Auditing Google's Search Headlines as a Potential Gateway to Misleading Content: Evidence from the 2020 US Election

Himanshu Zade[1], Morgan Wack[1], Yuanrui Zhang, Kate Starbird, Ryan Calo, Jason Young and Jevin D. West

Abstract. The prevalence and spread of online misinformation during the 2020 U.S. Election served to perpetuate a false belief in widespread fraud. Though much research has focused on how social media platforms connected people to related rumors and conspiracies, less is known about the search engine pathways that linked users to news content with the potential to undermine trust in elections. In this paper, we present novel data related to the content of political headlines during the 2020 U.S. Election period. We scraped over 800k headlines from Google's search engine results pages (SERP) in response to 20 election-related keywords — 10 general (*e.g.*, 'Ballots') and 10 conspiratorial (*e.g.*, 'Voter fraud') — when searched from 20 cities across 18 states. We present results from qualitative coding of 5,600 headlines focused on the prevalence of delegitimizing information. Our results reveal videos (compared to stories, search results and advertisements) to be the most problematic in terms of exposing users to delegitimizing headlines. We also illustrate how headline-content varies when searching from a swing state, adopting a conspiratorial search keyword, or reading from media domains with higher political bias. We conclude with policy recommendations on data transparency that allow researchers to continue to monitor search engines during elections.

1 Introduction

Despite no evidence that widespread fraud occurred during the recent US elections (Cybersecurity & infrastructure security agency 2021; Saranac Hale Spencer 2020; Cybersecurity & infrastructure security agency 2022), as reiterated in testimony by former Attorney General Bill Barr (Thompson, Cheney, and Zoe 2022), there remains skepticism among the public about the legitimacy of the election results. Following the election nearly 65% of the Republican voters believed that the results of the 2020 U.S. General

1. These authors led and contributed equally to this research.

doi:xx.xxxx/xxxxxxx xxxx

Election were illegitimate (Pennycook and Rand 2021). During the 2018 midterm elections, voters who cast their votes using mail-in ballots were skeptical as to whether their votes would be counted correctly (Alvarez, Cao, and Li 2021). Though considerable effort has been spent studying how social media platforms served to connect people to conspiracies, rumors, and misinformation related to unsubstantiated voter fraud, less is known about how and what kind of political content is spread through search engines.

Search engines are the doors to information and news on the internet. In 2020, 65% of Americans used search engines as a primary source to gather news and information (Shearer 2021), of which Google has a global market share of over 90% (StatCounter 2021). As evidenced by 'election results' and 'coronavirus' constituting the top two search terms on Google in 2020 [2], search engines have a tremendous potential to provide access to critical information that can influence democratic discourse. This is particularly true during election periods — in particular, the 2020 U.S. General Election — when political polarization, COVID uncertainty, and demand for election information were all high (Kapferer 1987; Bordia and DiFonzo 2017; Starbird, Spiro, and Koltai 2020).

The 2020 U.S. election gave rise to several narratives that cast doubt on the legitimacy of the results. Several official organizations, including the *Cybersecurity and Infrastructure Security Agency*, have debunked these narratives and confirmed in December, 2020 that it was indeed a 'secure election' (Cybersecurity & infrastructure security agency 2021, 2022; Saranac Hale Spencer 2020). Despite the acknowledgment of confidence in the election by several Government officials and elected leaders, both Democratic and Republican (Brennan Center for Justice 2020), unproven and misleading election-related narratives were (and some remain) widely available online. The goal of this paper is to investigate *whether and potentially how Google served as a gateway to content that may have undermined trust in election processes, institutions, and results*. We conducted an audit of headlines appearing in Google's SERPs in response to several search terms before, during, and after the 2020 U.S. Election. Specifically, our research was guided by the following questions:

- *Question One:* How do the SERP verticals — search results, stories, videos and advertisements — differ in the amount of misleading content?
- *Question Two:* How does one's location in a specific city — split by population and party representation — change the kind of election content found in search results?
- *Question Three:* Do different search terms lead to different search result quality?
- *Question Four:* Which online news domains served as the most frequent gateways to content that may have undermined trust during the election period?

To answer these questions, we focused on news headlines from Google's SERP data (Figure 1). The headline of a news story is known to influence user interpretation of the story's content (Tannenbaum 1953) and impact its popularity (Rieis et al. 2015). We collected headlines using election-related search keywords as seen on Google's search engine across 20 locations spread throughout the U.S. Since Google does not officially support a search API, and other services do not support location-specific requests, we resorted to a third party paid service called SerpAPI (SerpApi 2020). This service allowed us to perform searches so that the results were associated with the locations of our 20 selected sites, rather than the results that Google would normally associate with the geographic location of our local IP address. Our collection of the data commenced prior to the election in early October and ran through mid-December 2020. We performed an extensive qualitative analysis of a random sample of 5,600 headlines from over 500k SERP search results, 242k SERP stories, 62k SERP videos, and 47k SERP advertisements

2. https://trends.google.com/trends/yis/2020/US/

to evaluate the potential of SERP data to undermine trust in the election. In addition to the analysis, we make the raw Google SERP data corresponding to election-related keywords across several disparate U.S. locations openly available to further analysis by other researchers [3].

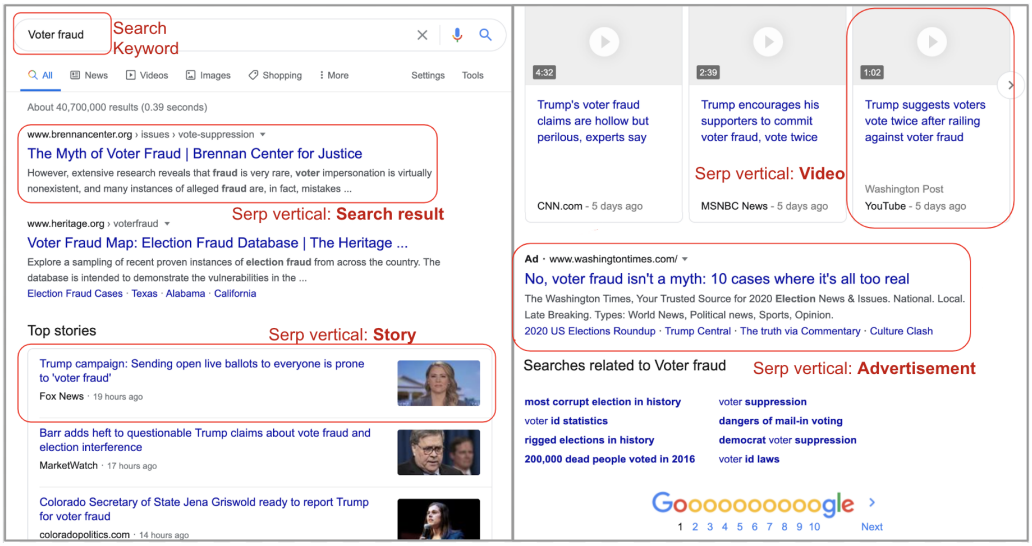

Figure 1: Two example screenshots of Google SERP data if a user were to search for 'voter fraud'. We collected the headlines and metadata for both the search results and all three top stories pictured here on the left. In addition, we also collected the headlines and metadata for all the three videos and the only advertisement pictured here on the right. Overall, we collected first ten search results, top ten stories, top ten videos and all the advertisements returned by the search engine in response to all keywords.

From these searches and our subsequent coding of reported headlines, we found that the headlines of the video content reported in our Google search engine homepage contained a disproportionate amount of undermining-trust content when compared to alternative SERP-verticals (search results, stories, and advertisements). Although swing states received more campaign advertisements than non-swing states, user's location generally did not moderate the quality of information served by search engine headlines. We also found that the headlines displayed on the homepage were more likely to undermine trust when searches included conspiratorial election-related terms (*e.g.*, 'Voter fraud', 'Rigged election', etc.) as opposed to general election-related terms (*e.g.*, 'Ballots', 'Where do I vote', etc.), as well as if the headlines were associated with media domains with a relatively more right-leaning bias. Upon investigating the mainstream media headlines specifically, we found that legacy news sites with large audiences like CNN and Fox News played an outsized role in delivering content with the potential to undermine trust. Finally, we present the topics that were the focus of trust-undermining and trust-imparting content across our coded sample.

Our work builds upon previous work that has emphasized the influence of online electoral content in altering perceptions about the legitimacy of the 2020 U.S. General Election. First, we present a novel dataset consisting of geographically and topically distinct search results presented by Google prior to, during, and following, November 3rd, 2020. Second, we developed a coding scheme for assessing the content of headlines in SERP data and its role in undermining trust that can serve as a template for future studies. Third, using our coding scheme, we analyze the political content likely presented to a large number of users on Google's platform before, during, and shortly after the election.

3. Data is available on the Open Source Foundation.

From this analysis, we identify the topics, domains, and search patterns of election-delegitimizing content. We conclude with recommendations focused on open, auditable, and anonymized data for investigating these research questions in future elections.

2 Background: 2020 Election Delegitimization

2.1 Significance of news headlines

News headlines, along with other forms of content collected in Google search results, play a critical role in conveying information and creating impressions. For news headlines specifically, a survey conducted nearly a century ago found that out of 375 people, 192 based their opinions about news from the reading and skimming of the headlines only (Emig 1928). The importance of headlines in the conveyance of content has persisted as news has shifted online. Psychologists have known that early impressions matter and that early biases affect users in what they learn in further impressions of the artifact (Digirolamo and Hintzman 1997). Based on analysis of about 70k headlines, Reis et. al confirmed that the sentiment of the headline could have a serious impact on how popular the story might become and the kind of discourse it encourages (Rieis et al. 2015). These projects have reiterated how headlines can serve as influential shortcuts for readers that can subsequently guide their interpretation of the news (Tannenbaum 1953).

Misleading content, even if misleading only to a small extent, can bias interpretation of events, such as elections. This is why they are often used to frame real world events in a particular light (Jamieson, Hardy, and Romer 2007; Liu et al. 2019). Framing strategies have often been employed—as was tracked during the 2004 Canadian Federal election—to select aspects of particular news stories that increase the salience of the writer or news source's chosen perspective (Andrew 2007). By inducing bias amongst readers, exposure to misleading headlines can limit the capacity of its audience to process corrected information, thereby impacting their memory and reasoning (Ecker et al. 2014). Complicating matters further, readers have a tendency to over-weight headlines that are consistent with their social and political attitudes (Beam 2014) while choosing to focus on headlines that they perceive to be true a priori (Edgerly et al. 2020), leaving readers vulnerable to misleading headlines that align with partisan values. The challenge posed by misleading headlines has been exacerbated by growing use of social media platforms, where headlines are often prominently displayed as a substitute for the actual content of the article (Gabielkov et al. 2016). In fact, there is little incentive for platforms to push users off the platform to the actual article. Despite these growing concerns, little is known about the role of the content of headlines appearing in different SERP verticals (*e.g.*, stories vs videos) during elections to undermine voter trust. This is the focus of our research.

2.2 Role of Google search in shaping user opinion

Google search is the most commonly used search engine (StatCounter 2021) and therefore the focus of numerous studies into search engine function and performance. A recent study found that Google fares better in limiting the promotion of conspiratorial results as well as the presentation of links to conspiracy-dedicated websites when compared with other search engines like Bing, DuckDuckGo, Yahoo and Yandex (Urman et al. 2021). Despite relatively higher resilience to conspiratorial content, concerns remain regarding bias evident in Google search results (Robertson et al. 2018). These potential biases are of concern to election integrity advocates, who have shown that

Google search engine has in the past privileged certain topics in its news-homepage (including a disproportionate presentation of articles detailing the 2016 Trump campaign over his challengers) (Diakopoulos et al. 2018)[4].

When investigating Google's role in shaping user attention to the news, Trielli et. al found a small skew towards the political left in Google search results (Trielli and Diakopoulos 2019). Although the diversity of the media sources varied by topic, a small fraction of the media contributed about 50% of the overall suggestions in the top stories. Similarly, recent research has found that small number of sources contributed majority of the stories about 2020 U.S. Presidential election coverage on the Google SERPs (Kawakami, Umarova, and Mustafaraj 2020). Epstein et al. (Epstein and Robertson 2015) showed that a search engine manipulation effect (SEME) — *i.e.*, influencing user behavior through manipulation of search results by search engine providers — can impact the outcomes of elections. Voting preferences can be strongly influenced in favor of a candidate (20% or more) by showing biased search results biased towards a particular candidate (Epstein and Robertson 2015; Spenkuch and Toniatti 2016).

Search engines can impact user perception about credibility of the news not only through the selection of stories (and sources) on the homepage, but also through the rankings in which these stories appear. A higher position in the ranking of a (SERP vertical) story impacts user decisions more, even if it is less relevant to the topic of the user's search, than another story that appears at a lower rank (Pan et al. 2007). Researchers have also questioned the role played in the presentation of content across different information modalities including text, stories, videos across several platforms. Though recent work has suggested that video content may not be as persuasive as was once feared, users tend to believe in a video more easily than in text (Wittenberg et al. 2021). Given the increased prevalence of video-based misinformation, there is a shared belief amongst researchers that the real extent of persuasiveness of videos might diverge in real settings that are not lab-controlled. For example, when comparing the role of text versus video modality within messaging apps, researchers found that users process videos superficially and tend to more influenced by it compared to text (Sundar, Molina, and Cho 2021). Based on this result, we compare the different SERP verticals — *e.g.*, news, stories, search results and ads — in our study.

2.3 Auditing as a method to trace mis/disinformation

Algorithms of platforms like Twitter and Reddit facilitate amplification of problematic content by bringing more attention of the users to problematic content (Fernández, Bellogín, and Cantador 2021; Shepherd 2020). Researchers have employed auditing mechanisms to investigate such role of algorithms. Audits have shown how YouTube deploys algorithms with the potential to lure people down conspiracy 'rabbit holes' by continuously suggesting related content (Rodriguez 2018; Albright 2018; Hussein, Juneja, and Mitra 2020). Auditing techniques have found that even e-commerce platforms like Amazon can promote a filter bubble effect, where users who browsed anti-vaccination content on the the platform received relatively more suggestions promoting similar content than those who did not (Juneja and Mitra 2021).

Researchers have expressed hope in the use of auditing method to witness and understand why some of the unwanted platform behaviors occur (Simko et al. 2021). For example, prior investigation focused on understanding Google search engine's behavior has shown that searching for specific queries that have limited authoritative information (*i.e.*, data voids) can lead to easy discoverability of conspiratorial websites (Bradshaw

4. Diakopoulos et. al found that during the 2016 U.S. elections, Google News had 941 indexed articles about Trump, 710 about Clinton, and 630 about Sanders (Diakopoulos et al. 2018)

2019). In order to expand our understanding of how search engines can lead users to misleading content, we conduct an audit of Google SERP data focused on election content during the 2020 electoral period.

3 Data Collection

Search terms: To conduct our analysis, we generated a list of election-related search terms in October of 2020 (see Table 1). These terms were used to assess differences in headlines related to different SERP verticals: search results, news, advertisements, and videos. We split our terms into two distinct categories. The first category of search terms aimed to capture the results produced when searching for general election-related content. This included terms such as "presidential election" as well as common election questions such as "where do I vote". We also included a second category of terms targeting electoral conspiracies identified across existing misinformation narratives. This list was designed to mimic potential searches focused on issues related to the legitimacy of election processes and results. As the list was developed in advance of the election in September, it was informed by prior political controversies and online rumors and does not include terms related to conspiracies such as *Sharpiegate*, which only became relevant after election day[5]. As such, it was comprised of both general conspiratorial phrases such as "election fraud" and "stolen election" as well as more specific actions such as "voter fraud" and "ballot dumping".

General Terms	Conspiratorial Terms
Election results	Rigged election
Ballots	Late ballots
How do I vote	Voter fraud
Where do I vote	Voter intimidation
Mail-in voting	Election fraud
My ballot	Electoral fraud
Absentee ballot	Stolen election
Presidential election	Ballot harvesting
Vote by post	Ballot dumping
Vote	Mail dumping

Table 1: Election-related search terms fed into Google's search engine. Ten of the search terms were general election terms, and the other ten terms were terms linked to conspiracies related to the 2020 U.S. Presidential Election.

Search locations: Google customizes its search results based on geographic location (Rogers 2013). The results of a search for the terms 'election results' in Los Angeles, for example, could be different than the results of the same search in Topeka, Kansas. These differences can, in turn, shape geographic differences in how individuals think about and behave, since search results can both prime audiences to think about certain issues and frame how they think about those issues (Zook and Graham 2007). However, the exact relationship between search customization and local understandings of emerging news events remains understudied (Ballatore, Graham, and Sen 2017). To contribute in this area, we developed a purposive sampling approach to collect search results across

5. The 2020 *Sharpiegate* conspiracy, which claimed that sharpies were being deliberately distributed to Republican voters in order to invalidate their votes, is distinct from the prior controversy related to Donald Trump's use of a sharpie on a weather map displaying the trajectory of Hurricane Dorian in 2019.

locations in the US that varied by region and degree of urbanization. Social scientists have long explored how shared economies and cultural traditions produce regional socio-political identities, and urban-rural divides have emerged as an even more salient variable in shaping current partisan politics in the US (Gimpel et al. 2020).

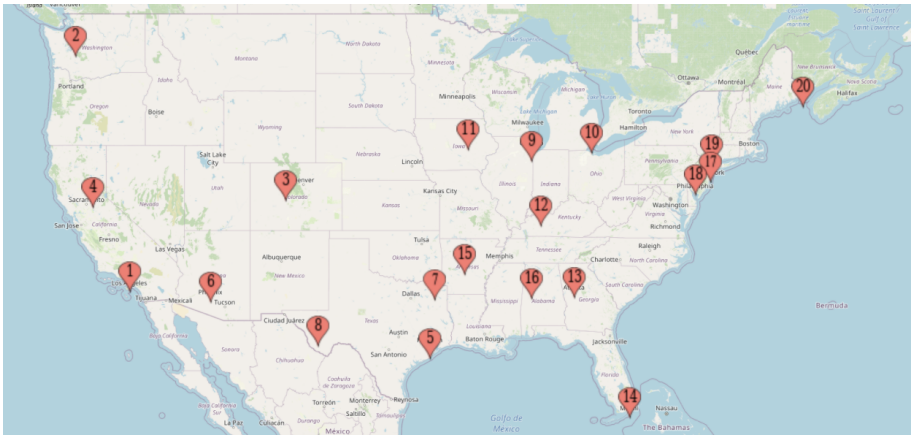

Figure 2: Geographic spread of the 20 cities across which we scraped the Google search results for search terms enlisted in Table 1.

MapID	City, State	Size	Swing	MapID	City, State	Size	Swing
1	Los Angeles, CA	UA	N	11	Cedar Falls, IA	UC	N
2	Seattle, WA	UA	N	12	Santa Claus, IN	RA	N
3	Vail, CO	UC	N	13	Atlanta, GA	UA	Y
4	Grass Valley, OR	RA	N	14	Miami, FL	UA	Y
5	Houston, TX	UA	N	15	Morrilton, AR	UC	N
6	Phoenix, AZ	UA	Y	16	Berry, AL	RA	N
7	Clarksville, TX	UC	N	17	New York, NY	UA	N
8	Fort Davis, TX	RA	N	18	Philadelphia, PA	UA	Y
9	Chicago, IL	UA	N	19	Poughkeepsie, NY	UC	N
10	Detroit, MI	UA	Y	20	Eastport, ME	RA	Y

Table 2: Our data collection includes Google SERP data as rendered in these 20 cities spread across 18 states in the USA. *UA* refers to urban areas, *UC* refers to urban clusters, and *RA* refers to rural areas. *Y* or *N* refers to whether it was a swing state, or not.

To select locations, we first divided the U.S. into Northeast, Southeast, Southwest, Midwest, and West Regions, drawing on a common five region classification schema (National Geographic 2009). Within each of those regions, the project identified four locations that represented varying levels of urbanization. Here we used the U.S. Census Bureau's classification of locations as urbanized areas (UAs) of 50,000 or more people; urban clusters (UCs) with populations between 50,000 and 2,500; and rural locations that have a population under 2,500 (US Census Bureau 2010). For each region we chose two UAs, one UC, and one rural location. We chose to over-represent UAs because there tends to be more election-related news and activity related to more densely populated locations, allowing us to better examine regional differences across these larger markets. However, we also attempted to select UAs within each region such that they also varied in size, with one containing a population of several million and the other a population close to

one million. The researchers selected specific locations within this framework, based upon their knowledge of interesting news having emerged from those locations as shown in Figure 2. Our hope was that this would produce a richer dataset. We also strove to select locations that were diverse in regards to partisan political orientation. In many instances, our first choice rural locations were not found within the API that we were using for data collection. In these instances, we chose a nearby city that could be found within the API. This process resulted in 20 locations, as enlisted in Table 2.

Search service: Google does not officially support any search API and other search-services do not allow easy access to location-specific SERP data. While we had access to a white-listed IP address to crawl unlimited Google SERP data, this data would have reflected SERP results as seen from that specific location. To accommodate location as a factor in SERP-related audits, prior research resorted either to using browser-based plugins (Robertson et al. 2018) (limiting the data collection to queries adopted by select users at specific times), or to making data requests from multiple locations with unique IP-addresses (Mustafaraj, Lurie, and Devine 2020) (limiting the scalability to only a couple of unique locations). To overcome these limitations, we used the *SerpApi* platform (SerpApi 2020) to search for keywords of our choice mentioned in Table 1 at regular intervals each day and fetched the corresponding Google search results as it would be seen at the twenty unique locations 2.

SerpApi is a real-time API to scrape Google SERP data with an option to choose a specific search location (out of the available choices) without any adjustments based on the location of researcher's IP-address. We did a preliminary check for generic keywords like "School", "Cafe" and "Museum" to confirm the location specific differences of the SERP data returned by the API. We were able to observe similar variations even for the election-related search keywords that we used in this research. For example, searching for the keyword "Vote" at the same time and day returned a headline "How to vote the new way in L.A. (in 2020)" when we specified *Los Angeles, California* as the location, but returned "How to Vote In Colorado" when we specified *Vail, Colorado* as the location. We used the paid version of the API to make about fifty thousand unique searches per day. We make this data openly available for future research projects[6].

Search schedule: We intended to scrape the SERP data several times a day to capture news headlines soon after they are released by different media sources. As we begin the collection, we collected data four times everyday (3:00, 9:00, 15:00, 21:00 EST) between October 5th and October 29th, 2020. Later, we reduced this frequency to three times a day (00:00, 08:00, 16:00 EST) from October 30th to December 3rd, 2020 to fit within the constraint of fifty thousand allowed searches based on our service subscription and required searches for a related research project. Even with the reduced daily frequency, we were able to capture news headlines in the morning, evening and late night (EST) as intended.

Overall collection: For every search, we collected first ten search results, top ten news stories, top ten videos and all the advertisements returned by the search engine in response to a search keyword, which is more than the information rendered on the Google SERPs as seen by the user and illustrated in Figure 1. It included the headlines of all the components and corresponding attributes like website-link, domains, date and time of publishing (for videos and stories) etc. as seen on the Google search engine. Overall, our initial collection consisted of 56,763 unique location specific keyword searches. Across these searches, we collected about 47k advertisements, 500k search results, 240k stories and 66k videos.

Given that higher ranked results are known to influence user decisions (Joachims et

6. We have made the data available on the Open Source Foundation.

al. 2007; Brooks 2004; Lorigo et al. 2008), we decided to focus on the top five search results, top three news stories, top three videos, and all included advertisements. Focusing on the higher ranked results across varying SERP verticals — comprising of 485,805 results — allowed us to inspect headlines that were more influential towards impacting user opinions. These contained 47k advertisements (same as before since we always considered all the advertisements), 283k search results, 242k stories and 36k videos.

For each combination of *search keyword and search location*, we now had either three or four SERPs per day depending on the frequency of collection during that time. For each of that combination, we then randomly selected one SERP per day to make the data sample size more manageable and ensure even distribution of headlines across the duration of 2 months. This reduced our sample to 174,511 total headlines including about 14k advertisements, 97k search results, 41k stories and 20k videos. Table 3 summarizes the steps that we took to filter the sample of headlines.

Step#	Procedure	Resultant data sample
Step 1	Collected SERPs (10 search results, 10 stories, 10 videos and all ads) for 20 search keywords (Table 1) as seen at 20 locations (Table 2) several times a day using *SerpAPI*.	56,763 unique location specific SERPs; About 47k ads, 500k search results, 240k stories and 66k videos.
Step 2	To focus on data that easily appears on SERPs without any extra user-clicks, we selected the top 5 search results, top 3 stories, top 3 videos, and all ads.	About 47k ads, 283k search results, 242k stories and 36k videos.
Step 3	For each combination of *search keyword and search location*, we randomly chose exactly one SERP per day.	About 14k ads, 97k search results, 41k stories and 20k videos. Summary statistics in Table 4.
Step 4	Using stratified random sampling technique, we split the Oct-Dec 2020 duration into four 2-week long periods and selected 50 SERP headlines per combination of location type (2 urban areas, 1 urban cluster, 1 rural area), SERP vertical type (result, stories, videos, ads) and search term type (general, conspiratorial).	1,600 stories, 1,600 videos and 1,600 searches across 4 time periods and 800 ads across the first 2 time periods; out of the 5,600 SERP headlines (as per power analysis), we qualitatively coded 2,438 unique ones.

Table 3: Step-wise illustration of how we sampled the headlines in our SERP data to make it suitable for qualitative coding.

Filtered collection for qualitative coding: After we collected the data, we assigned a label and coded each headline into different categories. Although same headline could appear multiple times in our data—*e.g.*, the headline "Voter Fraud Map: Election Fraud Database" appeared once in relation to Atlanta (Georgia) and then in relation to Cedar Falls (Iowa)—we only coded unique headlines. To filter the 174,511 headlines and generate a set that is small enough for manual-coding but large enough to allow the use of inferential statistics, we conducted a power analysis using the G-power tool (Faul et al. 2007). Given that the assigned codes served as the outcome variables, we chose a two-tailed a priori analysis for the z-test family suitable for logistic regression and discovered that we need a sample size of 5,408 headlines—assuming a minimal effect size corresponding to odds ratio of 1.1 with about 80 percent power.

Median number of headlines per day across 20 locations.

Search Keyword	*Search Results* (Top 5)	*Stories* (Top 3)	*Videos* (Top 3)	*Advertisements* (All)
Absentee ballot	100	54	0	55.5
Ballot dumping	100	0	3	1
Ballot harvesting	100	57.5	13.15	3.5
Ballots	100	60	45	38.5
Election fraud	100	60	6	44
Election results	100	60	28.5	22
Electoral fraud	100	60	0	25
How do I vote	100	0	0	32.5
Late ballots	100	56.5	0	14
Mail dumping	100	0	57	0
Mail-in voting	100	60	25.5	51
My ballot	100	28.5	0	14
Presidential election	100	60	51	19.5
Rigged election	100	60	15	55.5
Stolen election	100	55	1.5	20
Vote	100	60	30	25.5
Vote by post	100	0	0	25.5
Voter fraud	100	60	24	57
Voter intimidation	100	54	9	3
Where do I vote	100	0	0	27.5

Table 4: Summary statistics of SERP data separated by SERP verticals and search keywords. Given the skewed nature of the data — *e.g.*, while searching for "Electoral fraud" returned a maximum of 75 ads (October 5, 2020), searching for "Ballot dumping" only returned a maximum of 3 ads (October 8, 2020) across different locations — we report the median measure as our choice of summary statistic. A median score of 0 indicates a relatively lesser (but non-zero) number of headlines for the corresponding keyword.

To ensure that the data evenly represented the different search terms, search locations and information modalities, but was not biased either by the time or the day when it was scraped, we opted for a stratified random sample. We split our timeline into 4 two-week long periods *Oct. 5-19, Oct. 20-Nov 3, Nov. 4-18, and Nov. 19-Dec 3* such that each period contributed evenly to our sample. We next set out to select 50 search instances per combination of city type (2 urban areas, 1 urban cluster, 1 rural area), SERP vertical type (result, stories, videos, ads) and search term type (general, conspiratorial) — thus, selecting 1,600 headlines for each of the 4 time periods that will overall exceed the sample size of 5,408 as suggested by the power analysis. Unfortunately, we could not fetch 1,600 headlines from advertisements since: (1) there were no advertisements for any of the issue-specific terms in the third and fourth time period after November 4 and (2) Google did not surface 50 advertisements per city type even for the regular search terms. To overcome this asymmetry, we only collected ads for the first and second time period. Our data-sample thus consisted of 1,600 stories, 1,600 videos and 1,600 searches across 4 time periods and 800 ads across the first 2 time periods.

These 5,600 headlines selected through a stratified randomly sampling were not necessarily unique. For example, the headlines "Mail carrier arrested for dumping mail" and

"Including ballots USPS employee arrested, accused of dumping mail" showed up the most—61 and 59 times respectively—at different locations and/or in different search batches within our sampled set. We then coded the unique 2,438 headlines out of this set using the codebook described below.

4 Coding Scheme

In designing the coding scheme, initial data was first analyzed during a two-week exploratory period. During this time, we spoke with journalists and researchers regarding the headline construction process, including discussion of best journalistic practices related to the dissemination and presentation of online content. These included an emphasis on the centering of facticity through the use of keywords associated with falsehoods (*e.g.*, "misinformation", "false accusations", "misleading"), avoiding the spotlighting of problematic groups, focusing headlines on impact rather than eventizing aberrations or anecdotes, and ensuring that headlines are well-watched with the content of the related article rather then solely matching on prominent terms. In addition to providing additional insights such as these to help inform the coding scheme, the preliminary period also allowed us to simplify the primary categories contained in our coding scheme.

Informed by this preliminary process, we developed the coding scheme around a central "Stance" category, which was used to categorize headlines based on their potential impact on search engine users' trust in the election's legitimacy. Once this central variable was in place, we trained three coders to differentiate between various codes on this dimension, which sought broadly to answer the question:

If voters were to have read this headline on the day it was captured, how (if at all) could it have affected their perception of the integrity of the 2020 U.S. Election's processes, institutions, and results?

Eventually the "Stance" category was narrowed to focus on three central codes: *Sows doubt*, *Imparts trust*, and *Provides information*. The shortened definitions of these separate codes were finalized as follows:

- **Sows Doubt:** The headline has the potential to lower voter trust in the election's integrity.

  Example: *"Allegheny County ballot contractor accused of sending out late ballots in other counties"*

- **Imparts Trust:** The headline has the potential to improve voter trust in the election's integrity.

  Example: *"Barr says he hasn't seen fraud that could affect the election outcome"*

- **Provides Information:** The headline is not likely to alter voter trust in the election's integrity.

  Example: *"Biden projected to win Georgia, Trump projected to win North Carolina"*

In addition to these three central codes, two more codes were added to the "Stance" category. The *Campaign Ad* code was used to identify content that might appear in SERP verticals like search results or videos but were merely a promotional campaign in nature. The *Other* code was included to separate headlines which did not pertain to the election at all.

Based on insight from the initial exploratory period which illustrated that the headlines

coded as either *Imparts Trust* or *Sows Doubt* could be further divided to discern headlines actively spotlighting or emphasizing issues related to the election's legitimacy. For example, while one subset of headlines was coded solely as "Sows doubt" (or "Imparts trust"), denoting its potential to reduce (or impart) trust in election integrity — *e.g.*, "Poll worker accused…", "voters are concerned…" — a second subset appeared constructed specifically to undermine (or bolster) perceptions of its integrity — *e.g.*, "Voter Fraud Map: Where to find evidence…", "6M+ votes shifted by big tech…". To capture this crucial difference, the "Promotion" category (which involved a binary code) was developed to augment the "Stance" categorization. Collectively, the two categories are reported in tandem to ensure that we identify not only headlines that promote distrust, but also those that promote trust in the election's legitimacy.

In cases where we assigned the "Stance" as *Imparts Trust*, the "Promotion" category was used to identify headlines that deliberately attempted to build trust in the integrity of the election among readers. Similarly, where the "Stance" was coded as *Sows Doubt*, the "Promotion" category was used to identify headlines that appeared to be deliberately aimed at undermining perceptions of the election's integrity. Definitions and examples of headlines that we determined to promote distrusting content are presented below:

- **"Promotion" + "Sows Doubt" = "Promotes Distrust"):** the headline is *actively* reducing voter trust in the election's integrity

  Example: This accounts for differences in headlines discussing topics that may undermine trust in the election, such as *"Voters fear voter suppression in the build-up to the election"*, and headlines that push these narratives, such as *"Guns, lies and ballots set on fire: This is voter suppression in 2020".*

- **"Promotion" + "Imparts Trust"= "Promotes Trust"):** the headline is *actively* improving voter trust in the election's integrity

  Example: This accounts for differences in headlines discussing topics that may improve trust in the election, such as *"Ohio county officials shoot down Trump claim of 'rigged election' "*, and headlines that push narratives to improve trust, such as *"Election fraud claims are baseless"*

Content coded both as *Promotion* and *Sows Doubt*) is the closest to matching our conception of content with the potential to undermine trust in the election. As such, we used this subset as the basis for the primary analyses included here. Additional categories were included in the coding scheme, but remain peripheral to the central analyses discussed in this paper. These are discussed further in Appendix A.

4.1 Coding Process

Once the coding categories had been finalized, a subset of 200 of the 5,600 selected headlines were used as a practice set to test out the final coding scheme on real data. Once each coder had completed their coding of the initial set, the lead researcher on the project went through each disagreement individually with all three coders to identify issues in the coding scheme to ensure consistency across the coders before moving on to the full set. Most of the discrepancies resulted from differences in each coder's knowledge of the conspiracies that had proliferated online during the election period, which resulted in more knowledgeable coders correctly identifying headlines coding these narratives as "Promotes Distrust". Less knowledgeable coders were subsequently given a longer list of common conspiratorial narratives to review.

Once the coding scheme was finalized and the coders felt confident in their ability to discern between the codes in each of the categories, the data was organized in de-

scending order based on the frequency of headline appearance in the database. This resulted in the collection of 492 headlines which occurred more than two times each in the database. All the three coders coded them as a final check to ensure shared understanding of the coding scheme. After arbitrating any coding conflicts and determining enough consistency across coding, the team then proceeded to code the entire primary headline dataset.

The unique 2,438 headlines were randomized and each coder was given 2/3 of the headlines to code, resulting in each headline being coded twice by two different coders. After the first two coders finalized their coding, we found that our coders shared an *almost perfect* understanding of the "Stance" and "Promotion" categories as indicated by a Cohen's Kappa of 0.78 and 0.9 respectively (Landis and Koch 1977). Any disagreements between the first two coders were then arbitrated by a neutral third coder[7].

5 Results

5.1 R1: SERP vertical type

Our analyses show strong correlations between specific SERP verticals and the frequency of headlines that promoted distrust in the election's integrity ("Promotes Distrust"). Specifically, as seen in Figure 3, videos during the period were more likely to contain undermining content than other SERP verticals by a wide margin. This relationship persists both with headlines that serve to sow doubt in the credibility of the election and also among the more concerning content that promotes, rather than simply discusses or mentions, similar content.

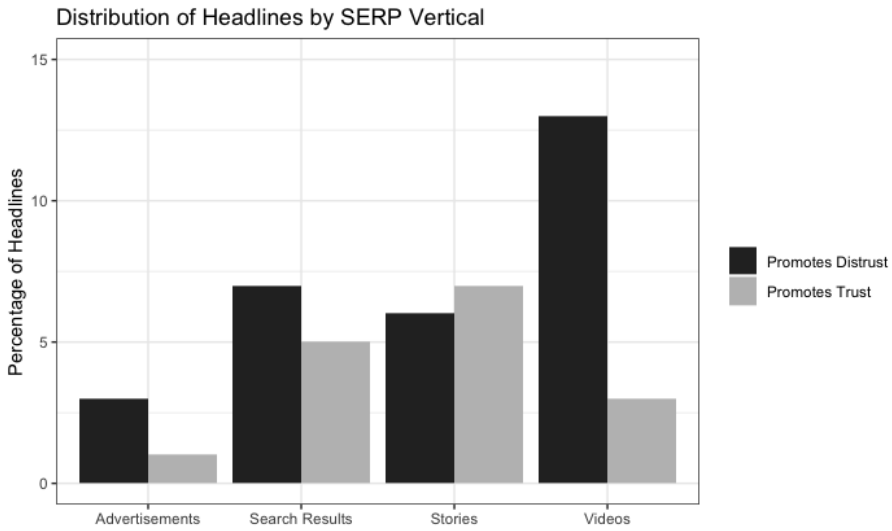

Figure 3: Percentage of coded headlines that promoted trust and distrust in the integrity of the election.

We ran a multinomial regression analysis by modeling the SERP vertical type (advertisement, search results, top stories, videos) to compare the extent to which headlines promoting distrust and promoting trust were identified in headlines related to videos

7. Overall, the final agreement rates ranged from 75% to 99%—corresponding to a Cohen's Kappa of 0.69 and 0.99—suggesting shared understanding across all the coding categories. We have included the inter-coder reliability measures across all the categories in Section B.

and top stories. We found that the odds of a video having a headline containing content with the potential to undermine trust were almost three times greater than headlines associated with a story (Table 5). Moreover, top stories were about three times as likely to promote a trust-imparting headline than videos 5, suggesting that video headlines contained both disproportionately high amounts of content that promoted distrust as well as low amounts of content that promoted trust.

SERP vertical type	Odds ratio	CI [95%]	p-value
Sowing doubt and promoting it (Yes, No; reference: No)			
(Intercept)	0.049^*	[0.036, 0.069]	$< .001$
Searches	2.213^*	[1.536, 3.186]	$< .001$
Stories	1.874^*	[1.294, 2.713]	$< .001$
Videos	5.472^*	[3.867, 7.741]	$< .001$
Imparting trust and promoting it (Yes, No; reference: No)			
(Intercept)	0.009^*	[0.004, 0.019]	$< .001$
Searches	6.991^*	[3.227, 15.144]	$< .001$
Stories	8.755^*	[4.062, 18.869]	$< .001$
Videos	2.905^*	[1.295, 6.514]	$< .001$

Table 5: Odds ratios for *Sowing doubt and promoting it* and *Imparting trust and promoting it* through different information modalities of searches, stories and videos (over campaign ads) calculated using logistic regression.

Figure 3 illustrates how top stories were the most common channel for the promotion of content that served to enhance readers' trust in the integrity of the election. When compared with other modalities such as videos, ads and search results, stories were the only SERP verticals with more headlines that imparted trust than headlines that sowed doubt throughout the sample. When viewed longitudinally in Figure 4, we see that this discrepancy between SERP vertical type and trusted content was more prominent in the post-election period. We found an increase in the post-election videos with headlines like "Dominion whistleblower says she didn't see a single vote cast for.."[8] and "ELECTORAL FRAUD: Where To Find The Evidence | Rudy Giuliani | Ep. 89"[9]

Overall, though the focus in the pre-election period was primarily on the role of trust-undermining advertisements(Zeng et al. 2021), video content appears to have been a far more challenging issue in the production of content that cast doubt on the election's legitimacy. Moreover, given that videos are more difficult to monitor due to the challenges associated with tracking in-video content and graphics (Nakov et al. 2021; Jalli 2021; Bradshaw et al. 2020), we believe that videos could be more delegitimizing beyond these headline differentials. For example, our data included videos with titles such as "LIVE 2020 Presidential Election Results", which, though coded as *Provides Information*, was found to be projecting false election results. Further research is needed to determine the scope of the use of misleading headlines to mask controversial in-video content and to capture deliberate efforts to evade censoring through the deployment of innocuous headlines (Moran, Grasso, and Koltai 2022).

8. The entire title read as "Dominion whistleblower says she didn't see a single vote cast for Donald Trump in her 27 hour shift" and directed users to a YouTube video that can still be accessed online as of April 30, 2022.

9. This video was later removed from YouTube for violating its community guidelines.

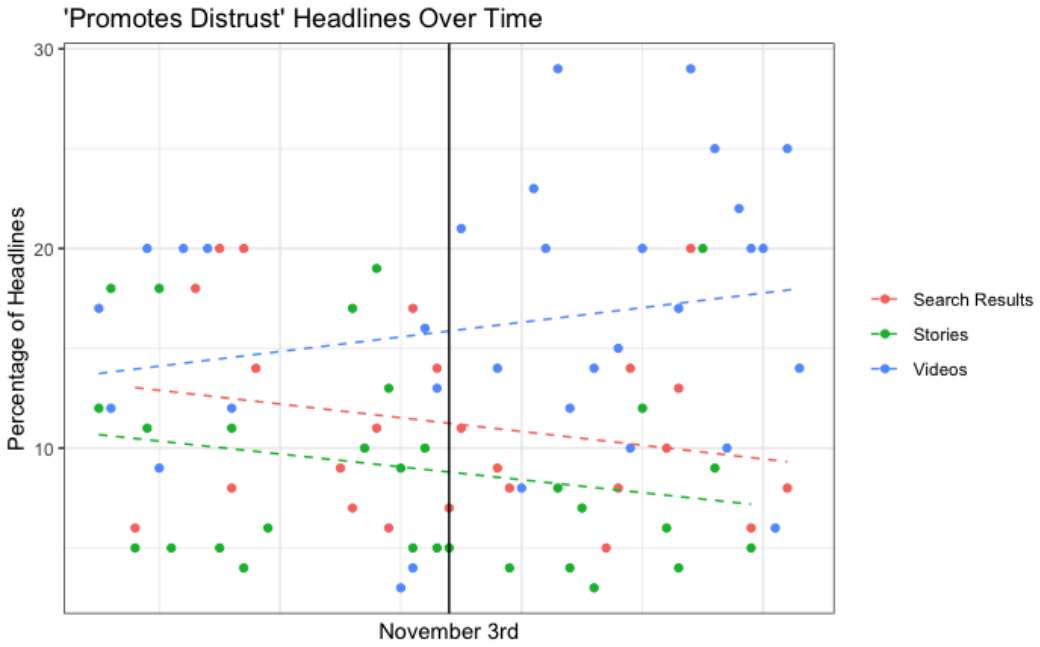

Figure 4: The percentage of headlines per day (Y axis) from SERP data that promoted distrust over the duration of data collection (X axis) from October 5 to December 3, 2020. Content in headlines of SERP-videos promoted increasingly more distrust than SERP-stories in the post-election period after November 03, 2020.

5.2 R2: Geographic location

In our analysis of geographic trends, our subset of election-related Google headlines provides evidence of both effective stewardship as well as concerning patterns of distribution of delegitimizing political content. Our coding—to our surprise—did not identify any differences in the kind of content based on any combination of the "Stance" and "Promotion" categories that was served to cities based on their sizes (*i.e.*, whether we classified the city or location as an urban area, urban cluster or rural area as specified in Table 2). We suspect this to have happened since search engine platforms may not find it useful to personalize the results for smaller regions with a population of a few thousand people, especially when the news involves topics about the national election.

One difference between the swing states and non swing states that stood out was the amount of campaign ads that emerged in the search engine home page. A multinomial regression analysis indicated that the odds of a campaign ad (compared to merely providing information) occurring in a swing state was almost twice that of a non-swing state. Figure 5 illustrates how these campaign ads almost always occurred through the advertising—and hence paid—SERP vertical in swing states as opposed to non-swing states. We found a similar pattern when investigating the difference across the electoral vote with red states having more campaign ads than the blue states.

5.3 R3: Search terms

Moreover, while differences in political content were small across cities, focusing on conspiratorial search terms often led to politically biased and more frequent misleading search results. As previously noted in Table 2, our search terms consisted of two groups — one that focused on ordinary election terms and another that focused on conspiratorial topics. By using these two types of election terms as predictors, our models suggested

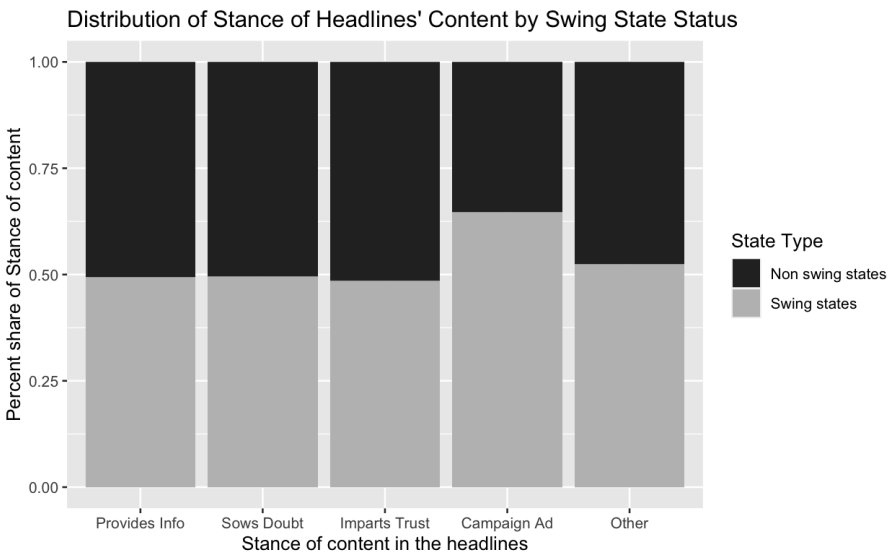

Figure 5: Searches made from swing states (total 6 locations in our collection) during the 2020 Presidential election returned relatively higher percent share of campaign-based advertisements as compared to searched made from non-swing states (total 14 in our collection).

that a headline containing content that promoted distrust in the election were about six times more likely to appear on Google SERPs when a user actively searched for a conspiratorial topic than when compared to use of more general searches about election topics (Table 6). Figure 6 shows the number of trust undermining headlines that appear on the homepage of Google search given the various search terms inspected in this study. Headlines promoting distrust in the election increased considerably when we conducted searches based on conspiratorial terms.

Search term type	**Odds ratio**	**CI [95%]**	**p-value**
Sowing doubt and promoting it (Yes, No; reference: No)			
(Intercept)	0.252^*	[0.229, 0.276]	$< .001$
General search term	0.167^*	[0.135, 0.206]	$< .001$

Table 6: Odds ratios for *Sowing doubt and promoting it* when searching for general election-related terms as compared to conspiratorial election-related terms (described in Table 1) calculated using logistic regression.

Searching for specific terms during the election period did return content that promoted distrust in the election, but the rates were much higher for the conspiratorial terms. That is, individuals who sought out narratives that discussed potential issues with the election were not always directed away from delegitimizing content. This is not surprising, given that Google's business model emphasizes its ability to deliver the content most likely of interest to end-users. However, it does place more emphasis on the process by which this content is selected and delivered (*e.g.*, tagging, labels, etc.). For individuals searching for general election terms and questions, which likely included a far greater share of Google's

users[10], our data suggests these users were subjected to fewer headlines containing content with the potential to undermine their trust in the election. Given this distinction, we see this as some evidence of successful limitation of pathways to delegitimizing content.

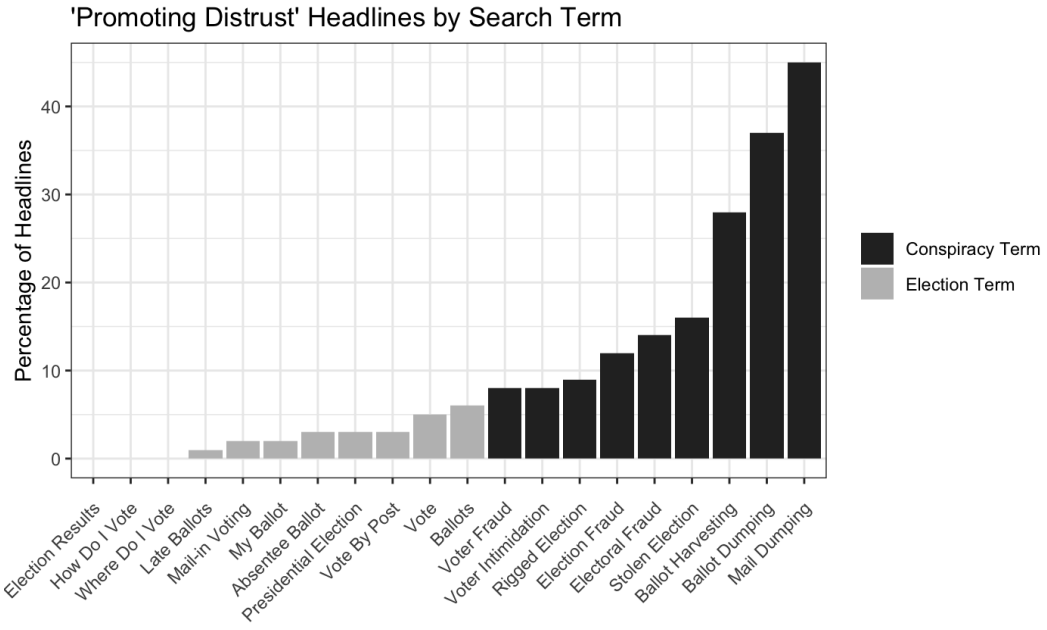

Figure 6: Frequency of doubt-sowing headlines given various search terms. Conspiratorial search terms that actively look for election-related issues served more delegitimizing content than the general search terms.

5.4 R4: Media domains

In addition to differences across SERP vertical type and location, our coding also revealed differences in the presentation of content across distinct news domains — with a specific focus on partisan outlets. We employed the media *bias* and media *reliability scores* from Ad Fontes Media (6.0) (Ad Fontes Media 2020)—a choice based on recent research that also needed interpreting media bias of online news sources (Huszár et al. 2022; Brooks and Porter 2020; Baranauskas 2022; Zhao et al. 2020)—as predictors for investigating the effect of media partisanship onto the kind of content that was served by these domains. As per these measures, a bias-score of +21.29 for *OANN* and -18.12 for *Democracy Now!* indicated the partisan-right and partisan-left, respectively, in our data.

Consistent with expectations, our models indicated that with every unit increase in the bias of a media domain (implying higher right-leaning bias), the likelihood of a headline's content that challenge the integrity of election by sowing doubt (relative to mere providing information) increased significantly by roughly 5.3% (Table 7). This trend continued when we accounted for how some media sources promoted the headlines that were served to delegitimize the election's integrity; every unit increase in the bias scores of a media source (*i.e.*, increasing right-leaning bias) could result in 2.6% higher chance

10. According to Google trends, only one query "Newsmax election results"—that we believe might have displayed some delegitimizing content—appeared in the top 25 rising search queries on Google's search engine in the same time period as our collection; most other queries involved phrases like "election results", "Presidential election", "where do I vote", "who is winning" which resonate with the general search terms that we used.

of it promoting content with the potential to undermine trust in the election and about 4% lower chance promoting content that reinstated public trust in the election (Table 8). Our models indicated no such effect for media reliability scores. Although *AdFontes Media v6.0* data only accounts for 44% of the unique headlines from our sampled data, results indicate the severity of damage that partisan media could cause — by promoting debunked content in mainstream information channels like search engines — towards public faith in democratic processes.

Type of stance	**Odds ratio**	**CI [95%]**	**p-value**
Campaign Ad (Intercept)	0.005^*	[0.006, 0.006]	$< .001$
Campaign Ad	0.877	[0.674, 1.142]	$.331$
Imparts trust (Intercept)	0.454	[0.139, 1.485]	$.191$
Imparts trust	1.002	[0.983, 1.022]	$.829$
Sows doubt (Intercept)	2.476	[0.931, 6.585]	$.069$
Sows doubt	1.053^*	[1.036, 1.071]	$< .001$
Other (Intercept)	0.031^*	[0.001, 0.341]	0.004
Other	1.016	[0.977, 1.057]	0.421

Table 7: Odds ratios for the different *Stance* type (relative to *Providing information*) reported for every unit increase in the media bias score taken from *AdFontes Media (5.0)* (Ad Fontes Media 2020) calculated using logistic regression.

	Odds ratio	**CI [95%]**	**p-value**
Sowing distrust and promoting it (Yes, No; reference: No)			
(Intercept)	0.259^*	[0.079, 0.844]	$.025$
Bias score	1.026^*	[1.005, 1.048]	$.014$
Imparting trust and promoting it (Yes, No; reference: No)			
(Intercept)	0.133^*	[0.026, 0.674]	$< .015$
Bias score	0.961^*	[0.935, 0.989]	$< .005$

Table 8: Odds ratios for *Sowing doubt and promoting it* and *Imparting trust and promoting it* with every unit increase in the media bias score taken from *AdFontes Media (5.0)* (Ad Fontes Media 2020) calculated using logistic regression.

Moreover, many of the domains associated with the presentation of content that promoted narratives with the potential to undermine electoral integrity with the highest frequency were affiliated with hyper-partisan outlets when examined both by the total and frequency of concerning posts. Looking first at total headlines coded as promoting sows doubt narratives [11], we find that while this included less reputable sites such as the Chinese language site *NTD* (see Figure 7), the list also included activist organizations such as *Rigged*, which, though perhaps well-intentioned, promoted ads with headlines that served to undermine trust in the election [12]. More alarmingly, several prominent legacy news sites, including CNN and Fox News, also rank toward the top of total articles with these dual designations.

11. For both total and frequency calculations only domains with more than three headlines appearing in the coding dataset were included in the plots.

12. Users of the Google search engine were shown advertisements titled “The Voter Suppression Playbook - Watch ‘Rigged’ for Free” upon searching for the keywords “rigged election” or “stolen election”, and upon clicking directed to the following url: https://www.riggedthefilm.com/

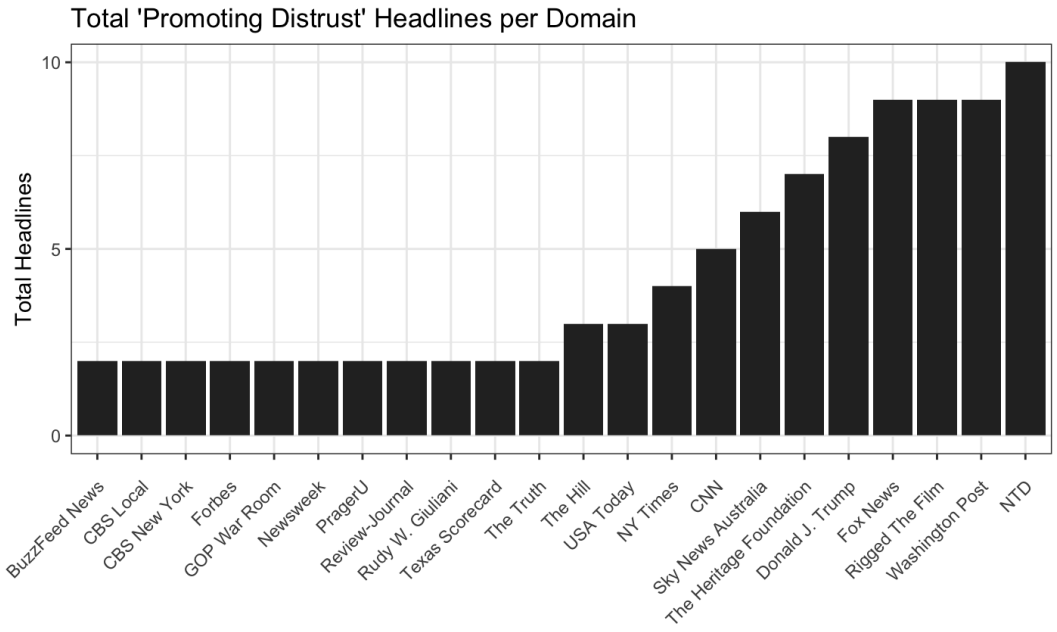

Figure 7: Included are the media domains (X axis) that promoted the most number of headlines (Y axis) with delegitimizing content.

However, when looking at percentages (Figure 8) rather than totals for sites like *CNN* and *Fox News*, their comparable rankings stand out less. While this is encouraging, taken together these outputs serve as a reminder that due to the much greater quantity of information put out by many legacy news organizations, even a small share of concerning articles can play an outsized influence in delivering content with the potential to undermine trust in the election to the public.

Based on a review of the headlines associated with the organizations that had the highest percentage of promoting doubt designations, emphasis on electoral fraud appeared to be the most common strategy to address the election's integrity and/or validity. In all, our preliminary foray into domain analysis should serve to initiate further examination of the sources of content with the potential to undermine trust in the election across both legacy and partisan media outlets. As with our media bias analyses, the narrow scope of our work here should serve only to draw broad conclusions regarding the types of headlines deployed across distinct forms of media groups rather than to identify specific domains for critique.

6 Discussion

6.1 Reflecting on Google search engine as a gateway to 2020 US election

Through this research, we found that Google SERPs do serve some concerning content, but primarily when users searched for conspiratorial terms or through the *videos* SERP-vertical. However, for searches based on general election terms, Google did a relatively good job of surfacing relevant content without leading users towards misleading arguments that negatively impacted civic trust in the election processes. Given the diversity of public opinion—sometimes at odds—across different regions in the United States, it can be challenging to deliver information that caters to the public interest yet steer clear of

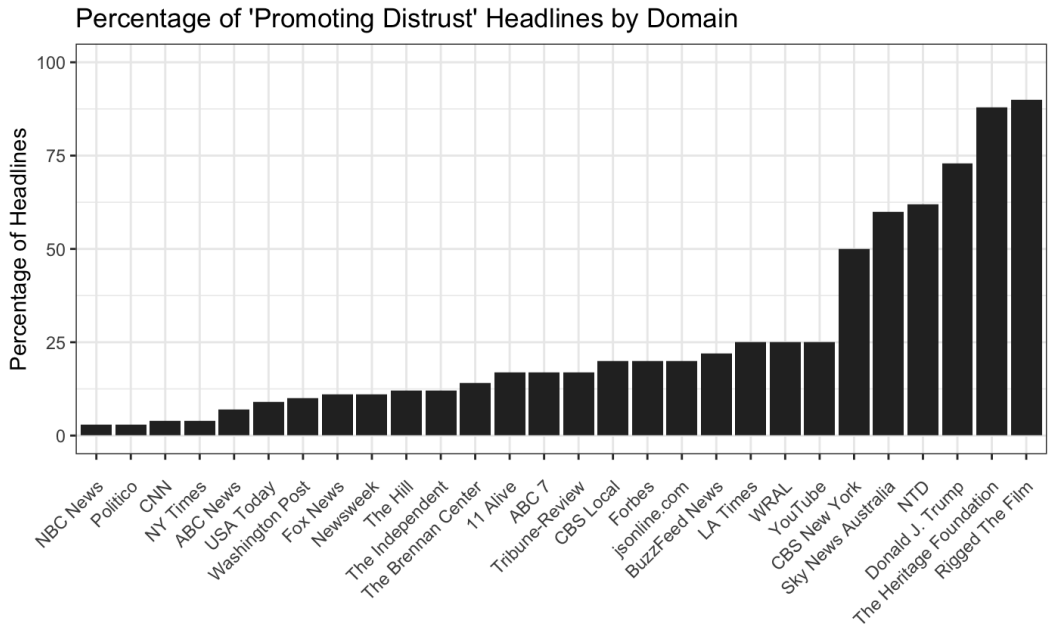

Figure 8: Percent share of unique headlines (Y axis) per media domain (X axis) that promoted delegitimizing content.

any regional biases. We were pleased to find no evidence that the search engine created information bubbles catering to any regional bias. The proportion of trust-undermining to trust-imparting content served in swing states was also similar to the proportion in non-swing states. We believe that Google offers at least some customization of SERP general results based on one's location to ensure such relative indifference to users' location when using the search engine.

One tricky area that remains rife with opportunity for discussion is what to serve users when they actively search for it. Our analysis demonstrates that Google offers more access to controversial content when users actively search for it. For example, headlines like "The biggest election fraud story you haven't heard about..." only showed up when users searched for the keyword "election fraud". This can be concerning given that SERP headlines are known to offer its users more partisan cues as compared to the original webpage (Hu et al. 2019). However, controversial content did not surface when users searched for more general keywords like "election results" or "Presidential election". More encouragingly, we found that during the 2020 US Election, individuals who searched for general election terms, issues, and questions — which we believe to be the dominant set of users — were largely shielded from headlines that could undermine electoral trust.

Several researchers have looked into *advertisement* as a medium of information that serves misleading content – political ads in particular (Zeng et al. 2021; Kreiss and McGregor 2019). While we found that campaign-based ads occurred more in our data, these were mostly placed by activism-based organizations, such as ACLU [13] and Winred [14]. These ads did not seem harmful to perceptions of election integrity. Some other ads that our coding suggested to have misleading content in their headlines occurred evenly

13. The American Civil Liberties Union (aclu.org) is a nonprofit organization to safeguard human rights and liberties.
14. Winred.com: Winred is the official online fundraising platform supporting the GOP.

across geographic locations[15].

Though ads were comparatively less of a concern in our audit, the video headlines served to be a notable pathway to content with the potential to undermine trust [16]. Given that videos are more difficult for fact-checkers to check for misleading perspectives, it might bypass their scrutiny, compared to the text modality. We did not have access to the usage of the videos, but it could be that videos are also more viewed. Prior research already points out that people tend to believe more in the information they see in a video than what they read through text (Wittenberg et al. 2021). In addition, plenty of studies have shown that clickbait is successful for a reason – people click on it (Scacco and Muddiman 2016; Bhowmik et al. 2019) At present, researchers believe that Google's tools for stopping video-based misinformation seen on the *Youtube* platform are only partially effective (Hussein, Juneja, and Mitra 2020; Donzelli et al. 2018). With the possibility of Google including more videos from a broader range of platforms like TikTok and Instagram on its homepage, more work will be needed to monitor this kind of content. This is of particular concern given the ability of creators to exploit the difficulties of monitoring videos to disguise content by evoking innocuous headlines, as occurred during the 2020 U.S. Election (Tenbarge 2020). Our current audit is unable to track these additional issues.

In terms of actionable responses, though we acknowledge the challenges associated with video content management, we find that simple headlines can narrow auditors' focus onto concerning content without necessitating the designation of resources necessary to sample all election-related videos randomly. While this does not solve the issue of videos using deliberately vague or misleading headlines to hide controversial content, it could be used to limit the mainstream influence of similar videos by ensuring that they remain in the periphery without showing in the results of users searching for general election concepts, terms, and questions. However, nothing in our audit suggests that censorship should be promoted as a central strategy of search engines in managing political and politically-adjacent content.

6.2 Designing future election-based audits

The included analyses were enabled by the strategic collection of data around the 2020 US Election. Future analyses can build on these results in several ways.

First, with additional resources, we can refine the coding scheme and build it out to address a broader range of issues, topics, and concepts. Although we developed a rigorous coding scheme to make sense of the news headlines, we utilized only those codes in this research that focus on headline content with the potential to undermine trust as a compromise that allows for a longitudinal peek under the curtain while keeping the work manageable. With collaborative efforts of the search engines themselves, it may be possible to capture and categorize similar headlines in real-time and match headlines with the associated content of each SERP vertical type. For instance, by pairing potentially undermining headlines with the nature of the underlying video content, it might be possible to generate a more nuanced understanding of the pathways connecting users to political content and generate knowledge of how these components intersect. Further investment in post hoc coding may also enable differentiation between types of potentially

15. We identified a couple of advertisements in September, 2020 — prior to the period of data collection that we analyzed within the scope of this paper — that we believed contained delegitimizing content. These ads were taken down soon after we reported them to Google. We suspect that including these ads in the collection might have impacted the reported findings.

16. Given the possibility that Google tends to overwrite about 33.4% (Pecanek 2021), it poses a question if Google's rewriting could play had any role in altering the trust-undermining or trust-imparting potential of SERP headlines.

undermining content, such as misleading content and outright false content.

Second, future audits can inform the priorities of search engine staff during election periods. While Washington DC and Silicon Valley have given much emphasis to the content linked to political advertisements, our audit suggests that when compared to other modalities, advertisements may not be the primary pathway from which users encounter content with the potential to undermine electoral trust. Further research into the different sources of content promotion should allow search engines to allocate resources more efficiently across their networks.

Third, we recommend that auditing reports should be thorough, comprehensible and easily accessible to different stakeholders so they can contribute in meaningful ways towards safeguarding the trust in election processes. For instance, although Google publishes a list of the political ads hosted on the search engine as the "Google transparency report on political advertising", the vague criteria of what constitutes a political ad and the limited information it requests about an ad publisher make it easy to circumvent the report's scope. For example, our extended data collection contained ads from "protectthevote.org" (paid by the Republican National Committee) that appeared in the transparency report, but ads from "protectmyvote.org" did not seem to fit Google's criteria of political advertising[17]. In addition, platforms should make search engine data available to researchers so they can serve as independent third-party auditors and help monitor the health of these information environments. This research was possible since we paid a third party for the API-access which is a financial hurdle and a caveat that might impact the quality of the data. We acknowledge that when collecting data through different sources, it is important to protect people's privacy and believe that the data that we accessed poses less threat to user privacy than by collecting SERP data through browser-plugins (*e.g.*, (Robertson et al. 2018)). We believe that law should require search engine platforms to provide researchers with access to anonymized data as nothing in Section 230 or the First Amendment stands in the way of such transparency.

Previous audits on search engines like Google search have discovered several insights into how these platforms can shape public opinion, especially around critical topics like elections (Hu et al. 2019; Mustafaraj, Lurie, and Devine 2020; Trielli and Diakopoulos 2019, 2022; Diakopoulos et al. 2018; Robertson et al. 2018). Our research adds to this body of literature on how Google fared in delivering election-related news to its users across America in 2020. In addition, we curate a dataset consisting of google search engine results — 47k advertisements, 500k main search results, 240k news stories, and 66k videos — and make it publicly available[18] to facilitate the discovery of more insights about the 2020 US elections as pictured through the agency of search engines.

17. As reported in the article (https://www.washingtonpost.com/technology/2020/08/28/google-ads-mail-voting/), Google took five days before they removed the "protectmyvote.org" ads from their platform after its discovery.

18. URL to be added later.

References

Ad Fontes Media. 2020. "Media Bias Chart 6.0."

Albright, Jonathan. 2018. "Untrue-Tube: Monetizing Misery and Disinformation" (February). https://d1gi.medium.com/untrue-tube-monetizing-misery-and-disinformation-388c4786cc3d.

Alvarez, R Michael, Jian Cao, and Yimeng Li. 2021. "Voting Experiences, Perceptions of Fraud, and Voter Confidence." *Social Science Quarterly* 102 (4): 1225–38.

Andrew, Blake C. 2007. "Media-generated shortcuts: Do newspaper headlines present another roadblock for low-information rationality?" *Harvard International Journal of Press/Politics* 12 (2): 24–43.

Ballatore, Andrea, Mark Graham, and Shilad Sen. 2017. "Digital hegemonies: the localness of search engine results." *Annals of the American Association of Geographers* 107 (5): 1194–215.

Baranauskas, Andrew J. 2022. "News media and public attitudes toward the protests of 2020: An examination of the mediating role of perceived protester violence." *Criminology & Public Policy* 21 (1): 107–23.

Beam, Michael A. 2014. "Automating the news: How personalized news recommender system design choices impact news reception." *Communication Research* 41 (8): 1019–41.

Bhowmik, Sima, Md Main Uddin Rony, Md Mahfuzul Haque, Kristen Alley Swain, and Naeemul Hassan. 2019. "Examining the Role of Clickbait Headlines to Engage Readers with Reliable Health-related Information." *arXiv preprint arXiv:1911.11214.*

Bordia, Prashant, and Nicholas DiFonzo. 2017. "Psychological motivations in rumor spread." In *Rumor mills,* 87–102. Routledge.

Bradshaw, Samantha. 2019. "Disinformation optimised: gaming search engine algorithms to amplify junk news." *Internet policy review* 8 (4): 1–24.

Bradshaw, Samantha, David Thiel, Carly Miller, and Renee DiResta. 2020. "Election Delegitimization: Coming to you Live" (November). https://www.eipartnership.net/rapid-response/election-delegitimization-coming-to-you-live.

Brennan Center for Justice. 2020. "It's Official: The Election Was Secure" (December). https://www.brennancenter.org/our-work/research-reports/its-official-election-was-secure.

Brooks, Heather Z, and Mason A Porter. 2020. "A model for the influence of media on the ideology of content in online social networks." *Physical Review Research* 2 (2): 023041.

Brooks, Nico. 2004. "The Atlas rank report: How search engine rank impacts traffic." *Insights, Atlas Institute Digital Marketing.*

Cybersecurity & infrastructure security agency. 2021. "Election security rumor vs. reality." *Published online* (November). https://www.cisa.gov/rumorcontrol.

———. 2022. "Election infrastructure security." *Published online,* https://www.cisa.gov/election-security.

Diakopoulos, Nicholas, Daniel Trielli, Jennifer Stark, and Sean Mussenden. 2018. "I vote for — how search informs our choice of candidate." *Digital Dominance: The Power of Google, Amazon, Facebook, and Apple, M. Moore and D. Tambini (Eds.)* 22.

Digirolamo, Gregory J, and Douglas L Hintzman. 1997. "First impressions are lasting impressions: A primacy effect in memory for repetitions." *Psychonomic Bulletin & Review* 4 (1): 121–24.

Donzelli, Gabriele, Giacomo Palomba, Ileana Federigi, Francesco Aquino, Lorenzo Cioni, Marco Verani, Annalaura Carducci, and Pierluigi Lopalco. 2018. "Misinformation on vaccination: A quantitative analysis of YouTube videos." *Human vaccines & immunotherapeutics* 14 (7): 1654–59.

Ecker, Ullrich KH, Stephan Lewandowsky, Ee Pin Chang, and Rekha Pillai. 2014. "The effects of subtle misinformation in news headlines." *Journal of experimental psychology: applied* 20 (4): 323.

Edgerly, Stephanie, Rachel R Mourão, Esther Thorson, and Samuel M Tham. 2020. "When do audiences verify? How perceptions about message and source influence audience verification of news headlines." *Journalism & Mass Communication Quarterly* 97 (1): 52–71.

Emig, Elmer. 1928. "The connotation of newspaper headlines." *Journalism Quarterly* 4 (4): 53–59.

Epstein, Robert, and Ronald E Robertson. 2015. "The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections." *Proceedings of the National Academy of Sciences* 112 (33): E4512–E4521.

Faul, Franz, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. "G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences." *Behavior research methods* 39 (2): 175–91.

Fernández, Miriam, Alejandro Bellogín, and Iván Cantador. 2021. "Analysing the Effect of Recommendation Algorithms on the Amplification of Misinformation." *arXiv preprint arXiv:2103.14748.*

Gabielkov, Maksym, Arthi Ramachandran, Augustin Chaintreau, and Arnaud Legout. 2016. "Social clicks: What and who gets read on Twitter?" In *Proceedings of the 2016 ACM SIGMETRICS international conference on measurement and modeling of computer science,* 179–92.

Gimpel, James G, Nathan Lovin, Bryant Moy, and Andrew Reeves. 2020. "The urban-rural gulf in American political behavior." *Political behavior* 42 (4): 1343–68.

Hu, Desheng, Shan Jiang, Ronald E. Robertson, and Christo Wilson. 2019. "Auditing the partisanship of Google search snippets." In *The World Wide Web Conference,* 693–704.

Hussein, Eslam, Prerna Juneja, and Tanushree Mitra. 2020. "Measuring misinformation in video search platforms: An audit study on YouTube." *Proceedings of the ACM on Human-Computer Interaction* 4 (CSCW1): 1–27.

Huszár, Ferenc, Sofia Ira Ktena, Conor O'Brien, Luca Belli, Andrew Schlaikjer, and Moritz Hardt. 2022. "Algorithmic amplification of politics on Twitter." *Proceedings of the National Academy of Sciences* 119 (1): e2025334119.

Jalli, Nuurrianti. 2021. "'Mission impossible?': tracking political misinformation and disinformation on TikTok" (December). https://theconversation.com/mission-impossible-tracking-political-misinformation-and-disinformation-on-tiktok-173247.

Jamieson, Kathleen Hall, Bruce W Hardy, and Daniel Romer. 2007. "The effectiveness of the press in serving the needs of American democracy."

Joachims, Thorsten, Laura Granka, Bing Pan, Helene Hembrooke, Filip Radlinski, and Geri Gay. 2007. "Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search." *ACM Transactions on Information Systems (TOIS)* 25 (2): 7–es.

Juneja, Prerna, and Tanushree Mitra. 2021. "Auditing e-commerce platforms for algorithmically curated vaccine misinformation." In *Proceedings of the 2021 chi conference on human factors in computing systems,* 1–27.

Kapferer, Jean-Noël. 1987. *Rumeurs: le plus vieux média du monde.* Editions du seuil.

Kawakami, Anna, Khonzodakhon Umarova, and Eni Mustafaraj. 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,* 14:868–77.

Kreiss, Daniel, and Shannon C McGregor. 2019. "The "arbiters of what our voters see": Facebook and Google's struggle with policy, process, and enforcement around political advertising." *Political Communication* 36 (4): 499–522.

Landis, J Richard, and Gary G Koch. 1977. "The measurement of observer agreement for categorical data." *biometrics,* 159–74.

Liu, Siyi, Lei Guo, Kate Mays, Margrit Betke, and Derry Tanti Wijaya. 2019. "Detecting frames in news headlines and its application to analyzing news framing trends surrounding US gun violence." In *Proceedings of the 23rd conference on computational natural language learning (CoNLL).*

Lorigo, Lori, Maya Haridasan, Hrönn Brynjarsdóttir, Ling Xia, Thorsten Joachims, Geri Gay, Laura Granka, Fabio Pellacini, and Bing Pan. 2008. "Eye tracking and online search: Lessons learned and challenges ahead." *Journal of the American Society for Information Science and Technology* 59 (7): 1041–52.

Moran, Rachel, Izzi Grasso, and Kolina Koltai. 2022. "Folk Theories of Avoiding Content Moderation: How Vaccine-Opposed Influencers Amplify Vaccine Opposition on Instagram." *Social Media + Society.*

Mustafaraj, Eni, Emma Lurie, and Claire Devine. 2020. "The case for voter-centered audits of search engines during political elections." In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency,* 559–69.

Nakov, Preslav, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021. "Automated fact-checking for assisting human fact-checkers." *arXiv preprint arXiv:2103.07769.*

National Geographic. 2009. *United States Regions.* https://www.nationalgeographic.org/maps/united-states-regions/.

Pan, Bing, Helene Hembrooke, Thorsten Joachims, Lori Lorigo, Geri Gay, and Laura Granka. 2007. "In Google we trust: Users' decisions on rank, position, and relevance." *Journal of computer-mediated communication* 12 (3): 801–23.

Pecanek, Michal. 2021. "6 Important Insights About Title Tags (953,276 Pages Studied)." November. https://ahrefs.com/blog/title-tags-study/.

Pennycook, Gordon, and David G Rand. 2021. "Examining false beliefs about voter fraud in the wake of the 2020 Presidential Election." *The Harvard Kennedy School Misinformation Review.*

Rieis, Julio, Fabrício de Souza, Pedro Vaz de Melo, Raquel Prates, Haewoon Kwak, and Jisun An. 2015. "Breaking the news: First impressions matter on online news." In *Proceedings of the International AAAI Conference on Web and Social Media,* 9:357–66. 1.

Robertson, Ronald E, Shan Jiang, Kenneth Joseph, Lisa Friedland, David Lazer, and Christo Wilson. 2018. "Auditing partisan audience bias within google search." *Proceedings of the ACM on Human-Computer Interaction* 2 (CSCW): 1–22.

Rodriguez, Ashley. 2018. "YouTube's algorithms can drag you down a rabbit hole of conspiracies, researcher finds" (February). https://qz.com/1215937/research-youtubes-algorithms-lead-down-a-rabbit-hole-of-conspiracies/.

Rogers, Richard. 2013. *Digital methods.* MIT press.

Saranac Hale Spencer. 2020. "Nine Election Fraud Claims, None Credible" (December). https://www.factcheck.org/2020/12/nine-election-fraud-claims-none-credible/.

Scacco, Joshua M, and Ashley Muddiman. 2016. "Investigating the influence of "clickbait" news headlines." *Engaging News Project Report.*

SerpApi. 2020. *Serpapi: Google Search Api. Available at https://serpapi.com/*. Accessed June 30, 2021.

Shearer, Elisa. 2021. "More than eight-in-ten Americans get news from digital devices." *Pew Research Center,* https://www.pewresearch.org/fact-tank/2021/01/12/more-than-eight-in-ten-americans-get-news-from-digital-devices/.

Shepherd, Ryan P. 2020. "Gaming Reddit's Algorithm: r/the_donald, Amplification, and the Rhetoric of Sorting." *Computers and Composition* 56:102572.

Simko, Jakub, Matus Tomlein, Branislav Pecher, Robert Moro, Ivan Srba, Elena Stefancova, Andrea Hrckova, Michal Kompan, Juraj Podrouzek, and Maria Bielikova. 2021. "Towards Continuous Automatic Audits of Social Media Adaptive Behavior and its Role in Misinformation Spreading." In *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization,* 411–14.

Spenkuch, Jörg L, and David Toniatti. 2016. "Political advertising and election outcomes." *Kilts Center for Marketing at Chicago Booth–Nielsen Dataset Paper Series,* 1–46.

Starbird, Kate, Emma S. Spiro, and Kolina Koltai. 2020. "Misinformation, Crisis, and Public Health—Reviewing the Literature V1." Edited by MediaWell Social Science Research Council (June).

StatCounter. 2021. *Search Engine Market Share Worldwide.* https://gs.statcounter.com/search-engine-market-share.

Sundar, S. Shyam, Maria D Molina, and Eugene Cho. 2021. "Seeing is believing: Is video modality more powerful in spreading fake news via online messaging apps?" *Journal of Computer-Mediated Communication* 26 (6): 301–19.

Tannenbaum, Percy H. 1953. "The effect of headlines on the interpretation of news stories." *Journalism Quarterly* 30 (2): 189–97.

Tenbarge, Kat. 2020. *YouTube channels made money off of fake election results livestreams with thousands of viewers. Available at https://www.insider.com/youtube-fake-election-results-livestreams-monetized-misinformation-2020-11*, November.

Thompson, Bennie, Liz Cheney, and Lofgren Zoe. 2022. "Thompson, Cheney, & Lofgren Opening Statements at Select Committee Hearing." *Social Media + Society* (June). https://january6th.house.gov/news/press-releases/thompson-cheney-lofgren-opening-statements-june-13th-select-committee-hearing.

Trielli, Daniel, and Nicholas Diakopoulos. 2019. "Search as news curator: The role of Google in shaping attention to news information." In *Proceedings of the 2019 CHI Conference on human factors in computing systems,* 1–15.

———. 2022. "Partisan search behavior and Google results in the 2018 US midterm elections." *Information, Communication & Society* 25 (1): 145–61.

Urman, Aleksandra, Mykola Makhortykh, Roberto Ulloa, and Juhi Kulshrestha. 2021. "Where the Earth is flat and 9/11 is an inside job: A comparative algorithm audit of conspiratorial information in web search results." *arXiv preprint arXiv:2112.01278.*

US Census Bureau. 2010. *2010 Census Urban and Rural Classification and Urban Area Criteria.* https://www.census.gov/programs-surveys/geography/guidance/geo-areas/urban-rural/2010-urban-rural.html.

Wittenberg, Chloe, Ben M Tappin, Adam J Berinsky, and David G Rand. 2021. "The (minimal) persuasive advantage of political video over text." *Proceedings of the National Academy of Sciences* 118 (47).

Zeng, Eric, Miranda Wei, Theo Gregersen, Tadayoshi Kohno, and Franziska Roesner. 2021. "Polls, clickbait, and commemorative $2 bills: problematic political advertising on news and media websites around the 2020 US elections." In *Proceedings of the 21st ACM Internet Measurement Conference,* 507–25.

Zhao, Erfei, Qiao Wu, Eileen M Crimmins, and Jennifer A Ailshire. 2020. "Media trust and infection mitigating behaviours during the COVID-19 pandemic in the USA." *BMJ global health* 5 (10): e003323.

Zook, Matthew A, and Mark Graham. 2007. "Mapping DigiPlace: geocoded Internet data and the representation of place." *Environment and Planning B: Planning and Design* 34 (3): 466–82.

Authors

Affiliations: University of Washington (UW); Center for an Informed Public (CIP); Human-Centered Design and Engineering (HCDE); Political Science (PS); School of Law (SL); Information School (IS).

Himanshu Zade is a Ph.D. candidate at UW in the HCDE department and in CIP.

Morgan Wack is a Ph.D. candidate at UW in the PS department and in CIP.

Yuanrui Zhang is a B.S. student at UW in IS and a research assistant at CIP.

Kate Starbird is an Associate Professor at UW in the HCDE department and in CIP.

Ryan Calo is a Lane Powell & D. Wayne Gittinger Endowed Professorship Professor at UW in SL and faculty at CIP.

Jason Young is a Senior Research Scientist and Affiliate Assistant Professor at UW in IS and in CIP.

Jevin D. West is an Associate Professor at UW in IS and in CIP.

Acknowledgements

We reserve additional gratitude to recognize the contributions of Michelle Weng, Kei Hartley, and Isla Wisemore.

Data Availability Statement

We have made the data available on the Open Source Foundation.

7 Funding Statement

The research was supported through funding from the John S. and James L. Knight Foundation, the UW Center for an Informed Public, The William and Flora Hewlett Foundation and Craig Newmark Philanthropies. Jevin D. West and Kate Starbird acknowledge the support of the National Science Foundation (award no. 2027792). Kate Starbird acknowledges the support of the National Science Foundation (NSF CAREER award no. 1749815).

Ethical Standards

This research used data that was publicly available on Google's search engine and accessed through the SerpAPI (SerpApi 2020). The data that we analyzed and make publicly available does not contain any identifiers to an individual or to a group of people.

Keywords

Election misinformation; Misleading content; Google SERP; Search engine; Headlines; Audit.

Appendices

Appendix A: Additional Coding Categories

During the coding process, it was decided that the brief structure of most headlines led to considerable uncertainty regarding the potential impact of certain posts. To capture this uncertainty, in addition to the primary coding categories of "Stance" and "Promotion", an "Ambiguous" category was introduced to denote instances where it was unclear from the headline how readers would react, with two plausible interpretations possible that would lead to different assessments of trust. For example, a headline which sarcastically insists that there was copious amounts of election fraud could be read in a straightforward manner, resulting in reduced trust, or in a sarcastic tone. The "Ambiguous" category was optional and only needed to be used when applicable.

Another additional category "Topic" was added during the planning phase to capture in broad strokes the issue being discussed by the headline. The list of codes in this category included *Mail-in Voting, In-Person Voting, Voting Machines, Public Perceptions, Misinformation, Voter Suppression/Intimidation, Ballot Harvesting, Election Info/Procedures, Election Processes/Results*. The final two topics were catch-all inclusions which were used only when one of the first seven specific topics were not applicable. To differentiate between the two, the first code *Election Info/Procedures* was used when the headline in question detailed information pertaining to the election or surrounding events. The second code *Election Processes/Results* was used to identify headlines which related directly to the outcome of the election or issues that could impact the outcome of the election which did not fall into the more specific topical themes.

The next two categories, "Attribution" and "Claimant", were coded in tandem. "Attribution" is a binary code denoting whether or not the headline in question attributed the contained information to a particular individual, entity, or source. If this was coded *Yes*, denoting an attribution within the headline, then the "Claimant" category was used to identify or approximate the best categorization of individual, entity, or source. This included everything from specific partisan supporters to public officials, media members, and election workers, as well as either Biden or Trump.

The "Subject" category was added late on in the process to provide additional information not captured by codes in either the "Topic" or "Claimant" category. This was developed by the coders as a representation of the *heroes or villains* of the headline in question. While the "Topic" category focused on the theme, the "Subject" focused on the individual or the group involved in that action. For instance, if a headline read: "Media Group X: Notable Democratic politician accused of ballot harvesting in Minnesota", the topic would be ballot harvesting, where as the subject would be the Democratic politician in question (here "Media Group X" would be considered the claimant).

Four additional categories were included to provide context for specific headlines. These included "Specific Event", "Legal Claim", "Leading Question", and "Fact Check". Each of these is a binary code indicating whether or not a specific headline involves any of these specific issues.

Appendix B: Inter-coder reliability

Table 9 describes the inter-coder reliability amongst the three coders and the final agreement rates amongst them for each category. Given that the first five primary

Coding category	Cohen's Kappa (IRR)	Percent agreement
Stance	0.78	82%
Promotion	0.90	92%
Topic	0.69	75%
Subject	0.74	79%
Attribution	0.95	96%
Fact Check	0.99	99%
Leading Question	0.99	99%
Specific Event	0.86	89%

Table 9: Category-wise agreement amongst the three coders.

categories included more than three possible codes, these reported results provide strong evidence that are final codes reflect strongly related responses of our coders. As per the standard for interpreting kappa-scores Landis and Koch 1977, our coders shared a *substantial* or *almost perfect* understanding of the codes and its employment on the data.