Leveraging Language Models for Temporal Political Bias Analysis

Clayton Carlson, Ryan Diaz, Charlie Rapheal, Sanjali Roy
Syntax Errors
University of Minnesota, Twin Cities
{carl6640, diaz0329, raphe011, roy00142}@umn.edu

Abstract

Political bias detection in news media has been a well-studied problem in natural language processing, and research in this area has focused on bias identification at different levels of granularity, building datasets to identify different types of political bias, and the reasoning of language models behind bias classification decisions. However, language models have not yet been used to analyze the dynamics of political bias over time. In this project, we use a fine-tuned political bias classification model known as PoliticalBiasBERT on a dataset of web-scraped news articles and transcripts to identify political bias trends over time. Important world news events in this time will serve as "benchmark events", and bias in news sources will be analyzed before and after these events to capture any trends or possible effects thereof. We discuss our findings concerning temporal political bias patterns, and also analyze the classification decisions of our model via an interpretability tool. The code for this project can be found at https://github.com/RyangDiaz/temporal-political-bias-lm.

1 Introduction

In a world that runs on information, it is important to fully understand how our news sources exhibit bias in order to be well-informed citizens. Political bias detection has been extensively analyzed from the perspective of natural language processing methods, and much effort has been put into detecting these types of biases in real-world news texts on the individual level. However, it is also relevant to analyze how political biases in news media has changed *over time*, so that we may get a richer perspective on how global events can shape the political landscape.

Our project aims to provide insight into the research question: "Can existing state-of-the-art methods for detecting article-level political biases in textual media be used for performing temporal bias analysis concerning real-world political events?". We would like to see whether or not language models can reflect the same level of insight on how political bias from different news sources has changed over time as human-produced meta-analyses. We also want to see how 'benchmark events' in the news have changed the political landscape, and we want to see if language models can capture changes that humans have seen anecdotally.

We believe that our work will help enable all citizens who participate in civic activities such as voting to gain a more nuanced view of the media they are consuming and make more informed political decisions. We also hope that this study can help supplement human-produced analyses of political bias trends and provide a different view on a social issue from a data-driven perspective.

2 Related Work

2.1 Political Bias Detection

Political bias detection in text has been done in varying scales of granularity, such as sentence-level, article-level, and media-level bias. Baly et al. (2020) specifically focuses on article-level political bias classification (left-leaning, right-leaning, or center-leaning), and highlights some challenges and possible solutions to this domain. They discuss the challenges of obtaining fair and accurate bias annotations at the article level, highlighting the importance of websites such as allsides.com that provide these article-level annotations with the goal of balanced labels. As discussed in their study, article-level bias classification runs the risk of the model overfitting to the media source rather than analyzing the article itself, and presents several techniques to "de-bias" these models such as adversarial adaptation and triplet loss pretraining.

Others have worked on creating sufficient and comprehensive ways for LLMs to detect bias. Fan et al. (2019) created a dataset called BASIL (Bias Annotation Spans on the Informational Level) to also cover informational bias, which they define as factual information that is used to sway the reader one way or another, as opposed to lexical bias, which is the use of wording to sway the reader. They found that informational bias is actually more prevalent than lexical bias, but at the same time they did not have great success fine-tuning BERT to identify information bias. This means that in the future, works with LLMs identifying bias should keep in mind what kind of bias the LLM is finding.

Large language models have also recently been used to help improve the quality and explainability of bias detection. Lin et al. (2024) leverages LLMs by classifying biases using the top matched bias indicators, approaching the bias classification problem from a recommender system standpoint rather than from a pure classification perspective. These bias indicator vectors represent more fine-grained explanations for why bias in a text was classified the way that it was, dramatically improving the interpretability of these systems and aiding future human annotators in creating datasets geared towards recognizing bias. These indicators are generated using LLMs and stored in a vector database; text is classified by querying the database with a generated summary, and the top indicator vectors it matches with are used for bias classification. The system exhibits a high degree of adaptability to out-of-distribution data compared to methods that fine-tune on a specific dataset.

2.2 Temporal Political Bias Analysis

Some work on temporal political bias analysis has been done in fields other than computer science. Grayson and Greene (2019)'s paper found cyclical trends in Reddit datasets following benchmark events like the 2020 U.S. presidential election and the 2022 U.S. midterm election, but they did this using a statistical technique called Seasonal and Trend decomposition using Loess (STL), not LLMs. A political science paper utilized the Stanford Cable TV News Analyzer, a dataset of videos from many U.S. news networks from 2010 and 2021, and manually correlated screen-time of certain political actors with their campaign donation behavior to find trends in bias over time (Kim et al., 2022). This approach focused more on visual data and does not use LLMs.

3 Methods

3.1 Data Collection via Web Scraping

We used Python web scraping libraries, specifically BeautifulSoup4, to scrape various news websites for transcripts and articles. The news sources we used were CNN, FOX, and The New York Times. We scraped articles and transcripts in the 'Politics' and 'U.S.' sections of these sources. Web scraping CNN and FOX required us to create several custom Python scripts for each form of content. These websites use inconsistent conventions for accessing their transcripts. CNN, for example, has 4 conventions for where the body of a transcript is located.

3.2 The PoliticalBiasBERT Model

The model we used for political bias classification is known as "PoliticalBiasBERT", and is based on the work of Baly et al. (2020). PoliticalBiasBERT is a classification model with a BERT (Devlin et al., 2019) backbone fine-tuned on a dataset of news articles with both expert and crowd-sourced political bias annotations. For each input article, PoliticalBiasBERT outputs three softmax classification scores corresponding to a left-leaning, center-leaning, or right-leaning political bias.

Before applying PoliticalBiasBERT on the scraped articles, we elected to run a "sanity check" to verify its apparent abilities. Running the model on the same test set used in Baly et al. (2020) gave us an accuracy of around 71.2%, which closely aligns with the 72% accuracy reported in the paper that corresponds to the BERT model's performance on the media-based dataset split after all de-biasing techniques had been applied.

After verifying its functionality, we used PoliticalBiasBERT to perform inference on all of the articles and transcripts that we had previously scraped. For each dataset of articles or transcripts, we collected the output softmax scores for the individual datapoints across the timespan of the dataset. We averaged these scores for each month and used these score averages for analysis.

3.3 Interpreting Political Bias Classifications

To allow for some interpretation of PoliticalBiasBERT's classifications, we make use of Local Interpretable Model-Agnostic Explanations (LIME) (Ribeiro et al., 2016). LIME generates "explanations" of an individual datapoint by perturbing instances of the datapoint (individual words in this case) and learning a sparse linear model local to these perturbations while treating the underlying global model as a black box. In this way, we can determine which words in the input contribute the most to each of the three classes and gain some insight from them. Figure 1 shows an example LIME explanation.

3.4 Hypothesis of Study

We hypothesize that the output of the PoliticalBiasBERT model will reflect commonly held beliefs on shifts in the political bias landscape with respect to certain politically significant “benchmark events". Specifically, we believe that news media sources known for having a certain political bias will become more polarized as either left-leaning or right-leaning when a major political event occurs, while periods of low activity have a more center-leaning characteristic. The scale of our dataset and our use of a model specifically fine-tuned for political bias classification will enable us to investigate this claim.

3.5 Novelty

Much work has been done with LLMs on individual articles and news outlets, but an analysis of bias over time has not been conducted. Social scientists have performed such temporal analyses, but have not leveraged LLMs in the process. In this way, our approach is novel, as we combine the former application with the latter method.

3.6 Challenges and Limitations

The main limiting factor for our analysis was the quantity of data. Many news websites, such as the New York Times, Associated Press, and The Wall Street Journal have measures in place to prevent mass scraping of their content. These websites have developer programs, which we tried to take advantage of but ultimately did not qualify for. This dramatically limited the amount of data we could acquire, which limited our ability to compare several news sources against one another. CNN and FOX News were straightforward to web scrape, but FOX deletes its transcripts regularly. We discovered that the Wayback Machine has most of the transcripts we were looking for, but scraping it was beyond the scope of the project. The New York Times’s developer program maintains a simplified version available to the public, which allows us to get summaries of every article. Running summaries on PoliticalBiasBERT is less reliable than running the whole article, but the summaries tend to line up with the ratings of the actual content.

Additionally, the classification analysis using LIME that we performed was limited in scope due to computation limitations. LIME must learn a linear model for each datapoint it is explaining, and the explainability power of this model is determined by the number of perturbations it is trained on, which can be affected by memory constraints.

4 Experiment Design and Results

Measuring the success of our analysis boiled down to a two part conclusion: being able to project the temporal data as it crosses time in an interpretable way, and deriving patterns and conclusions from that depiction. We collected data from multiple sources and use PoliticalBiasBERT to analyze scores across time from those sources. Our analysis centers on two main datasets, and we use the results from our experiments on these datasets to illustrate our key findings.

4.1 Analysis of Benchmark Events in the NYT Dataset

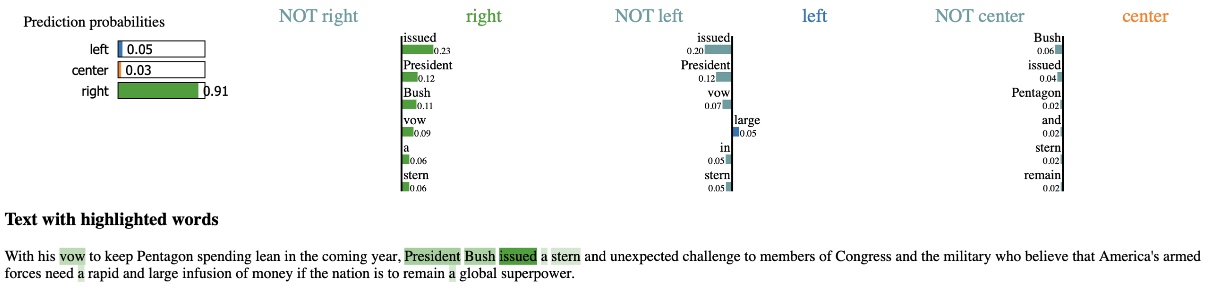
The most successful example of our pipeline is from the New York Times (NYT) Article summary data. This data was significant, as it had the most longevity and generated interpretable softmax scores back to 1998, making it ideal for analysis.

4.1.1 Experiment Results

We characterized the success of this analysis by taking the interpretable data and drawing conclusions about the patterns of bias for each benchmark event. Using our results from the NYT article summaries shown in Figure 2, we can see that a similar pattern emerges after each benchmark event. Following each event, there is a surge in the center softmax score and therefore a decrease in the left and right scores. This is especially noticeable after the 9/11 Attacks, the 2008 Market Crash, and the 2016 election.

4.1.2 Discussion and Analysis

We hypothesize that data becomes more center-leaning after these benchmark events. Left and right scores decrease together, rather than individually, suggesting that the data becomes more centered. We hypothesize that as eventful things happen, news tends to focus on the factual, unbiased aspects of the event.



With his vow to keep Pentagon spending lean in the coming year, President Bush issued a stern and unexpected challenge to members of Congress and the military who believe that America's armed forces need a rapid and large infusion of money if the nation is to remain a global superpower.

Figure 1: An example explanation produced by LIME for one of the summaries in our New York Times dataset.

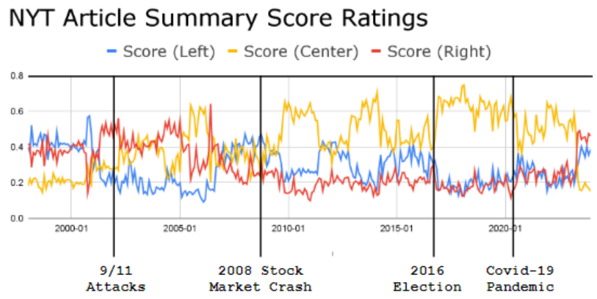


Figure 2: NYT Article Summary Scores from 1998 to 2024

To quantify this claim, we looked into the correlation of these softmax scores. The Left, Center, and Right scores sum to 1, so if the Center Score increases, Left and Right should have a negative correlation. We split the averaged Left and Right scores based on if they occurred before or after the most recent benchmark event. The scores from 0-6 months after the event were compiled into one dataset, and the remaining scores into another. This allows analysis of the correlation between the Left and Right metrics directly after the events and compares it to their correlation at other times. The results support our hypothesis. The correlation between scores not following a benchmark event are **-0.795**. This is expected, as these numbers should always have opposite reactions to each other. However, the correlation of the scores 6 months following benchmark events to be **-0.103**. This suggests Left and Right scores change similarly after benchmark events.

The NYT article summaries is the only dataset complete enough to draw such a conclusion. To further verify this claim would require seeing this pattern with other sources, as well as sources that are transcript text, not summaries. This analysis is outside the scope of this project, as we were unable to get access to additional data from other sources.

4.2 Analysis of Model Behavior in the CNN Live at Dawn Dataset

The second major dataset that we ran inference on was our CNN Live at Dawn dataset. This dataset contained web-scraped transcripts from the CNN Live at Dawn show spanning from the year 2000 to the year 2005.

4.2.1 Experiment Results

Despite having access to a similar time frame of data as the NYT summaries, our model returned scores that were seemingly uninterpretable past mid-2002. Around that time, each of the softmax scores becomes extreme. The scores (typically in the 0.3 and 0.5 range) shot up to above 0.98 for the politically left and below 0.02 for the politically center and right scores. The reported monthly averages stay in that range for the remaining air time of Live at Dawn. This suggested some significant change in the model's scoring which would adversely affect our analysis.

4.2.2 Discussion and Analysis

We have a few hypotheses for the reason behind this sudden change. Certain words in the show transcripts were labeled as biased simply due to overfitting, and don't accurately reflect the bias of the transcript. This includes terms like the speakers' names and phrases like 'CNN.' It would also explain why the NYT data didn't have this problem, because those are summaries, and wouldn't include specifics like who is authoring what is being said, and phrases like "New York Times." Another one of our hypotheses is that mid-2002 was around when the US began planning to war on Iraq. CNN being a left-leaning news source would have reported on this in a politically charged manner, and the model is potentially be reflecting that.

When further analyzing the patterns behind our results for the CNN Live at Dawn dataset, we produced LIME explanations on some of the tran-

Unfiltered CNN Text

“JOHN BERMAN, CNN HOST: That does it for us. CNN Tonight with Don Lemon starts now..."

Filtered CNN Text

“JOHN BERMAN, News Network HOST: That does it for us. News Network Tonight with Don Lemon starts now..."

Figure 3: An example showcasing the substitution of the term “CNN" with a more neutral term in our CNN dataset

scripts and found that the phrase “CNN" contributed significantly to the model’s overwhelmingly left-leaning classification. This is an issue, since the phrase “CNN" mostly only appears in parts of the transcripts that are not relevant to the main topic of discussion, such as when introducing news anchors or the name of a show. Furthermore, this result provides contradictory evidence against Baly et al. (2020)’s claims of de-biasing PoliticalBiasBERT against overfitting to the source of the media rather than making a decision based on the content of the article. Of course, this may be due to the different format of our data: PoliticalBiasBERT had been trained on published news articles, whereas the CNN Live at Dawn dataset consisted of audio transcripts. However, this poses a problem for our purposes of temporal political bias analysis.

To investigate the impact of this phrase, we decided to rerun PoliticalBiasBERT on a “filtered" version of the CNN datasets, where each instance of the phrase “CNN" was replaced with a neutral phrase (see Figure 3). Individual LIME explanations on filtered data showed a remarkably different pattern after the data was filtered (see Figure 4), representing the large effect that the phrase “CNN" was having on the model’s decisions. On a macroscopic scale, Figure 5 shows the dramatic effect on the outputs of the models across the entire CNN Live at Dawn dataset. The same time period from around mid-2002 onwards that exhibited left-leaning classifications with extremely high confidence now shows more balanced classifications, though left-leaning classifications still retain the majority. These results showcase a key finding about the PoliticalBiasBERT model, as its decisions are very susceptible to single words that appear in an innocuous context.

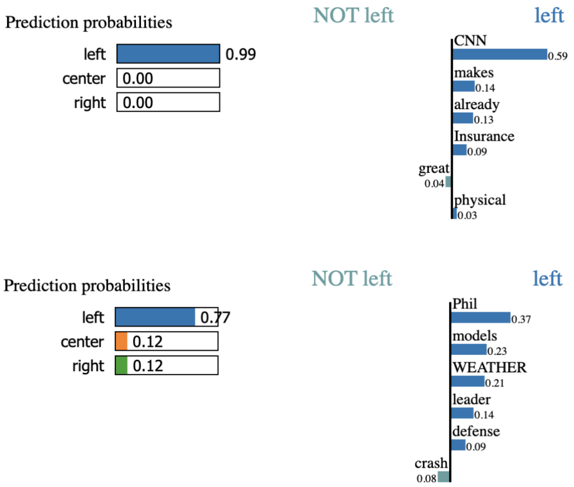
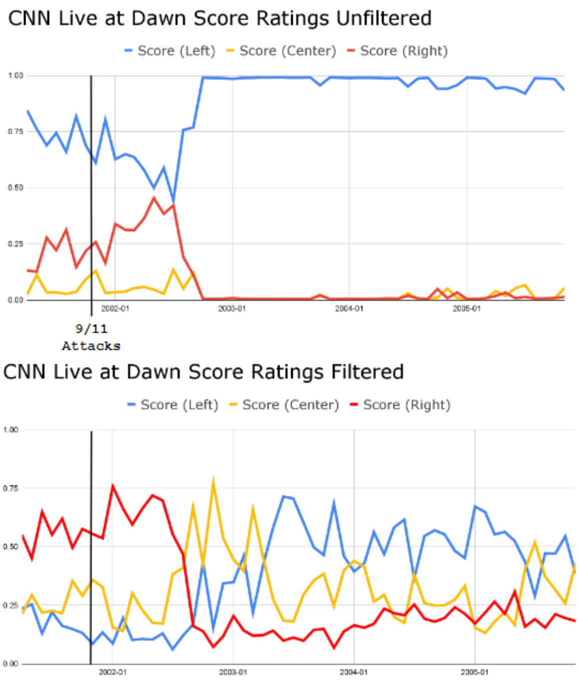


Figure 4: LIME explanations for an unfiltered datapoint from the “CNN: Live at Dawn" dataset (top) versus the same datapoint after being filtered (bottom)


CNN Live at Dawn Score Ratings Unfiltered

CNN Live at Dawn Score Ratings Filtered

Figure 5: A comparison of CNN Live at Dawn Scores before (top) and after (bottom) LIME Filtering

5 Conclusion

5.1 Replicability and Dataset Significance

Assuming another party can acquire the relevant transcripts, our results should be easy to replicate. Since this is not published work, we do not know of others whom our work has inspired, but we hope we have opened a new gateway to conduct further temporal analysis of news networks and anchors using LLMs since we did not find studies that did that before. Our extensive datasets of web-scraped articles can serve as a starting point for those who may want to perform a larger-scale analysis of political bias in multiple media networks.

5.2 Ethics of Study

There is more benefit from doing our study than harm. We intend to point out how the political bias of different news networks changes over time and after benchmark events, so at most, our project helps highlight the bias of news networks. People consume news to become informed, influencing their stance on various social and political issues. Transparently seeing how different news networks are biased and how events affect that bias can help to inform readers better.

5.3 Discussion of Study

Our study reveals the limitations of PoliticalBiasBERT. It's efficient at bias classification, but the words it chose to make those decisions were questionable, as we saw through the LIME interpretability tool. Our study reveals a need for stronger human-AI collaboration if a similar analysis is to be done properly in the future.

5.4 Future Work

We focused on American news written by American news networks surrounding American benchmark events. This is the context we understand best. There are many directions for future work, such as observing temporal trends in global news and comparing these to American news. An article written about the same news event from the same network in two different languages (i.e. CNN en Español vs CNN) may have a different bias classification from the model – would that be due to authorship, or the model's ability to handle the two languages?

References

Ramy Baly, Giovanni Da San Martino, James Glass, and Preslav Nakov. 2020. We can detect your bias: Predicting the political ideology of news articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4982–4991, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Lisa Fan, Marshall White, Eva Sharma, Ruisi Su, Prafulla Kumar Choubey, Ruihong Huang, and Lu Wang. 2019. In plain sight: Media bias through the lens of factual reporting. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6343–6349, Hong Kong, China. Association for Computational Linguistics.

Siobhan Grayson and Derek Greene. 2019. Temporal analysis of reddit networks via role embeddings.

Eunji Kim, Yphtach Lelkes, and Joshua McCrain. 2022. Measuring dynamic media bias. volume 119, page e2202197119. National Acad Sciences.

Luyang Lin, Lingzhi Wang, Xiaoyan Zhao, Jing Li, and Kam-Fai Wong. 2024. IndiVec: An exploration of leveraging large language models for media bias detection with fine-grained bias indicators. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1038–1050, St. Julian's, Malta. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.