frequency/

Spinde, T., Jeggle, C., Haupt, M., Gaissmaier, W., & Giese, H. (2022). How do we raise media bias awareness effectively? effects of visualizations to communicate bias. *PLOS ONE*, *17*(4). https://doi.org/10.1371/journal.pone.0266204

Strengthen Media Literacy to Win the Fight Against Misinformation (SSIR). (n.d.). (C) 2005-2024. https://ssir.org/articles/entry/strengthen_media_literacy_to_win_the_fight_against_misinformation

Suryaningrum, K. M. (2023). Comparison of the TF-IDF method with the count vectorizer to classify hate speech. *Engineering, Mathematics and Computer Science (EMACS) Journal*, *5*(2), 79–83. https://doi.org/10.21512/emacsjournal.v5i2.9978

Tajfel, H., & Turner, J. (1979). An integrative theory of intergroup conflict. *Organizational Identity*, 33–47. https://doi.org/10.1093/oso/9780199269464.003.0005

Tyson, A. (2018, November 14). *America's polarized views of trump follow years of growing political partisanship*. Pew Research Center. https://www.pewresearch.org/short-reads/2018/11/14/americas-polarized-views-of-trump-follow-years-of-growing-political-partisanship/

Vayansky, I., & Kumar, S. A. P. (2020a). A review of Topic modeling methods. *Information Systems*, *94*, 101582. https://doi.org/10.1016/j.is.2020.101582

Wegge, M., & Klinger, R. (2023). Topic Bias in Emotion Classification. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2312.09043

What drives media slant? evidence from U.S. Daily Newspapers. (2010). *Econometrica*, *78*(1), 35–71. https://doi.org/10.3982/ecta7195

Xia, P., Zhang, L., & Li, F. (2015). Learning similarity with cosine similarity ensemble. *Information Sciences*, *307*, 39–52. https://doi.org/10.1016/j.ins.2015.02.024

ACKNOWLEDGEMENTS

We would like to thank our instructor, Musa Malik, for creating an interactive and engaging learning environment during lectures. His guidance enabled us to incorporate lecture theories into our research paper effectively. We also appreciate his overall feedback with coding, especially during the initial stages of choosing models and addressing complications that arose later. We also want to thank our teaching assistant, Sovannie Len, for her valuable insight into the research paper writing process. We also would like to thank Brittany Wheeler, our other teaching assistant, for providing us help with programming and the paper writing process. Lastly, we would like to thank Dr. Lina Kim, Dr. Teresa Holden, and the UCSB Summer Research Academies' staff for making the research conducted possible.

AUTHOR CONTRIBUTION STATEMENT

T.B. worked on the code for topic modeling with BERT and LDA, sentiment analysis with BERT, and their corresponding graphs for data visualization and statistics. T.B. also contributed to the abstract, introduction, wrote the entire methods section (excluding inductive paragraph), results, conclusion, and creating the capstone seminar presentation. Additionally, T.B. found the dataset for Trump's rally speeches and collected the liberally-aligned articles. A.P. worked on the code for topic modeling using Top2Vec, along with its corresponding graphs for data visualization and statistics. A.P. also contributed to the abstract, introduction, inductive reasoning of the methods section, results, conclusion, and creating the capstone seminar presentation. A.P. cleaned the article data by extracting content from websites into text files. M.P. worked on the introduction, creating the capstone seminar presentation, and proofreading the paper. M.P. also collected the conservatively-aligned articles.

APPENDIX

Figure 1

Similarity Matrix of Cosine Similarity Scores for A Single Comparison

	L_1	L_2	L_3	L_4	L_5
T_1	$\cos(T_1, L_1)$	$\cos(T_1, L_2)$	$\cos(T_1, L_3)$	$\cos(T_1, L_4)$	$\cos(T_1, L_5)$
T_2	$\cos(T_2, L_1)$	$\cos(T_2, L_2)$	$\cos(T_2, L_3)$	$\cos(T_2, L_4)$	$\cos(T_2, L_5)$
T_3	$\cos(T_3, L_1)$	$\cos(T_3, L_2)$	$\cos(T_3, L_3)$	$\cos(T_3, L_4)$	$\cos(T_3, L_5)$
T_4	$\cos(T_4, L_1)$	$\cos(T_4, L_2)$	$\cos(T_4, L_3)$	$\cos(T_4, L_4)$	$\cos(T_4, L_5)$
T_5	$\cos(T_5, L_1)$	$\cos(T_5, L_2)$	$\cos(T_5, L_3)$	$\cos(T_5, L_4)$	$\cos(T_5, L_5)$

Note. T_i represents topics from Trump's rally speech, where i ranges from 1 to 5. L_j represents topics from liberal articles, where j ranges from 1 to 5. $\cos(T_i, L_j)$ is the cosine similarity between the i-th topic from Trump's speech and the j-th topic from liberal articles.

Figure 2

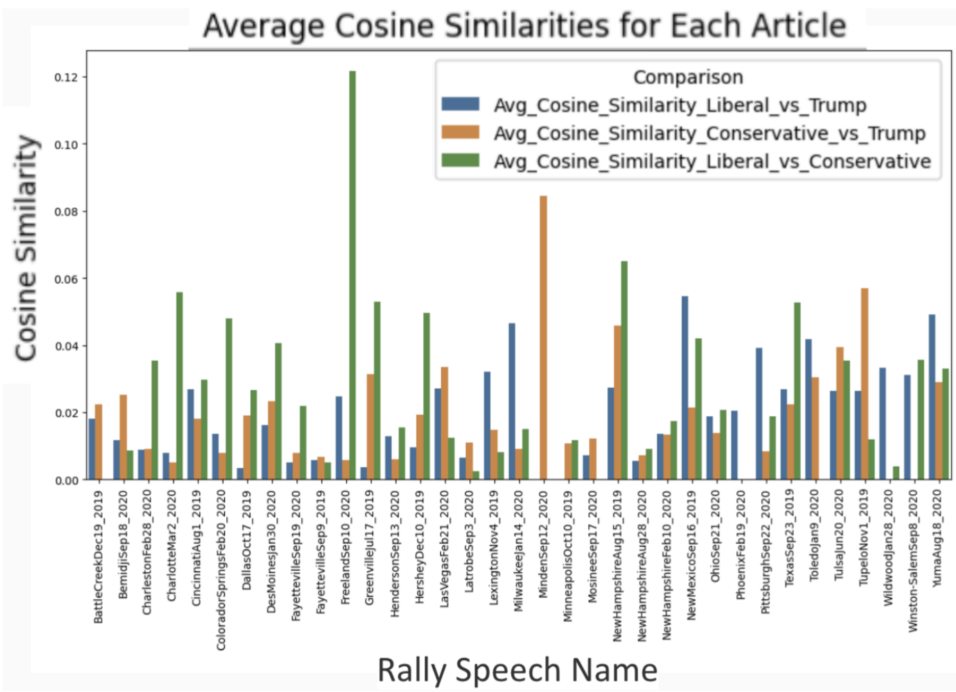


Figure 3

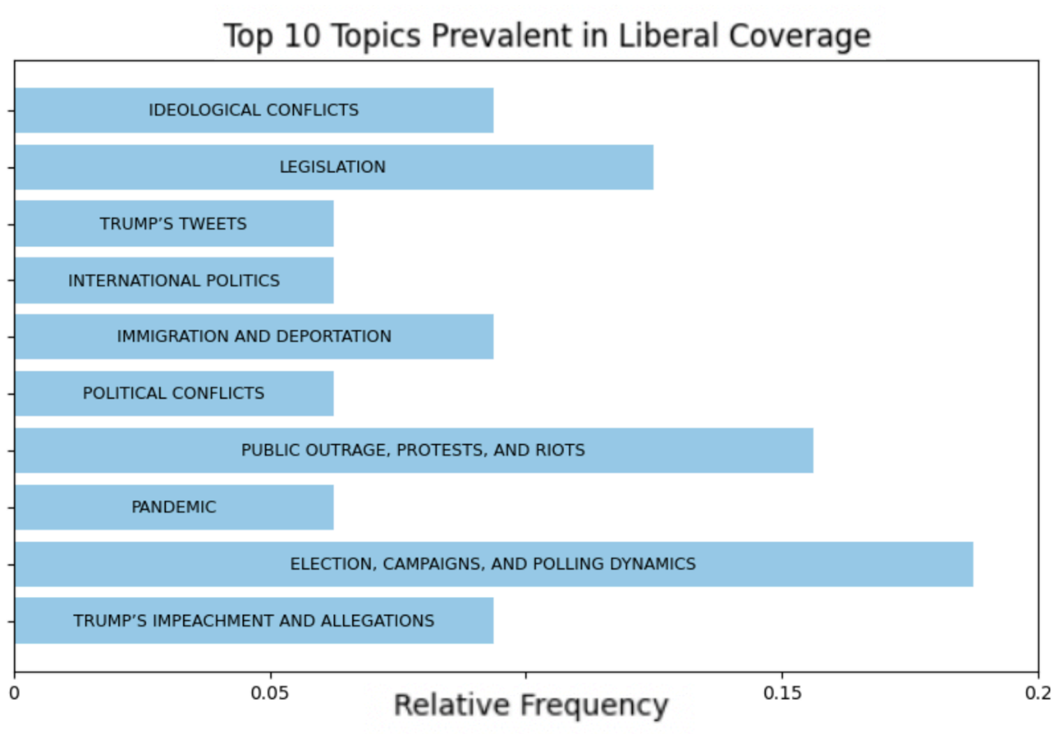


Figure 4

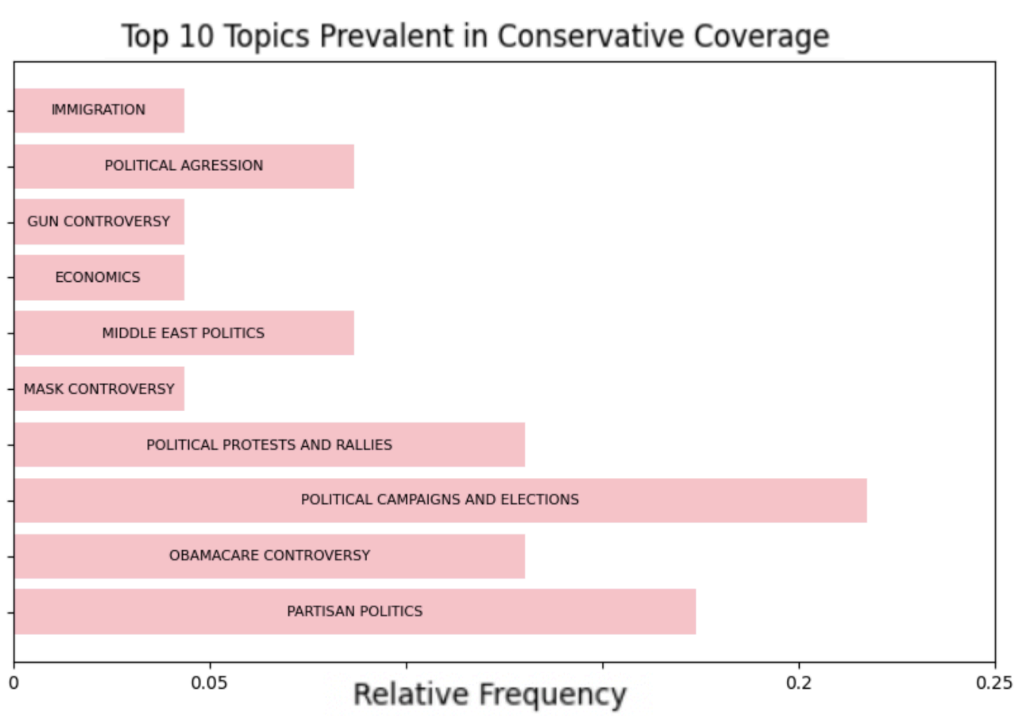


Figure 5

Model of Compound Score Formula

$$C = \begin{cases} s, & \text{if } \ell = \text{POS} \\ -s, & \text{if } \ell = \text{NEG} \\ \text{undefined}, & \text{if } \ell = \text{NEU} \end{cases}$$

$$I(\ell) = \begin{cases} 1, & \text{if } \ell \in \{\text{POS}, \text{NEG}\} \\ 0, & \text{if } \ell = \text{NEU} \end{cases}$$

$$C = I(\ell) \cdot \begin{cases} s, & \text{if } \ell = \text{POS} \\ -s, & \text{if } \ell = \text{NEG} \end{cases}$$

$$S \leftarrow S \cup \{C\} \text{ if } I(\ell) = 1$$

Note. Given a sentiment result with a label ℓ and score s, we define the compound score C based on the sentiment label ℓ. If the label is positive (POS), C equals the sentiment score s. If the label is negative (NEG), C equals the negative sentiment score $-s$. For neutral (NEU) labels, C is undefined. We introduce an indicator function $I(\ell)$ to differentiate between these cases. The function $I(\ell)$ is 1 for positive or negative labels and 0 for neutral labels. Using this indicator function, the compound score C is expressed as $C = I(\ell) \cdot s$ for positive labels and $C = I(\ell) \cdot (-s)$ for negative labels. If $I(\ell) = 1$, indicating a positive or negative label, the compound score C is appended to the sentiment scores list S.

Figure 6

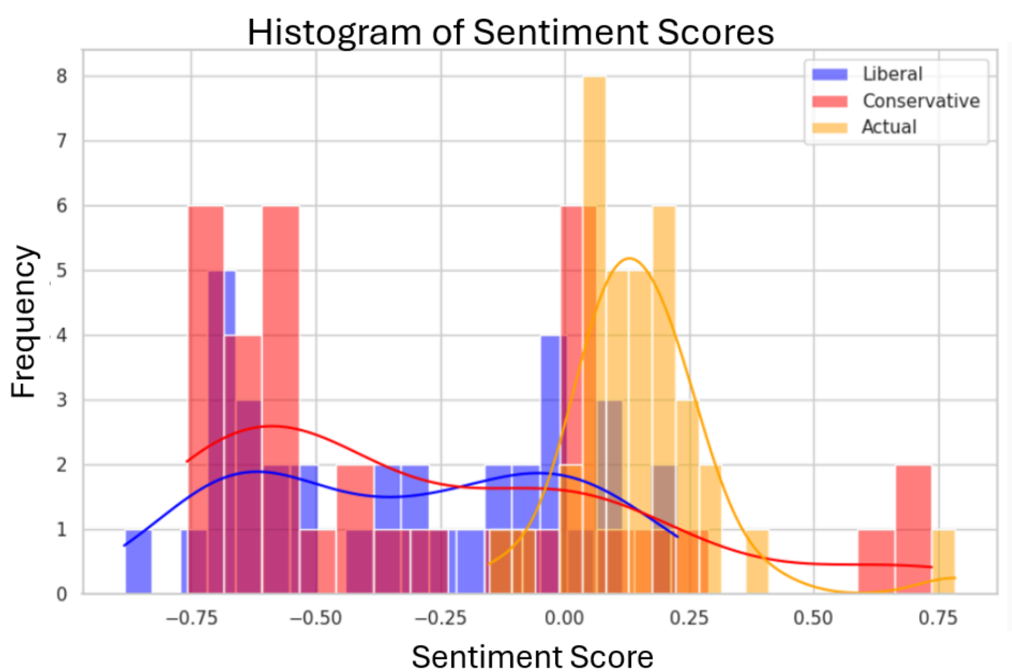
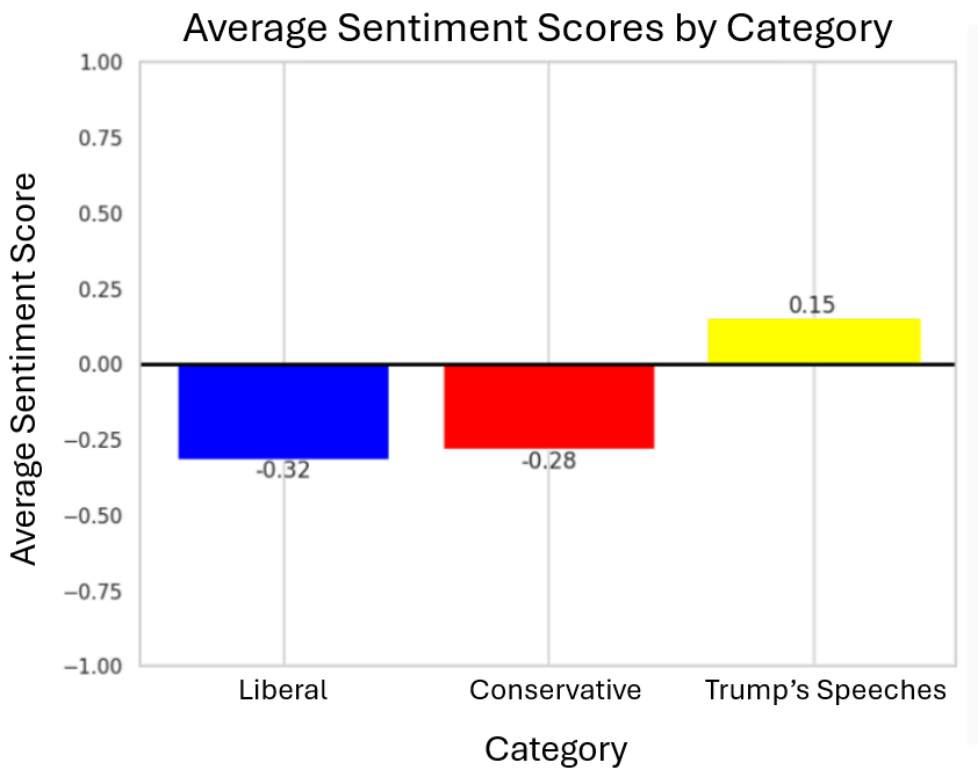


Figure 7


Average Sentiment Scores by Category
Average Sentiment Score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
0.15
-0.32
-0.28
Liberal
Conservative
Trump's Speeches
Category