Proceedings of the Eighteenth International AAAI Conference on Web and Social Media (ICWSM 2024)

Online News Coverage of Critical Race Theory Controversies: A Dataset of Annotated Headlines

Anna Lieb[1, 2], **Maneesh Arora**[2], **Eni Mustafaraj**[1]

[1]Department of Computer Science, Wellesley College, MA, USA
[2]Department of Political Science, Wellesley College, MA, USA
anna.lieb, maneesh.arora, eni.mustafaraj@wellesley.edu

Abstract

In this paper, we introduce an annotated dataset of 11,704 unique U.S. news headlines related to critical race theory and its controversies from August 2020 through December 2022. Annotations generated by GPT-4 specify the headline stance and the primary actor in the headline. GPT-4 annotations performed well on the validation dataset, with weighted average F-scores of 0.8339 for headline stance annotations and 0.7625 for primary actor annotations. Along with the annotated headlines and URLs to the full article, we augment the dataset with metrics that are relevant to future research on political polarization, news frame analysis, and regional news coverage. The dataset includes partisan audience bias scores by news source domain, tags for mentions of U.S. states in the article body, and exposure and engagement metrics for articles shared on Reddit. Among other preliminary descriptive analyses, we find that the most frequent headline stance in our dataset is anti-CRT (43.06%), and the most frequent primary actor is political influencers (56.56%). This paper describes the data collection methodology, preliminary descriptive analysis, and possible uses of the dataset for future research in political science, computational social sciences, and natural language processing. Our dataset and replication code is available to access on Zenodo at zenodo.org/doi/10.5281/zenodo.10516190

Introduction

Starting in September 2020, conservative American political figures popularized the term "critical race theory" (CRT) to describe educational curricula and employee training on racial inequality in the United States. Between 2020-2022, CRT remained one of many polarizing political controversies related to racial identity in the U.S. (Ray and Gibbons 2021). News media has played a significant role in shaping the narrative surrounding CRT controversy (Pollock et al. 2022). In fact, some journalists have traced the emergence of CRT controversy to a single Fox News television broadcast, in which conservative commentator Christopher Rufo spoke about the issue on "Tucker Carlson Tonight" and caught the attention of U.S. President Donald Trump.[1]

[1]newyorker.com/news/annals-of-inquiry/how-a-conservative-activist-invented-the-conflict-over-critical-race-theory

Political communications research has previously established that news media issue framing — in other words, how news media emphasizes certain aspects of an issue — can influence political attitudes and public understandings of an issue (Iyengar 1994; Chong and Druckman 2007). Inspired by the ways that news media has shaped the public narrative surrounding CRT, we developed the dataset presented in this paper to aid future interdisciplinary research on online news coverage of CRT controversies.

A main contribution of this dataset is not only a set of online article headlines and URLs, but also a novel set of headline frame annotations that we generated for each article headline. The headline frames are specific to CRT and include two aspects: headline stance and primary actor. We prompted a pre-trained large language model (OpenAI's GPT-4) to generate the stance and primary actor labels for each headline in our dataset. To supplement the news headlines and annotations, we augmented the dataset with additional preexisting metrics that we believe will be helpful for future research. A summary of all features in our dataset is provided in the following section.

Summary of Dataset Features

Our dataset includes 11,704 unique news article headlines published as 18,686 unique URLs at 1,684 unique web domains from August 2020 through December 2022. To further characterize the landscape of online news coverage of CRT during this period, we augmented the dataset with a series of metrics that may be useful for future researchers who are interested in tracking online news reporting on U.S. social issues. Our dataset includes the following metrics:

GDELT articles and their metadata. The dataset includes data from the GDELT Project DOC 2.0 API: news article headlines, URL(s) to the article, publication date, and the web domain of the URL. Data is first sorted chronologically by date, and then alphabetically by article title.

Mentions of U.S. states. For each headline in the dataset, we track the state-specific search result lists in which the headline appears. 49 columns correspond to mentions of a U.S. state in the article (one column for each state, excluding Washington). A state's column value is `0` if the state is not mentioned, and `>0` if the state is mentioned. If the state is mentioned, then the value counts repetitions; it represents the number of times the same headline appeared in search

Primary actor	Actor role	Example actors
Educational practitioner	Deliver instruction to students	schools, universities, school districts, teachers, professors, school administration
Political influencer	Represent the political interests and/or policy preferences of the people	governors, school boards, political commentators, the president, political figures
Impacted actor	Participate in the school system without direct influence over policies or educational practices	Students, parents, voters
None /Other	Does not fit any of the other categories	Headlines that do not identify any primary actor. For example: "What is CRT?"
Headline stance	**Stance description**	**Example headlines**
Anti-CRT	The headline appears to favor CRT bans and/or make alarmist claims about threats posed by CRT	"The Left Assault on Racist Math Continues: DeSantis Rejects CRT-Riddled Textbooks"
Defending CRT	The headline appears to support CRT and/or oppose restrictions on CRT-related curricula	"Guest commentary: Legislation on critical race theory isn't a good idea"
Neutral	The headline is impartial. It reports on news events without favoring one viewpoint or the other.	"Idaho House And Senate Pass Bill Banning Critical Race Theory In Schools"

Table 1: CRT news frame labeling scheme with two aspects: primary actor (4 levels) and headline stance (3 levels). The news frame annotation scheme was developed inductively through multiple rounds of collaborative qualitative analysis.

results corresponding to a different URL.

CRT issue frames. All of the CRT-relevant headlines in our dataset are annotated with CRT issue frames. These annotations describe the how the given headline reports on CRT controversies. The issue frames are defined with two aspects: (1) primary actor (educational practitioner, political influencer, impacted actor, or other) and (2) headline stance (anti-CRT, defending CRT, and neutral). The issue frames were developed inductively by human labelers according to the process described in the Methods section and in Table 2. Their values and definitions are provided in Table 1.

Partisan audience bias scores. The data is further augmented with partisan audience bias (PAB) scores, which indicate whether the domain of the URL is more likely to be shared by Democrats or Republicans. The PAB scores were developed by Robertson et al. (2018) based on web domains shared by Democrats and Republicans on Twitter. The PAB scores range from -1 to 1, where a score of -1 means that a web domain is shared exclusively by Democrats, a score of 1 means that a domain is exclusively shared by Republicans, and a score of 0 indicates that the domain is shared equally by Democrats and Republicans (Robertson et al. 2018).

Reddit engagement data. Using Pushshift Reddit dumps, we collected 6,715 Reddit posts that contain a URL from our dataset of CRT-relevant news URLs. We saved the submission date, submission title, subreddit name and ID, number of subreddit subscribers, number of comments, score (number of upvotes minus number of downvotes), and upvote ratio for each post. This Reddit data may be useful for computational social science work that investigates how CRT news circulated in the Reddit community.

Motivation and Potential Uses

By publishing this annotated CRT news headlines dataset, we hope that researchers from multiple disciplines will be able to use the data for future research.

Political scientists, for example, might be interested in using the dataset to answer questions about the ways that racial identity issues emerge in U.S. news media. With the combination of metrics available in this dataset, political scientists could explore research questions like: How does CRT controversy spread across different U.S. states? Which U.S. states are "trendsetters" in terms of policy action or national narrative-building? Are Republican-leaning articles (according to PAB) more likely to feature "grassroots" activists (like students and parents) or "top-down" political executives (like presidents and governors)?

Our preliminary results suggest that U.S. states with nationally-prominent Republican leadership drove initial CRT coverage; that headlines tended to have language that favors an anti-CRT viewpoint, especially in early CRT coverage; and that headlines largely focused on elite-driven "top-down" political events instead of grassroots actors like parents and students, even in the earliest mentions of CRT. Further research using this dataset could contribute to political science research areas including political polarization, news media framing theory, and partisan agenda-setting.

Meanwhile, researchers in computational social science may be interested in using the dataset to investigate questions related to the distribution of CRT articles in online communities. For example, which types of news articles (ie. PAB scores, headline stance, primary actor) tend to be shared more frequently on Reddit? Do news stories that mention certain U.S. states gain more traction on digital platforms? Computational social scientists could use this dataset to contribute to research areas related to polarized online communities and distribution of news sources online.

Finally, natural language processing research could also benefit from the findings of this paper. As high-performance large language models have become increasingly accessible over the past year, various researchers have been testing its capabilities for qualitative analysis (Xiao et al. 2023; Zhang et al. 2023). Our dataset builds on this work by similarly using a large language model to generate news frame an-

notations. This paper also relates to previous work on NLP methods for issue framing detection in online news articles (Field et al. 2018; Kwak, An, and Ahn 2020). This task also relates more broadly to mainstream NLP tasks like stance detection and semantic role labeling, since the news framing annotations are assigned by the two subfactors "headline stance" and "primary actor." Future natural language processing work could also make use of the URLs provided in the dataset to access full text versions of news articles to complement the headline-based and URL-based data provided in our dataset.

Related Work

Critical Race Theory Controversy

Decades before it entered the American political mainstream, the term "critical race theory" (CRT) originated from an academic movement that emphasizes how racial inequality persists in the United States due to systemic racism embedded in American institutions such as the criminal justice system, labor market, and housing market (Crenshaw et al. 1995). Starting in September 2020, conservative American political figures appropriated the term to describe educational curricula and employee training on racial inequality.

Opponents of CRT have made it clear that they aim to roll back any progress made on battling racial inequality and other forms of oppression. Christopher Rufo, a conservative activist and a primary architect of the anti-CRT movement, tweeted: "We have successfully frozen their brand—'critical race theory'—into the public conversation and are steadily driving up negative perceptions. We will eventually turn it toxic, as we put all of the various cultural insanities under that brand category."[2]

Following conservative activists' pleas to "abolish critical race theory," President Donald Trump issued an executive order to limit the federal government's promotion of "divisive concepts" related to race and sex.[3] Though the Biden administration rescinded the executive order a few months later, it had already become a rallying cry for Republican politicians and racial conservatives around the country. Indeed, since Trump's executive order, almost 800 anti-CRT measures have been introduced by government entities at the local, state, and national level (Alexander et al. 2023).

Political science scholarship demonstrates that racial inequality, a primary mobilizing issue of racial conservatives, has transformed throughout history, adapting to new political climates (Siegel 1997; Mendelberg 2001; Alexander 2010; Haney-López 2014). Anti-CRT is the latest form of anti-Blackness that emerged from backlash to the "racial reckoning" of summer 2020.[4] Conservatives have positioned the anti-CRT movement in opposition to social issues ranging from Black Lives Matter protests, LGBTQ+ school organizations, transgender rights, and diversity training.[5]

The dataset presented in this paper can provide valuable insight into news media coverage of the anti-CRT movement. The dataset can be used to show how the issue spread throughout states and local municipalities, what frames were used, and who the primary actors were. Also, this analysis can be replicated on the raw dataset of school headlines to investigate other issues simultaneously taking place in American schools.

The GDELT Project

To establish a large dataset of online news articles related to critical race theory in the United States, we collected raw data from the GDELT (Global Data on Events, Location and Tone) Project.[6] The GDELT Project monitors print, broadcast, and web news media in over 100 languages from countries around the world (Leetaru and Schrodt 2013). Although the GDELT Project data contain some offline sources, previous analyses have shown that over 99.9% of news articles on GDELT are collected from the web (Kwak and An 2016). To create our dataset we used the GDELT Events Database 2.0 DOC API, which provides free and unrestricted access to a database of news archives starting from January 1, 2017 with updates every 15 minutes[7].

We considered other methods for collecting online news data, such as scraping news websites directly or news aggregators like Google News. However, a major drawback of this approach is that it limits our search to news published in the most recent days or weeks. Since we hoped to track the emergence of critical race theory news coverage from its inception, it was important for our dataset to include robust historical data. Additionally, previous research has shown that the GDELT dataset includes more detailed metadata, wider coverage, and more extensive database of documents compared to similar widely-used datasets, such as the ICEWS (Integrated Conflict Early Warning System) dataset and the Event Registry platform (Ward et al. 2013; Arva et al. 2013; Kwak and An 2016).

Methods for Data Collection

Our dataset is based on a set of news articles collected from the GDELT Events Database. After collecting news articles and their basic metadata, we tagged the articles for mentions of U.S. states and determined their relevance to CRT. Then, we augmented the data with headline frame labels (stance and actor), partisan audience bias scores, and engagement metrics from Reddit.

Querying the GDELT DOC 2.0 API

First, we conducted a series of GDELT DOC 2.0 API requests to collect the headline, URL, date, and web domain

[2]https://twitter.com/realchrisrufo/status/1371540368714428416

[3]https://www.vox.com/2020/9/24/21451220/critical-race-theory-diversity-training-trump

[4]https://www.npr.org/2020/08/16/902179773/summer-of-racial-reckoning-the-match-lit

[5]https://www.heritage.org/civil-rights/report/critical-race-theory-the-new-intolerance-and-its-grip-america

[6]According to the GDELT Project Terms of Use, GDELT datasets may be redistributed with attribution in any form: https://www.gdeltproject.org/about.html

[7]https://blog.gdeltproject.org/gdelt-doc-2-0-api-debuts/

of articles about critical race theory controversy. Since conservative commentators brought critical race theory controversy to mainstream media starting in September of 2020, we searched for articles in the date range from August 1 2020 through December 31 2022. We further restricted our search to articles published in the United States in English. To cast a wide net, we ran a series of queries with the keyword "school," with additional queries to track mentions of states in the U.S. By initially collecting broadly, we can normalize future results by total school-related coverage. We collected a maximum of 100 articles per day per state for the following keyword searches (the keyword(s) can appear in the article headline or article text):

- "school"
- "school" and a U.S. state name[8]
- "critical race theory" or "CRT"

After collecting data on articles that mentioned "school," we tagged articles with a binary relevance variable depending on whether the article headline contained "critical race theory" or "CRT".[9] By tagging relevance this way, we can analyze the frequency of CRT-relevant headlines as a proportion of total school-related coverage. We defined relevance this way as a "first pass" to isolate articles whose headlines directly mention CRT, rather than articles that may mention CRT less prominently in the article body. Based on qualitative review, we found that this was a reasonably strict rule that mostly prevented false positives (i.e. falsely tagging irrelevant articles as relevant). However, it likely led to false negatives, which we did not systematically diagnose. Note that some false positives were removed later, after clustering analysis (described below).

Counting U.S. State Mentions and Repetition

We combined the data from multiple searches in two ways: (1) compiling results by unique headlines, and (2) compiling results by unique URLs.

In case (1), we created a dataset where each row represents a unique headline that appears at least once in any GDELT search result. The dataset has a column for each U.S. state, so that each headline can be tagged by the number of times it appears in state-specific results. For example, consider an article that mentions "Florida" and "California" in its headline or body. We would expect this article to appear once in the "Florida" keyword search results and once in the "California" keyword search results, such that the headline's entry would have `FL_count = 1` and `CA_count = 1`. This indicates that the article headline appears exactly once in each state result list. However, some articles are published multiple times at different URLs. Consider the case that this headline is published once at a Florida NPR[10] web domain, and again at a California NPR domain. In this case, the headline is repeated in our results so `FL_count = 2` and `CA_count = 2`.

For this dataset, the majority of headlines appeared exactly once in the result list. 19.2% of headlines repeated once in the same search result list, and only 6.7% of headlines repeated two or more times. The most-repeated article headline had 83 distinct URLs for the same headline: "Opponents of critical race theory seek to flip school boards." This headline was from the Associated Press, which shows that repetition may indicate syndication. However, future researchers should be cautious about assigning significance to repeated publication. Previous research suggests that the number of documents from a source in the GDELT database does not correlate with the web traffic of the source (Kwak and An 2016), so we cannot make conclusions about salience or spread of these news sources based on URL repetition alone.

In case (2), we created a dataset where each row represents a unique URL that appears at least once in any GDELT search result. With the same method as described above, each URL was tagged for mentions by U.S. state. This dataset (organized by unique URL) may be more useful than the other dataset (organized by unique headline) when researchers aim to analyze news sources by web domain or access the full article text at the URL, rather than primarily analyzing headlines.

Phase 1: Exploratory sample	7 lab members independently reviewed a random sample of 12 CRT-related headlines and reported their findings in a group discussion.
Phase 2: Define distinct frames	4 lab members independently labeled a random sample of 10 CRT-related headlines based on labels generated in Phase 1. Labelers reported feedback in a group discussion.
Phase 3: Develop final labeling scheme	2 lab members independently coded 150 headlines and reviewed conflicts to create a final consensus labeled dataset.

Table 2: Process for developing CRT-specific issue frames inductively.

Headline Frame Detection with GPT-4

In political communications, news framing describes various ways that different news media can report on the same issue (Chong and Druckman 2007). Previous research shows that news framing of a particular issue can significantly influence public attitudes and understandings of an issue (Iyengar 1994; Sophie Lecheler and Hänggli 2015). After collecting 11,704 unique headlines about CRT, we worked to develop a scheme for identifying the news frames that are present in the CRT headlines.

We developed CRT-specific news frames in three phases, which are outlined in Table 2. This process led to frame labels along two different dimensions: primary actor and headline stance.[11] These labels are defined in Table 1.

[8] We conducted searches for 49 of 50 states, excluding Washington state due to confusion with Washington, D.C.

[9] The relevance rule matched headlines that included ("critical" and "race") or ("race" and "theory") or "CRT"

[10] NPR - National Public Radio, operates local stations across the country, which share the news coverage among them.

[11] We also developed two other frames related to the key action

Cluster description	Example headline
Cluster 17: Gaming monitors and cathode ray tube screens	"What to look for in a CRT monitor : The ultimate guide for retro gamers"
Cluster 21: Personal finance and Credit Risk Transfer	"Freddie Mac CRT Program Reports 2021 Issuance Of Nearly $20B"
Cluster 80: Cardiac Resynchronization Therapy medical device	"Leadless CRT Can Deliver Left Bundle Branch Area Pacing for Cardiac Resynchronization"
Cluster 82: CRT stock ticker	"Cross Timbers Royalty Trust (NYSE : CRT) Raises Dividend to $0.16 Per Share"

Table 3: Irrelevant headlines identified by unsupervised clustering algorithm. The use of CRT refers to acronyms for other concepts.

After inductively developing CRT issue frames for the headlines, two lab members labeled a ground truth dataset of 50 randomly-selected headlines.[12] Due to concern that the small validation dataset would not be representative of the headlines in the dataset, the two lab members labeled an additional 100 headlines. However, these 100 headlines were not chosen randomly. Instead, in an effort to select a representative sample of headlines, 100 headlines were selected with the following process:

1. Convert unique headlines in the dataset to vector representations using Sentence-BERT (SBERT). SBERT is a modification of the pretrained BERT language model, and so it produces semantically meaningful sentence embeddings (Reimers and Gurevych 2019).
2. Perform unsupervised K-means clustering on the SBERT vectors, with the number of clusters = 100.
3. Calculate the headline at the center of each cluster and add it to the validation dataset.

This process resulted in 100 headlines that were, in theory, semantically distinct and broadly representative of semantic groupings in the headline dataset. An additional use for these headline clusterings was identifying irrelevant headlines that had not been caught in the initial rule-based relevance filter. The labelers determined that four of the 100 clusters were not relevant to CRT. In total, 120 headlines assigned to these clusters were removed. A summary of the removed headlines is available in Table 3.

After the two labelers had annotated all 150 validation examples with frame labels, we calculated inter-rater reliability for the headline stance and primary actor labels. We found substantial agreement between raters for both frame labels (Landis and Koch 1977), although reliability was slightly higher for the headline stance label compared to the primary actor label (Cohen's kappa = 0.742 and 0.673, respectively). To create one unified validation dataset, the labelers met in person to resolve conflicting labels between their ratings. This resulted in a dataset of consensus labels, which could be used as a ground truth validation set.

Primary actor	% validation data	class-wise F-score
Educational practitioner	17.99	0.8148
Political influencer	50.36	0.8993
Impacted actor	8.63	0.8696
None /other	23.02	0.6923
Headline stance	**% validation data**	**class-wise F-score**
Anti-CRT	28.78	0.7200
Defending CRT	7.19	0.6923
Neutral	64.03	0.7894

Table 4: GPT-4 performance for predicting primary actor and headline stance labels. The weighted average F-score for primary actor was 0.8339. The weighted average F-score for headline stance was 0.7625.

To predict frame labels on the 11,704 headlines, we prompted GPT-4, a state-of-the-art large language model developed by OpenAI (OpenAI et al. 2023). We prompted the model using natural language prompts that were as close as possible to the prompts given to the human labelers, with definitions consistent with Table 1 above.[13] In contrast with traditional unsupervised NLP methods like topic modeling and clustering, the GPT-4 zero-shot classification method incorporates high levels of human oversight by leveraging the human-validated labeling scheme and class definitions. Additionally, in contrast with traditional supervised NLP methods like neural networks and Naive Bayes classifiers, GPT-4 does not require a large corpus of training examples to achieve high performance.

The OpenAI API sets rate limits on GPT-4 and charges users by token usage. Therefore, it was relatively expensive and time-consuming to generate the dataset. Our dataset labeling cost about $118 and took about 5 days to generate, even when requests were spread across multiple API keys.

We assessed model performance by comparing the GPT-generated results with our human-labeled consensus dataset (excluding headlines from irrelevant clusters). We found that GPT-4 performed well for predicting both headline stance and primary actor, with weighted average F-scores of 0.8339 and 0.7625 respectively. Theoretically, we could iterate on the GPT-4 prompting method to improve these F-scores. However, since the negative impact for misclassification is relatively low and the cost (in terms of both time and money) of iterative GPT-4 usage is relatively high, we determined that F-scores greater than 0.75 are acceptable. More detailed performance metrics are presented in Table 4.

taking place in the headline. However, interrater reliability and GPT-4 label performance were both lower for these categories, thus, they are not included in the dataset.

[12]CAPS lab members were paid $17.00 per hour for their work.

[13]GPT-4 prompts are available in our supplemental code materials, in the get_labels.py file.

Matching Partisan Audience Bias Scores

The dataset of CRT-relevant news article URLs was augmented with scores to indicate the partisan audience bias of each URL. Partisan audience bias (PAB) scores were assigned based on the URL domain, which indicates the host news source (for example, the URL https://www.foxnews.com/politics/texas has the domain foxnews.com). For each domain in our dataset, we matched the domain to the PAB scores developed by Robertson et al (2018).[14] The majority of the domains in our dataset (65.1%) matched with domains available in the PAB scores dataset. However, 6,530 URLs (34.9%) are missing PAB scores.

Engagement Metrics from Reddit

To augment our dataset with information that might point toward the level of interest that these articles encountered in the social web, we make use of the Pushshift Reddit dataset (2005-2022) (Baumgartner et al. 2020), which is currently archived at `archive.org`. Since our period of interest, Aug 2020-Dec 2022, falls within the timeframe in which Pushshift was operating with Reddit's approval, we were able to download the submission dumps for all 27 months. We exhaustively searched all Reddit submissions (ie. Reddit posts) and referenced the URL field against our list of article URLs. For the Reddit submission URLs that were included in our GDELT dataset, we saved the submission date, submission title, subreddit name and ID, number of subreddit subscribers, number of comments, score (number of upvotes minus number of downvotes), and upvote ratio.

We found a total of 6,715 Reddit submissions that contained a URL from our CRT news dataset. These submissions matched with 2,612 unique GDELT news article URLs, or about 14.0% of the total number of news URLs in the dataset. Matched Reddit submissions were most commonly posted in news aggregator subreddits (like r/TheNewsFeed and r/BreitbartNews) and conservative political subreddits (like r/Conservative and r/Republican). The average score of the Reddit submissions (number of upvotes minus number of downvotes) was about 67. No user-specific information is part of our analysis or dataset. Additionally, no text from the submissions or the comments is used.

Preliminary Analysis & Findings

The dataset presented in this paper could be used to answer a variety of research questions in the computational social sciences. In this section, we provide a preliminary descriptive analysis to illustrate key features of the dataset and demonstrate its potential for future research.

Most Common Headline Frames

Based on the headline frame annotations generated by GPT-4, the most common headline stance in our dataset was anti-CRT (43.06% of headlines), followed by neutral (37.36%) and defending CRT (19.57%). When we plot the headline stance occurrence over time (not shown), we find that anti-CRT stance remained the dominant class until July 2021, after which neutral and anti-CRT stances became about equally common. This suggests that CRT news headlines, especially those that reported on CRT before July 2021, tend to have language that favors an anti-CRT viewpoint.

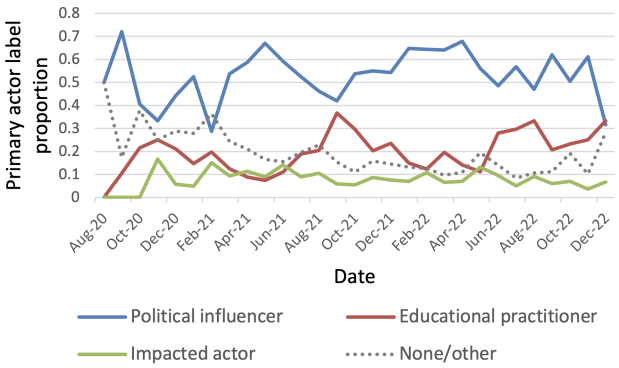

Figure 1: Primary actor labels for CRT headlines (as a proportion of total headline labels) over time. Political influencers were the most frequent primary actor across all time intervals, while impacted actors were the least frequent.

Additionally, the most common primary actor in our dataset was political influencers (56.56% of headlines), followed by educational practitioners (17.99%), none / other actor (16.29%), and impacted actors (9.15%). This suggests that the majority of CRT news headlines tend to feature actors like legislatures, governors, and political commentators. Notably, the proportion of headlines with impacted actors (ie. students and parents) as the primary actor was lowest, making up less than 10 percent of the headlines. Broadly, these trends held over time (see Figure 1). This indicates that CRT is more commonly framed as a "top-down" political issue, and is less commonly framed as a grassroots political issue, among articles in our dataset. In other words, the headline-worthy events in the anti-CRT movement do not seem to occur at a grassroots level.

This paper does not diagnose headlines that were labeled as "none / other" for the primary actor category. However, future work could attempt to characterize the headlines and primary actors in this category.

Validating our Results

Although it is difficult to establish a "ground truth" for news coverage data, we can compare our GDELT Project data with other data sources to feel more confident that our GDELT results capture actual news trends. Figure 2 charts CRT news coverage over time (from August 2020 to December 2022), measured as a proportion of overall U.S. news articles mentioning schools. The chart shows two identifiable peaks: first, the largest peak in June-July 2021, and then a smaller secondary peak in November 2021.

These peaks in CRT controversy are also reflected in similar datasets over this time period. For example, in addition to their Internet News Archive, the GDELT Project main-

[14]The Partisan Audience Bias Score dataset is available in the public domain under a CC0 1.0 License. See https://doi.org/10.7910/DVN/QAN5VX

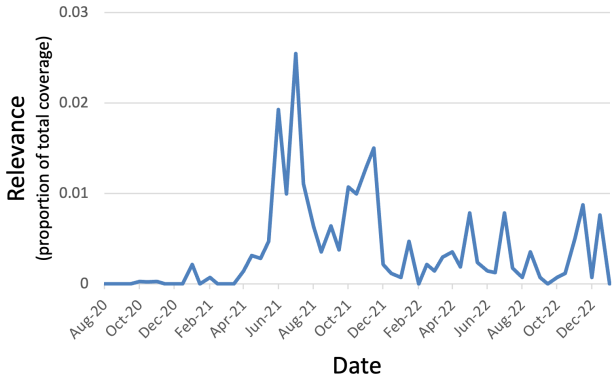

Figure 2: Headlines from our news dataset relevant to CRT (as a proportion of "school" related articles) over time. Data source: GDELT DOC 2.0 API

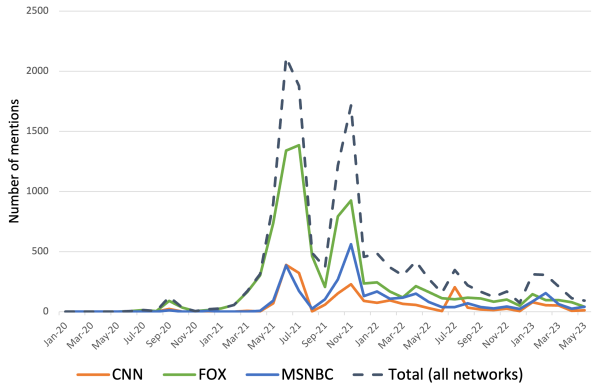

Figure 3: Number of 15-second cable news clips per month mentioning CRT, as a subset of 24-hour news clip coverage. Data source: GDELT Television News Archive, api.gdeltproject.org/api/v2/summary/summary?d=iatv

tains a TV News Archive.[15] The TV News Archive contains all 15-second clips from 24-hour coverage of major American cable news networks since 2009. Compared to the Internet News Archive, the TV News Archive is more limited in scope, but we can be more confident that the dataset reliably captures *all* coverage for a given news network in a specified time period. Figure 3 shows that the two peaks in CRT-relevant coverage for TV news match up with the peaks for internet news: it shows a large peak in June 2021 and a secondary peak in November 2021.

Finally, these two distinct peaks are also visible in Google Trends data. Google Trends analyzes Google users' web searches to determine the popularity of searches in a certain time period.[16] We looked at Google Trends data for the search term "critical race theory" in the same time period as our dataset, as shown in Figure 4. The search volume is normalized and measured on a scale from 0-100,

[15]https://api.gdeltproject.org/api/v2/summary/summary?d=iatv

[16]https://newsinitiative.withgoogle.com/resources/trainings/google-trends-understanding-the-data/

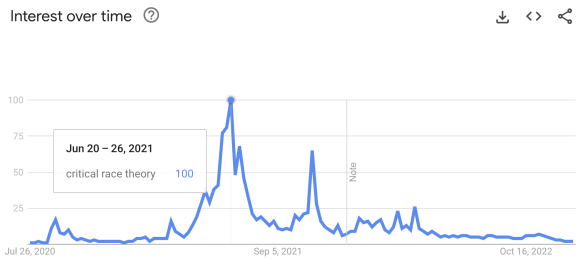

Figure 4: Google search queries for "critical race theory" (as a percentage of peak search frequency) over time. Data source: Google Trends, https://trends.google.com/trends/

where 100 represents the peak search volume. We find that the Google search data roughly follows the same trend as both our dataset (based on GDELT online news data) and television news coverage of CRT. The similarities between changes in political salience of CRT over time, as measured with different metrics and depicted through Figures 2, 3, and 4, supports the validity of our news article dataset.

Identifying Trendsetter States and State-Specific Events

One unique element of critical race theory controversy compared to other national political issues is that many CRT-related political actions took place at the state level. For example, in states all around the country, governors issued executive orders, state legislatures introduced new legislation, and local school boards passed CRT policies (Alexander et al. 2023). To get a full picture of CRT news coverage across the U.S., it is essential that researchers have access to state-specific data. Our dataset includes tags by mention of 49 states (excluding Washington) for this purpose. The following analysis demonstrates how article data tagged by state can help us explain trends in CRT news coverage over time. However, there are many more potential uses for this regional metadata. Political scientists may consider investigating the relationship between CRT online news coverage and other state variables such as state demographics, state-specific CRT legislation, and state political leadership.

To start exploring the key states whose news coverage may have shaped CRT controversies in the U.S., we can look more closely at the results from the two most striking peaks in CRT-related news in our combined national dataset. As mentioned in the previous subsection, these peaks took place around June-July 2021 and October-November 2021.

First, let's investigate the first peak, which appears in June-July 2021. To isolate CRT news coverage trends for the most salient states, we identified the states with the five highest rates of CRT-relevant news in the time period: Florida, Oklahoma, Virginia, Idaho, and Montana. After visualizing the coverage trends in these states,[17] we found that the initial shape of CRT news coverage trends in these states was

[17]Visualizations for this trend, and any other state-specific trend, can be generated using code in the Zenodo repository.

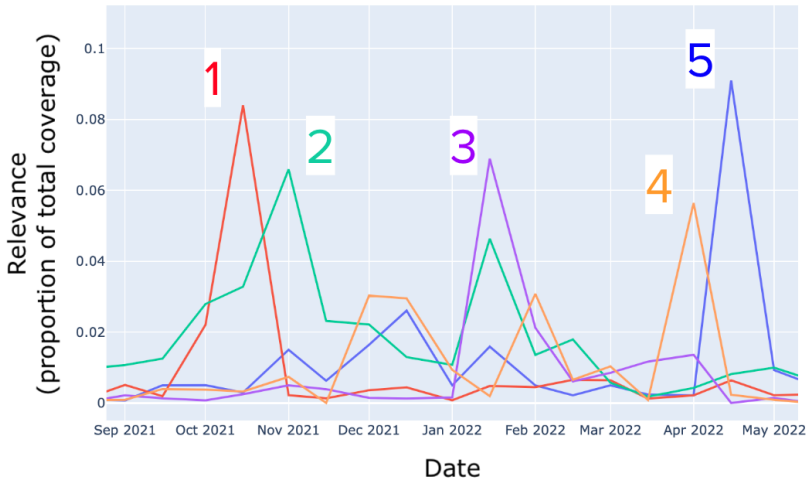

Figure 5: The second major peak in CRT news coverage (as a proportion of total school-related coverage), illustrated with five states that had the highest proportion of CRT articles in the time period. We identified events in the top five states that correspond to spikes in coverage: (1) Oklahoma ACLU files lawsuit against ban on CRT; (2) Debate over CRT played a large role in Glenn Youngkin's Virginia gubernatorial campaign; (3) Mississippi Senators walked out of the state legislature to protest CRT ban; (4) South Dakota Governor Kristi Noem issues an anti-CRT executive order; (5) Florida Department of Education rejects math textbooks over CRT concerns.

relatively unified. Notably, Idaho has the first spike starting as early as April 2021, while Florida had a higher peak relevance proportion than the other states. This is consistent with real-world events, since Idaho was one of the first states to pass a bill to ban teaching about critical race theory in classrooms, and was also one of only two states whose legislation explicitly used the term "critical race theory" (Ray and Gibbons 2021). Additionally, Florida governor Ron DeSantis' status as a nationally prominent conservative leader likely contributed to the high rates of Florida-related CRT news in this time period.

Next, let's investigate the second peak, which appears in October-November 2021. Based on the zoomed-in view of coverage by state shown in Figure 5, we notice that the state-specific peaks in this time period appear to be less unified than in the first peak. To isolate the most salient states, we identified the states with the five highest rates of CRT-relevant news in the time period: Florida, Oklahoma, Virginia, Mississippi, and South Dakota. By conducting a qualitative analysis of the headlines for each state in this time period, we can make inferences about key events that shaped the national conversation surrounding CRT. After manually reviewing headlines from the dataset, we identified events in the top five states that correspond to spikes in coverage for the given period. The results are shown in Figure 5.

Sources with Republican Audience Bias Dominated Initial CRT Coverage

The URLs in our database were collected from a variety of online news sources across 1,684 unique web domains. Based on the formulation of the partisan audience bias score metric (Robertson et al. 2018), along with other publications

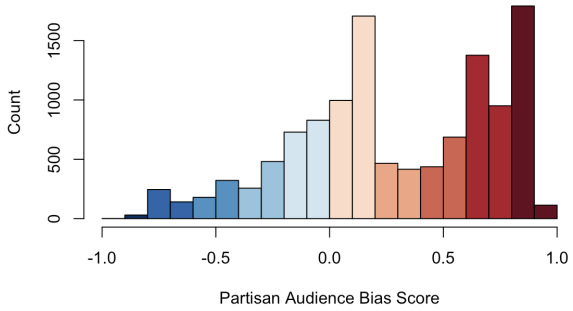

Figure 6: Distribution of partisan audience bias scores in our CRT news dataset. The distribution is highly skewed toward domains that tend to be shared by Republicans.

that use PAB scores (Kawakami, Umarova, and Mustafaraj 2020; Perreault et al. 2023), one might expect that the PAB scores for URLs in our dataset would be relatively normally distributed in the range from -1 to 1.

However, this is not what we observe. Instead, the distribution of PAB scores was highly skewed toward domains shared by Republicans (see Figure 6). Additionally, when we plot the median PAB score at bimonthly intervals (see Figure 7), the median PAB score remained positive across all intervals. Even more notably, the median PAB score was much higher in the first year of the date range (September 2020 through September 2021) compared to later years. This suggests that a majority of the URLs in our dataset come from sources that are most likely to be shared by Republi-

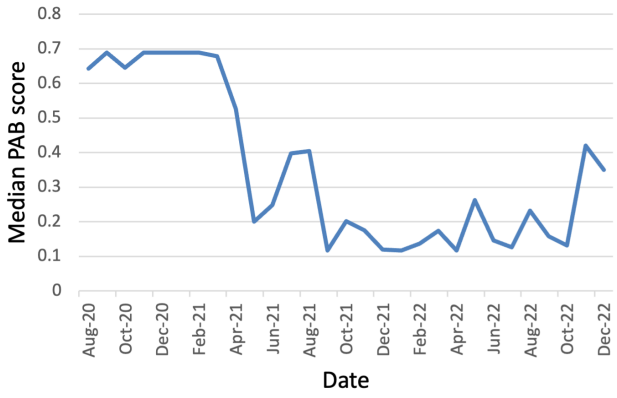

Figure 7: Median partisan audience bias (PAB) score over time. Median scores remained positive across all bimonthly intervals, and was highest in the first year of available data.

cans, and that Republican domains were especially overrepresented in the first year of news coverage of CRT. This is consistent with our finding that language with an anti-CRT stance was most common in headlines published before July 2021. Additionally, it aligns with our previous understandings about the emergence of CRT controversy, which was largely driven by conservative political figures.

Limitations & Ethical Considerations

Previous publications have explored limitations of the GDELT repository, including problems related to GDELT's over-collection of non-news documents, lack of methods for pruning false positive articles, and difficulties matching up GDELT news archives with other similar datasets to demonstrate validity (Ward et al. 2013; Kwak and An 2016). Researchers who hope to use this dataset should carefully consider their use case and its tolerance for false positives and missed news articles. Users should also note that offensive content could exist in the GDELT news headline data. Since our dataset specifically targets articles about CRT, a political issue that plays on longstanding racial tensions in the U.S., racially discriminatory language may appear in the data.

Additionally, many headlines in our dataset are missing partisan audience bias scores, or do not match any Reddit posts from the Pushshift Reddit dataset. In particular, the small number of Reddit posts that matched with headline URLs may limit future research that aims to use Reddit data. We did not conduct a thorough review of the 6,530 URLs (34.9%) with missing PAB scores, so future analysis should consider the possibility that the PAB scores are missing not at random. Since the original PAB scores dataset contains 19,022 domains collected from websites shared on Twitter, we predict that the domains without PAB scores are from more obscure sources that are less commonly shared online.

This dataset does not contain any personally identifiable information and primarily relies on publicly accessible news data that was produced with the intention of its publication on the web. Therefore, we do not foresee major misuses or negative societal impacts of this work. To abide by FAIR Data Principles, we have published a Zenodo repository that includes our dataset, code, and supplemental data required to reproduce our results. The repository is Findable using the DOI URL https://doi.org/10.5281/zenodo.10516191, and it includes documentation and file types that are intended to support the Accessible, Interoperable, and Re-usable principles of FAIRness.

One limitation in our dataset's FAIRness is that the protocol for reproducing GPT-4 annotations is not completely Accessible. Since GPT-4 is a closed-source model managed by OpenAI, we have some concern about the explainability and reproducibility of our GPT-4 generated results. Previous research indicates that GPT-3 has systematic left-leaning political bias, which may also hold for GPT-4 and influence the annotation results (Motoki, Pinho Neto, and Rodrigues 2024). Additionally, reproducibility is not accessible to those who do not have the time or budget to make requests to GPT-4 with the OpenAI API pricing and rate limits. Although this limitation is a detriment to FAIRness, this also demonstrates the value of this dataset, which allows redistribution and re-use of the results from our GPT requests.

Conclusion

This paper presents a novel dataset of annotated U.S. news headlines related to critical race theory that were published online from August 2020 through December 2022. In addition to news headlines, URLs, and tags by state gathered from the GDELT Project database, the dataset includes news frame annotations generated by GPT-4. These news frame annotations identify the headline stance (anti-CRT, defending CRT, or neutral) and primary actor (political influencer, educational practitioner, impacted actor, or none / other actor) in each headline. The dataset is further augmented with partisan audience bias scores to indicate the Democratic or Republican audience bias of the news article URL domain. Reddit engagement data was also added for URLs that appear in Reddit posts from Pushshift dumps. From a preliminary analysis of our dataset, we found that the timeline chart of CRT-relevant coverage in our dataset generally followed similar trends as CRT coverage data from cable TV networks and Google Trends data. Overall, we found that the partisan audience bias scores of the URLs leaned Republican, and that the most common news headline frame annotations were an anti-CRT stance and had a political primary actor. We hope that this dataset can be used for future research contributions in political science, computational social science, and natural language processing.

Acknowledgments

We are grateful for feedback and support from student members of the Computational Analysis for Political Science (CAPS) Lab at Wellesley College, especially CAPS members Tarishi Gupta and Ariel McGee for their labeling work. This research was partially supported by the Jerome A. Schiff Fellowship and startup research funds awarded by Wellesley College.

References

Alexander, M. 2010. The Color of Justice. *The new Jim crow: Mass incarceration in the age of colorblindness*, 59–62.

Alexander, T. N.; Clark, L. B.; Flores-Ganley, I.; Harris, C.; Kohli, J.; McLelland, L.; Moody, P.; Powell, N.; Reinhard, K.; Smith, M.; and Zatz, N. 2023. CRT Forward Tracking Project.

Arva, B.; Beieler, J.; Fisher, B.; Lara, G.; Schrodt, P. A.; Song, W.; Sowell, M.; and Stehle, S. 2013. Improving Forecasts of International Events of Interest.

Baumgartner, J.; Zannettou, S.; Keegan, B.; Squire, M.; and Blackburn, J. 2020. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, 830–839.

Chong, D.; and Druckman, J. N. 2007. Framing Theory. *Annual Review of Political Science*, 10(1): 103–126.

Crenshaw, K.; Gotanda, N.; Peller, G.; and Thomas, K. 1995. *Critical race theory: The key writings that formed the movement*. The New Press.

Field, A.; Kliger, D.; Wintner, S.; Pan, J.; Jurafsky, D.; and Tsvetkov, Y. 2018. Framing and Agenda-setting in Russian News: a Computational Analysis of Intricate Political Strategies. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 3570–3580. Brussels, Belgium: Association for Computational Linguistics.

Haney-López, I. 2014. *Dog whistle politics: How coded racial appeals have reinvented racism and wrecked the middle class*. Oxford University Press.

Iyengar, S. 1994. *Is anyone responsible?: How television frames political issues*. University of Chicago Press.

Kawakami, A.; Umarova, K.; and Mustafaraj, E. 2020. The Media Coverage of the 2020 US Presidential Election Candidates through the Lens of Google's Top Stories. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, 868–877.

Kwak, H.; and An, J. 2016. Two Tales of the World: Comparison of Widely Used World News Datasets GDELT and EventRegistry. *Proceedings of the International AAAI Conference on Web and Social Media*, 10(1): 619–622. Number: 1.

Kwak, H.; An, J.; and Ahn, Y.-Y. 2020. A Systematic Media Frame Analysis of 1.5 Million New York Times Articles from 2000 to 2017. In *Proceedings of the 12th ACM Conference on Web Science*, WebSci '20, 305–314. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-7989-2.

Landis, J. R.; and Koch, G. G. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1): 159–174.

Leetaru, K.; and Schrodt, P. A. 2013. GDELT: Global data on events, location, and tone, 1979–2012. In *ISA annual convention*, volume 2, 1–49. Citeseer.

Mendelberg, T. 2001. *The race card: Campaign strategy, implicit messages, and the norm of equality*. Princeton University Press.

Motoki, F.; Pinho Neto, V.; and Rodrigues, V. 2024. More human than human: Measuring ChatGPT political bias. *Public Choice*, 198(1): 3–23.

OpenAI; Achiam, . J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; (...); and Zoph, B. 2023. GPT-4 Technical Report. arXiv:2303.08774.

Perreault, B.; Dau, L.; Wintner, A.; and Mustafaraj, E. 2023. Capturing the Aftermath of the Dobbs v. Jackson Women's Health Organization Decision in Google Search Results across the US. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, 1063–1072.

Pollock, M.; Rogers, J.; Kwako, A.; Matschiner, A.; Kendall, R.; Bingener, C.; Reece, E.; Kennedy, B.; and Howard, J. 2022. The Conflict Campaign: Exploring Local Experiences of the Campaign to Ban "Critical Race Theory" in Public K–12 Education in the U.S., 2020–2021.

Ray, R.; and Gibbons, A. 2021. Why are states banning critical race theory?

Reimers, N.; and Gurevych, I. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Robertson, R. E.; Jiang, S.; Joseph, K.; Friedland, L.; Lazer, D.; and Wilson, C. 2018. Auditing Partisan Audience Bias within Google Search. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW).

Siegel, R. 1997. Why equal protection no longer protects: The evolving forms of status-enforcing state action. *Stanford law review*, 1111–1148.

Sophie Lecheler, A. R. S., Mario Keer; and Hänggli, R. 2015. The Effects of Repetitive News Framing on Political Opinions over Time. *Communication Monographs*, 82(3): 339–358.

Ward, M. D.; Beger, A.; Cutler, J.; Dickenson, M.; Dorff, C.; and Radford, B. 2013. Comparing GDELT and ICEWS event data. *Analysis*, 21(1): 267–297.

Xiao, Z.; Yuan, X.; Liao, Q. V.; Abdelghani, R.; and Oudeyer, P.-Y. 2023. Supporting Qualitative Analysis with Large Language Models: Combining Codebook with GPT-3 for Deductive Coding. In *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces*, IUI '23 Companion, 75–78. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701078.

Zhang, H.; Wu, C.; Xie, J.; Kim, C.; and Carroll, J. M. 2023. QualiGPT: GPT as an easy-to-use tool for qualitative coding. arXiv:2310.07061.

Paper Ethics Checklist

1. For most authors...

(a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes, this dataset primarily involves online news articles that were intended for publication on the internet. Data was augmented with other publicly accessible datasets and output generated from GPT-4, and does not contain any personal data. The research includes some Reddit data but does not include any identifiers for individual users. See Limitations and Ethical Considerations section.

(b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes, the abstract and introduction focus on description of the dataset, collection methods, and potential for future use, not substantial experimental findings.

(c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, see "Methods for Data Collection" and "Motivation and Potential Uses"

(d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? NA

(e) Did you describe the limitations of your work? Yes, see Limitations section.

(f) Did you discuss any potential negative societal impacts of your work? Yes, see Limitations section.

(g) Did you discuss any potential misuse of your work? Yes, see Limitations section.

(h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, the Zenodo repository includes data and code for documentation and reproducibility.

(i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes

2. Additionally, if your study involves hypotheses testing...

(a) Did you clearly state the assumptions underlying all theoretical results? NA

(b) Have you provided justifications for all theoretical results? NA

(c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA

(d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA

(e) Did you address potential biases or limitations in your theoretical framework? NA

(f) Have you related your theoretical results to the existing literature in social science? NA

(g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA

3. Additionally, if you are including theoretical proofs...

(a) Did you state the full set of assumptions of all theoretical results? NA

(b) Did you include complete proofs of all theoretical results? NA

4. Additionally, if you ran machine learning experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes, a URL to the Zenodo dataset with replication data and code is available as a footnote on the first page.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes, all scripts are included in the replication data and code repository: https://zenodo.org/doi/10.5281/zenodo.10516190

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? No, our procedure queried a large language model with temperature set to zero, which leads to deterministic output. Therefore, error bars are not necessary.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes, in the methods section

(e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes, in the methods section

(f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Yes, in the methods section

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...

(a) If your work uses existing assets, did you cite the creators? Yes, cited creators and/or data sources for GDELT Project data, partisan audience bias scores, and Reddit data

(b) Did you mention the license of the assets? Yes, for GDELT data and partisan audience bias scores

(c) Did you include any new assets in the supplemental material or as a URL? Yes, it is available in the data and code replication Zenodo repository

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? No, the dataset does not contain any personal data

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes, in the Methods section for Reddit metrics and in the Limitations section

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? Yes, in Limitations section

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? No. This dataset paper, along with the Zenodo repository with replication code and data, addresses the main motivations, uses, ethical concerns, and limitations of the dataset.

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
   (a) Did you include the full text of instructions given to participants and screenshots? Yes, these materials are provided in the replication data and code Zenodo repository.
   (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? Yes, in the methods section.
   (d) Did you discuss how data is stored, shared, and deidentified? Data is stored and shared in the replication data and code Zenodo repository. Deidentification was not necessary.