ChatGPT as a Tool for News Summarization and Bias Detection: An Empirical Analysis

Trevor Harrington and **Qiwei Men** and **Jerrin Wofford**
The Ohio State University

Abstract

ChatGPT and other Large Language Models (LLMs) can be used to accurately summarize and gauge the bias of a large number of articles in a relatively short period of time. However, the summary and bias classification need to be accurate so that readers have a proper understanding of the article. Here, we use Information Coverage (IC), BERT score, TF-IDF, and cosine similarity to determine the accuracy of ChatGPT's summaries and bias classification. BERT scores for the generated summaries were relatively high, suggesting relatively high similarity to the human-generated summaries. However, the IC was significantly lower. ChatGPT was much more likely to classify a central or weak leaning to the bias. While summaries produced by ChatGPT included major points in the article, the summarization process often shifts biases toward adjacent, less extreme categories.

1 Introduction

Media consumption is the primary way people stay up-to-date on current events. TV, podcasts, and news articles all provide different perspectives on similar topics. Despite being one of the oldest forms of supplying the public with information, news articles are still one of the largest media consumed. However, many articles cover the same topic and are often written with varying degrees of bias. Reading articles and gaining a holistic view of the subject can become extremely time-consuming. In today's fast-paced society, not everyone has the time to read multiple articles. The burgeoning AI industry, specifically large language models (LLMs), has allowed for article summary generation [1,2]. Using these tools, a reader can glean information more quickly, much like how an abstract is used.

Media bias involves the selective presentation of information that aligns with a particular viewpoint, potentially shaping public perceptions of events or issues [3]. The majority of Americans believe that mass media organizations exhibit bias [4]. Extensive research has shown that media bias could facilitate the spread of misinformation, undermining decision-making processes and eroding trust in news sources [3,5]. Numerous approaches have been developed to detect media bias, ranging from manual content analysis conducted by human evaluators (e.g., [6, 7, 8, 9]) to computational methods leveraging machine learning and natural language processing techniques [10, 11, 12, 13].

The advancements in large language models (LLMs) provide a novel approach for summarizing news and detecting media bias by analyzing patterns, tone, and content in text. However, their accuracy and efficiency remain under-explored and require further evaluation. In this study, we use tools such as BERT and TF-IDF, as well as comparing the generated summaries to those produced by people, the quality and informational coverage of the generated summaries were tested and validated. These measures will be judged against three claims: (1) ChatGPT-4o mini is able to summarize an article with 70% accuracy or more, determined by the Information Coverage (IC) of the article present in the summary and the BERT score between the generated summary and a human summary, (2) ChatGPT-4o mini is able to identify the media bias within the article, and (3) ChatGPT-4o mini is able to preserve the bias from the article in the article summary with 70% accuracy.

2 Experiment Setup

This section outlines the experimental setup, detailing the processes of data collection and preparation for the experiment. It also describes the design of the experiment aimed at verifying three key claims presented in the previous section: (1) a comparison between human performance and ChatGPT in news summarization, (2) a comparison between the source bias class and ChatGPT detected bias class,

and (3) a comparison between bias detection results from original articles and their ChatGPT-generated summaries. Additionally, topics such as prompt engineering and [specific methodological considerations] will be discussed. These components are critical for understanding the experimental procedure, validating the results, and ensuring the study's replicability.

2.1 Data

The analysis in this paper utilizes two datasets to verify three key claims. The first dataset is a subset of the BBC News Summarization dataset, originally created by Greene and Cunningham in 2006 [14]. This dataset was derived from a categorization dataset comprising 2,225 documents from the BBC News website, spanning five topical areas from 2004–2005, each paired with a human-written summary. For this study, 100 articles from the politics category were randomly selected to test the hypothesis in Claim 1.

The second dataset was independently collected and consisted of 400 news articles from 41 media outlets (Table 3), representing the five bias categories as rated by AllSides (Figure 1). Most articles were obtained from the politics or policy sections of these outlets' websites and primarily reported on events from early December. To ensure consistency for the experiment, articles shorter than 200 words or longer than 1,500 words were excluded. From the initial dataset of 657 articles, 80 articles were selected for each bias category to test the hypothesis in Claims 2 and Claim 3.

2.2 Claim 1: News Summarization Analysis

To evaluate Claim 1, which posits that the article summary generated by ChatGPT maintains an 'accuracy' of above 70%, we analyzed the Information Coverage of the summary using TF-IDF on the article and the summary. Then we use the cosine similarity of the two vectors to produce an Information Coverage or IC score. Furthermore, it is desirable to know how similar the human summary and the generated summary are quantitatively. To illustrate this we also produced a BERT score of the two summaries, which creates embeddings using a transformer and then compares these embeddings. This shows how close the two summaries are in form and meaning. These two methods allow us to determine how much information from the article is in the generated summary and if it is similar to how a human would write it.

Figure 1: AllSides MediaBias Chart Version 8

2.3 Claim 2: Bias Detection Analysis

To evaluate the hypothesis in Claim 2, which posits that ChatGPT can accurately determine the bias of an article, we compared the bias classification by ChatGPT to the bias of the source that produced the article, as defined by 'AllSides'(Figure 1). In order to derive some reasoning from ChatGPT for the classification decision, we asked for an explanation, and by running TF-IDF on the resulting explanations, we can determine the words the model finds to be indicators of bias.

To visualize the TF-IDF analysis results of the bias explanations provided by ChatGPT, we utilized Principal Component Analysis (PCA) to reduce the high-dimensional TF-IDF features to a 2D space. Each row in the dataset represents a bias explanation, and its corresponding TF-IDF scores were calculated to form the input features. PCA was then applied to project these features into two principal components, which serve as the x and y axes of the plot. The data points were color-coded by the AI-detected bias categories ("Left," "Lean Left," "Center," "Lean Right," and "Right") to facilitate visual differentiation and clustering analysis.

2.4 Claim 3: Bias Deviation Analysis

To evaluate the hypothesis in Claim 3, which posits that ChatGPT preserves the bias present in the original articles during summarization, we analyzes the

bias distribution across five categories—Left, Lean Left, Center, Lean Right, and Right. The bias detection results from the original articles will serve as a baseline, as these articles provide complete information and context. By comparing the bias distribution in the original articles with that in the summaries generated by ChatGPT, we aim to identify any deviations or shifts introduced during the summarization process. Further, the analysis will examine whether shifts in bias distribution differ across the five categories using a confusion matrix. Specifically, we will examine if certain bias groups are more prone to shifts or show distinctive patterns in how their ideological alignment is preserved or altered during summarization.

2.5 Prompt Engineering

The prompt given to ChatGPT 4o-mini for the summarization task was: "Here is a news article, summarize the article in less than {length} words while still maintaining information." Where {length} is the total word count of the human summary and {quotes} is the article to summarize.

The prompt given to ChatGPT 4o-mini for the bias detection task was: "Here is a news article, try to classify the article based on media bias into one of 5 discrete classes: "Left", "Lean Left", "Center", "Lean Right", "Right". Put this classification on the first line with no added parts. Give an explanation on a new line for this classification in under 100 words and do not use new line characters in the explanation. Quotes: {quotes}" Where {quotes} is the article the model is checking for bias.

3 Results

3.1 News Summarization

The results for claim 1 are shown in the table and figure below. The table describes the average BERT precision, recall, and F1, which is derived by treating the human summary as 'ground truth'. The result shows that the human and AI-generated summaries are contextually similar in form and meaning.

The evaluation of the 100 pairs of human-generated and AI-generated summaries yielded a mean BERT score of 0.873 across precision, recall, and F1-score, as shown in Table 1, which indicates a high level of similarity between the two types of summaries, supporting our hypothesis that AI-generated summaries are closely aligned with human-generated ones, and suggesting that AI models can achieve human-like performance in summary generation under the given experimental conditions.

Mean Precision	Mean Recall	Mean F1
0.8728	**0.8734**	**0.8731**

Table 1: The average Bert precision, recall, and F1 of the generated summaries compared to the human summary baseline.

The distribution plot of information coverage reveals significant differences between human and AI-generated summaries. Human summaries, represented by the blue curve, exhibit a sharp and consistent peak near 0.9, indicating a high level of information retention from the original text. In contrast, AI-generated summaries, represented by the orange curve, display a broader and less concentrated distribution, with a peak of around 0.7. The lower peak and broader spread suggest that AI-generated summaries generally underperform in retaining the original content's key information.

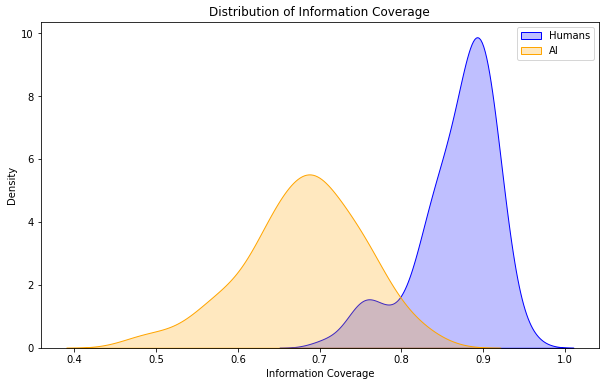

Figure 2: Distribution of the Infomation Coverage of Human and AI Summaries

3.2 Bias Detection

The confusion matrix in Figure 3 depicts the AI-detected bias and the source bias of the articles. It provides insights into how likely the model is to predict an article as a certain class, given the source of the article.

The "Center" bias shows acceptable accuracy, with 40.95% of articles originally classified by ground truth as "Center" actually being classified as "Center" by the model. However, this is due to the model favoring "Lean Left" (21.25%) or "Lean Right" (35.00%), indicating that the model has difficulty distinguishing the three classes.

In the “Lean Left” category, 47.5% of the articles have an accurate bias classification. However, 30% shift toward “Center,” and 16.25% shift to “Lean Right,” suggesting that the model pulls the "Lean Left" class to the right. Notably, no articles were classified as "Right".

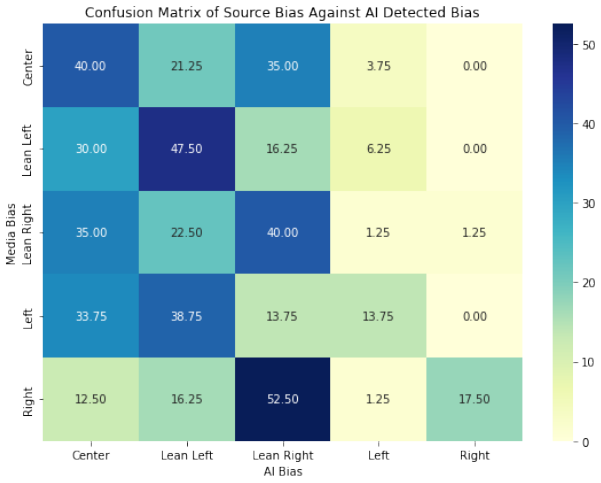

Figure 3: Confusion Matrix of Source Bias Against AI Detected Bias

For “Lean Right,” 40.0% of articles maintain their ’ground truth’ classification. However, 35.00% shift to “Center,” indicating that the model has a strong preference for neutrality, and 22.5% switch to "Lean Left" showing that the model has some difficulty with distinguishing the three center classes. Small proportions also shift to “Right” (1.25%) or “Left” (1.25%).

Articles classified as “Left” show low consistency, with only 13.75% being classified by ChatGPT as "Left". A significant proportion, 38.75%, shifts toward “Lean Left,” and 33.75% to "Center". Softening strongly left-leaning content toward a more moderate stance.

Articles classified as “Right” also show low consistency, with only 17.50% being classified by ChatGPT as "Right". Once again, a significant proportion, 52.50%, shifts toward “Lean Right.” However, this class had the lowest number of "Center" classifications, with "Lean Left" and "Left" having similarly low numbers.

The use of TF-IDF to determine words that strongly predict a certain class was mostly unsuccessful. By using the model provided, all five classes contain similar words, as seen in Figure 4. This shows that these words are not particularly strong indicators for any class, despite their high values. Furthermore, outside of the top words, all other words had small TF-IDF values, showing that they are not strong indicators for prediction.

Overall, the model often favored ’center’ classifications over more extreme classifications, but with the advantage of never straying too far from the source bias classifications.

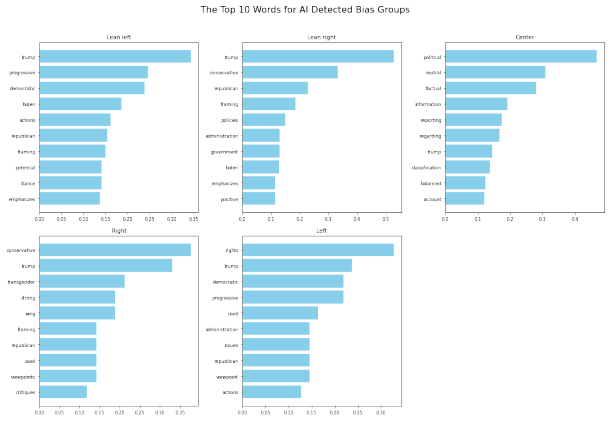

Figure 4: Top 10 TF-IDF Words for Each Bias Group: Insights from AI-Generated Bias Explanations

The scatter plot (Figure 5) shows distinct clustering patterns for certain bias categories. Notably, explanations classified as “Center” form a clearly defined cluster on the right side of the plot, suggesting a high degree of homogeneity in the TF-IDF features for this group. In contrast, points representing “Lean Left,” “Lean Right,” and “Right” overlap significantly on the left side of the plot, indicating less distinct separation in their textual features. Explanations labeled as “Left” are sparse and distributed among the other groups, showing no clear clustering pattern.

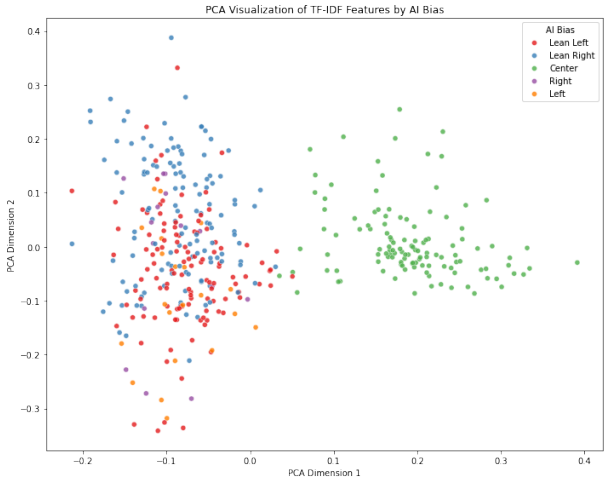

Figure 5: PCA Visualization of TF-IDF Features by AI Bias

These findings suggest that while some bias categories, such as "Center," exhibit consistent linguistic characteristics, others, particularly "Lean Left" and "Lean Right," share overlapping features. This could reflect the nuanced and less polarized nature of these groups, as well as potential challenges in distinguishing these biases solely based on textual features. Further investigation is needed to refine the prompt engineering and get more meaningful explanations to improve differentiation among these overlapping categories.

3.3 Bias Deviation

The comparison of bias distributions across the three variables—AllSides Source Bias, AI Bias Classification, and AI Summary Bias Classification (Figure 6)—reveals distinct shifts in how bias is represented during AI classification and summarization. Notably, articles categorized under the "Center" bias are significantly overrepresented in both AI classifications (original articles and summaries) compared to the AllSides source bias. This trend suggests a potential tendency of the AI system to shift content towards a neutral stance, either as part of its inherent bias or due to its approach to content processing.

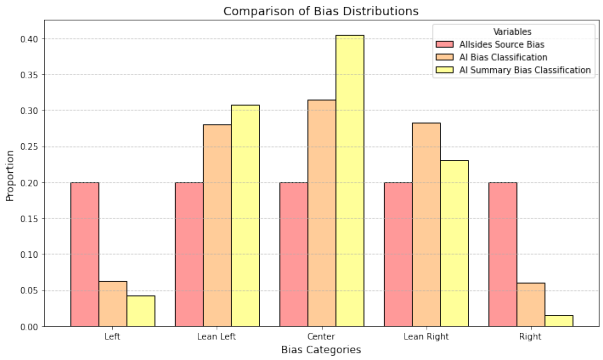

Figure 6: Comparison of Bias Distributions

Conversely, articles originally categorized as "Left" or "Lean Left" appear to be underrepresented in AI classifications. This indicates that the AI may systematically dilute left-leaning perspectives during classification or summarization. In contrast, there is a modest increase in the representation of "Lean Right" biases in AI classifications, suggesting that certain right-leaning perspectives may be emphasized more prominently by the AI. However, articles categorized as "Right" exhibit notable underrepresentation, signaling challenges in accurately preserving or detecting strong right-leaning biases.

Interestingly, the bias distribution for AI summaries closely aligns with the distribution for AI classifications of the original articles. This alignment indicates that the summarization process does not significantly alter the bias detected by AI. However, the overall shifts in bias representation highlight a potential bias drift in AI systems, where the content tends to converge toward a more centrist position, potentially at the expense of accurately reflecting the original bias distribution.

The confusion matrix illustrates the shifts in bias classification between original articles and their summaries as detected by AI. It provides insights into how the summarization process affects the preservation of bias.

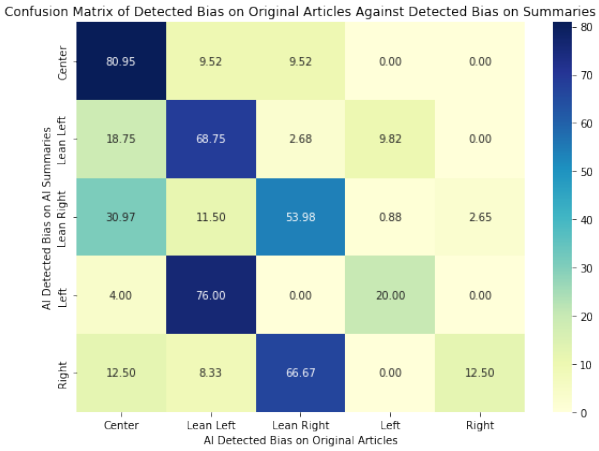

Figure 7: Confusion Matrix of Detected Bias on Original Articles Against Detected Bias on Summaries

The "Center" bias shows strong consistency, with 80.95% of articles originally classified as "Center" maintaining their classification in summaries. However, a small proportion shifts to "Lean Left" (9.52%) or "Lean Right" (9.52%), indicating a slight tendency for summaries to shift toward adjacent biases.

In the "Lean Left" category, 68.75% of the articles retain their bias classification in summaries. However, 18.75% shift toward "Center," and 9.82% shift to "Left," suggesting that some content gravitates toward a more centrist or strongly left-leaning stance during summarization.

For "Lean Right," 53.98% of articles maintain their original classification. However, 30.97% shift to "Center," indicating a notable tendency for summaries to move toward neutrality. Smaller proportions also shift to "Right" (2.65%) or "Lean Left" (11.50%).

Articles classified as "Left" show low consistency, with only 20.00% retaining their original classification. A significant proportion, 76.00%, shifts toward "Lean Left," suggesting a tendency for summaries to soften strongly left-leaning content toward a more moderate stance.

Similarly, "Right" articles exhibit even lower retention of their original classification, with only 12.50% maintaining the "Right" label. A majority, 66.67%, shifts toward "Lean Right," indicating a similar moderation trend. Additionally, another 12.50% shift further toward "Lean Right," highlighting a potential over-representation of slightly moderated biases in AI summaries.

4 Conclusions

This research illustrates that the ChatGPT-4o mini model is adequate for the task of summarizing. On average, the generated summary is similar to a human summary, measured by the cosine similarity of BERT embeddings. This is contrasted by the model's poor performance in Information Coverage, showing worse results on average when compared to human summarization.

The use of ChatGPT-4o mini for media bias detection often pushes the classification of the article to the center. For example, both left-leaning and right-leaning articles are often classified as 'center' or unbiased articles.

Furthermore, the process of summarization often shifts biases toward adjacent, less extreme categories, particularly for "Left" to "Lean Left" and "Right" to "Lean Right." This indicates a moderation effect in the summarization process, which softens strong biases and moving content toward neutrality. Articles classified as "Left" or "Right" exhibit low retention rates for their original classifications, with the majority shifting toward "Lean Left" and "Lean Right," respectively. This highlights the challenge of preserving strongly biased content in AI-generated summaries. Biases classified as "Center," "Lean Left," and "Lean Right" show moderate consistency, However, "Lean Left" and "Lean Right" categories frequently shift toward "Center," suggesting a slight neutralization tendency for these biases.

These findings emphasize the systematic bias drift introduced during AI summarization, particularly for strongly polarized content. Understanding and mitigating this drift is crucial to improving the fidelity of AI summaries while preserving the original bias representation. Future work should explore strategies to address these shifts while maintaining the integrity of the original content.

5 Limitations

The primary limitation was a lack of diverse data. The BBC dataset was extensive but not diverse. This is due to a lack of article datasets with accompanying human summaries. Furthermore, the summaries generated were judged automatically using Information Coverage and similarity to the human summary. Ideally, some subsections of our generated summaries should have been judged by humans. The second dataset also covered a short time frame, which may be insufficient to illustrate the performance of the model's bias detection. Furthermore, the bias labels provided in the data set are based on the political leanings of the news outlets, which may not be representative of the articles. The use of the commercially available ChatGPT-4o mini without any tuning using downstream data likely severely hampered its performance as a summarizer and bias detector. Ideally, a dedicated model should be used for this task. In a future iteration of this project, the model would be tuned for article summarization to improve the overall output of the model.

References

[1] Zhang, X., Tian, Z., & Liu, T., "A Comprehensive Evaluation of Large Language Models for News Summarization," arXiv preprint, arXiv:2301.13848, 2023. [Online]. Available: https://arxiv.org/abs/2301.13848

[2] Basyal, S., & Sanghvi, A., "Performance Comparison of LLMs in Text Summarization: MPT-7b-instruct, Falcon-7b-instruct, and ChatGPT," arXiv preprint, arXiv:2310.10449, 2023. [Online]. Available: https://arxiv.org/abs/2310.10449

[3] Wessel, M., Horych, T., Ruas, T., Aizawa, A., Gipp, B., & Spinde, T. Introducing MBIB–the first Media Bias Identification Benchmark Task and Dataset Collection. 2023, arXiv preprint arXiv:2304.13148.

[4] Puglisi, R., & Snyder Jr, J. M. Empirical studies of media bias. 2015, Handbook media Econ. 1, 647-667.

[5] Hamborg, F., Donnay, K., & Gipp, B. Automated identification of media bias in news articles:

an interdisciplinary literature review. 2019, Inter. J. Digital Libraries, 20(4), 391-415.

[6] Groseclose, T., & Milyo, J. A measure of media bias. 2015, The Quarterly J. Economics, 120(4), 1191-1237.

[7] Papacharissi, Z., & de Fatima Oliveira, M. News frames terrorism: A comparative analysis of frames employed in terrorism coverage in US and UK newspapers. 2008 The Inter. J. press, 13(1), 52-74. [8] Smith, J., McCarthy, J. D., McPhail, C., & Augustyn, B. From protest to agenda building: Description bias in media coverage of protest events in Washington, DC. 2001, Social Forces, 79(4), 1397-1423.

[9] Van der Pas, D. J., & Aaldering, L. Gender differences in political media coverage: A meta-analysis.2020 J. Communication, 70(1), 114-143

[10] D'Alonzo, S., & Tegmark, M. Machine-learning media bias. 2022 Plos one, 17(8), e0271947.

[11] Spinde, T., Krieger, J. D., Ruas, T., Mitrović, J., Götz-Hahn, F., Aizawa, A., & Gipp, B. Exploiting transformer-based multitask learning for the detection of media bias in news articles. 2022, Inter. Conf. Virtual Event, 225-235.

[12] Gangula, R. R. R., Duggenpudi, S. R., & Mamidi, R. Detecting political bias in news articles

using headline attention. 2019 Anal. Interpret. Neur. Net. NLP 77-84. [13] Thota, A., Tilak, P., Ahluwalia, S., & Lohia, N. Fake news detection: a deep learning approach. 2018, SMU Data Science Review, 1(3), 10

[14] D. Greene and P. Cunningham, "Practical solutions to the problem of diagonal dominance in kernel document clustering," in Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, USA, Jun. 2006, pp. 377–384.

A Appendix

Table 2: List of sources of the media bias detection dataset

Bias	Source	Count
Center	CNBC	20
	Forbes	12
	News Nation	4
	News Week	15
	Reuters	11
	SAN	16
	The Hill	2
Lean Left	ABC	18
	Bloomberg	3
	Business Insider	9
	CBS	9
	CNN	6
	NBC	7
	NPR	15
	NYTimes	12
	TIME	12
Lean Right	Just The News	2
	National Review	18
	New York Post	13
	TheDispatch	1
	TheEpicTimes	4
	The Free Press	9
	Washington Examiner	14
	Washington Times	19
Left	AP	31
	Alter Net	20
	Democracy Now	1
	Huffpost	8
	MSNBC	3
	Mother Jones	1
	SLATE	7
	The Guardian	9
Right	Blaze	4
	Breitbart	18
	CBN	5
	Daily Caller	3
	Daily Wire	24
	FOX	2
	Federalist	14
	IJR	1
	News Max	9