Proceedings of the Fifteenth International AAAI Conference on Web and Social Media (ICWSM 2021)

# COVID-19 Coverage By Cable and Broadcast Networks

**Ceren Budak**[1], **Ashley Muddiman**[2], **Yujin Kim**[3], **Caroline C. Murray**[3], **Natalie J. Stroud**[3]
[1] University of Michigan
[2] University of Kansas
[3] University of Texas at Austin
cbudak@umich.edu, ashley.muddiman@ku.edu, yujin.kim@utexas.edu,
caroline.murray@austin.utexas.edu, tstroud@austin.utexas.edu

## Abstract

In this paper, we present a dataset of COVID-19 coverage by cable and broadcast news networks. Our dataset, which spans the time period between January 21, 2020 and June 12, 2020, includes 44,643 transcript paragraphs that are manually labeled according to their relevance to COVID-19 and 486,068 paragraphs that are further labeled using supervised classifiers. We further provide descriptive analysis that shows differences in the degree to which networks covered the pandemic and how the content of this coverage varied. Our distinctive phrase analysis also suggests that cable news networks, particularly Fox News and MSNBC, are politicizing COVID-19. This dataset can be leveraged to model and characterize the role cable and broadcast news networks play in shaping COVID-19 attitudes and behaviors, as well as how the coverage was related to external events (e.g. the number of COVID-19 cases), coverage in other media (e.g. newspapers), and COVID-19 conversations on social media (e.g. Twitter). The COVID-19 cable and broadcast news dataset is publicly available to the research community, and can be accessed at https://doi.org/10.7910/DVN/LWMYAD.

## Introduction

Americans overwhelmingly consume news from cable and broadcast news networks (Allen et al. 2020). Therefore, these networks arguably play an out-sized role in shaping our national conversations and public opinion and behaviors (Iyengar and Kinder 1987; McCombs and Valenzuela 2020). Despite the significant role these networks play, most recent large-scale studies of news media focus on online news (Bode et al. 2020; Guo and Vargo 2020; Budak 2019), with a smaller number of notable exceptions that focus on cable and broadcast news (Muddiman, Stroud, and McCombs 2014; Feldman et al. 2012; Cadorette, Savitz, and Cockerill 2018; Allen et al. 2020; Nassar 2020) There are numerous reasons for this trend, one of which is the relative ease with which online news can be gathered and processed compared to cable and broadcast news. This paper aims to help close this data accessibility gap by providing a dataset on cable and broadcast news coverage on a timely topic—COVID-19.

The need to understand cable and broadcast news coverage is even more pronounced in the case of COVID-19. Given the novel nature of the disease, both the public's information needs and the consequences of a misinformed public make it particularly important to examine the role cable and broadcast networks play in this news ecosystem. Indeed, early work suggests a link between consuming news from certain cable networks and conspiratorial beliefs about the virus (Jamieson and Albarracin 2020) and undesirable COVID-19 disease outcomes (Bursztyn et al. 2020) and behaviors (Simonov et al. 2020). Polling from the Pew Research Center in March of 2020 also found that those relying on Fox News believed that media coverage exaggerated the risks of COVID-19 (Jurkowitz and Mitchell 2020). These studies highlight the importance of paying more careful attention to the content shared by cable and broadcast networks. This motivates our project, which builds on the preliminary studies by analyzing all broadcast and cable news programs aired on weekday evenings during the beginning of the pandemic in the U.S. and by sharing the collected data to aid future research.

We introduce a new dataset that includes COVID-19 coverage by cable and broadcast news networks. We first gather a large dataset of cable and broadcast news transcripts. We sample random paragraphs stratified across networks. These paragraphs are labeled by experts as being related to COVID-19 or not. We then use these labels to build various classifiers. Our best performing classifier is a BERT language model, with an F1-score of 0.873. Finally, we provide some high level analysis of all content predicted to be related to COVID-19. Our analysis reveals that (i.) networks vary significantly in the attention they paid to the pandemic, (ii.) the language varied across networks, with the dissimilarity being most pronounced across cable networks. Finally, an inspection of the distinct phrases across cable networks suggests that (iii.) cable news networks are politicizing COVID-19. These findings have significant implications. In short, our findings suggest that people's exposure to COVID-19 coverage—both in terms of the amount and nature—varies depending on the news source they select.

## Related Work and Background

**COVID-19 and News** Information needs associated with COVID-19 motivated a significant body of research in the last year. Here, we summarize the scholarship focused on the interaction between the pandemic and news coverage.

Survey studies highlight one important reason for inspecting cable and broadcast news content. For instance, (Mitchell et al. 2020) show that (i.) the U.S. population is paying close attention to news about the pandemic, (ii.) Democrats and Independents are more engaged with news on this topic, and (iii.) news diets tie closely to COVID-19 attitudes and behaviors (e.g. whether plans changed for Thanksgiving because of the coronavirus outbreak). In particular for (iii.), Republicans who only used Fox News or talk radio as a major source for news were much less likely to have changed their Thanksgivings plans. The effect of one's news diet was less significant for Democrats. Another nationally representative survey study leads to a similar conclusion (Jamieson and Albarracin 2020). Jamieson and Albarracin find that accurate beliefs about COVID-19 correlated with print media consumption—controlling for respondents' political party and mainstream broadcast media use (e.g., NBC News). In contrast, conservative media use (e.g., Fox News) correlated with belief in conspiracy theories. Relatedly, (Bursztyn et al. 2020) show, through an instrumental variable approach, that areas with greater exposure to two particular Fox News shows downplaying the threat of COVID-19 experienced a greater number of cases and deaths. In another study using Nielsen data and cell phone data, (Simonov et al. 2020) find that Fox News viewership reduced compliance with stay-at-home orders. These studies highlight the (destructive) role cable news can play in controlling the spread of the pandemic.

Although not directly linked to cable and broadcast news coverage, work by (Green et al. 2020) provides further insights into how partisanship shapes COVID-19 narratives. By examining tweets by the current members of the U.S. House and Senate, they show that Democrats discussed the pandemic more frequently. They also show differences in content: Democrats emphasized public health threats and concerns about American workers. In comparison, Republicans placed greater emphasis on China and businesses. Our descriptive analysis suggests a similar partisan trend for cable and broadcast news. We note that (Green et al. 2020) rely on dictionaries of words to identify COVID-19 content. We believe that the dataset we share can be used by future researchers interested in similar research questions to build classifiers and identify a potentially more complete set of content to analyze.

Finally, while a number of studies make the connection between news consumption and COVID-19 attitudes and behaviors (Mitchell et al. 2020; Jamieson and Albarracin 2020; Bursztyn et al. 2020; Jurkowitz and Mitchell 2020; Simonov et al. 2020), these studies cannot tell exactly what in news coverage led to these outcomes. Such an analysis necessarily involves making sense of the *content* of such news. We aim to aid research in this space by sharing our COVID-19 TV coverage dataset.

**Broader Scholarship on Cable and Broadcast News** It is also important to put our study in the larger context of scholarship on cable and broadcast news. Research has uncovered distinct patterns of coverage among Fox News, CNN, and MSNBC (Budak, Goel, and Rao 2016; Stroud 2011) and among broadcast networks and Fox News (Groeling 2008). Individuals' opinions about political issues (Hoewe et al. 2020; Muddiman, Stroud, and McCombs 2014), their belief in incorrect information (Arsenault and Castells 2006), and their partisan attitudes (Stroud 2011) are related to the cable news outlets they watch. These differences are pronounced in coverage of scientific issues. For instance, several studies demonstrate that MSNBC and CNN cover climate change in substantially different ways from Fox News (Feldman et al. 2012; Cadorette, Savitz, and Cockerill 2018), and exposure to these different networks affects audiences' trust in scientists and perceptions of climate change (Hmielowski et al. 2014; Feldman et al. 2012). We expect the same pattern to occur for COVID-19 coverage. Preliminary research suggests that news audiences for the various cable networks hold different opinions about COVID-19 (Jurkowitz and Mitchell 2020; Mitchell et al. 2020). Exploring coverage differences across networks will help scholars begin to understand whether these differences in belief are due to content differences across network or differences in audience characteristics. Characterizing such coverage, adding information about widely-viewed broadcast news coverage, and providing data for future research is what motivates this paper.

## Transcript Extraction and Processing

We use Nexis Uni to gather transcripts for the following programs on weeknights between January 21, 2020, the day of the first confirmed case of the coronavirus in the United States, and June 12, 2020, right after the country passed two million cases.

1. CNN: Anderson Cooper 360 Degrees, The Lead with Jake Tapper, Cuomo Prime Time, Erin Burnett OutFront, CNN Tonight, and The Situation Room
2. Fox News: The Five, Special Report with Bret Baier, The Story with Martha MacCallum, Tucker Carlson Tonight, Hannity, Ingraham Angle, and Fox News @ Night
3. MSNBC: MTP Daily, The Beat with Ari Melber, All in with Chris Hayes, The Rachel Maddow Show, The Last Word with Lawrence O'Donnell, 11th Hour with Brian Williams, and Hardball
4. ABC World News Tonight
5. CBS Evening News
6. NBC Nightly News

These programs cover all prime-time nightly news across the cable and broadcast networks. All transcripts are provided to LexisNexis directly from the publishers of each network. The transcripts are formatted by the publishers, who break each transcript into paragraphs. LexisNexis is one of the most widely used news archives in the social sciences (Deacon 2007). LexisNexis provides academics access to full-text news transcripts and articles (as well as other data such as legal and business publications) through Nexis Uni. We accessed Nexis Uni through our university libraries and manually searched for transcripts to include in the dataset.

There are a total of 486,068 paragraphs across 4,589 transcripts from the aforementioned programs in the specified time window. We searched by ”Publication” (i.e. the news program title) and ”Source” (i.e. the network name) between the dates of January 21 and June 12, 2020. We then checked that we had transcript content for each network from 5pm to midnight on each weekday in the timeframe[1]. Any timeslots that had no transcripts in the database are not included in the dataset. The percentage of all paragraphs accounted for by ABC, CBS, NBC, MSNBC, CNN, and Fox News are 2.3%, 2.5%, 2.7%, 28.4%, 31.9%, and 32.1% respectively. Broadcast news accounts for fewer paragraphs because it airs for only one hour, whereas the cable outlets aired for the full seven hours.

## Data Labeling Process

Our next goal, upon gathering all news transcripts, was to filter out content unrelated to COVID-19. To achieve this goal, we first performed human coding of the transcripts to create training and test datasets. Trained coders classified each paragraph in a transcript as belonging to one of the following three categories:

1. Directly related to COVID-19 (that is, the paragraph included words directly related to COVID-19, including the health, political, economic, and other implications of the disease),
2. Indirectly related to COVID-19 (that is, the paragraph did not include words that directly identified COVID-19, but the context of the transcripts made it clear that the health, political, economic, or other implications of the disease were being discussed)
3. Not related to COVID-19.

Following best practices of manual content analysis (Krippendorff 2018), we trained two coders to identify COVID-19 coverage until they consistently agreed about the category of each paragraph. The two coders manually coded the transcripts for 52 news broadcasts, which were selected from each network (transcripts for two to three broadcasts from each program and two to three broadcasts for each week of content in the dataset), totaling 12,298 paragraphs. Reliability was high (Direct COVID-19: Krippendorff’s alpha = .87; Indirect COVID-19: Krippendorff’s alpha = .85). After establishing reliability, a single coder manually classified the transcripts from an additional 214 broadcasts. These transcripts included one randomly selected transcript a week from two programs on each network and a randomly selected broadcast for each month for the remaining programs. This resulted in a total of 44,643 manually labeled paragraphs.

| Model | Precision | Recall | F1 Score | Accuracy |
|---|---|---|---|---|
| Logistic regression | 0.745 | 0.784 | 0.764 | 0.826 |
| Extra-Trees | 0.729 | 0.618 | 0.669 | 0.809 |
| XGBoost | 0.822 | 0.783 | 0.802 | 0.839 |
| DistilBERT | 0.889 | 0.865 | 0.873 | 0.897 |

Table 1: Test set performance across models.

## Automated Detection of COVID-19 Content

We rely on supervised machine learning techniques to classify all transcript paragraphs according to their relevance to COVID-19. Below, we describe our pipeline and results.

**Data Preparation** The original transcripts include certain program level meta-data, such as an introductory section with program, network, and date information. The transcripts also include the name of the speaker for conversation turns and text that denotes ad breaks. Such data were removed using automated scripts.

The cleaned transcript data was split into training and test sets approximating an 80-20 split. Given the conversational nature of the transcripts, consecutive paragraphs being observed in both train and test sets could lead to a misleadingly high performance. In order to minimize potential data leakage we ensured that all paragraphs from a given transcript were included in only the training or test sets. This resulted in a training set of 40,867 and a test set of 8,562 paragraphs.

**Models** We use the following models to perform the classification: (i.) Logistic Regression, (ii.) ExtraTrees, (iii.) XGBoost, (iv.) DistilBERT.

All models perform binary classification. We collapse direct and indirect mentions of COVID-19 to a single positive class to carry out this task. We use tf-idf weighting for the first three models, a common data preparation step for conventional machine learning approaches. Grid search was used for parameter tuning. Finally, given the advantages observed in recent work, we also used a pretrained language model (Devlin et al. 2018) and tuned it for our specific task. We relied on the DistilBERT (distilbert-base-uncased) model[2] which provides comparable performance to BERT despite being a smaller and lighter model (Sanh et al. 2019). The training data was further divided into randomized training and validation sets using a 90/10 split. We tuned the hyperparameters through cross validation and found that a model trained using two epochs, a learning rate of 5.00E-05, and a batch size of 32 provided the best performance.

**Results** The overall performance achieved by each model is given in Table 1. As is common in recent work, we see a significant improvement by using a pre-trained language model.

[1] Two episodes of the CBS Evening News that aired on weekend dates (March 14 and 15) and two Jake Tapper episodes (449 COVID-related paragraphs total) were unintentionally included in the dataset that the training data was sampled from. We ran robustness tests and found no substantive differences in our results when these were not included in downstream analyses.

[2] https://huggingface.co/distilbert-base-uncased

| Network | Precision | Recall | F1 Score | Accuracy |
|---|---|---|---|---|
| ABC | 1.000 | 0.869 | 0.930 | 0.879 |
| CBS | 0.805 | 0.755 | 0.779 | 0.851 |
| CNN | 0.922 | 0.852 | 0.885 | 0.901 |
| Fox News | 0.865 | 0.872 | 0.868 | 0.893 |
| MSNBC | 0.913 | 0.865 | 0.889 | 0.900 |
| NBC | 0.935 | 0.864 | 0.899 | 0.884 |

Table 2: DistilBERT Test set performance across networks.

We further examine how the best performing model (DistilBERT) test labels are distributed across the three categories used by the human coders. We observe that all content labeled as directly related to COVID-19 by human coders are correctly labeled as COVID-19 related. Of the 2,631 test set paragraphs that are labeled as indirectly related by human coders, 79% are correctly labeled. Finally, out of 4,640 paragraphs labeled as not related to COVID-19 by human coders, 93% are correctly labeled. In summary, and as expected, the classifier performs significantly better for content explicitly related to COVID-19.

While the overall accuracy of the classifier is high, it is important to inspect whether the performance is high across networks since much of the analysis we aim to perform here relate to comparisons across networks. Thus, we perform error analysis to answer this question. The results are given in Table 2. There are no clear patterns of broadcast/cable or ideological bias. We do, however, observe some meaningful differences. For instance, ABC has the highest F1 score while CBS has the lowest. Regardless, even the lowest performing network has high accuracy and an acceptable F1-score.

## Data Description

We provide descriptive analysis to determine (i.) the attention each network paid to the pandemic in its coverage and how this changed over time, (ii.) the degree to which coverage content differed across networks, and (iii.) the phrases that help explain the divergence in content observed across networks.

### Prevalence of COVID-19 Coverage

We begin our descriptive analysis by examining the degree to which different networks covered COVID-19 and how that behavior changed over time. Figure 1 gives a high level overview. We observe that approximately 42% of paragraphs in our dataset pertain to COVID-19. However, there is significant variance across networks. Overall, there is a stronger emphasis on the disease by broadcast networks–most notably by NBC which dedicated 51% of its coverage, measured through the paragraphs throughout the time frame, to COVID-19. The lowest attention was paid by Fox News, which dedicated 36% of coverage to COVID-19. Interestingly, MSNBC has the second lowest frequency contrary to the general belief that left-leaning networks covered COVID-19 more. This could be due to a more conversational nature of coverage in cable news, leading to a generally lower percentage, even compared to more centrist broadcast networks.

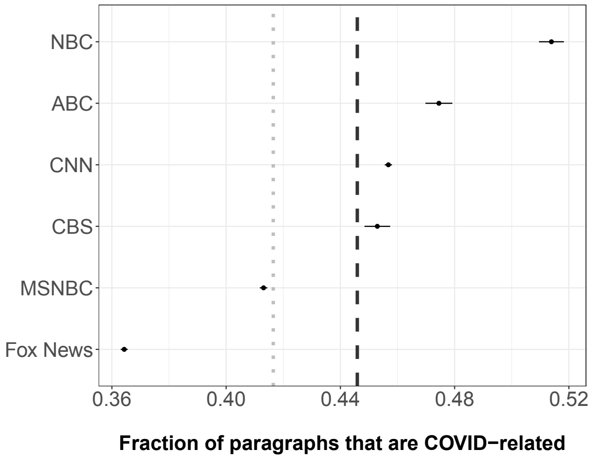


Figure 1: Fraction of paragraphs across networks that are COVID-19 related: This plot shows, for each network, the fraction of paragraphs that are COVID-19 related. Error bars represent two standard errors. We provide two other data points: the light gray dotted line gives the average across all paragraphs, irrespective of network and the gray dashed line gives the unweighted average of fractions across networks. The former is lower due to there being a larger number of paragraphs in our dataset from cable networks, compared to broadcast networks.

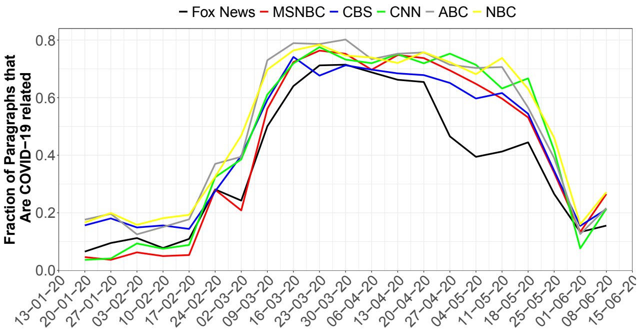


Figure 2: Fraction of paragraphs across networks that are COVID-19 related over time: This plot shows, for each network, the fraction of paragraphs that are COVID-19 related on a weekly basis.

We next examine how this coverage changed over time. Figure 2 presents the results and shows that coverage changed in a similar pattern across the networks. Coverage began to increase across all outlets in late February, with increasing U.S. cases and the discovery of community spread of COVID-19. It peaked in April when the number of cases surpassed 100,000 in the U.S. The attention dropped for Fox News earlier than the other networks, in late April. All networks dropped their attention in May as the media started focusing on racial justice protests following the killing of

George Floyd. In short, while there are significant differences across outlets in how much they focused on the pandemic, the temporal patterns were generally consistent.

## COVID-19 Language Similarity Across Networks

The previous section highlights the first order differences across networks. We see that broadcast networks and CNN have a stronger COVID-19 focus and Fox News is the network with the lowest attention on the subject. Here, we examine the *content* of COVID-19 coverage, filtering out all paragraphs predicted to be irrelevant to the pandemic.

**Program-level language similarity** We start by examining language similarity across all programs and all networks. We do so by the clustering of programs as follows: We construct a tf-idf based vector-representation for each network, removing stop-words and words that occur in fewer than ten COVID-19 related paragraphs. We define the distance between two networks based on their 1) cosine similarity and 2) Euclidean distance and use hierarchical clustering to cluster programs. The results are consistent for the two distance measures. For brevity, here we only present the results for cosine distance (in Figure 3). We observe that programs from the same network generally cluster together. All MSNBC programs are more similar to other MSNBC programs compared to programs from the other networks. The patterns observed for CNN and Fox News are generally consistent with this finding. The one exception is "The Five" on Fox News, which is closer to CNN programs than other Fox News programs. This may be due to the nature of this specific program–"The Five" hosts have more balanced partisan leanings compared to the other Fox News programs. Finally, we observe that broadcast network programs cluster together.

**Network-level language similarity** Our finding above suggests that analyzing COVID-19 content at the network level is justified. As such, the analysis for the rest of this paper is provided at the network level. To determine the language similarity across networks, one can start by inspecting Figure 3. We observe that broadcast networks cluster together, suggesting that their coverage is most alike. Cable news networks are less similar to each other, as given by the y-axis of the dendrogram that indicates cosine distance. Here, we augment this analysis by grouping all content from the same network into a single document and computing distance in two distinct ways: (i) Average KL-divergence and (ii) Distinctive phrase analysis using log odds with informative Dirichlet priors (Monroe, Colaresi, and Quinn 2008).

*Average KL Divergence* We start by representing language used in each network as a probability distribution. To do so, we first remove stop words and infrequent words (words that are seen in fewer than ten COVID-19 related paragraphs overall). We then use TF-IDF weighting to convert the set of words and their frequencies into a vector space representation. Let $P_i$ denote the probability distribution for network $i$ and let $X$ denote the vocabulary of words. We then use symmetric (average) KL divergence to compute the language dissimilarity between two networks $i$ and $j$ as follows:

$$D_{KL}(P_i||P_j) + D_{KL}(P_j||P_i) \tag{1}$$

where

$$D_{KL}(P_i||P_j) = \sum_{x \in X} P_i(x) log(\frac{P_i(x)}{P_j(x)}) \tag{2}$$

KL-divergence and its variants have been used extensively in NLP research to define text similarity (Dagan, Lee, and Pereira 1999; Carpineto et al. 2001; Bigi 2003).

*Distinctive Phrase Frequency* While average KL-divergence is commonly used in the literature to measure language similarity, it is not guaranteed to correspond to human perception of language similarity. Therefore, to provide further evidence of language (dis)similarity, we use a second approach. Here, we first identify distinctive words when comparing two networks $i$ and $j$. In other words, we find the set of words that are significantly more common in $i$ compared to $j$, and the set of words that are significantly more common in $j$ compared to $i$. The fraction of all words this set accounts for, the more the language of the two networks diverge from each other.

To find distinctive words when comparing two networks, we rely on log odds with informative Dirichlet priors (Monroe, Colaresi, and Quinn 2008). This approach commonly outperforms simpler methods such as computing differences in frequencies, ratio of frequencies, or the log odds ratio (Monroe, Colaresi, and Quinn 2008; Jurafsky et al. 2014). Given two networks $i$ and $j$, we estimate the log odds with Dirichlet priors as follows:

$$\delta_w^{(i-j)} = \log \frac{y_w^i + \alpha_w}{n^i + \alpha_0 - y_w^i - \alpha_w} - \log \frac{y_w^j + \alpha_w}{n^j + \alpha_0 - y_w^j - \alpha_w} \tag{3}$$

where $n^i$ ($n^j$) is the size of corpus for network $i$ ($j$), $y_w^i$ ( $y_w^j$) is the word count of $w$ in network $i$ (resp. $j$), $\alpha_0$ is the size of the background corpus, and $\alpha_w$ is the word count of $w$ in the background corpus. This formula, while similar to simple log-odds, makes one important adjustment—it essentially shrinks the estimates towards the prior by adding the corresponding values from the background corpus. We use all COVID-19 related paragraphs across all networks to define the background corpus. Further, to account for uncertainty for rare words, variance of this measure is computed as:

$$\sigma^2(\delta_w^{(i-j)}) \approx \frac{1}{(y_w^i + \alpha_w)} + \frac{1}{(y_w^j + \alpha_w)} \tag{4}$$

and the $Z$-score is calculated by normalizing the log-odds using the variance as follows:

$$Z = \frac{\delta_w^{(i-j)}}{\sqrt{\sigma^2(\delta_w^{(i-j)})}} \tag{5}$$

The larger the Z-score is, the more distinctly common the phrase is for network $i$ compared to network $j$. Similarly, the smaller (with negative values) the Z-score, the more

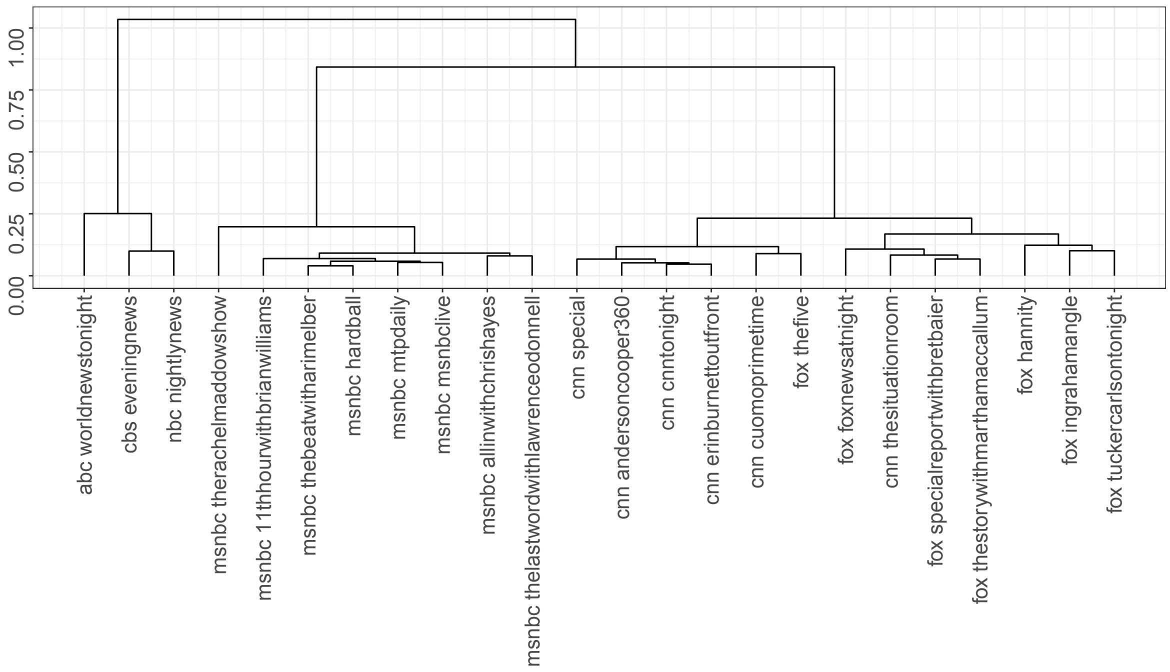


Figure 3: Clustering of Programs according to language similarity: The dendrogram is produced using hierarchical clustering using cosine distance. The plot summarizes the clustering of programs according to language similarity. The more similar two programs are to each other, the smaller number of steps it takes to connect them in this dendrogram.

distinctly common the phrase is for network $j$. The values greater than two and smaller than negative two are statistically significant.

Next, we measured the distance between networks $i$ and $j$ as: $\frac{(n_{distinct,i,k}+n_{distinct,j,k})}{n_k}$, where $n_k$ is the number of phrases that occur at least $k$ times across the two networks, and $n_{distinct,i,k}$ ($n_{distinct,j,k}$) is the number of words that are (1) distinct to corpus $i$ ($j$) and (2) occur at least $k$ times across the two networks. Phrases with Z-scores greater than two are distinct to network $i$ and Z-scores less than negative two are distinct to network $j$. This measure captures the fraction of popular words that are distinct to either one of the networks. We set $k = 100$ in our main analysis but varying this measure does not substantively change the findings.

The results for the two measurement approaches are given in Figure 4. Both approaches lead to the same high level conclusion: the highest similarity is observed between broadcast networks. Cable-cable and cable-broadcast networks exhibit higher dissimilarity. There are however, distinctions worth noting for the patterns observed in Figures 4a and 4b. The divergence between cable-cable pairs are more apparent when using distinct phrase frequency as a measure of language dissimilarity.

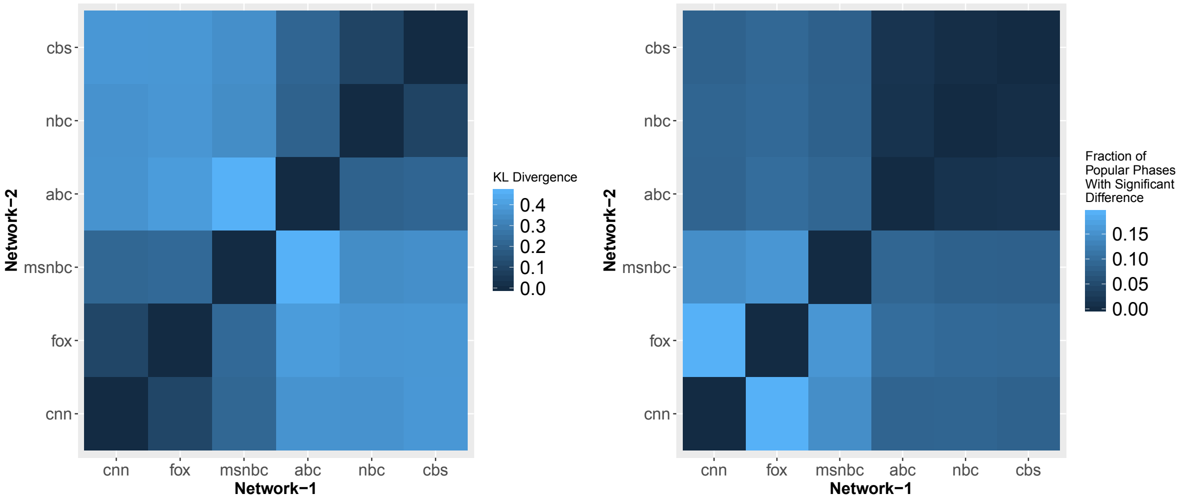
## Distinctive Phrase Analysis

In this section, we aim to identify the nature of phrases that lead to the dissimilarity observed across the news networks. Given the homogeneity observed across broadcast networks, here we focus on the following comparisons: 1) cable vs. broadcast, 2) Fox vs. CNN, 3) Fox vs. MSNBC, and 4) MSNBC vs. CNN. The results are shared in Table 3[3].

There are a number of noteworthy patterns. First, comparisons between the cable and broadcast networks show that various disease related words (e.g. hospital, death toll, covid19) and places related to the spread, or lack thereof, (e.g. flight, home, cruise) are more common in broadcast networks compared to cable news.

Comparisons across the cable networks reveal other important patterns. For instance, Fox News is more commonly using words related to China, wedge issues (abortion, racist), and unproven treatments (hydroxychloroquine) when compared to MSNBC. MSNBC, on the other hand, uses words related to the federal government response (e.g. federal government, white house) and COVID-19 scale and capacity (e.g. testing, cases). Both the Fox News vs CNN and CNN vs. MSNBC comparisons suggest that CNN has fewer references to politicians compared to Fox News and MSNBC. However, these politicians tend to be Democratic when comparing CNN to Fox News and Republican when comparing it to MSNBC. In other words, we see that Fox News is mentioning the Democrats more and MSNBC is mentioning the Republicans more. Comparing CNN to the other two cable networks reveals a higher tendency by CNN to use health related words. Overall, this analysis suggests that the cable news networks, especially Fox News and MSNBC, are

[3] Names of networks (e.g., CNN) and network employees (e.g., Jesse Waters) are removed. These phrases are naturally unique to the corresponding network.



(a) KL-divergence.

(b) Distinct Phrase Frequency

Figure 4: Language Dissimilarity Across Networks Measured Through Two Approaches.

politicizing COVID-19 in their coverage.

## Ethics and FAIRness

The gathered data are transcripts of national cable and broadcast networks. As such, these data contain newsworthy information that is unlikely to pose privacy concerns. In addition, this information has been publicly broadcast. In addition to the low privacy and security concerns, gathering and examining such data has significant benefits to society given the role news organizations play in informing the public. The importance of this role is magnified given the informational needs associated with a novel disease. The gathered data and our analyses bring to light important biases in coverage across cable and broadcast networks.

Our dataset also conforms to FAIR principles. In particular, the dataset is *findable* since it is shared publicly via https://doi.org/10.7910/DVN/LWMYAD. This dataset is also *accessible* given the global availability of the data and the format used (CSV), which allows a broad range of researchers to access/use it. This file format also makes the dataset *interoperable* given that most programming languages have built-in libraries to process this file format. Finally, the dataset is supplemented with a readme file explaining data files in detail to aid with the re-use of the dataset.

## Potential Uses of Data

In this paper, we introduce a dataset on COVID-19 coverage by broadcast and cable news networks. While the dataset can be used advance knowledge broadly in media studies, politics, and communication, it has implications for social media and web research as well. First, the dataset can be used to identify COVID-19 related social media content. Past work on COVID-19 and social media use dictionaries to identify COVID-19 related content (Green et al. 2020; Jiang et al. 2020). Our labeled data can help build robust models through transfer learning. Such an approach can lead to a reduction in the amount of social media data required for manual labeling. Second, social media conversations do not exist in a vacuum. Our dataset can be used to examine how COVID-19 related information flows from traditional media to social media. Past work has considered the role of polarization in COVID-19 social media behavior focusing on political elites (Green et al. 2020) and citizens (Jiang et al. 2020). Combining these data with our news media data can provide insights into the channels through which polarization affects COVID-19 conversations. Further, scholars have examined how information flows from social to mainstream media and vice versa (Guo and Vargo 2020), another project that could be advanced by these data.

## Conclusion

Research suggests that cable and broadcast news play a significant role in affecting public opinion and behavior (Iyengar and Kinder 1987; McCombs and Valenzuela 2020). The way they cover a health issue like COVID-19 stands to shape opinions and attitudes that have direct real-world consequences. This coverage can lead to informed citizens or to polarized audiences that approach the disease in partisan ways. The coverage can also affect (and be affected by) social media conversations of COVID-19. Therefore, it is crucial to create reliable datasets to examine the coverage of this pandemic across news networks. However, the research community currently lacks a large labeled dataset of COVID-19 cable and broadcast news coverage, hindering research efforts in this space. The aim of our study was to provide such a dataset and provide descriptive analysis to characterize it for interested readers.

We gathered transcripts of prime time programs from cable and broadcast networks, built classifiers to detect COVID-19 coverage, and provided descriptive analysis to shed light on cable and broadcast coverage of COVID-19. Our analysis reveals that (i.) cable and broadcast networks

| Comparison | Unique To | Phrases |
|---|---|---|
| Cable vs. Broadcast | Cable | go, think, know, thing, people, want, thats, talk, well, mean, really, look, yes, dont, get, lot, right, would, way, dr, youre, actually, theyre, need, sort, happen, see, ok, let, something |
|  | Broadcast | tonight, hundred, na, thousand, new, passenger, nearly, toll, doctor, image, test positive, covid19, grow, child, patient, family, evening, alarm, tom, food, hospital, flight, death toll, home, ten, ship, cruise, new york, york, inside |
| Fox vs. MSNBC | Fox | china, chinese, medium, wuhan, communist, travel ban, ban, democrat, ok, travel, well, communist party, america, president trump, racist, hydroxychloroquine, american, lockdown, covid19, chinese government, mob, chinese communist, yes, pelosi, good, abortion, drug, tonight, border, shutdown |
|  | MSNBC | test, donald, donald trump, plant, sort, public, case, public health, white house, federal, federal government, white, health care, term, health, county, kind, place, testing, care, katy, quote, today, capacity, report, prison, need, reporting, state, im curious |
| Fox vs. CNN | Fox | china, chinese, democrat, medium, america, american, lockdown, communist, bill, covid19, biden, travel ban, ban, pelosi, joe, business, communist party, money, wuhan, racist, economy, nancy, chinese communist, mob, lie, party, abortion, senator, shutdown, left |
|  | CNN | test, know, wear, mean, reopen, outfront, wear mask, mask, cnns, sort, still, white house, white, dr, president, vaccine, get test, testing, hear, youre, case, tell cnn, obviously, erica, president say, contact, say, symptom, fauci, town hall |
| CNN vs. MSNBC | CNN | know, say, outfront, cnns, yes, wear, mean, wear mask, coronavirus, tell cnn, study, reopen, mask, town hall, obviously, ok, erica, break news, youre, kid, question, hall, global town, well, president say, virus, mean know, dr, trial, reiner |
|  | MSNBC | donald, donald trump, plant, trump, republican, crisis, prison, government, quote, health care, facility, epidemic, today, bill, katy, senator, care, vote, election, federal, worker, senate, report, kind, view, mcconnell, meat, meatpacking, iowa, trump administration |

Table 3: Words that distinguish pairs of networks. For each pair, we identify top-30 words that are most distinctly popular for the first and second network in pair, in the order of their uniqueness. Uniqueness of word is measured through log-likelihood with informative Dirichlet priors as given in Equation 5. For instance "donald" has the lowest (negative) z-score among all phrases with at least a hundred occurrences across CNN and MSNBC. This means that this word is most unique to MSNBC, when comparing the two networks. We use a higher threshold of three hundred when comparing cable and broadcast given increased document size.

vary significantly in the amount of coverage they dedicated to the pandemic, with broadcast networks covering it most and Fox News covering it least, (ii.) the language varied across networks, with the dissimilarity being most pronounced across cable networks. Finally, an inspection of the distinct phrases across cable networks suggests that (iii.) cable news networks are politicizing COVID-19.

## Acknowledgments

Funding provided by Mark Gibson, CEO, Capital Markets, JLL Americas to the Center for Media Engagement at the Moody College of Communication at The University of Texas at Austin.

## References

Allen, J.; Howland, B.; Mobius, M.; Rothschild, D.; and Watts, D. J. 2020. Evaluating the fake news problem at the scale of the information ecosystem. *Science Advances* 6(14): eaay3539.

Arsenault, A.; and Castells, M. 2006. Conquering the minds, conquering Iraq: The social production of misinformation in the United States–a case study. *Information, Communication & Society* 9(3): 284–307.

Bigi, B. 2003. Using Kullback-Leibler distance for text categorization. In *European Conference on Information Retrieval*, 305–319. Springer.

Bode, L.; Budak, C.; Ladd, J. M.; Newport, F.; Pasek, J.; Singh, L. O.; Soroka, S. N.; and Traugott, M. W. 2020. *Words that matter: How the news and social media shaped the 2016 Presidential campaign*. Brookings Institution Press.

Budak, C. 2019. What happened? the spread of fake news publisher content during the 2016 us presidential election. In *The World Wide Web Conference*, 139–150.

Budak, C.; Goel, S.; and Rao, J. M. 2016. Fair and balanced? quantifying media bias through crowdsourced content analysis. *Public Opinion Quarterly* 80(S1): 250–271.

Bursztyn, L.; Rao, A.; Roth, C.; and Yanagizawa-Drott, D. 2020. Misinformation during a pandemic. *University of Chicago, Becker Friedman Institute for Economics Working Paper* .

Cadorette, J.; Savitz, R.; and Cockerill, K. 2018. Good and bad news: Climate science affirmation and cable news coverage. *Environmental Practice* 20(4): 104–111.

Carpineto, C.; De Mori, R.; Romano, G.; and Bigi, B. 2001. An information-theoretic approach to automatic query expansion. *ACM Transactions on Information Systems (TOIS)* 19(1): 1–27.

Dagan, I.; Lee, L.; and Pereira, F. C. 1999. Similarity-based models of word cooccurrence probabilities. *Machine learning* 34(1-3): 43–69.

Deacon, D. 2007. Yesterday's papers and today's technology: Digital newspaper archives and 'push button'content analysis. *European journal of communication* 22(1): 5–25.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* .

Feldman, L.; Maibach, E. W.; Roser-Renouf, C.; and Leiserowitz, A. 2012. Climate on cable: The nature and impact of global warming coverage on Fox News, CNN, and MSNBC. *The International Journal of Press/Politics* 17(1): 3–31.

Green, J.; Edgerton, J.; Naftel, D.; Shoub, K.; and Cranmer, S. J. 2020. Elusive consensus: Polarization in elite communication on the COVID-19 pandemic. *Science Advances* 6(28).

Groeling, T. 2008. Who's the fairest of them all? An empirical test for partisan bias on ABC, CBS, NBC, and Fox News. *Presidential Studies Quarterly* 38(4): 631–657.

Guo, L.; and Vargo, C. 2020. "Fake news" and emerging online media ecosystem: An integrated intermedia agenda-setting analysis of the 2016 US presidential election. *Communication Research* 47(2): 178–200.

Hmielowski, J. D.; Feldman, L.; Myers, T. A.; Leiserowitz, A.; and Maibach, E. 2014. An attack on science? Media use, trust in scientists, and perceptions of global warming. *Public Understanding of Science* 23(7): 866–883.

Hoewe, J.; Peacock, C.; Kim, B.; and Barnidge, M. 2020. The Relationship Between Fox News Use and Americans' Policy Preferences Regarding Refugees and Immigrants. *International Journal of Communication* 14: 21.

Iyengar, S.; and Kinder, D. R. 1987. News that matters: Agenda-setting and priming in a television age. *News that Matters: Agenda-Setting and Priming in a Television Age* .

Jamieson, K. H.; and Albarracin, D. 2020. The Relation between Media Consumption and Misinformation at the Outset of the SARS-CoV-2 Pandemic in the US. *The Harvard Kennedy School Misinformation Review* .

Jiang, J.; Chen, E.; Yan, S.; Lerman, K.; and Ferrara, E. 2020. Political polarization drives online conversations about COVID-19 in the United States. *Human Behavior and Emerging Technologies* 2(3): 200–211.

Jurafsky, D.; Chahuneau, V.; Routledge, B. R.; and Smith, N. A. 2014. Narrative framing of consumer sentiment in online restaurant reviews. *First Monday* 19(4).

Jurkowitz, M.; and Mitchell, A. 2020. Cable TV and COVID-19: How Americans perceive the outbreak and view media coverage differ by main news source. *Pew Research Center* .

Krippendorff, K. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

McCombs, M.; and Valenzuela, S. 2020. *Setting the agenda: Mass media and public opinion*. John Wiley & Sons.

Mitchell, A.; Jurkowitz, M.; Oliphant, J. B.; and Shearer, E. 2020. Americans' attention to news about the coronavirus pandemic remains steady in November as cases surge. *Pew Research Center, Journalism and Media* .

Monroe, B. L.; Colaresi, M. P.; and Quinn, K. M. 2008. Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis* 16(4): 372–403.

Muddiman, A.; Stroud, N. J.; and McCombs, M. 2014. Media fragmentation, attribute agenda setting, and political opinions about Iraq. *Journal of Broadcasting & Electronic Media* 58(2): 215–233.

Nassar, R. 2020. Framing Refugees: The Impact of Religious Frames on US Partisans and Consumers of Cable News Media. *Political Communication* 1–19.

Sanh, V.; Debut, L.; Chaumond, J.; and Wolf, T. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108* .

Simonov, A.; Sacher, S. K.; Dubé, J.-P. H.; and Biswas, S. 2020. The persuasive effect of fox news: non-compliance with social distancing during the covid-19 pandemic. Technical report, National Bureau of Economic Research.

Stroud, N. J. 2011. *Niche news: The politics of news choice*. Oxford University Press on Demand.