Manvika Satish
Bella Boulais
Professor Kearns
CSC 480-03
December 5, 2024

Detecting Political Bias with AI: Comparing Bias Detection Tools

ABSTRACT

This project investigates the effectiveness of AI-driven tools in detecting and categorizing political bias in news and social media content. By leveraging models such as GPT-4, VADER, and PoliticalBiasBERT, we analyze their ability to identify linguistic patterns and emotional tones that signal left-leaning, center, or right-leaning biases. Using a curated dataset of political articles, social media posts, and candidate statements, we evaluate the accuracy and limitations of these tools in politically sensitive contexts. Our findings reveal that while these tools excel at detecting overt biases, they struggle with nuanced or neutral content. We recommend combining multiple models, refining training data, and incorporating human oversight to enhance the reliability and interpretability of AI in political bias detection. This study provides valuable insights into the potential and challenges of using AI for promoting fairness and transparency in political discourse.

INTRODUCTION

In an era where political discourse heavily influences public opinion, the ability to detect and categorize political bias in media and social platforms has become increasingly important. This project investigates the effectiveness of AI-driven tools in identifying political bias across news articles and social media content. By comparing multiple models, including GPT-4, VADER, and Hugging Face Transformers, we aim to evaluate their accuracy, transparency, and interpretability in detecting biases. Using a curated dataset of political texts—spanning news sources, candidate statements, and social media posts—we will assess the strengths and limitations of these tools in politically sensitive contexts.

SOCIAL PROBLEM ADDRESSED

How accurate are AI-driven bias detection tools in identifying and categorizing political bias in news and social media content?

AI TOOLS AND TECHNOLOGY

OpenAI GPT-4

In our project, we use OpenAI's GPT-4, accessed via ChatGPT, to evaluate its ability to identify political bias in texts through prompt-driven analyses. By crafting specific prompts, we guide GPT-4 to scrutinize tone, bias, and language choices within a dataset of news articles and social media posts. These prompts aim to detect political leanings, evaluate emotional tone, and highlight language that may influence public perception, systematically assessing whether a text is left-leaning, right-leaning, or neutral. Comparative analysis prompts further allow us to explore how biases vary between sources and over time. This approach tests GPT-4's capability to uncover both explicit and implicit biases, providing insights into its effectiveness as a tool for detecting political bias in politically sensitive contexts.

General Analysis Prompts

1. **Basic Bias Detection**

Prompt: *"Analyze the following text for political bias. Identify any language that may suggest a leaning towards particular political ideologies or parties."*

Summary of Findings: The analysis by GPT-4 demonstrates its capability to detect potential political bias in news articles by identifying specific linguistic cues and framing techniques that align with particular ideological leanings. For the left-leaning articles, GPT-4 noted biased elements such as derogatory language, selective emphasis, and negative portrayal of one political group while highlighting the success of the opposing group. In the right-leaning articles, GPT-4 recognized biased cues in emotive language, repetitive negative incidents, and the strategic placement of information to influence perception.

2. **Tone and Sentiment Analysis**

Prompt: *"Evaluate the tone and sentiment of this text. Does the language used suggest a positive, negative, or neutral view towards the subjects discussed? Elaborate."*

Summary of Findings: OpenAI's GPT-4 demonstrates a nuanced ability to evaluate tone and sentiment in the provided text samples. The model effectively identifies a neutral-to-negative

tone in the left-leaning articles, emphasizing criticism towards Republican politicians, and a negative sentiment in the right-leaning articles, focusing on Democrats' interaction with the press. GPT-4's analysis highlights the use of emotionally charged language and framing techniques that suggest biases in both texts, pointing towards its capability in uncovering subtle biases within political content. Though it does seem to be more left-leaning.

3. **Comparative Bias Analysis**

Prompt: *"Compare the political bias in these two texts. Which text shows a more pronounced political bias, and in what way?"*

Summary of Findings: The output from GPT-4 comparing the political biases in texts from left and right-leaning sources shows that the AI identifies some right-leaning texts' portrayal as more overtly biased due to its emotionally charged language and framing of the GOP in a negative light. In contrast, left-leaning documents suggest less pronounced bias. This ability to discern nuanced differences in how bias manifests in the texts indicates that GPT-4 can effectively uncover biases in news content.

4. **Detailed Language Use Analysis**

Prompt: *"Describe the use of emotionally charged language, euphemisms, or dysphemisms in this text. How might these language choices affect the perceived bias of the content?"*

Summary of Findings: GPT-4 effectively identifies and describes the use of emotionally charged language, euphemisms, and dysphemisms in articles from different sources like CNN and FOX News, highlighting how these language choices contribute to perceived bias in the content. It provides detailed examples of specific terms and phrases used in each article that carry implicit connotations, enhancing the reader's perception of bias toward or against political figures and actions.

Specific Prompts for Different Political Orientations

1. **Left-Leaning Bias Detection**

Prompt: *"Identify and explain any left-leaning bias in the following text. What specific phrases or rhetoric indicate this bias?"*

Summary of Findings: The output from OpenAI GPT-4 effectively identifies several elements indicating a left-leaning bias in the given text, such as negative phraseology directed at Republicans, a focus on Republican failures, a resilient portrayal of Democrats, negative

portrayals of individual Republicans, and an imbalance in contextual reporting between the two parties.

2. **Right-Leaning Bias Detection**

Prompt: *"Identify and explain any right-leaning bias in the following text. What specific phrases or rhetoric indicate this bias?"*

Summary of Findings: GPT-4 demonstrates a strong ability to uncover subtle biases in text by analyzing linguistic choices, framing, and rhetorical patterns. In most cases, it identifies right-leaning bias through the use of negative imagery, dramatic descriptions, and repeated emphasis on themes of avoidance and non-transparency regarding President Biden.

3. **Neutral Bias Assessment**

Prompt: *"Does this text maintain political neutrality? Please provide examples from the text that support your analysis."*

Summary of Findings: GPT-4 demonstrates a strong ability to identify subtle biases in text by analyzing language, tone, and framing, as seen in its assessment of the article. It highlights emotionally charged terms, selective presentation of data, and the portrayal of policies in a potentially negative light, providing concrete examples to support its analysis. Appears to be more left-leaning, but is able to detect that this article is relatively neutral.

In-Depth Evaluation Prompts

1. **Contextual Bias Analysis**

Prompt: *"Considering the political context of [Insert Year/Event – varies based on document], analyze this text for any implicit or explicit biases. How might the context influence the presentation of information?"*

Summary of Findings: GPT-4 demonstrates a strong ability to identify and categorize both explicit and implicit biases by analyzing language, context, and framing within politically charged texts. It effectively highlights key aspects, such as the explicit racial dynamics, emotive language, and systemic issues emphasized during notable political movements and events, taking into account who is in office, showing sensitivity to the sociopolitical climate. However, its analysis relies heavily on widely accepted narratives and does not critically challenge these narratives, suggesting it may sometimes reflect rather than independently assess societal biases.

2. **Impact of Bias on Public Perception**

Prompt: *"Assess how the bias in this text could influence public opinion about the topics discussed. What long-term impacts might this bias have if the text is widely accepted as factual?"*
Summary of Findings: GPT-4 demonstrates strong analytical capabilities in uncovering biases by identifying how narratives in texts can shape public opinion and highlighting potential long-term impacts of such biases. In the provided examples, it effectively analyzes the framing of misinformation about Democrats and Republicans, noting the risks of increased cynicism, polarization, normalization of misinformation, and erosion of trust in media or democratic institutions.

3. **Bias in Reporting Style**

Prompt: *"Analyze the reporting style of this article. How does the structure, language, and presentation contribute to or mitigate the political bias?"*
Summary of Findings: GPT-4 demonstrates strong capabilities in identifying structural, linguistic, and presentation-based biases in news articles by analyzing language tone, narrative focus, and framing. In the left-leaning articles, it highlights a subtle left-leaning bias through emotionally charged language and selective emphasis on Republican absences, while the right-leaning articles' right-leaning bias is revealed through conflict-driven narratives and negative connotations toward President Biden. The analysis shows GPT-4's potential in uncovering nuanced biases, but its accuracy depends on how well it contextualizes these elements and avoids overinterpreting isolated details, making it a promising but not flawless tool for bias detection.

VADER

VADER (Valence Aware Dictionary and sEntiment Reasoner) is an open-source sentiment analysis tool designed to work effectively with social media text, although it is also applicable to larger datasets. VADER scores text on four sentiment categories: positive, neutral, negative, and compound. This scoring system helps highlight emotionally charged or biased language, which can be crucial for detecting potential bias in texts such as political data.

Model Overview

VADER is a sentiment analysis tool designed to detect emotional tones in text, making it useful for identifying political bias. It uses a sentiment lexicon, where each word is assigned a score

from -1 (negative) to 1 (positive). The model also adjusts sentiment scores based on context, such as punctuation, capitalization, and negations (e.g., "not good" lowers sentiment). VADER calculates an overall polarity score for the text, which ranges from -1 (very negative) to 1 (very positive). All we had to do now was transfer our data into text files and send it through our Vader program!

Sample Output

For our analysis, the articles from different media outlets yielded the most interesting results! We assessed sources such as CNN, Fox News, and Forbes, to assess the emotional tone and potentially detect any bias in their content. Below are the results from the sentiment analysis:

1. Absences by Trump's Senate pals help Democrats confirm Biden judges

2. As Biden dodges press on Brazil trip, frustrated reporters resort to holding up signs with questions | Fox News

3. https://www.forbes.com/sites/stuartanderson/2021/02/01/the-story-of-how-trump-officials-tried-to-end-h-1b-visas/

Observations

The sentiment analysis results were unexpected. As shown in our samples, the CNN article received a notably negative sentiment score, consistent with its critical tone towards Republicans. The Fox News article, on the other hand, exhibited a positive sentiment score, reflecting a more favorable stance. The Forbes article also carried a negative sentiment, primarily due to its critical portrayal of the Trump administration. Interestingly, across all analyzed articles, neutral sentiment scores were rare, indicating that political reporting often skews toward emotional extremes. Even articles from media outlets labeled close to "neutral" (AdFonte) often used emotionally charged language, highlighting the challenge of separating sentiment from

political leanings in both content and sources. VADER provides valuable insights into the emotional tone of political texts and can reveal biased language across different political orientations. However, its results are influenced by the emotional content of the material, making it a useful but imperfect tool for detecting political bias. Additional research would be needed to separate bias in the reporting from bias inherent in the subjects being reported. Specifically, if you were trying to determine if a media company was biased, additional research would be needed to separate the events and people being reported on from the actual reporting.

Hugging Face Transformers (Pre-trained BERT model)

For our project, we utilize the Hugging Face Transformers library to enhance our analysis of political bias in text. Specifically, we employ the pre-trained model "bucketresearch/politicalBiasBERT," which is designed to classify input text as left-, center-, or right-leaning. This model provides a quantitative measure of bias by assigning probabilities to each political category, enabling us to infer potential leanings based on the language and framing used in the text. For example, articles with strong left-leaning probabilities might use language that aligns with progressive ideologies, while those classified as right-leaning may feature phrasing associated with conservative perspectives. To integrate this model into our methodology, we stripped documents into .txt files and ran them through the model to detect patterns of political bias. By analyzing the classification results, we systematically examine news articles and other textual content to identify biases, enriching our understanding of how political narratives are shaped and represented across different sources.

Model Overview

The PoliticalBiasBERT model is a fine-tuned version of BERT designed specifically to classify text based on political bias. It works by processing input text through a transformer-based architecture, which uses self-attention mechanisms to understand the context and relationships between words. The model assigns probabilities to three predefined classes: left, center, and right. To use the model, the text is first tokenized into numerical representations using a tokenizer compatible with BERT. These tokenized inputs are then passed through the model, which outputs logits (unprocessed scores) for each class. These logits are converted into probabilities using a softmax function, allowing the model to predict the most likely bias

category based on the input text. The fine-tuning process involved training the model on datasets containing politically biased text, enabling it to recognize linguistic patterns and cues associated with different political ideologies. This setup makes PoliticalBiasBERT a valuable tool for analyzing bias in news articles, social media content, and other textual sources.

Sample Output

Document: Absences by Trump's Senate pals help Democrats confirm Biden judges

CNN Article, Left-leaning (https://adfontesmedia.com/interactive-media-bias-chart/)

Probabilities Format: [Left, Center, Right]

```
Text: A ##bs ##ence ##s by Trump ' s Senate p ##als help Democrats confirm B ##iden judges By Tier ##ney S...
Probabilities: [0.9615755081176758, 0.021505387499928474, 0.016919072717428207]
Predicted Class: Left
Text: said , while arguing that Democrats would have still confirmed the nominee - even if all Republicans...
Probabilities: [0.9632146954536438, 0.005643416196107864, 0.03114188462495804]
Predicted Class: Left
Text: shows the high stakes with which every judges ##hip is viewed . When Trump re ##take ##s the White H...
Probabilities: [0.9557483196258545, 0.007961812429130077, 0.03628985211253166]
Predicted Class: Left
Overall Predicted Class: Left
Overall Probabilities: [0.9601795077323914, 0.011703538708388805, 0.02811693586409092]
```

Observations

The Hugging Face BERT model performs well in detecting clear political bias, particularly when the text has overtly left- or right-leaning language. For example, it confidently and accurately classified texts like "Absences by Trump's Senate pals…" and "Misleading GOP videos…" with over 95% certainty. However, it struggles with neutral or nuanced content, often misclassifying center-leaning texts as right-leaning, such as in "The Story of How Trump Officials Tried to End H-1B Visas." This indicates the model may rely too heavily on features tied to right-leaning discourse, possibly due to imbalances in the training data.The model also shows variability in confidence, with some classifications highly certain and others more evenly split across categories, reflecting uncertainty in ambiguous cases. When processing multi-chunk texts, errors in aggregating probabilities can further skew results. While the model excels with texts

containing explicit bias cues, it struggles with balanced or conflicting perspectives, indicating a dependence on direct linguistic signals. Overall, the model is effective for detecting strong bias but needs improvements in handling neutrality and subtle framing to be reliable for real-world political bias detection.

ANALYSIS

Our analysis involved applying three AI tools to a dataset of political articles and social media content. GPT-4 was tested with a set of well-defined prompts, while the VADER and BERT just required setting up the data for it to run. From these models, we were able to evaluate their ability to detect bias.

- OpenAI GPT-4: GPT-4 demonstrated a strong ability to detect subtle political bias in text. Through specific prompts, it identified language, tone, and framing techniques that signaled left-leaning and right-leaning biases. GPT-4 also excelled in comparative analysis, discerning nuanced differences in bias across texts. However, GPT-4's analysis was occasionally skewed toward a left-leaning perspective, reflecting potential biases in its own training data.
- VADER: VADER, proved useful in highlighting emotionally charged language, which is often indicative of political bias. However, VADER's analysis may miss more subtle forms of bias that are not expressed through emotion but through framing or ideological slant. Thus, VADER is useful for spotting emotionally charged language in political content, but it may need to be combined with other methods to fully address political bias. Further evaluation from the user would be needed in order to explain the biases in the data.
- Hugging Face PoliticalBiasBERT: This pre-trained model classifies text as left-, center-, or right-leaning. By using the model's output probabilities, we were able to identify political leanings in various texts. While the model provided a quantitative measure of political bias, its predictions were not always accurate, particularly in cases where texts contained mixed political tones or were more nuanced.

NEXT STEPS

Moving forward, we propose the following next steps for improving AI-driven political bias detection:

1. Combine AI Models: To address the limitations of individual tools, future studies could integrate GPT-4, VADER, and PoliticalBiasBERT to create a more robust and reliable system for detecting political bias across various media sources.
2. Refine Training Data: Bias in AI models can often reflect biases in their training data. Efforts to diversify and refine training data, particularly for models like GPT-4 and PoliticalBiasBERT, could help mitigate the inherent biases present in these systems.
3. Expand to More Diverse Sources: Future investigations should include a broader range of media sources, including international outlets and independent media, to assess the generalizability of AI-driven bias detection tools.
4. Human-AI Collaboration: Given the complexities of political bias, human oversight will continue to play an important role in interpreting the results of AI tools. Combining AI analysis with human judgment could lead to more nuanced and accurate bias assessments.

SUMMARY

This study examines the capabilities and limitations of AI-driven tools in detecting political bias in news and social media content. While tools like GPT-4, VADER, and PoliticalBiasBERT show promise in identifying and categorizing bias, they each have strengths and weaknesses. GPT-4 excels in nuanced bias detection and tone analysis, VADER provides valuable sentiment insights, and PoliticalBiasBERT offers a quantitative measure of political leanings. However, none of these tools can independently offer a foolproof solution to detecting political bias. As AI continues to evolve, combining multiple models and refining their training data will be crucial for improving the accuracy and reliability of political bias detection tools.

REFERENCES & DATA

A&E Television Networks. (n.d.). *George Floyd is killed by a police officer, igniting historic protests*. History.com. https://www.history.com/this-day-in-history/george-floyd-killed-by-police-officer

Anderson, S. (2024, February 20). *The story of how trump officials tried to end H-1B visas*. Forbes. https://www.forbes.com/sites/stuartanderson/2021/02/01/the-story-of-how-trump-officials-tried-to-end-h-1b-visas/

Eubanks, A. (2020, July 1). *Political tweets*. Kaggle. https://www.kaggle.com/datasets/khanradcoder/political-tweets/data

Interactive Media Bias Chart. Ad Fontes Media. (2024, October 9). https://adfontesmedia.com/interactive-media-bias-chart/

Margolis, A., & Fox News. (2024, November 3). *Trump campaign clarifies after candidate jokes about shooting "through the fake news" in Pennsylvania*. Fox News. https://www.foxnews.com/politics/trump-campaign-clarifies-after-candidate-jokes-about-shooting-through-fake-news-pennsylvania

NBCUniversal News Group. (2024, June 20). *Misleading GOP videos of Biden are going viral. the fact-checks have trouble keeping up.* NBCNews.com. https://www.nbcnews.com/politics/2024-election/misleading-gop-videos-biden-viral-fact-checks-rcna133316

Panreck, H., & Fox News. (2024, November 20). *As Biden Dodges press on Brazil trip, frustrated reporters resort to holding up signs with questions*. Fox News. https://www.foxnews.com/media/biden-dodges-press-brazil-trip-frustrated-reporters-resort-holding-up-signs-questions

Sneed, T. (2024, November 20). *Absences by Trump's Senate pals help Democrats confirm Biden judges | CNN politics*. CNN. https://www.cnn.com/2024/11/20/politics/judges-trump-biden-missing-senators/index.html

Wikimedia Foundation. (2024b, December 3). *Donald Trump*. Wikipedia. https://en.wikipedia.org/wiki/Donald_Trump

Wikimedia Foundation. (2024c, December 4). *Joe Biden*. Wikipedia. https://en.wikipedia.org/wiki/Joe_Biden