Analyzing Media Bias: A Text Analytics Approach to Distinguishing Liberal and Conservative News Content Based on Headline

Author: Alabhya Dahal

ABSTRACT

News agencies in the USA exhibit significant polarization, with distinct leanings toward the left or right. These leanings manifest in differing values across socio-political, economic, and cultural spheres. For example, left-leaning agencies tend to advocate for progressive social values, a socialist economic approach, and tighter government controls. In contrast, right-leaning agencies favor conservative values, a capitalist economy, and less government intervention.

In this study, I collected headlines from prominent news agencies with known political leanings. On the left side, I included The New York Times and CNN, while on the right side, I included Fox News and NewsMax. The study aims to analyze the text of these headlines and determine the similarity scores between left and right-leaning agencies on the same issue or news headline.

Using the Word2Vec model, I compared the similarities between the headlines of each news agency on a specific headline. The results were varied but provided some evidence of the tendencies that text analytics tools could capture in terms of similarity. For instance, among nine headlines, three headlines having the highest similarities were between left-leaning agencies, CNN compared to NYT, and one was between right-leaning agencies, Fox compared to Newsmax. Moreover, for the nine headlines, eight headlines with the least similar scores were when compared between left and right-leaning agencies.

The final data is presented in ANNEX 1.

INTRODUCTION

Like in many parts of the world, news agencies often have affiliations with political agendas, social issues, and economic agendas. However, due to the bipartisan nature of US politics, the media landscape in the US tends to be more polarized into two distinct schools of thought. While media outlets often present themselves as non-partisan and independent, the content they provide can make it challenging to fully accept such claims. There has been evidence

of media bias, such as in presidential elections, where news has been generated in a way to support a particular pool of thought. Weatherly, J.N., Petros, T.V., Christopherson, K.M., & Haugen, E.N. (2007) conducted a study on the perception of political bias between CNN and Fox News. The study found evidence of bias in both news agencies.

In this research, I undertook primary data collection to gather news content from four leading news outlets in the US: The New York Times (NYT), CNN, Fox News, and Newsmax. Historically, NYT and CNN have been perceived as leaning towards the left in their reporting, whereas Fox News and Newsmax are recognized for their right-leaning perspectives (Pew Research Center, 2014).

LITERATURE REVIEW

Billie S Anderson(2023) in his study employs a text mining clustering approach to analyze a large number of COVID-19-related abstracts, aiming to assist researchers in efficiently identifying relevant articles amidst the rapidly growing volume of research and publications on the headline. The text clustering analysis reveals emerging research themes, with a focus on clinical management, risk factors, and the impact of underlying conditions such as obesity and heart disease. The study suggests practical implications, including the potential to reduce the time researchers spend searching for pertinent articles, by incorporating the clustering methodology into a search engine. Future work could involve developing predictive models to categorize research articles based on cluster-derived themes, comparing clustering results of abstracts with full articles, and assessing the effectiveness of using only keywords for clustering to streamline the research process further. While this study is not based on news content, it still gives an idea of how text analytics can be used in analyzing data from information found in the internet.

Hendrickx & Pakvis (2022) in their paper presents a structured literature review of news content analyses published in peer-reviewed journals between 2001 and 2020, analyzing 2,909 papers based on their abstracts for platform and location diversity. Key findings include a persistent focus on newspapers despite declining circulation, a strong preference for U.S. and Western European content, and increasing representation of African and Asian studies. The study highlights the enduring relevance of news content analysis in social sciences, with an emphasis on integrating automated and manual coding processes. It also

acknowledges the importance of monitoring democratic progress and the availability of a free press worldwide. The study is limited to abstract analysis and does not delve into conceptual or theoretical advancements in news content analysis.

Cornell research highlights that word lists, or "seeds," used to measure bias in online texts often contain inherent biases and stereotypes, potentially skewing findings. For example, using the seed term "mom" in gender-related text analysis can lead to biased results. Doctoral student Maria Antoniak, in her paper "Bad Seeds: Evaluating Lexical Methods for Bias Measurement," emphasizes the need for a critical evaluation of the measurement tools themselves to identify biases coded in models and datasets. The study, underscores that even tools designed for bias detection can have biases, especially when seed terms are undocumented or hidden in the code. It calls for researchers in digital humanities and natural language processing to carefully investigate and test seed sets for bias to ensure trustworthy results.

RESEARCH GAP

Text analytics has been widely utilized in sentiment analysis, large language models (LLMs), and natural language processing (NLP), but its applications extend into media and social science as well. By analyzing unstructured data, text analytics can help identify biases and perceived perceptions, enabling better access to and understanding of information. This approach holds significant potential for uncovering insights in media content and social science research, contributing to a more nuanced understanding of societal dynamics and communication.

DATA SOURCE AND METHODOLOGY

In this study, I conducted primary data collection by compiling similar news from four different platforms. To ensure that the news pieces were discussing the same event, I read each piece and collected news from the same dates. Furthermore, I focused only on the news headlines, excluding any subhead lines or the content of the news. I also avoided opinion pieces, as they do not necessarily reflect the same events. In total, I compiled nine different news headlines from four agencies.

For the text analytics process, I conducted preprocessing on all the headlines. Instead of using stop words, I applied lemmatization to the text. This approach was chosen because words like 'not' and 'very' can have a significant impact on the meaning of news

headlines, and lemmatization helps retain the essential meaning of words while reducing them to their base or root form. Further, tokenization and lowercase were applied.

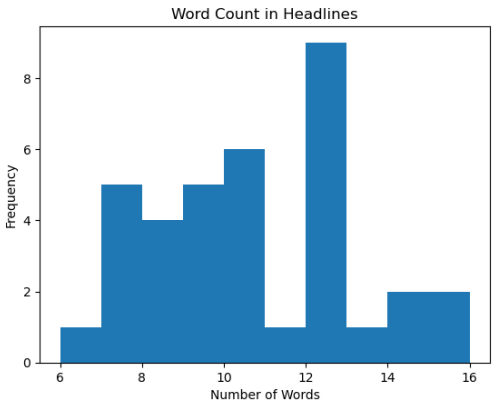
After the text cleaning process, I conducted a similarity analysis between the sentences to assess the level of similarity within each headline. I had a total of nine different headlines from four distinct news agencies. The similarity analysis involved comparing each headline with those from the other agencies to evaluate the degree of similarity between them. This approach allowed me to identify patterns and differences in how each agency reported on the same event or headline.

I utilized the Word2Vec model and cosine similarity to calculate similarity scores between sentences. I developed a function to preprocess each sentence, transforming it into a new vector representation. These preprocessed words were then evaluated for cosine similarity, leveraging the vector values provided by Word2Vec. All resulting scores were recorded for further analysis. Additionally, I ranked the scores within their respective headline clusters to compare the reporting of news from left-leaning and right-leaning sources on the same headlines. This approach allowed me to quantitatively assess the similarity in reporting between different news agencies.

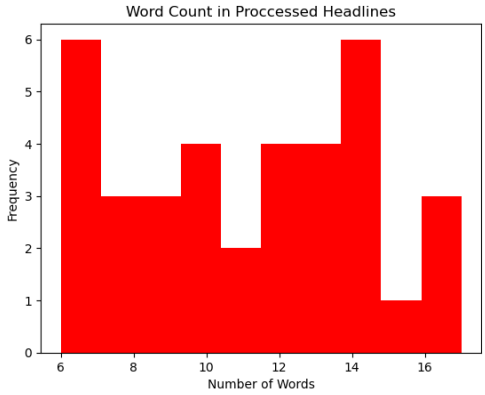
ANALYTICS AND FINDINGS

Analyzing the Text

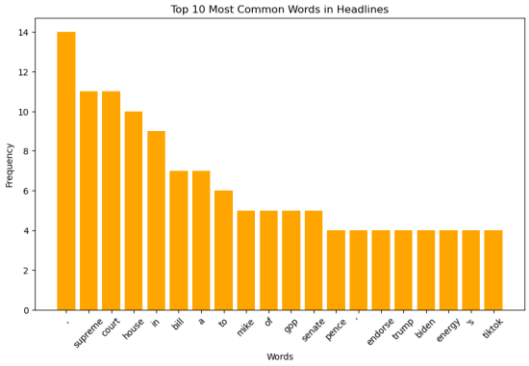
The initial analysis I conducted involved counting the frequency of words in each news headline and presenting the results in a histogram. The histogram below illustrates the distribution of word counts across the headlines, with a significant portion having 13 words. There is no distinct pattern or distribution observable in this data.



After pre-processing the words, the distribution differs a bit. Again, there is lack of any particular trends, however the words are now more concentrated in six to eight words.



The most used words in the list of these headlines are as follows:



It's notable that words such as 'supreme,' 'court,' 'house,' and 'in' frequently appear in news headlines, reflecting their relevance to key headlines. Interestingly, the most common word is '`,' which likely emerged due to the decision not to use stopwords in order to retain the full context of the headlines. This choice can result in the inclusion of such characters and common words in the analysis. If '`' is not meaningful, it may be worth considering its removal during preprocessing to focus on more substantive words.

Similarity Score Analytics

Now that we have completed the basic text analytics, let's examine the findings.

First, I compare how left news agencies compare to each other. That is, how headlines from New York Times and CNN scores similarity scores between one another.

Headline	Agency1	Agency2	Similarity	Rank in the headline	Agency_Group
1	NYT	CNN	0.777658	1	left
2	NYT	CNN	0.657365	5	left
3	NYT	CNN	0.799955	2	left
4	NYT	CNN	0.560019	5	left
5	NYT	CNN	0.860911	1	left
6	NYT	CNN	0.72714	3	left
7	NYT	CNN	0.823158	1	left
8	NYT	CNN	0.614401	4	left
9	CNN	NYT	0.665846	5	left

Table 1: Left Agency Similarity Scores

From this table, it is evident that The New York Times (NYT) and CNN have quite high similarity scores for most of the headlines. The three instances of rank one indicates that left-leaning agencies had the best similarity scores among the respective headlines. The descriptive statistics of the similarities between left-leaning agencies are as follows.

Count	Mean	Std	Min	Max	Median
9	0.72	0.1	0.56	0.86	0.73

Table 2: Left Agency Descriptive Analytics

From the table you can see that the maximum similarities between left agencies is 0.86 and mean is 0.72. The standard deviation is very small of 0.1.

Now, I compare how right news agencies like Fox and Newsmax compare with each other.

Headline	**Agency1**	**Agency2**	**Similarity**	**Rank in the headline**	**Agency_Group**
1	Fox	Newsmax	0.721333	3	right
2	Fox	Newsmax	0.880733	1	right
3	Newsmax	Fox	0.714163	4	right
4	Newsmax	Fox	0.539558	6	right
5	Fox	Newsmax	0.441741	5	right
6	Newsmax	Fox	0.691917	5	right
7	Fox	Newsmax	0.786937	5	right
8	Fox	Newsmax	0.634364	2	right
9	Newsmax	Fox	0.706774	3	right

Table 3: Right Agency Similarity Scores

Although not as high as the left-leaning agencies, right-leaning agencies also exhibit decent similarity scores. However, only one headline ranks at the top in similarity scores among right-leaning agencies. Surprisingly, some headlines have very low scores as well and are ranked the least. While we might expect these news agencies to align closely, it is possible that variations in wording and phrasing may have contributed to the lower scores.

Count	Mean	Std	Min	Max	Median
9	0.67	0.13	0.44	0.88	0.7

Table 4: Left Agency Descriptive Analytics

The descriptive statistics reveal a mean similarity score of 0.67 across the right-leaning agencies, which is lower than the mean score of 0.72 observed for the left-leaning agencies. Additionally, there is a slightly higher standard deviation of 0.13 for the right-leaning agencies. The maximum similarity score for the right-leaning agencies is 0.88, but the minimum score is notably lower at 0.44.

Now, finally, I compare the similarities between the right and the left leaning news agencies. The results are as follows:

Topic	Agency1	Agency2	Similarity	Rank	Agency_Gr
1	NYT	Fox	0.739639	2	left vs right
1	NYT	NewsMax	0.680377	5	left vs right
1	CNN	Fox	0.673966	6	left vs right
1	CNN	NewsMax	0.717856	4	left vs right
2	NYT	Fox	0.714895	4	left vs right
2	NYT	NewsMax	0.559088	6	left vs right
2	Fox	CNN	0.827044	2	left vs right
2	CNN	NewsMax	0.816033	3	left vs right
3	NYT	NewsMax	0.769083	3	left vs right
3	NYT	Fox	0.625241	6	left vs right
3	NewsMax	CNN	0.802617	1	left vs right
3	Fox	CNN	0.650247	5	left vs right
4	NYT	NewsMax	0.623406	2	left vs right
4	NYT	Fox	0.737745	1	left vs right
4	NewsMax	CNN	0.584192	4	left vs right
4	CNN	Fox	0.612556	3	left vs right
5	NYT	Fox	0.441692	6	left vs right
5	NYT	NewsMax	0.858804	2	left vs right
5	CNN	Fox	0.470321	4	left vs right
5	CNN	NewsMax	0.816306	3	left vs right
6	NYT	NewsMax	0.751799	2	left vs right
6	NYT	Fox	0.683209	6	left vs right
6	CNN	NewsMax	0.711991	4	left vs right
6	CNN	Fox	0.911476	1	left vs right
7	NYT	Fox	0.766453	6	left vs right
7	NYT	NewsMax	0.787154	4	left vs right
7	Fox	CNN	0.807746	3	left vs right
7	NewsMax	CNN	0.812389	2	left vs right
8	NYT	Fox	0.550194	6	left vs right
8	NYT	NewsMax	0.615476	3	left vs right
8	CNN	Fox	0.589803	5	left vs right
8	CNN	NewsMax	0.637792	1	left vs right
9	CNN	NewsMax	0.725708	2	left vs right
9	CNN	Fox	0.765137	1	left vs right
9	NYT	NewsMax	0.667726	4	left vs right
9	NYT	Fox	0.562983	6	left vs right

Table 5: Left-Right Agency Similarity Scores

The comparison between right-leaning and left-leaning agencies offers a wealth of information and shows a diverse distribution of similarity scores. There are some surprising findings, such as the high similarity score of 0.91 between CNN (left) and Fox (right) for news headline 6, which is the highest similarity score across all combinations. However, in general, CNN and Fox tend to have lower similarity scores compared to other combinations. For example, for headline 5, the similarity score between CNN and Fox is only 0.47, and for headline 8, it is 0.58. These variations highlight the complexity of the relationship between the reporting styles of different news agencies.

Count	Mean	Std	Min	Max	Median
36	0.69	0.11	0.44	0.91	0.7

Table 6: Right Agency Descriptive Analytics

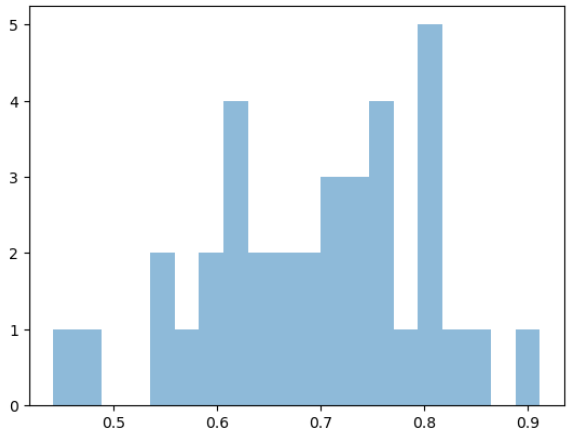


Table 7: Left_Right Similarity Distribution

Agency wise comparison

When directly comparing one left-leaning agency with one right-leaning agency, the results are particularly interesting. For example, the direct comparison between The New York Times (NYT) and Fox News shows the following results:

Headline	Agency1	Agency2	Similarity	Rank
1	NYT	Fox	0.74	2
2	NYT	Fox	0.71	4
3	NYT	Fox	0.63	6
4	NYT	Fox	0.74	1
5	NYT	Fox	0.44	6
6	NYT	Fox	0.68	6
7	NYT	Fox	0.77	6
8	NYT	Fox	0.55	6
9	NYT	Fox	0.56	6

Table 8: NYT vs Fox Similarity Scores

The comparison between Fox News and The New York Times (NYT) reveals the most contrasting results among the pairs. Out of the nine different headlines analyzed, six received the lowest rank in terms of similarity between these two agencies. Surprisingly, headline 5 achieved the top rank among these pairs, but even then, the similarity score was 0.74, which is not particularly high considering that the best-ranking results typically range from 0.8 to 0.9. This indicates a significant divergence in the reporting styles and content between Fox News and The New York Times.

CONCLUSION

While the overall results are varied and not entirely conclusive, there is evidence that text analytics can capture differences in media reporting. This project is limited to the use of news headlines only, which are typically shorter and less varied than the full news articles. However, if this project were to be scaled up to include more extensive text analysis, there is a strong indication that text analytics and similarity scores could effectively differentiate between left-leaning and right-leaning news agencies.

The major finding from this project is that The New York Times (NYT) and Fox News have divergent views, which can be captured through text analytics. The headlines collected for this analysis were not randomly selected, but they were not specifically targeted either. Therefore, this result provides some indication that the text analytics method is effectively capturing the differences in media output between these two news agencies.

Another finding from this project is that when analyzing only headlines, words without specific bias, such as 'supreme,' 'court,' or the name of a prominent figure like 'Trump,' may constitute a substantial proportion of the corpus. This can make it challenging for the algorithm to differentiate between two headlines. However, this issue should be mitigated if the analysis is scaled up to include secondary headings or the full

content of the news articles, which would provide a richer and more varied context for the text analytics model to work with.

A potential application of this type of analytical project is its replication to compare smaller news agencies, whose biases may not be as obvious or well-established. Additionally, beyond assessing the bias of entire agencies, this approach can also be used to analyze how the views of different authors align or diverge, even within the same publication or agency.

To conclude, this study demonstrates the potential of text analytics to capture differences in media output, particularly between prominent news agencies like The New York Times and Fox News. While the analysis of headlines alone presents certain limitations, scaling up the corpus to include more extensive text could enhance the model's ability to distinguish between varying perspectives. This project lays the groundwork for further exploration into media bias and offers a methodological framework that can be adapted to explore subtler biases in smaller news outlets or among individual authors.

REFERENCES

DiPietro, Louis. "Words Used in Text-Mining Research Carry Bias, Study Finds | Cornell Chronicle," October 18, 2021. https://news.cornell.edu/stories/2021/10/words-used-text-mining-research-carry-bias-study-finds.

Hendrickx, Jonathan, and Michael Pakvis. "News Content Analyses in the 21st Century A Structured Literature Review." *Media & Jornalismo*, December 12, 2022. https://doi.org/10.14195/2183-5462_41_7.

Matsa, Amy Mitchell, Jeffrey Gottfried, Jocelyn Kiley and Katerina Eva. "Section 1: Media Sources: Distinct Favorites Emerge on the Left and Right." *Pew Research Center's Journalism Project* (blog), October 21, 2014. https://www.pewresearch.org/journalism/2014/10/21/section-1-media-sources-distinct-favorites-emerge-on-the-left-and-right/.

S Anderson, Billie. "Using Text Mining to Glean Insights from COVID-19 Literature," April 2023. https://doi.org/10.1177/01655515211001661.

Weatherly, Jeffrey N., Thomas V. Petros, Kimberly M. Christopherson, and Erin N. Haugen. "Perceptions of Political Bias in the Headlines of Two Major News Organizations." *Harvard International Journal of Press/Politics* 12, no. 2 (April 1, 2007): 91–104. https://doi.org/10.1177/1081180X07299804.

Headline	Agency1	Agency2	Similarity	Rank Within Each Headline	Agency_Group
1	NYT	CNN	0.777658	1	left
1	NYT	Fox	0.739639	2	left vs right
1	NYT	NewsMax	0.680377	5	left vs right
1	CNN	Fox	0.673966	6	left vs right
1	CNN	NewsMax	0.717856	4	left vs right
1	Fox	NewsMax	0.721333	3	right
2	NYT	Fox	0.714895	4	left vs right
2	NYT	CNN	0.657365	5	left
2	NYT	NewsMax	0.559088	6	left vs right
2	Fox	CNN	0.827044	2	left vs right
2	Fox	NewsMax	0.880733	1	right
2	CNN	NewsMax	0.816033	3	left vs right
3	NYT	NewsMax	0.769083	3	left vs right
3	NYT	Fox	0.625241	6	left vs right
3	NYT	CNN	0.799955	2	left
3	NewsMax	Fox	0.714163	4	right
3	NewsMax	CNN	0.802617	1	left vs right
3	Fox	CNN	0.650247	5	left vs right
4	NYT	NewsMax	0.623406	2	left vs right
4	NYT	CNN	0.560019	5	left
4	NYT	Fox	0.737745	1	left vs right
4	NewsMax	CNN	0.584192	4	left vs right
4	NewsMax	Fox	0.539558	6	right
4	CNN	Fox	0.612556	3	left vs right

5	NYT	CNN	0.860911	1	left
5	NYT	Fox	0.441692	6	left vs right
5	NYT	NewsMax	0.858804	2	left vs right
5	CNN	Fox	0.470321	4	left vs right
5	CNN	NewsMax	0.816306	3	left vs right
5	Fox	NewsMax	0.441741	5	right
6	NYT	CNN	0.72714	3	left
6	NYT	NewsMax	0.751799	2	left vs right
6	NYT	Fox	0.683209	6	left vs right
6	CNN	NewsMax	0.711991	4	left vs right
6	CNN	Fox	0.911476	1	left vs right
6	NewsMax	Fox	0.691917	5	right
7	NYT	Fox	0.766453	6	left vs right
7	NYT	NewsMax	0.787154	4	left vs right
7	NYT	CNN	0.823158	1	left
7	Fox	NewsMax	0.786937	5	right
7	Fox	CNN	0.807746	3	left vs right
7	NewsMax	CNN	0.812389	2	left vs right
8	NYT	CNN	0.614401	4	left
8	NYT	Fox	0.550194	6	left vs right
8	NYT	NewsMax	0.615476	3	left vs right
8	CNN	Fox	0.589803	5	left vs right
8	CNN	NewsMax	0.637792	1	left vs right
8	Fox	NewsMax	0.634364	2	right
9	CNN	NYT	0.665846	5	left
9	CNN	NewsMax	0.725708	2	left vs right
9	CNN	Fox	0.765137	1	left vs right

9	NYT	NewsMax	0.667726	4	left vs right
9	NYT	Fox	0.562983	6	left vs right
9	NewsMax	Fox	0.706774	3	right