The Media Coverage of the 2020 US Presidential Election Candidates through the Lens of Google's Top Stories

Anna Kawakami, Khonzodakhon Umarova, Eni Mustafaraj
Department of Computer Science
Wellesley College, Wellesley, MA
{akawakam, kumarova, emustafaraj}@wellesley.edu

Abstract

Choosing the political party nominees, who will appear on the ballot for the US presidency, is a long process that starts two years before the general election. The news media plays a particular role in this process by continuously covering the state of the race. How can this news coverage be characterized? Given that there are thousands of news organizations, but each of us is exposed to only a few of them, we might be missing most of it. Online news aggregators, which aggregate news stories from a multitude of news sources and perspectives, could provide an important lens for the analysis. One such aggregator is Google's Top stories, a recent addition to Google's search result page. For the duration of 2019, we have collected the news headlines that Google Top stories has displayed for 30 candidates of both US political parties. Our dataset contains 79,903 news story URLs published by 2,168 unique news sources. Our analysis indicates that despite this large number of news sources, there is a very skewed distribution of where the Top stories are originating, with a very small number of sources contributing the majority of stories. We are sharing our dataset[1] so that other researchers can answer questions related to algorithmic curation of news as well as media agenda setting in the context of political elections.

Introduction

When the terrorist attacks of 9/11 happened, Krishna Bharat,[2] a Google engineer, understood that Google's search engine was missing something important, namely, the ability to respond to public's interest for unfolding events by displaying breaking news articles as part of its search results. This realization gave birth to Google News,[3] an automated system that continuously monitors thousands of online news websites around the world and indexes what they are publishing on any given topic. Since its inception in 2002, Google News has been and remains a separate product, distinct from Google Search. The latter has since incorporated news headlines in a paragraph titled "In the News." However, since late 2016,[4] the more prominent feature of Top stories has become a fixture of many search pages. Figure 1 depicts an example of the Top stories panel in Google's search engine results page (SERP), containing "fresh" headlines about Amy Klobuchar, one of the Democratic candidates running to become the Democratic nominee for the 2020 US Presidential Election.

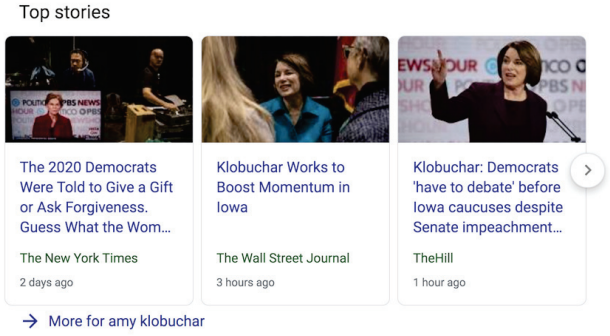

Figure 1: Top stories panel that appeared on Google search for the query "Amy Klobuchar" on Dec 22, 2019, 12:52 EST. The panel contains up to ten stories, published at different times close to the query date.

Such a prominent display of news stories, especially for individuals who are already known to the public (as most presidential candidates are), focuses the attention on the news. Therefore, the news stories that appear in Top stories can affect what the readers take away from the event, especially when the headlines are several words long, obviating the need to click and read the whole story. Given the potential to frame a current event, it is natural to ask: which news stories are being chosen to appear in Top stories and why? This question has become more important recently, due to the political climate in the United States, where online platforms such as Google, Facebook and Twitter have come under attack for a variety of reasons. A recurring criticism toward Google is that its news aggregators such as Google News favor so-called "left-leaning" news sources,[5] a criti-

[1] https://doi.org/10.7910/DVN/0ZLHOK

[2] https://en.wikipedia.org/wiki/Krishna_Bharat

[3] http://mediashift.org/2010/02/google-news-to-publishers-lets-make-love-not-war035/

[4] https://searchengineland.com/google-replaces-news-box-top-stories-desktop-264993

[5] https://pjmedia.com/trending/google-search-results-show-

cism repeated by President Trump as well.[6] Proving or disproving such claims is challenging, since Google algorithms are complex black boxes not open to external scrutiny.

In the recent years, the new field of algorithm audits (Sandvig et al. 2014) has provided an opportunity to systematically study the issues of fairness or bias in algorithmic decision making systems. Several auditing studies of search engines have been published in the past years, with some of them focusing explicitly on political or election related search results (Robertson, Lazer, and Wilson 2018; Diakopoulos et al. 2018; Metaxas and Pruksachatkun 2017; Metaxa et al. 2019). Our data collection is part of this trend in auditing online platforms, but with a more dedicated focus on their role for shaping news exposure, a process known as algorithmic curation of news (Trielli and Diakopoulos 2019; Thurman et al. 2019; Trevisan et al. 2018; Bandy and Diakopoulos 2019).

Using the established scraping method of search engine audits, we have created a dataset,[7] that is novel on two axes:

- **Duration:** Our dataset covers an entire year of search results for 30 political candidates in the 2020 US elections. This is the longest period of any of published studies in search engine audits.
- **Frequency:** Our dataset contains between 4-12 daily measurements, with the intention to observe the freshness or stickiness of news stories over time. This aspect of Google's Top stories has not been previously studied.

Findings

The primary goal of this paper is to introduce a new dataset that can be used to study algorithmic news curation, as well as media agenda setting and issue framing during presidential elections. Additionally, we provide a first (but limited) analysis on these research questions, whose results point to the need for more in-depth analysis, which we hope the release of the dataset will enable.

Not all news sources are equal. Our dataset contains 2,168 news sources, however, one third of all articles are produced by only eight news publishers. Is the dominance of these news sources explained by their journalistic output or by other factors? We start to answer this question by analyzing the output of New York Times and Politico, but future research should look at a larger number of news outlets. To better understand algorithmic news curation, we need to quantify what content is available to choose from.

Not all candidates are equal. In our data collection we followed 30 candidates (with some limitations) who are running to become their party nominee for the 2020 US Presidential Election. Leaving aside President Trump, whose coverage is extensive due to his current office, the dataset suggests that media picked its favorite candidates as early as May 2019, and these happen to be the four front-runners as of January 2020: Biden, Warren, Sanders, and Buttigieg.

pervasive-anti-trump-anti-conservative-bias/

[6]https://www.nytimes.com/2018/09/05/technology/google-trump-bias.html

[7]https://doi.org/10.7910/DVN/0ZLHOK

Table 1: Statistics on the stories collected for each candidate. Star (*) next to candidate name indicates that as of January 9, 2020, the candidate is no longer a presidential candidate. `Total moments` refers to the number of measurements that contained top stories. The `% Missed Collection` is calculated toward the expected possible collections, 3,032 moments. For most of the candidates, we missed only about 5% of the measurements, due to technical failure.

Candidate	Unique Articles	Total Moments	No Top Stories	% Missed Collections
Donald Trump	16956	2855	15	0.05
Joe Biden	8536	2859	6	0.06
Elizabeth Warren	7284	2867	0	0.05
Bernie Sanders	7100	2867	6	0.05
Pete Buttigieg	5205	2832	25	0.06
Kamala Harris*	4999	2860	4	0.06
Beto O Rourke*	3985	2862	11	0.05
Cory Booker	3435	2856	17	0.05
Tulsi Gabbard	2514	2828	28	0.06
Amy Klobuchar	2481	2871	3	0.05
Michael Bloomberg	2108	2245	616	0.06
Kirsten Gillibrand	2064	2099	765	0.06
Jay Inslee*	2047	2508	359	0.05
Julian Castro	2013	2808	57	0.06
Andrew Yang	2011	2513	9	0.17
Bill de Blasio*	1946	2106	45	0.29
Tom Steyer	1460	2691	166	0.06
John Hickenlooper*	1445	2178	687	0.06
Steve Bullock*	1337	2608	245	0.06
Marianne Williamson	1170	2423	79	0.17
Tim Ryan*	1112	2192	664	0.06
Seth Moulton*	932	1714	1143	0.06
Eric Swalwell*	901	957	136	0.64
Howard Schultz*	895	952	1915	0.05
Michael Bennet	676	2125	27	0.29
John Delaney	598	1761	391	0.29
Bill Weld	445	2006	92	0.31
Mike Gravel*	288	898	1515	0.2

Data Collection Process

To collect Top stories, we capture the Google Search result page for a list of candidate names, multiple times a day. The details of this process are described in the following.

The Candidates

In December 2018, we created a list of political figures that were being rumored of having 2020 presidential aspirations, based on two sources: the opinion poll aggregator FiveThirtyEight website,[8] which had compiled a list of 31 names; and the sports betting website OddsShark[9] which was taking bets on 54 names. Combining the two lists gave us 68 names. Not all these names belonged to politicians; famous people such as Michelle Obama and Mark Zuckerberg were included too. Unfortunately, the list didn't include Andrew Yang (not a politician) and other political figures who were less popular. We collected data for all 68 names in the first six months, but reduced the list to 30 names, once there was more clarity about who was seriously running. Out of these 30 candidates, we have complete data for 22 candidates. For

[8]October 11, 2018, https://fivethirtyeight.com/features/whos-behaving-like-a-2020-presidential-candidate/

[9]https://www.oddsshark.com/other/2020-usa-presidential-odds-futures

8 others, the data collection started later. Table 1 shows the names of candidates ordered by the number of stories in our dataset, as well as the amount of data available or missing for each candidate.

Automated Google Searches

The raw data we collected are HTML pages, namely, the first result page of Google Search for the name of each candidate. We use a custom-made Python script that opens a new instance of the Chrome browser in a blank-state (no user history or cookies), enters the name of the candidate in the search box, waits until the page is loaded and stores it as an HTML file. This script runs on the same computer (fixed IP address), at specific time intervals. From December 15, 2018 to June 23, 2019, the data collection happened four times a day: at 6am, 12pm, 6pm and 12am. From June 24, 2019 to December 30, 2019, the collection has happened 12 times a day, at even hours (2am, 4am, etc.). Theoretically, this leads to 3,032 expected measurement moments. Given that sometimes hardware and software fail, we have a 95% collection success rate for the majority of the candidates.

Other Auditing Variables

It is fair to ask whether the data collected in the above-described way is what would be collected in any everyday situation, thus, we discuss below some variables that might be affecting the auditing process.

Does Location Matter? If you have ever googled for phrases such as "weather" or "pizza", you are familiar with the concept of search results personalization based on location. What if Top stories are different in different locations? To test this hypothesis, we replicated our data collection in a second, random location (1000 miles apart from ours), using the same identical setup (but two different computers). From Dec 29, 2019 to Jan 3, 2020, we collected data at the exact same time, several times during the day, for 30 candidate names. At the end, we had 70 time moments and 976 pairs of Top stories. Our comparison analysis indicated that 100% of our Top stories pairs have complete overlap, something that was not true for the organic search results. Although we cannot claim that this will be the case for every location, given that this location pairing was chosen randomly, the test provides some assurance that the location in which the data was collected does not introduce variability in our dataset when it comes to Top stories.

Query Formulation. Our searches use as queries the candidates' full names (first and last name), e.g., Bernie Sanders. Given the known preference of users for short query phrases, it is likely that many users search with last names, such as Sanders or Klobuchar. Do the Top stories results differ in that case? For a period of five months (July 27, 2019 - Dec 15, 2019), we collected search result pages using both the full names and the last names of the top candidates. We then compared the two sets of the unique URLs extracted from the two sets of pages by calculating the Jaccard coefficient to measure the similarity of the sets. Andrew Yang has the highest similarity, 0.81, while Donald Trump, the lowest, 0.38. While for Kamala Harris, who has the second lowest score, 0.42, the diversity of results is due to her very common last name, which attracts stories about other people too, that is not the case for President Trump. We compared the number of unique news sources for each set and found 315 sources for the full name and 275 sources for the last name. All 275 sources were included in the bigger set. What this shows is that the competition for the name "Trump" is more fierce than for "Donald Trump," and the Top stories algorithm is even more selective.

In this paper, we focus our analysis only on the dataset that was collected using full names of the candidates. We are continuing to collect top stories for last names only through 2020. Our recent research on voter's query formulation (Mustafaraj, Lurie, and Devine 2020) has found that they often don't use the names of candidates in isolation, but write embedded queries like "trump impeachment" or "warren dna". We are considering using Google Trends to gather such phrases in order to expand our data collection.

Device Type. Our dataset was collected using a standard laptop, but users' behavior has changed over the past years, with more people using mobile devices to read news online. The most recent Pew study[10] on this question indicates that 57% use mobile devices to read news compared to 30% using a desktop/laptop. This move toward mobile devices shapes how platforms present information. Initially, Google created and introduced the Top stories element for mobile devices[11] and only ten months later included it on desktop devices.[12] Additionally, Google continues innovation on the design of Top stories for mobile devices often including 2-3 panels of top stories with different topical headings. Given these changes in user practices and Google Top stories reorganization, future work needs to implement ways to perform automated data collection for mobile devices as well.

Dataset Statistics

In total, we collected Top stories for 30 candidates in the 2020 US presidential election, including 28 Democratic candidates and 2 Republican candidates. By the end of the data collection process in December 2019, 17 of these 30 candidates were still running. All together, 2,168 unique news sources were included in the dataset with one or more articles about at least one of the 30 candidates. 79,903 of those articles had a unique URL, which we used to determine unique stories. Counts are summarized in Table 2.

Distribution of Articles by Candidate

Table 1 contains statistics on the story counts per candidate for the complete dataset collection. The Total Moments column provides the total number of distinct times (day and

[10] https://www.pewresearch.org/fact-tank/2019/11/19/americans-favor-mobile-devices-over-desktops-and-laptops-for-getting-news/

[11] https://searchengineland.com/amp-top-stories-now-live-243314

[12] https://searchengineland.com/google-replaces-news-box-top-stories-desktop-264993

Table 2: Statistics on the data collected. `Candidate Status` indicates whether the candidate was still in the presidential race on January 9, 2020.

Data Type	Amount
Candidates	Democratic (28), Republican (2)
Candidate Status	Running (17), Dropped (13)
Unique Sources	2,168
Total Story Collection	588,112
Unique Stories	79,903 (13.59%)
Unique Stories with Candidate First or Last Name in Title	76,782 (88.93%)

time) that our method extracted Top stories for that candidate from the search page. As is often the case with long-running measurements, we didn't have a 100% success rate of data collection. On some occasions, technical difficulties (hardware shutdown, internet disruption, Google blocking) prevented the collection. In others, we had not yet included the candidate in our list of names to search, because we did not know about them. Additionally, Table 1 reports numbers for a "No Top Stories" state, which occurred when the search page didn't contain an element with the stories. This happened more frequently for candidates who received less media coverage, but occasionally for popular candidates too, most likely due to Google's A/B tests for search results. The table calculates the missed collection percentage using the expected number of date/time moments in which the collection should have happened (3,032 moments).

In Table 2, we present more general statistics about the dataset. Note that we also distinguish between unique stories across all candidates and the aggregate number of unique stories per candidate. In the latter, we count unique article URLs for each candidate separately. Since occasionally the Top stories panel does not provide candidate-specific articles, we double-count identical articles for different candidates, for the purpose of making candidate-level comparisons. For instance, on June 6, 2019, the article "2020 Democratic candidates vow to undo Trump's immigration actions"[13] was a Top Story for Bernie Sanders, Jay Inslee, and Julian Castro. In Table 1, there are a total of 86,335 unique stories after aggregating over all candidates. We use this total number to calculate the number of unique stories with candidate's names in the title (88.93%). Finally, because we collected data every few hours, the Total Story Collection refers to all occurrences of the unique news stories in the collection. For many less-covered candidates, a story remained for several weeks in Top stories; however, for a candidate like Trump, stories were replaced every few hours. This explains why Trump has double the amount of stories (16,956) compared to Joe Biden, with 8,536 stories.

Licensing: Same Article, Different News Source

1,509 of the collected Top stories were identical articles, identified by having identical titles, but published by different sources. Thus, the number of unique articles counted by having a unique title differs from the number of unique articles counted by having a unique URL. For example, an article for Bernie Sanders, "Bernie Sanders faces questions about political future" from AP News[14] differs from the same article published by Star Advertiser[15] only by their URL. This is the result of news licensing, which makes it possible for less-resourced news outlets to publish articles distributed by news agencies like AP news, Reuters, etc. For our analysis, we consider licensed articles to be distinct when published by different sources, due to our focus on the choice of news sources by Google's algorithm.

Analysis

There are many questions we can answer with this dataset, but for the moment we are focusing on two research questions that examine two different kinds of preferences: a) does the Google Top stories algorithm prefer some news sources more than others; b) do news publications prefer some candidates over others? We explore these two questions in the following.

RQ1: Which News Sources does the Top Stories Algorithm Prefer?

Of the 2,168 news outlets in our dataset, 741 (34.18%) were only represented by one top story for the entire data collection of all candidates. This suggests that these news sources are rarely favored by the Top stories algorithm and their appearance in the panel may be seen more as an exception. On the other hand, we found a heavy concentration of a few sources that published thousands of the articles that appeared in Top stories. We discuss in the following such prolific news sources.

News Source Tier Distributions To quantify the concentration of news sources, we ranked the 2,168 sources by the number of unique Top stories articles they published. We then separated the sources into three tiers: Upper Tier, Middle Tier, and Lower Tier. As shown in Table 4, sources from each tier cumulatively published approximately 1/3 of the total number of unique stories in our entire data collection. In other words, within our dataset of unique articles, the 8 most frequently featured news sources (0.37%) published about the same number of stories as the 2,112 least frequently featured news sources (97.42%). Figure 2 presents a box plot distribution of news sources presence in each tier.

What is known about these news sources? For each news source in the Upper Tier and Middle Tier, we gathered information on their year founded, main medium of publication, region of distribution, owner, Alexa Ranking, and the Partisan Audience Bias score (see (Robertson et al. 2018)). Details for the Upper Tier sources are summarized in Table 3. As one might expect, these sources are national media outlets, with the exception of the Washington Post and the NY Post. At first glance, there seems to be great variety in the

[13] https://www.latimes.com/politics/la-na-pol-democratic-presidential-candidates-2020-pasadena-immigration-20190531-story.html

[14] https://www.apnews.com/492900cdfa464d35b9e154e13c368de2

[15] https://www.staradvertiser.com/2019/01/11/breaking-news/bernie-sanders-faces-questions-about-political-future/

Table 3: Comparing the 8 Upper Tier News Sources. Region indicates whether sources are only locally vs. nationally distributed by their Main Medium. Partisan Audience Bias [-1,1] (see (Robertson et al. 2018)) assigns -1 to a purely democratic audience and +1 to a purely republican audience. The scores in between show a mixed audience.

Tier Rank	Publisher	Year Founded	Main Medium	Region	Owner	Alexa Ranking	Partisan Audience Bias	Unique Articles Count
1	The Hill	1994	newspaper	national	News Communications Inc: Capitol Hill Publishing	367	-0.06	5,423
2	CNN	1980	cable TV	national	AT&T: WarnerMedia	25	-0.12	4,392
3	Fox News	1996	cable TV	national	Fox Corporation: Fox News Group	66	0.61	3,817
4	Politico	2007	newspaper	national	Capitol News Company	364	-0.19	2,871
5	Washington Post	1877	newspaper	local	Nash Holdings	68	-0.23	2,600
6	Washington Examiner	2005	magazine	national	Clarity Media Group: Media DC	1194	0.54	2,493
7	NY Post	1801	newspaper	local	News Corp	177	0.18	1,796
8	NY Times	1851	newspaper	national	The New York Times Company	29	-0.26	1,712

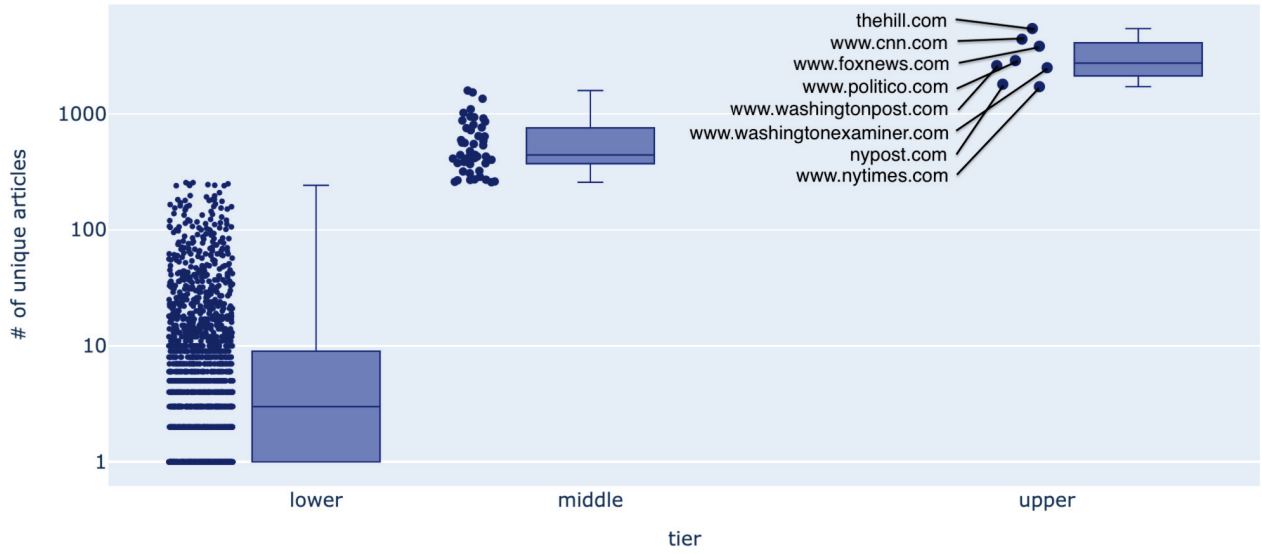

Figure 2: Three box plots for the number of unique articles from sources in the lower, middle, and upper tiers. We annotate data points, corresponding to sources in the upper tier. Notice the logarithmic scale for the Y axis. It is clear how the majority of news sources (lower-tier) are sparsely represented in the dataset, with their median value of about 5 stories over a one-year data collection.

Table 4: Statistics on each of the three News Source Tiers. Within each tier, sources cumulatively published approximately 1/3 of the unique Top story news articles we collected for all 30 candidates.

News Source Tier Level	# Sources	% Sources	# Articles	% Articles
Upper Tier	8	0.37%	25,104	31.42%
Middle Tier	48	2.21%	28,077	35.14%
Lower Tier	2,112	97.42%	26,722	33.44%
Total	2,168	100%	79,903	100%

political leanings of the sources. Perhaps surprisingly, there is also a large range of Alexa Rankings (a measure of web traffic), with 4/8 of the sources ranked in the top 100, 7/8 of the sources ranked in the top 400, and one source ranked lower than 1000. Below, we discuss these news source characteristics in more details.

Alexa Rankings Alexa Rankings[16] are computed by the web traffic analysis company Alexa Internet, Inc. For each domain on the web, the Alexa Rank computes its daily number of page views and unique visitors over the course of 3 months.[17] It then creates a ranking by comparing this number across all domains on the web. We provide the country-specific Alexa Ranking for the United States, which measures how a website ranks within this country.

The rankings provided are from December 30, 2019; however, Alexa Rankings are recalculated every day, so rankings may change from day to day. Additionally, sources with the same base domains are assigned the same Alexa Rankings, although we distinguish them as unique sources

[16] alexa.com/siteinfo
[17] alexa.com/about

in our dataset. For example, Yahoo (`yahoo.com`), Yahoo Finance (`finance.yahoo.com`), and Yahoo News (`news.yahoo.com`) are assigned the same Alexa Ranking. It is also important to remember that the Alexa Rankings are computed relative to all websites, not just news-specific websites, making some of the rankings unsuitable for source-to-source comparison.

Nevertheless, the Alexa Ranking can further characterize the top sources in the dataset. By considering the web activity of all global users in producing a ranking, Alexa Ranking enables us to notice possible discrepancies between the Top stories algorithm's news source preference versus actual web users' preferences. For the Upper Tier sources, The Hill, Politico, and Washington Examiner, have Alexa Rankings that are noticeably lower than the other sources. In fact, about 40% of the Middle Tier sources have Alexa Rankings higher than these 3 sources.

We conducted a Wilcoxon signed-rank test to compare Alexa Ranking to Tier Rank for (56) upper and middle tier sources from the dataset. With z=0.0367 (p-value=0.485) for a one-tailed test, we fail to reject the null hypothesis that the two medians are the same. Hence, even though they represent different measures to characterize sources in our dataset, the two rankings are consistent with one another. On one hand, this result indicates that overall, the most visited news websites on the web are also the ones from which Top stories draws the most often. On the other hand, the presence of Washington Examiner, a lowly-ranked website in terms of web traffic (Alexa rank = 1194), at position 6 of most frequent news sources in Top stories, indicates that there are other factors in place that allow for a non-popular news source to be ranked quite highly. Discovering such hidden factors remains an open research problem for the auditing community.

Partisan Audience Bias (Robertson et al. 2018) created the Partisan Audience Bias (PAB) dataset[18] from website links shared by real users on Twitter. Using voter registration records of US citizens with republican and democratic affiliations, the study identified 519,000 Twitter accounts matching these citizens. Over a certain period of time, a set of 113 million tweets by these accounts were collected and filtered to consider only the tweets with URLs. The URLs were processed to extract the second-level domain names. For example, `http://www.bbc.com/news/business-38686568` was converted into `bbc.com`. Then, to reduce noise, since 63% of links were shared only once, the authors kept only the domains that were shared more than 50 times (by different users). This led to a dataset of 19,022 sites. For each site, the authors calculated a bias score between -1 (a site shared only by democratic voters) to +1 (a site shared only by republican voters). Sites that get a bias score between -1 and +1 were shared by a mix of democratic and republican voters. For example, *The Wall Street Journal* had a score of 0.0106, signaling that it is a news source shared almost equally by both sides.

[18]Available at https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/QAN5VX

The PAB dataset contained the PAB score for 1,509 out of 2,168 Top stories news sources. Figure 3 shows their distribution across the political spectrum. One can observe that the majority of sources (ca. 80%) are right-of-center and left-of-center, with fewer sources in the extremes.

One thing to notice in the PAB scores in Table 3 is that two of the three right-leaning sources are the most partisan sources out of all eight Upper Tier sources. Although two of those three sources, *Fox News* and *NY Post*, appear to have different owners, a closer look into the Fox Corporation and News Corp show that they are actually both owned by the Murdoch Family.

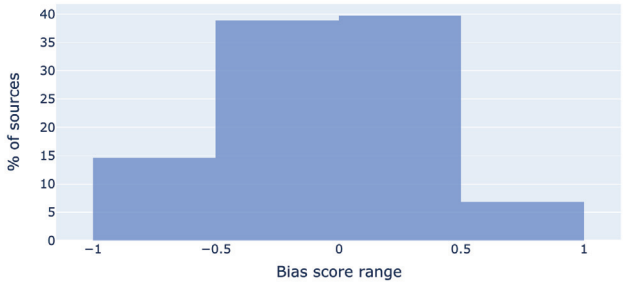

Figure 3: Distribution of Partisan Audience Bias scores for 1,509 sources in the Top stories dataset. Almost 80% of sources are equally divided in right-of-center and left-of-center, with smaller portion as far-right and far-left.

Relationship between news production and the algorithm One might explain the Top story tiers by arguing that some news organizations have more resources to publish a large number of articles than others, making their articles more likely to be selected by the Top stories algorithm. That is, the more stories an outlet produces, the more often their articles will show up in Top stories. This is a difficult hypothesis to test, because it requires access to the entire output of each news source. For the purpose of this paper, we only collect the entire news story output for two publications, *The New York Times* and *Politico*. In the following, we provide our findings.

- **The New York Times**. We used the New York Times Article Search API[19] to collect articles for 11 top candidates for the period Sep. 15 - Dec. 30, 2019. Then, we found the proportion of news stories returned by the NYTimes API to those present in our Top stories dataset during the same time period. Averaged over all the candidates, the proportion is 11% (std=5%). This number appeared to be low, thus, we explored the API results further. We learned that the API will return an article that has a mention of the candidate name anywhere in the body, even if the article is not predominantly about the candidate. When we restricted the number of the considered articles to include only the articles that contained the candidate name in the headline or article snippet, the proportion changed to the average

[19]https://developer.nytimes.com/docs/articlesearch-product/1/overview

of 78% (std=10%). That is, an article is most of the time chosen if the candidate name is in the title, and only rarely if that is not the case. One thing is clear, Google is not relying on the meta-information or the APIs of publishers to select the stories, as indicated by the large discrepancy between what NYTimes returns about a candidate and what Google chooses. Additionally, we found that on average, 12% (std=11%) of the chosen articles had the candidate name somewhere in the body, which means that not having a name in the title is not a reason for exclusion.

- **Politico.** Using Python scripts, we scraped articles on the category of politics[20] published by *Politico* between Sept 15 - Dec 30, 2019. In order to identify relevant articles, we considered articles that contain candidate names in the headlines (a total of 1,739 articles). Then, we compared them to stories that contain candidate names in the headlines from the Google Top stories dataset. Averaged over all the candidates, the percentage of Politico's articles from that period that appear in the Top stories of our dataset is 67% (std=35%). As expected, from Politico's articles without the candidate name in the headline, only 3.8% appear in Top stories.

These two examples suggest that the Top stories algorithm is able to read the content of an article and decide with some confidence whether to categorize it under a candidate name or not, instead of relying on meta-information from publishers. Further research could repeat this analysis for more sources to establish the connection between news production and algorithmic editorial choices of Top stories.

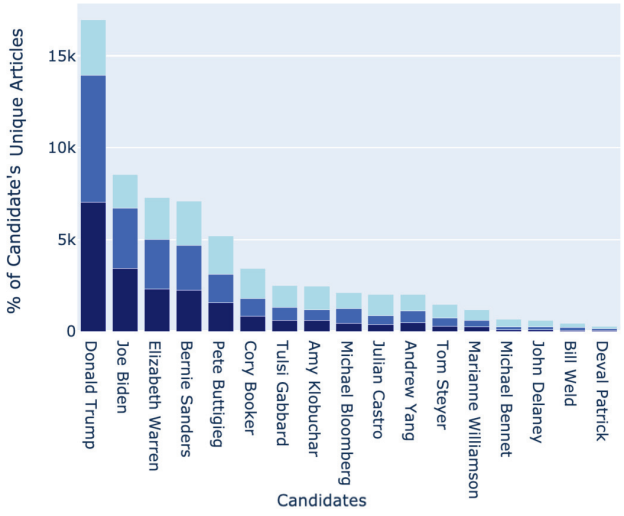

Figure 4: Each candidate's unique Top Story news articles, segmented by the proportion published by sources within each News Source Tier. Upper Tier, Middle Tier, and Lower Tier sources color-coded as dark blue, blue, and light blue, respectively.

[20]https://www.politico.com/politics

RQ2: Which Presidential Candidates do the News Sources Prefer?

It is common for political candidates to lament that they are not receiving enough coverage for their campaign from the media.[21] To what degree are their concerns of media coverage bias justified?

Volume of Media Coverage through Top Stories Some candidates are receiving more media coverage than others, as reflected through the varying amounts of unique top stories that each candidate received (refer to Table 1). How is the popularity of candidates associated with the types of sources that publish about them? In Figure 4, we segment the candidates' unique story counts by the percent published by upper, middle, or lower tier sources. At more than 40%, upper tier sources account for a disproportionate amount of Donald Trump and Joe Biden's articles, while the same sources account for 33% of Elizabeth Warren and Bernie Sanders' stories. On the other hand, candidates such as Cory Booker and Tulsi Gabbard have only 25% of their stories published by upper tier sources. Since upper tier sources account for 33% of the total number of unique stories, they appear to prefer publishing articles on some candidates more than others.

We also consider sources that frequently appear in the Top stories of each candidate. In Figure 5, we visualize the proportion of stories published by a given source within each candidates' dataset of unique Top stories, with candidates ordered by size of this dataset. As a result, the candidates in the top rows of the heat map have a larger percentage of new stories from most sources. Interestingly, Joe Biden's Top stories constitute 33.5% of the Breitbart News articles of our dataset. Additionally, 25.7% of Bloomberg's news articles appear in Michael Bloomberg's Top stories.

Stickiness of Top Stories We are also interested in quantifying how the freshness or stickiness of candidates' stories changes over the course of their campaigns. After key election events, do some candidates receive more new stories than others? For each day between January 1, 2019 and December 1, 2019, we calculated the cumulative number of new stories that each candidate query received. The results are shown in Figure 6. Candidates with steeper lines have less sticky top stories; in other words, their stories are switched out for new ones more frequently than candidates with flatter lines. In many cases, we can connect the stories' changes in stickiness to events that happened during the election. For example, when former Vice President Joe Biden and Mayor Pete Buttigig formally announced their presidential campaign on April 25, 2019 and April 15, 2019, respectively, there were sudden increases in the rate of new stories published about them. The four steepest lines correspond to the four front-runners as of January 2020.

One can consider the relationship between the stickiness and searchability, i.e. how frequently people search for the candidates. As a metric for searchability we consider Google

[21]https://www.theatlantic.com/politics/archive/2019/07/julian-castros-fight-attention/594679/

Figure 5: The ratio of articles from each source that appear in Top stories for candidates over the total number of articles by this source in our dataset. We consider 16 Democratic candidates who are still in the presidential race as of January 9, 2020 in addition to Republican candidate Bill Weld. As for the sources, we include top 30 with the largest article count. We can observe that from Breitbart News articles we see in Top stories, 33.5% appear when we search for Joe Biden.

Trends data. Figures 7 and 8 represent a side-by-side comparison of the two characteristics for candidate Kamala Harris. We can notice co-occurrence of spikes in both graphs, which often relate to events in the presidential campaign. Hence, further analysis of the relationship between the two characteristics is needed.

Future Research Questions

By providing longitudinal, year-long web data on the media coverage of 2020 presidential candidates, as selected by Google's Top stories algorithm, this dataset can provide opportunities to analyze the relationships between search engines, media, and democratic processes. Here we provide a few ideas on the kind of analyses this dataset could support for future research.

Media Agenda Setting

Previous research on search engine results pages during the 2016 US Presidential elections indicated that presidential nominees Hilary Clinton and Donald Trump received asymmetric media coverage on their agendas, with Clinton having a disproportionate amount of coverage on her scandals (Faris 2017). For the 2020 elections, on what aspects of the candidates' campaigns is the media focusing on? Are they talking more about the electability of women candidates or their political platforms? When they do write about less popular candidates, in what context do they do so? Although our dataset does not contain all stories written by all news sources, the fact that such stories are deemed as "Top stories" makes them nevertheless valuable in exploring how media sets the agenda for electoral coverage.

Balanced Inclusion of Partisan Sources

At least at the surface, it appears that the algorithm behind Google's Top stories is operating on the principle of the (now defunct) FCC Fairness Doctrine (Simmons 1978), which was a law in the United States from 1949-1987. According to this doctrine, media platforms (such as radio and TV broadcasters) had to give airtime to contrasting views on issues of public interest. That is, they needed to be neutral broadcasters and not take sides. Is Google trying to do the same with Top stories? Does its algorithm try to balance coverage by presenting stories from sources across the political spectrum? This dataset in combination with the PAB scores or another political bias metric can be used to consider questions related to the representation of left- and right-leaning political sources in the Top stories under a different light. Sources of what bias are primarily present overall and for each candidate individually? We already noticed a highly focused coverage by *Breitbart News* (a hyper-partisan, far-

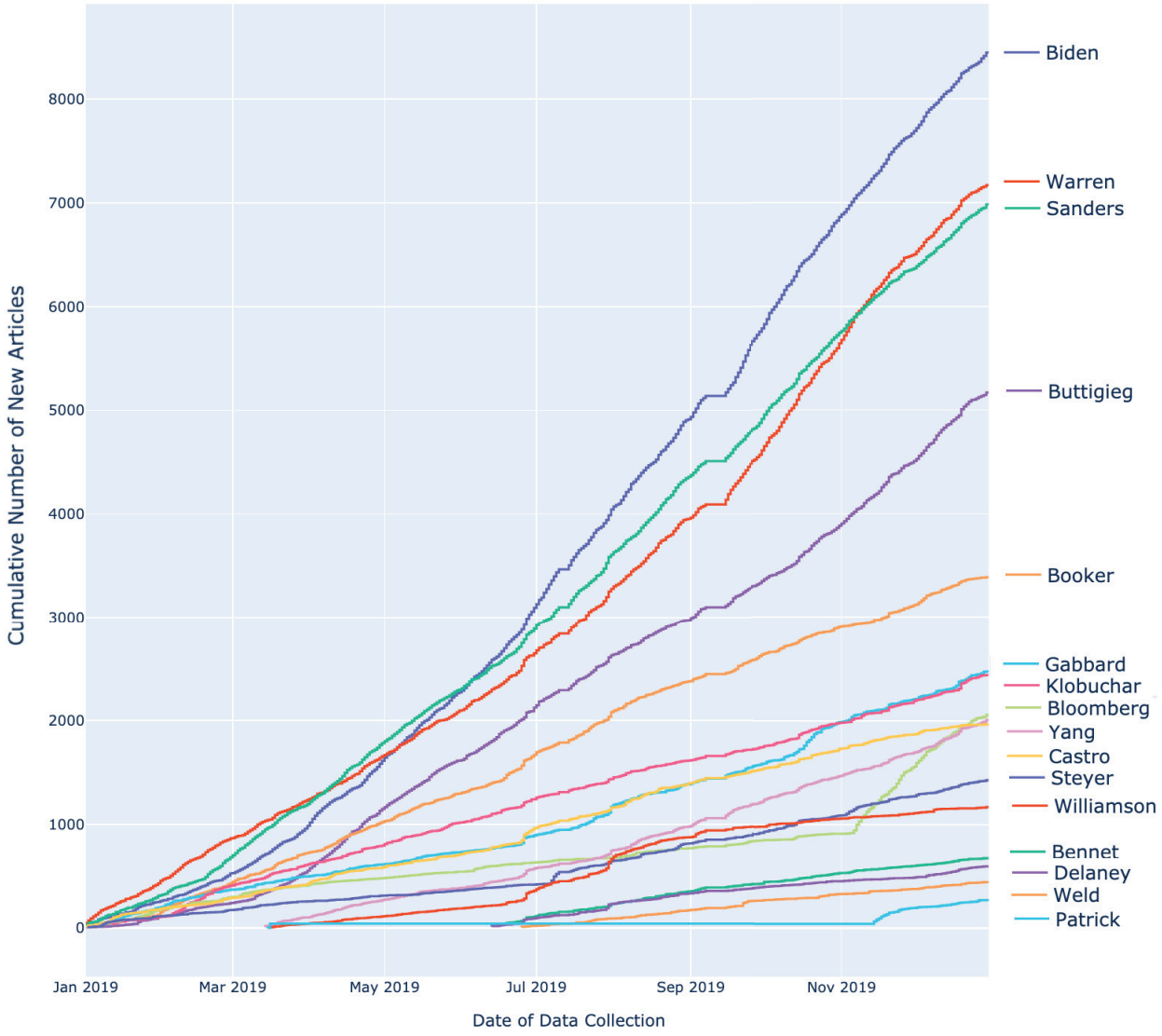

Figure 6: Trends in Top Story article "stickiness" over time. Low stickiness in a candidate's articles means that the Top Stories algorithm frequently replaces results for that candidate, as shown by steeper slopes in the graph (e.g. Joe Biden).

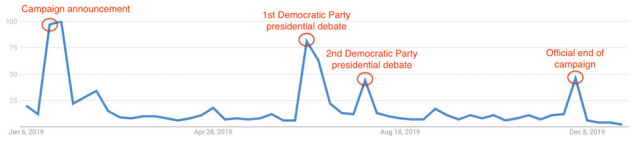

Figure 7: Google Trends plot representing search volume for Kamala Harris in 2019. We can observe spikes correlated to major events during her campaign, such as the running announcement, Democratic Party's primary debates, and dropping out of the presidential race.

right news source) on Joe Biden, but not on the other major candidates. As it is visible in Figure 3, there are fewer far-right (PAB ¿ 0.5) news sources than far-left ones. Do a few news sources like *The Washington Examiner* or *Breitbart News* get more often included in Top stories to compensate for this fact, while maintaining some overall balance of left and right?

Limitations

Despite our best efforts, this dataset is not perfect. It was collected on a single location, and although we claim location might not affect the composition of Top stories (based

Figure 8: The number of new/fresh articles in Top Stories for "Kamala Harris" query in each week of 2019. We can see that some (although not all) spikes/fluctuations in the frequency of Top stories updates also co-occur with presidential campaign events.

on a test we did with one random location), this claim needs to tested by systematically collecting data on multiple locations. We do not however claim that these Top stories are what all users might see, our aim was to remove variability introduced by user preferences and focus on the baseline choices of the algorithm. Finally, to characterize the overall media coverage of the candidates, access to all published news stories is needed. This might be possible by relying on databases like GDELT,[22] which we leave to future work.

Conclusion

This paper presents a novel dataset of news headlines about the candidates in the 2020 US Presidential Election, algorithmatically curated by Google's Top stories product and covering a one year period. Our major finding is that a very small percentage of news sources (2.6%) are responsible for 2/3 of the headlines in Top stories. A close analysis of the top eight sources that produce 1/3 of the stories indicates their political variety with some of them considered as center, slightly left-leaning, or strongly conservative (or far-right). Future analysis can look into how different candidates are treated by outlets that advance a certain worldview. Figure 5 already indicates that Breitbart was focused on Joe Biden, and Wall Street Journal on Elizabeth Warren.

Another finding is a confirmation of the candidates' worry that they don't get sufficient media coverage. Since early on in the race, the focus was on four candidates (Biden, Warren, Sanders, and Buttigieg), who by January 2020 were also the front runners in the race. While we cannot assign a causality direction to this observation, it is nevertheless important to point out. We hope that the public release of this dataset will lead to more research on algorithmic news curation and media agenda setting.

Acknowledgments

We are very grateful to Emma Lurie and to the Wellesley Cred Lab members for their continuous support. This project was partially funded by the National Science Foundation, through grant IIS 1751087.

[22] https://www.gdeltproject.org/

References

Bandy, J., and Diakopoulos, N. 2019. Auditing news curation systems: A case study examining algorithmic and editorial logic in apple news. *arXiv preprint arXiv:1908.00456*.

Diakopoulos, N.; Trielli, D.; Stark, J.; and Mussenden, S. 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.

Metaxa, D.; Park, J. S.; Landay, J. A.; and Hancock, J. 2019. Search media and elections: A longitudinal investigation of political search results. *Proc. ACM Hum.-Comput. Interact.* 3(CSCW).

Metaxas, P. T., and Pruksachatkun, Y. 2017. Manipulation of search engine results during the 2016 us congressional elections.

Mustafaraj, E.; Lurie, E.; and Devine, C. 2020. The case for voter-centered audits of search engines during political elections. In *Proc. of the 3rd Conference on Fairness, Accountability, and Transparency*, FAT* '20, 559–569. New York, NY: ACM.

Robertson, R. E.; Jiang, S.; Joseph, K.; Friedland, L.; Lazer, D.; and Wilson, C. 2018. Auditing partisan audience bias within google search. *Proceedings of the ACM on Human-Computer Interaction* 2(CSCW):148.

Robertson, R. E.; Lazer, D.; and Wilson, C. 2018. Auditing the personalization and composition of politically-related search engine results pages. In *Proceedings of the 2018 WWW*, 955–965.

Sandvig, C.; Hamilton, K.; Karahalios, K.; and Langbort, C. 2014. Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry* 22.

Simmons, S. J. 1978. *The fairness doctrine and the media*. University of California Press.

Thurman, N.; Moeller, J.; Helberger, N.; and Trilling, D. 2019. My friends, editors, algorithms, and i: Examining audience attitudes to news selection. *Digital Journalism* 1–23.

Trevisan, F.; Hoskins, A.; Oates, S.; and Mahlouly, D. 2018. The google voter: search engines and elections in the new media ecology. *Information, Communication & Society* 21(1):111–128.

Trielli, D., and Diakopoulos, N. 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*, 453. ACM.