Road to the White House: Analyzing the Relations between Mainstream and Social Media during the U.S. Presidential Primaries

Aaron Brookhouse∗
Michigan State University
brookho8@msu.edu

Tyler Derr∗
Vanderbilt University
tyler.derr@vanderbilt.edu

Hamid Karimi∗
Utah State University
hamid.karimi@usu.edu

H. Russell Bernard
Arizona State University
asuruss@asu.edu

Jiliang Tang
Michigan State University
tangjili@msu.edu

ABSTRACT
Information is crucial to the function of a democratic society where well-informed citizens can make rational political decisions. While in the past political entities primarily utilized newspapers and later radio and television to inform the public, the political arena has transformed into a more complex structure with the rise of the Internet and online social media. Now, more than ever, people express themselves online while mainstream news agencies attempt to utilize the power of the Internet to spread their articles as much as possible. To grasp the political coexistence of mainstream media and online social media, in this paper, we analyze these two sources of information in the context of the U.S. 2020 presidential election. In particular, we collected data during the 2020 Democratic Party presidential primaries pertaining to the candidates, and, by analyzing this data, we highlight similarities and differences between these two main types of sources, detect the potential impact they have on each other, and understand how this impact relationship can change over time.

CCS CONCEPTS
• Networks → Social media networks.

KEYWORDS
2020 Presidential Election, Mainstream News, Online Social Media

1 INTRODUCTION
The principal characteristic of a functional democratic society is to have well-informed citizens who are expected to gain useful and accurate information to make political decisions in a rational manner [33]. A natural question that then arises is where do citizens acquire information? In the past, the primary source of information was traditional unidirectional media such as newspapers or television, where “ordinary people” mainly were information consumers without access to a proper avenue to reflect their opinions. With the advent of the Internet in general and online social media in particular, however, we are now faced with a different and more complex scenario [9, 47]. On the one hand, mainstream news agencies have exploited mainly the power of the Internet and fully expanded their news diffusion operations [34]. But, on the other hand, thanks to readily accessible online social media platforms such as Twitter, a massive number of people can express their political opinions online [21, 42]. Further, the impact of a news source is beyond its readers. For example, online social media is shown to make individuals’ political opinions farther apart from each other, yet they are also linked to individuals with more exposure to the beliefs of people on the other side of the political spectrum [17]. Moreover, ideally, the news should only serve as a source of information that accurately reports the events of the world. However, in reality, it influences what topics people are interested in, how the topics are thought about, and, perhaps follow some sort of agenda [35], which is an area of research that has existed long before modern social media [3, 4, 38] and is commonly known as agenda-setting in communications research. It is essential to be aware of this when consuming any media.

As briefly discussed above, these two major sources of information profoundly impact the way people’s political ideology is shaped. However, in addition to this effect on people and politics, these two sources of information are not disjointed and have a complex relationship. More specifically, these two main sources of information (i.e., mainstream news outlets and online social media) interact and influence each other. Hence, a thorough analysis of the symbiosis of online social media and mainstream news is highly desired, and recent efforts have focused on this in other domains such as medical distrust [20]. This analysis is the topic of this paper and is discussed in the following.
To analyze news sources, we focus on a crucial political event: the 2020 U.S. presidential primaries/election. Hence, in this paper, we will explore the relations between different sources of information under the context of the 2020 U.S. presidential primaries/election from various perspectives. These sources include mainstream news, Twitter, and Google Trends. First, we perform an analysis of the data distributions and how their time series are correlated. Second, we investigate the correlations among the candidates in both news and Twitter data to better understand how the candidates are related in the two forms of media. In particular, on a given platform, we determine the mention frequency for two candidates. This helps us understand how candidates may have been perceived differently on different platforms. Third, we examine cross-media influence where we identify how one source influences the other. Since the interaction between sources is crucial, we further investigate the influence between Twitter and mainstream media from a causal perspective where we attempt to predict the trend in one source using the other and vice versa. Fourth, we strengthen our cross-media influence measure using the Granger Causality Test to quantify this relationship and test if one source adapts or reacts to changes in the other source. Fifth, we analyze how topics of interest are different across platforms and candidates. This helps determine what topics or issues are most deemed related to a given candidate on a platform. Finally, we investigate this association (i.e., candidates and topics) via both a candidate’s own posts and the public’s discussions about him or her.

Our main contributions are as follows:

- We present a comprehensive analysis of the two major information sources that affect the 2020 U.S. presidential primaries/election by focusing on Twitter and several mainstream news agencies.
- Each of our described analyses is presented from two perspectives with two goals. They provide direct insight into the candidates themselves as they relate to the news, public opinion, and the 2020 election. We also aim to highlight the differences between news and public opinion from a candidate-agnostic standpoint.
- We utilize methods to quantify nuanced ideas such as influence, causality, and topic mismatch between sources.

The rest of the paper is organized as follows. Next, in Section 2, we discuss related work. Section 3 provides details on our data collection process and provides an overview of the five analyses performed in the later sections of this work (i.e., Sections 4 through Section 8). Finally, in Section 9, we present a summarized conclusion of our findings and discuss future directions from this investigation.

2 RELATED WORK

Our work is related to streams of work related to investigating bias in media, social media’s impact on elections, and agenda-setting. In fact, many works have investigated media bias and how to detect it [19]. There has also been much research on search bias, which is built on the idea that when searching online, the results returned can be biased due to the data or the algorithms, including related to political science topics [27]. In regards to social media’s impact on elections, works have shown that Twitter commonly has spikes of conversation about topics recently covered in the news regarding presidential elections [46] and have even attempted to predict election outcomes based on social media [42]. There are also many works analyzing Twitter with respect to specific elections, such as one that analyzes the 2013 general election in Malaysia and how social media had an impact on the results [39], a similar work that analyzes the 2011 Singapore presidential election by using Twitter data [6], Twitter analysis of the 2012 South Korean presidential election [2], and an analysis of how members of the House of Representatives tweeted leading up to the 2012 U.S. election [15]. Regarding agenda-setting, this direction of research has long existed before modern social media [3, 4, 38], but has also recently been studied in online platforms such as Facebook [40] and Twitter [8, 40].

This is also related to “fake news,” [23, 24] which has had a heightened prevalence recently [30] and its impact on the 2016 presidential election has also been studied [1]. One concern is that it has been shown that fake news on Twitter can travel up to six times faster than real stories, and it is humans, not bots, who are the primary spreaders of this misinformation [45]. On the other hand, social media campaigns have had impressive results, shaping public discussions on many topics, and have had a large influence on public opinions as well as the news cycle [7, 28, 31]. It has also been shown through polls that people who align with a particular party are more prone to certain misbeliefs about recent events, and conservatives and liberals get their news from different sources [5]. In fact, most online news traffic is from people visiting the homepage of their favorite newspaper’s sites [17]. The final related area of work is that of agenda-setting [35], where there have been recent works focused on this topic related to today’s big data, and specifically online social media [16, 29, 44], as well as how to process social media streams in real time [46].

3 DATA

There are three main sources of data that we are analyzing: news, Twitter, and Google Trends. For the news, data were collected from Currents API,1 which provides information about articles such as title, URL, time published, and an excerpt of the article. To obtain the full texts of the articles, the Python Newspaper API was used.2

We collected the following news sources: Fox News, The American Conservative, The Washington Times, Associated Press, BBC News, CNBC, Reuters, The Hill, The Wall Street Journal, USA Today, ABC, NBC, CBS, CNN, The New York Times, The Washington Post, and Time. In addition, we intentionally collected a diverse set of news sources containing both conservative, liberal, and neutral sources, according to allsides.com. The Twitter data was collected using the Tweepy Python Streaming API.3 We downloaded tweets that mentioned any of the presidential candidates in real-time. We believe that our dataset offers sufficient representativeness because of a large number of collected tweets (i.e., over 160,000 daily tweets for our biggest user, 3,000 daily for our smallest). In addition, we collected Google Trends data directly through their online tool.4

We further process the Twitter and news data by extracting the following information: 1) the number of documents per day; 2) a

---

1https://currentsapi.services/
2https://newspaper.readthedocs.io/
3https://www.tweepy.org/
4https://trends.google.com/trends/
Table 1: Twitter and news data statistics.

| Candidate       | Elizabeth Warren | Bernie Sanders | Pete Buttigieg | Kamala Harris | Beto O’Rourke | Cory Booker | Andrew Yang | Amy Klobuchar | Tom Steyer | Donald Trump | Joe Biden |
|-----------------|------------------|----------------|---------------|--------------|--------------|-------------|-------------|---------------|------------|----------------|----------|
| **Twitter**     |                  |                |               |              |              |             |             |               |            |                |          |
| Average #TPD    | 53,286           | 127,213        | 31,938        | 16,808       | 3,086        | 4,526       | 37,585      | 12,020        | 4,194      | 160,905 (at 10%) | 138,923  |
| **News**        |                  |                |               |              |              |             |             |               |            |                |          |
| Average #APD    | 7.66             | 18.70          | 6.18          | 2.23         | 0.21         | 1.20        | 1.83        | 2.42          | 0.94       | 141.23         | 25.07    |
| Ratio           | 6,956            | 6,803          | 5,168         | 7,537        | 3,772        | 20,538      | 4,967       | 4,462         | 11,393     |                | 5,541    |

Figure 1: Visualizing the distribution of the documents related to each candidate per day.

The analysis presented in the rest of the paper (i.e., Sections 4 through Section 8) can be summarized as follows:

- Time Series Analysis (Section 4) - By plotting news, Twitter, and Google Trends data, it is clear that they show similar trends, particularly when interest for one candidate spikes rapidly in a short amount of time.
- Candidate Correlations (Section 5) - Correlations in Documents per day can show which candidates are often talked about at the same time. This can show how some candidates may be “lumped” together.
- Cross Media Influence (Section 6) - We will analyze how Pearson correlation changes when time series is lagged relative to each other. This can show when one media source is influencing or at least leading another one in trends.
- Granger Causality (Section 7) - Shows potential causality by seeing which time series is a better predictor of the other time series. The better one series can predict another, the more likely it is a cause.
- Topic Mismatch (Section 8) - Using a simple categorization technique, we demonstrate how some topics seem to have gathered more interest in either a candidate’s tweets, the news, or all of Twitter.

*For Donald Trump, only approximately 10% of tweets were saved given that his mentions were significantly more than the Democratic primary candidates. The real number of tweets per day is ten times this shown amount.*

point series of document publication times; and 3) political topics of the documents. In Table 1, we summarize the average amount of documents of each candidate per day, i.e., tweets per day (TPD) and articles per day (APD), for Twitter and news, respectively. One observation from Table 1 is that both Twitter and the news are quite variable in the number of documents appearing on average per candidate. Upon further investigation, we observed that the documents per day distributions were not following a normal distribution but had a long tail distribution in many cases. Thus, to better understand the Twitter and news document frequency, we visualize the data distributions in Figure 1. Note that different candidates follow different distributions of coverage between each other and between sources.
Table 2: Pearson Correlations between data sources for each candidate.

| Pairwise Relations | Elizabeth Warren | Bernie Sanders | Pete Buttigieg | Kamala Harris | Beto O'Rourke | Cory Booker | Andrew Yang | Amy Klobuchar | Tom Steyer | Donald Trump | Joe Biden | Avg. |
|--------------------|------------------|----------------|---------------|--------------|--------------|-------------|-------------|--------------|-----------|-------------|----------|------|
| News - Twitter     | .78              | .84            | .80           | .79          | .57          | .74         | .50         | .62          | .66       | .17         | .74      | .74  |
| News - Google Trends| .24              | .64            | .37           | -.01         | .47          | -.04        | .06         | .49          | .33       | -.02        | .30      | .26  |
| Twitter - Google Trends| .59             | .70            | .45           | -.07         | .21          | .01         | -.01        | .44          | .27       | .13         | .12      | .26  |

Figure 2: Candidates’ news, Twitter, and Google Trends data spanning the major points of the 2020 Democratic primaries. Vertical blue lines represent a candidate dropping out of the race, and red lines represent a Democratic primary debates.

4 TIME SERIES ANALYSIS

In Figure 1, we present the distributions of the Twitter and news data to observe the difference in trends per candidate in the two mediums. However, given that these data are inherently time series, in Figure 2 we also plot the candidates’ data over time from the three sources we have collected, namely news, Twitter, and Google Trends. Vertical lines are also plotted on the graph representing dates of importance. Red lines correspond to Democratic primary debates that took place over this time period. A number above a blue line corresponds to a candidate in the legend, indicating when that candidate dropped out of the race. Two candidates do not have a blue line on this chart, Beto O’Rourke, as he dropped out before the first date in this graph, and Joe Biden, as he did not drop out of the Democratic primary.

The first observation we make from Figure 2 is that news, Google Trends, and Twitter data follow a similar trend. For example, this is even more obvious when looking at the obvious spikes of Bernie Sanders, Joe Biden, and Elizabeth Warren. We observe that these are positioned around major points during the primaries, such as the seventh Democratic Party debate in the middle of January, Bernie Sanders’s early lead in February, Elizabeth Warren ending her campaign after unsatisfactory results in Super Tuesday along with Joe Biden taking the lead after receiving the endorsement of many candidates (such as Pete Buttigieg) who ended their campaigns before Super Tuesday, and finally Bernie Sanders’s ending his campaign in April of 2020. We decided to use Google Trends since they directly measure the frequency of search terms and are a reliable gauge of the public’s interest. Given that Google Trends shares similar trends with Twitter, the frequency of Tweets is justifiable to use as a representative measurement of the public’s interest. Others have also demonstrated that Google Trends and Twitter are good indicators of public interest [13]. To more precisely quantify and measure the level of similarity between the news, Twitter, and Google Trends, in Table 1 we present the pairwise Pearson correlation coefficients between these different sources at the candidate level and provide an averaged column across the candidates. The first thing to note is that, on average, news and Twitter are the two most correlated of the three data sources. The p-values for all news-Twitter correlations are significantly below 0.05, with the highest being Beto O’Rourke with a p-value of 4.27 x 10^-9. The next interesting thing is that while some candidates such as Amy Klobuchar and Bernie Sanders have relatively stable and higher correlations among all pairs, others such as Kamala Harris and Joe Biden do not. More specifically, we observe that both Kamala Harris and Joe Biden had a high correlation between news and Twitter, but at a near-zero correlation between Twitter and Google Trends (and very low for news and Google Trends as well). The highest correlation among all the three pairwise relations was Bernie Sanders. Last, we note that Donald Trump was significantly below average across all three pairwise data source correlations.

5 CANDIDATE CORRELATIONS

One type of influence that media can exert on an observer is to group candidates. This can impact how a viewer perceives the candidates. For example, if two candidates are often mentioned together as opposing each other, could a positive thought about one lead to a negative thought about the other? Alternatively, in the opposite case, two candidates often being mentioned together as agreeing may cause a viewer to think about one that they then
apply to the other automatically. This type of grouping of ideas decreases objectivity, as what one candidate does or says should not automatically impact perceptions of another candidate. To remain most objective, media sources should not consistently group candidates together. Instead, they should rely on information that pertains to each candidate separately without impressing upon the media viewer that two ideas are related. Also, it could potentially be a strategy to construct candidates designed to appeal to a certain smaller demographic and then by design have them later endorse (or oppose) another candidate.

We decided to measure this level of grouping of candidates by a media platform according to their correlations. Candidates were split into all possible pair combinations. Two figures were made for the candidate pair (X, Y), one for Twitter, and one for news. For each day, a data point is in the figure. The x-axis denotes the number of articles/tweets published about candidate X, and the y-axis denotes the same for candidate Y. The Pearson correlation is calculated for each pair of these correlations. For a case study, we chose the pair Elizabeth Warren and Bernie Sanders to visualize the difference in variance between Twitter and news because they show a strong correlation in both Twitter and news (seen in Figures 3a and 3b).

Some candidate pairs have stronger correlations than others—for some pairs, the correlation is stronger with the news data than with Twitter, while some instances are the other way around. Thus, for a comprehensive comparison, each pair of correlations is plotted in Figure 5, with the distribution of correlations plotted in Figure 4. For these correlation matrices, if the p-value for correlation were greater than .05, the correlation would be ignored and placed as a “*” in the heatmap. Figure 5, on the rightmost heatmap, shows the difference between the candidate correlations between news and Twitter, with blue indicating there is more correlation in the news and red indicating there is more correlation on Twitter. These averaged correlations are low, with 0.322 in the news and 0.267 in Twitter. This is because many candidates had very low correlation, while some had much higher (such as Elizabeth Warren and Bernie Sanders). While the news average correlation is higher by about 20%, a two-input t-test measures a p-value of .4, so there is not a significant difference in candidate grouping tendencies between news and Twitter sources overall (although we can visualize some differences in the distributions in Figure 4). Overall, this correlation can show which candidates are portrayed/perceived as related to each other. Here, we demonstrate that there is no tendency for news or Twitter to be more likely to group candidates than the other. The individual candidates may be grouped more by Twitter or news. However, this trend is not consistent among all candidate pairs. One future direction could be to model the positive and negative mentions of candidate pairs as a signed network [41] and calculate the node similarity between all candidate pairs [12].

6 CROSS-MEDIA INFLUENCE

In today’s world, every day, we have different forms of media influencing each other. This means that even if one is not directly observing a certain media platform, one is still likely susceptible to its influence on certain topics. For example, this suggests that people who do not use social media are still indirectly affected by what people say about the candidates. Also, people who do not read the mainstream news online are still indirectly impacted by what
it has to say about different candidates. Furthermore, it provides insights into whether public opinion (such as that found in social media) influences mainstream news, mainstream news influences public opinion or is bidirectional, and to what extent they influence each other. To do this, we quantify the influence of cross-media platforms.

More specifically, one way to determine which of two media sources is influencing the other more is to shift the data such that the time $t$ on one dataset will align with time $t - o$ of the other dataset, where $o$ is a defined offset in time. Then, to gain insights into the influence, we can vary the offset and calculate the Pearson correlation coefficient. In this way, the offset points at which the correlation coefficient is the greatest show where one source of media is influencing the other. If the coefficient, for example, is highest when data $x$ is leading data $y$, then it can be implied that data $x$ is influencing data $y$. We create heatmaps showing these correlations to better understand the effect over time and based on multiple offsets to analyze these relations. In Figure 6, we visualize this process and will next introduce it more formally.

First, we perform this investigation by binning the Twitter and news data into 5-minute bins denoted as $T = [T_0, T_1, \cdots, T_n]$ and $N = [N_0, N_1, \cdots, N_n]$, respectively, where we assume to have $n + 1$ 5-minute bins in our data. Note that in our data, the initial starting time is November 11, 2019, and the ending time is May 14, 2020. To better understand the potential influence between Twitter and news over time, we can localize two-week windows, i.e., $w = 2$ weeks, and obtain the correlation between them locally (as compared to the entire length of time), i.e., each row of the matrix in Figure 6. Then, we can run this window across our entire data over time, obtaining multiple views of the correlation (i.e., going from the top to bottom row of the matrix in Figure 6 with offset as 0 hours). While sliding the window of size $w$, we shift in intervals of $d = 12$ hours. However, it is also interesting to obtain the correlation when leading one data source before the other to uncover potential influence in the data. Hence, we utilize a set of offsets $O = \{-48hrs, -47hrs, \cdots, 48hrs\}$ that range from -48 hours to 48 hours in one-hour intervals, which are added to one of the data sources (e.g., news) while keeping the other one fixed (e.g., Twitter) (i.e., adding an offset is moving horizontally away from the matrix’s central column in Figure 6).

However, when attempting to zoom in to such a fine-grained view, it is likely that many hour intervals will not have news articles being created for some of the lesser discussed candidates. To alleviate this, we harness the exponential decay of the Hawkes Process (which is commonly applied to Twitter and news point series data [37]). This allows us to convert our discrete frequency series when tweets/articles are being created over time to a single continuous line through exponential decay. In this case, the publishing times of each tweet/article are fed into the Hawkes Process, and a
The Hawkes process helps smooth out these data that would otherwise be very oscillatory if values too quickly returned to zero. Below we present the Hawkes process:

\[
\text{Hawkes}(\text{news}(t)) = \sum_{i=0}^{\# \text{ events before time } t} e^{-r(t-t_e)}
\]

where \(i\) is the index of events (iterating from the first event to the most recent before time \(t\)), \(r\) is a decay constant calculated using the half-life being used, \(t_e\) is the time of event currently being summed, and \(t\) is time input to the Hawkes process. We let \(T^*\) and \(N^*\) denote the output of the Hawkes process over time when receiving \(T\) and \(N\) as input, respectively.

Thus, for a given two week window representing row \(i\) of the heatmap, that has a starting time \(s_i\), ending time \(e_i = s_i + w\), and offset \(o_j \in O\), we select the section of the Hawkes processed tweet sequence \(T^*(s_i, e_i)\) and news article sequence \(N^*(s_i + o_j, e_i + o_j)\) to discover the correlation \(h_{ij}\) for the cell \((i, j)\) in the heatmap, which is defined as follows:

\[
h_{ij} = \text{correlation}(T^*(s_i, e_i), N^*(s_i + o_j, e_i + o_j))
\]

which can be visualized in Figure 6 where the news articles are shifted according to offset \(o_j\) when aligned with the tweets for row \(i\). Note that the entire heatmap is filled in using the same process described above by varying \(i\) and \(j\).

Here we present the visualized heatmaps in Figures 7c, 7a, and 7d, for Andrew Yang, Donald Trump, and Elizabeth Warren. Note that to the left side of the figures, the news is leading Twitter, and on the right side, Twitter is leading the news. Looking at the figures from top to bottom follows the passage of time. Red indicates areas of higher correlation, white is no correlation, and blue is a negative correlation. These figures show how Twitter and news datasets relate to each other by visualizing how the correlation between them changes over time and with different lags.

Andrew Yang’s heatmap is shown in Figure 7c. It can be seen that there are high correlation sections and also sections that do not have a strong correlation no matter what the lag. Overall, it seems to be...
stronger correlations on the right side of the figure, indicating that on average, Andrew Yang’s Twitter base has a stronger influence on the news than the other way around. The darkest red spot on the heatmap is in mid-February when Andrew Yang dropped out of the race. Interestingly, this high correlation area is on the Twitter-leading side of the figure, indicating a large response to his announcement on Twitter before having a large news coverage response.

Donald Trump’s heatmap is shown in Figure 7a. There are segments of higher correlation with no shift, and also at exact increments of one day shifted. Donald Trump also has the most stable document publishing rates both on Twitter and news. This heatmap suggests a daily schedule of more popular times and less popular times that is repeated consistently and daily. This is why any shift seems to decrease the correlation quickly, and the correlation peaks are all in daily increments. It does not matter how far the data are shifted; as long as it is 24 hours, the daily schedule will line up, and the correlation will be at its maximum. This is supported by the fact that Trump’s Tweet frequency is unusually high compared to other presidents and has a tendency to dominate the news cycle.\(^6\)

Joe Biden’s heatmap, Figure 7b, shows a similar pattern. Elizabeth Warren’s heatmap is demonstrated in Figure 7d. We are not able to observe a clear trend as seen with Andrew Yang (i.e., Figure 7c), but it does seem to be leaning slightly in the news influence Twitter direction. There are many instances where the high correlation periods stretch far in both directions, indicating a period where the candidate is experiencing an unusually steady period of coverage. However, Warren has a few patches that exist only on the news leading side of the figure, indicating that overall news coverage seems to have a stronger influence on Twitter information than the other way around for Elizabeth Warren. We note that the dark red spot in April started around the time when she endorsed Joe Biden as the Democratic nominee.

7 GRANGER CAUSALITY

While the previous section analyzes areas of influence and how these areas change over time, it does not answer the more general question “Does news impact social media more, or vice versa?” The fine granularity of the previous analysis was helpful, but we perform a more general and grounded analysis.

More specifically, we harness the Granger Causality test [18] as a way to show potential causabilities between two sets of time series data, which has been heavily used for analyzing temporal data/events on Twitter [37]. The test is to see whether one time series (along with lags of that time series) helps predict the other time series. The Granger Causality is being applied to news and Twitter time series to explore the relationship between them. We ran the Hawkes process with a half-life equal to the average document period on each series to increase granularity. In other words, to establish a principled half-life value, if, on average, news articles were appearing every 1 hour, then this was the value used. The test was run with lags up to 24 hours for both news cause Twitter and Twitter cause news. In Figure 8 we show the resulting sum of squared residuals (SSR) based F-test results used in Granger Causality changes over the different time lags. We note that the average p-value for news influencing Twitter was 1.64 × 10\(^{-24}\), and the average p-value for Twitter influencing news was 1.53 × 10\(^{-12}\). Thus, although we have justification that Twitter does Granger-cause news and news does Granger-cause Twitter, we observe a significantly smaller p-value for news influencing Twitter, as well as a significantly higher F score. This suggests that according to this measure in our dataset, mainstream news has a larger impact on Twitter conversation than the other way around. While they both influence each other, in our dataset, we observe that overall, Twitter seems to influence the news less. However, we note that a more detailed analysis would be required here to provide a stronger claim.

Figure 8 also suggests that the influence of each of these news sources decreases rapidly. After five hours, it is just a fraction of what it was originally. Although news articles and Tweets are still read long after being published, Figure 8 shows an initial spike that one media source causes in the activity of the other happens within the first few hours.

8 TOPIC MISMATCH

As a campaigning strategy, U.S. presidential candidates usually concentrate their rhetoric around certain topics (e.g., Donald Trump on immigration [36] or Bernie Sanders on income and wealth inequality [14]). However, this does not mean that a particular medium perceives a candidate around the same topics. In other words, a candidate may be associated with some topics different from those he or she hinges on. This creates a mismatched perception of what is important to a candidate and how he or she is portrayed in the media. In this part, we attempt to quantify this topic mismatch for each candidate across three sources, namely Twitter, mainstream media news, and candidates’ Tweets.

First, we identify the topics associated with each candidate by using the list of topics curated by politico.com.\(^7\) Then, we convert each topic to a numerical representation using word2vec [32]. Similarly, each text in all three sources (i.e., Twitter, mainstream news, and candidates’ own Tweets) is converted to a numerical vector by taking the average of tokens’ word2vec vectors in the text. Next, we match each text to a topic. To do this, each text is matched to the closest relating topic whose vector has the highest cosine similarity to the text’s vector. To reduce the noise, we consider a minimum

\(^6\)https://www.washingtonpost.com/politics/2020/05/12/how-much-trumps-presidency-has-he-spent-tweeting/

\(^7\)https://www.politico.com/2020-election/candidates-views-on-the-issues/
cutoff threshold, and any text that did not meet this threshold in cosine similarity to any of the topics was discarded. The distribution of topics for general Twitter, the candidates’ Tweets, and mainstream news are shown in Figure 9.

Finally, we normalize each candidate’s data by calculating the percentage of the candidates’ documents in each category. This was done separately on each candidate to accommodate for the fact that some candidates had more documents than others. The distribution of topics for both general Twitter, the candidates’ Twitter accounts, and news are shown in Figures 9. In this figure, the darkness of each square represents the percentage of documents for the corresponding candidate that belongs to the corresponding topic. We make the below observations based on this figure.

First, we can observe some mismatches between the topics of interest for candidates themselves with those expressed on mainstream news and/or Twitter. Notably, for instance, gun control, while a topic of interest for Beto O’Rourke, is not a topic he tweets much about. This can be confirmed by his low value (bright color) for gun control in “Candidate Twitters”, while mainstream media (i.e., “news”) and Twitter have highly associated him with gun control. Next, for some topics, we see the consistency between different sources for all candidates. For instance, cybersecurity has not been expressed much by the candidates. Likewise, mainstream media and Twitter have hardly associated it with any candidate. We believe this is related to the nature of the topics as some topics are essentially less “controversial,” and consequently, less associated with the candidates.

Our last observation is that while comparing news and Twitter matching distribution in the first and second panel of Figure 9, respectively, in general, mainstream news shows more contrast. We attribute this more focused nature to news articles where a particular topic about a candidate is discussed in more depth compared to the informal and occasionally erratic Twitter posts.

9 CONCLUSION

In this paper, by exploring the influence that social media and the news have on each other, we offered a better understanding of how the different media sources interact with each other and their viewers when discussing the presidential candidates. In this regard, we visualized nuanced ideas such as cross-media influence, topic mismatching, and topic grouping through quantitative analyses. For example, the Granger Causality Test showed that news usually has a stronger impact on social media than social media has on the news. However, as the cross-media influence heatmaps revealed, this is not always the case for each candidate. Furthermore, we introduced several effective ways to illustrate different relationships between news, Twitter, and the candidates.

We believe our performed analyses contribute to the intersection of political science and online social media. For instance, it is exciting to demonstrate these ideas on the 2020 Presidential Primaries, while we believe these techniques could also be applied to analyze other political topics. For example, in the United States Congress related work has focused on the prediction and analysis of future votes [10, 11, 22] and we hypothesize that representative correlations would be related to their underlying social networks. Moreover, future work could analyze more social networks to explore if they exhibit similar or different patterns and explore how social media’s relationship with the news has changed over time by analyzing several years of data. Another interesting future work could expand on this by exploring why some candidates are correlated more on one platform or the other. Moreover, with proliferation of misinformation spreading across social media platforms [25, 26, 43], it would be interesting to investigate how online misinformation affect the discourse around a candidate and also whether misinformation’s effect permeates to mainstream news outlets.
ACKNOWLEDGMENTS

This research is supported by the National Science Foundation (NSF) under grant numbers IIS1714741, CNS1815636, IIS1845081, IIS1907704, DRL2025244, IIS1928278, IIS1955285, IOS107215, IOS2035472 and Army Research Office (ARO) under grant number W911NF-21-1-0198. We would also like to extend our special gratitude to Carole Bernard for revising this paper.

REFERENCES

[1] Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of economic perspectives* 31, 2 (2017), 21–36.

[2] Jung-Hwan Bae, Jiyeon Son, and Min Song. 2013. Analysis of twitter for 2012 South Korea presidential election by text mining techniques. *Journal of Intelligence and Information Systems* 19, 3 (2013), 141–156.

[3] Dan Berkowitz. [n.d.]. Who sets the media agenda? The ability of policymakers to determine news decisions. *Public opinion, the press, and public policy* 2 ([n.d.]), 81–102.

[4] Dan Bernhardt, Stefan Krasa, and Mattias Polborn. 2008. Political polarization and the electoral effects of media bias. *Journal of Public Economics* 92, 5–6 (2008), 1092–1104.

[5] Murphy Choy, Michelle LF Cheong, Ma Nang Laik, and Koo Ping Shung. 2011. A sentiment analysis of Singapore Presidential Election 2011 using Twitter data with census correction. *arXiv preprint arXiv:1108.5520* (2011).

[6] Derrick L Cogburn and Fatima K Espinoza-Vasquez. 2011. From networked to mainstream news: The information flow of the vaccine controversy in the digital era. *Science* 334, 6060 (2011), 259–262.

[7] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. *arXiv preprint arXiv:1310.4546* (2013).

[8] Terry Lee. 2019. The global rise of ‘fake news’ and the threat to democratic elections in the USA. *Public Administration and Policy* (2019), 1–23.

[9] Richard R Lau, Kathleen Rogers, and Jamel Love. 2020. Media Effects in the 2016 election: How racial and immigration attitudes, not economics, explain slumps in white voting. *Public Opinion Quarterly* 83, 1 (2019), 91–113.

[10] Marilyn Roberts and Maxwell McCombs. 1994. Agenda setting and political advertising: Origins of the news agenda. *Public opinion, the press, and public policy* 5, 2 ([n.d.]), 135–160.

[11] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359, 6380 (2018), 1146–1151.

[12] Hao Wang, Dogan Can, Abe Kazemzadeh, François Bar, and Shrikanth Narayanan. 2012. A System for Real-Time Twitter Sentiment Analysis of 2012 U.S. Presidential Election Cycle. In *Proceedings of the ACL 2012 System Demonstrations* (Jeju Island, Korea) (ACL ’12). Association for Computational Linguistics, USA, 115–120.

[13] Christopher Weare. 2002. The Internet and democracy: The causal links between technology and politics. *International Journal of Public Administration* 25, 5 (2002), 659–691.

[14] Nic Newman. 2011. Mainstream media and the distribution of news in the age of social media. (2011).

[15] Michael Parenti. 1995. *Inventing reality: The politics of news media*. St. Martin’s Press New York.

[16] Tyler T Renyi, Loren Collingwood, and Ali A Valenzuela. 2019. Vote switching in the 2016 election: How racial and immigration attitudes, not economics, explain slumps in white voting. *Public Opinion Quarterly* 83, 3 (2019), 91–113.

[17] Marjan-Andrei Rizoiu, Young Lee, Swapnil Mishra, and Lexing Xie. 2017. A tutorial on howeckes processes for events in social media. *arXiv preprint arXiv:1708.06401* (2017).

[18] Matt Egan. 2015. The Bernie Sanders plan to fix income inequality. https://money.cnn.com/2015/05/21/news/economy/bernie-sanders-income-inequality/2015/index.html Accessed: 2020-09-01.

[19] Elie Skogerb and Arne H Krumsvik. 2015. Newspapers, Facebook and Twitter: Intermedial agenda setting in local election campaigns. *Journalism Practice* 2 ([n. d.]), 350–366.

[20] Hamid Karimi, Courtland VanDam, Liyang Ye, and Jiliang Tang. 2018. End-to-end compromised account detection. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (ASONAM). IEEE, 314–321.

[21] Hamid Karimi, Tyler Derr, Aaron Brookhouse, and Jiliang Tang. 2019. Multi-factor congressional vote prediction. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 266–273.

[22] Hamid Karimi, Tyler Derr, Aaron Brookhouse, and Jiliang Tang. 2019. Multi-source multi-class fake news detection. In *Proceedings of the 27th international conference on computational semantics*. 1546–1557.

[23] Richard Karimi and Jiliang Tang. 2019. Learning Hierarchical Discourse-level Structure for Fake News Detection. In *Proceedings of NAACL-HLT*. 3342–3442.

[24] Hamid Karimi, Jiliang Tang, and Yane Li. 2018. Toward end-to-end deception detection in videos. In *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 1278–1283.

[25] Hamid Karimi, Courtland VanDam, Liyang Ye, and Jiliang Tang. 2018. End-to-end compromised account detection. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (ASONAM). IEEE, 314–321.

[26] Hamid Karimi, Courtland VanDam, Liyang Ye, and Jiliang Tang. 2018. End-to-end compromised account detection. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (ASONAM). IEEE, 314–321.

[27] Hamid Karimi, Courtland VanDam, Liyang Ye, and Jiliang Tang. 2018. End-to-end compromised account detection. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (ASONAM). IEEE, 314–321.
Visual-Meta Appendix

The data below is what we call Visual-Meta. It is an approach to add information about a document to the document itself, on the same level of the content (in style of BibTeX). It is very important to make clear that Visual-Meta is an approach more than a specific format and that it is based on wrappers. Anyone can make a custom wrapper for custom metadata and append it by specifying what it contains: for example @dublin-core or @rdfs.

The way we have encoded this data, and which we recommend you do for your own documents, is as follows:
When listing the names of the authors, they should be in the format 'last name', a comma, followed by 'first name' then 'middle name' whilst delimiting discrete authors with ('and') between author names, like this: Shakespeare, William and Engelbart, Douglas C.

Dates should be ISO 8601 compliant.

Every citable document will have an ID which we call 'vm-id'. It starts with the date and time the document's metadata/Visual-Meta was 'created' (in UTC), then max first 10 characters of document title.

To parse the Visual-Meta, reader software looks for Visual-Meta in the PDF by scanning the document from the end, for the tag @{visual-meta-end}. If this is found, the software then looks for @{visual-meta-start} and uses the data found between these tags. This was written September 2021. More information is available from https://visual-meta.info for as long as we can maintain the domain.

@{visual-meta-start}
@{visual-meta-header-start}
@visual-meta{version = {1.1},
generator = {ACM Hypertext 21},
organisation = {Association for Computing Machinery}, }
@{visual-meta-header-end}

@{visual-meta-bibtex-self-citation-start}
@inproceedings{10.1145/3465336.3475115,
  author = {Brookhouse, Aaron and Derr, Tyler and Karimi, Hamid and Bernard, H. Russell and Tang, Jiliang},
  title = {Road to the White House: Analysing the Relations Between Mainstream and Social Media During the U.S. Presidential Primaries},
  year = {2021},
  isbn = {978-1-4503-8551-0},
  publisher = {Association for Computing Machinery},
  address = {New York, NY, USA},
  url = {https://doi.org/10.1145/3465336.3475115},
  doi = {10.1145/3465336.3475115},
  abstract = {Information is crucial to the function of a democratic society where well-informed citizens can make rational political decisions. While in the past political entities primarily utilized newspapers and later radio and television to inform the public, the political arena has transformed into a more complex structure with the rise of the Internet and online social media. Now, more than ever, people express themselves online while mainstream news agencies attempt to utilize the power of the Internet to spread their articles as much as possible. To grasp the political coexistence of mainstream media and online social media, in this paper, we analyze these two sources of information in the context of the U.S. 2020 presidential election. In particular, we collected data during the 2020 Democratic Party presidential primaries pertaining to the candidates, and, by analyzing this data, we highlight similarities and differences between these two main types of sources, detect the potential impact they have on each other, and understand how this impact relationship can change over time.},
  numpages = {10},
  keywords = {2020 Presidential Election; Mainstream News; Online Social Media; News Source Correlation},
  location = {Virtual Event, USA},
  series = {HT '21},
  vm-id = {10.1145/3465336.3475115} }
@{visual-meta-bibtex-self-citation-end}

@{visual-meta-end}