## Distribution Agreement

Alex Fan April 9, 2021

# Red Feed vs. Blue Feed:

# Differences in COVID-19 Digital News Coverage based on Partisan Lean

by

Alex Fan

Dr. Lauren Klein

Adviser

The Department of Quantitative Theory and Methods

Dr. Lauren Klein

Adviser

Dr. Jinho Choi

Committee Member

Dr. Krzysztof Karbownik

Committee Member

2020

Red Feed vs. Blue Feed:

Differences in COVID-19 Digital News Coverage based on Partisan Lean

By

Alex Fan

Dr. Lauren Klein

Adviser

An abstract of

a thesis submitted to the Faculty of Emory College of Arts and Sciences

of Emory University in partial fulfillment

of the requirements of the degree of

Bachelor of Science with Honors

The Department of Quantitative Theory and Methods

2020

# Abstract

## Red Feed vs. Blue Feed:

## Differences in COVID-19 Digital News Coverage based on Partisan Lean

By Alex Fan

The rise of the COVID-19 Pandemic has sparked changes to the global landscape on both micro-political and macro-political levels. The ramifications of the pandemic, in many ways, has been a function of polarized rhetoric from traditional media institutions. Many studies have already examined the effects of polarized information on people's individual behavior in relation to the pandemic. However, not as much attention has been spent exploring the variation in content and contextual representations underlying the polarized information. This project contributes to the broader literature by constructing and leveraging a novel corpus of approximately 300,000 COVID-19 digital news articles that spans an entire year of the pandemic, representing the most comprehensive set of its kind to date. In addition, this project also created a novel hand-labeled dataset of approximately 1,500 articles, with classes that delineate an article's orientation toward micro-political or macro-political actions in combating COVID-19. Analysis of these datasets show a small shift in content based on the political lean of the news source, with right-leaning news sources spending a greater proportion of their content on state-level coverage than left-leaning news sources do. Moreover, right-leaning news sources tend to portray important public health measures such as mask-wearing and social distancing in a negative light, particularly by associating them with non-science. Overall, this work presents a new, more comprehensive angle of studying political media discourse in the context of a pandemic.

# Red Feed vs. Blue Feed:

# Differences in COVID-19 Digital News Coverage based on Partisan Lean

By

Alex Fan

Lauren Klein

Adviser

A thesis submitted to the Faculty of Emory College of Arts and Sciences

of Emory University in partial fulfillment

of the requirements of the degree of

Bachelor of Science with Honors

The Department of Quantitative Theory and Methods

2020

# Acknowledgements

I would like to thank my committee chair and advisor, Dr. Lauren Klein, for guiding me and helping me refine this project to its current state. Her continual encouragement and insightful comments, especially the more critical ones, really pushed me to critically think about these important research questions and ultimately brought this work to a higher level.

I would also like to thank the other members of the committee, Dr. Jinho Choir and Dr. Krzysztof Karbownik. Our discussions throughout the project and especially during the final defense brought to light new avenues of inquiry.

Finally, I would like to thank my family and friends for their mental and emotional support. As it turns out, conducting research during a pandemic is infinitely more difficult than during normal times, and without their presence (both physical and online) I would not have been able to see this project to its fruition.

# Contents

## Tables and Figures

## Introduction

The 2020 COVID-19 pandemic has generated large-scale sociopolitical change around the world. Many countries in the beginning of the pandemic started with lock downs and quarantines followed by calls for social distancing in public spaces. In the United States, those macro-political actions, however, did not translate into a public health victory. In fact, despite comprising only 4.25% of the world's population, the U.S. has contributed to 25% of total COVID-19 cases and 20% of COVID-19-related deaths (Johns Hopkins University ). Where did the U.S. go wrong with its public health response? One possible explanation is the dissemination of public health information and its interpolation by the polarized media landscape within the U.S.

Since the 1980s, various mediums of news have become politically polarized, with the trend particularly increasing in the past few decades (Martin and Yurukoglu 2017). This polarization has affected the distribution of information during the pandemic as well. Anecdotally, conservative news sources have been shown to provide contrary information about the pandemic in comparison to liberal news sources, either downplaying the severity of the pandemic or directly providing misinformation about the pandemic to their audience (Ingraham 2020). Multiple studies have shown that consumption of these conservative news sources have correlative (and in some cases, causative) relationships with a person's beliefs about COVID-19 and their actions in following public health measures (Burstzyn et al. 2020, Jamieson and Albaraccin 2020, Simonov et al. 2020).

This project seeks to contribute to this literature base by exploring different aspects of the digital news environment with respect to COVID-19 coverage. Most of the existing literature have not put as much focus on collating these digital news articles and examining their trends in a comprehensive manner (Hart, Chinn and Soroka 2020, Motta, Stecula and Farhart 2020). By combining a humanistic outlook with quantitative methodologies, this project will show that there are subtle, yet noticeable differences in how right-leaning news sources talk about COVID-19, particularly on a contextual level with elements of deflection and diversion away from the gravity of COVID-19. With clear avenues for improvements and extensions, this project will also generate further inquiry into the digital news landscape and its relation to these large-scale national crises, both for COVID-19 and beyond.

For the purposes of this study, a unique corpus of digital news articles related to COVID-19 over the course of the pandemic was constructed by querying a large news aggregation API. This corpus represents the largest, most comprehensive dataset of COVID-19-related digital news articles to date. Each article was extracted as a text file along with metadata information so that analytical techniques such as topic modeling and word embedding analysis could be applied to the corpus. With this corpus, the project will interrogate these research questions:

- What are the general topics covered in the corpus, and what, if any, marginal effect does political lean have on the proportion of news covering topics such as testing, vaccines, individual experiences etc?
- What, if any, marginal effect does political lean have on the news covering micro-political actions with respect to COVID-19 versus macro-political, governmental actions?
- How has the context around public health measures changed according to the political lean of the news source?

These questions are important because it will show how a polarized news environment disseminates information during a crisis when timely, factual information is of the highest priority. Knowing these differences among political lean, i.e. whether political lean affects coverage explicitly through topic proportions or implicitly through the contextual and linguistic differences, would also allow for corresponding variation in regulatory responses.

In summary, this paper's primary contributions are as follows:

- The creation of a new dataset of COVID-19-related digital news articles that contains nearly 300,000 articles from January 1, 2020 to December 31, 2020. Moreover, there are still additional data frames that can be processed, which could increase this number to more than one million articles.
- Topic model and word embedding analysis of the new corpus was used to show changes in the content and context of the articles based on their political lean. Specifically, right-leaning sources are shown to cover state-level content more than left-leaning sources do. Moreover, right-leaning sources tend to portray public health measures such as masking or social distancing with less gravity than their left-leaning or center counterparts.
- The annotation of 1,532 articles into classes based on their orientation toward micro-political or macro-political actions of combatting COVID-19. These annotations were then used to build a classifier that could identify micro-political versus macro-political news articles.

## Literature Review

Text mining has become more commonplace as big data analytical methods have allowed researchers to leverage large corpora for quantitative analysis, and the COVID-19 research space is no exception. One of the major points of inquiry is how Twitter, a rather unique space of the internet in terms of its velocity and connectivity, talks about COVID-19. Wicke and Bolognesi (2020), for example, examined COVID-19 discussions on twitter with respect to contextual framing, showing that discourse about COVID-19 bears a striking similarity to common wartime discourse. Other similar studies have also been conducted throughout the pandemic on Twitter-based datasets, including analysis on the perception of COVID-19 policies, the prevalence of COVID-19 conspiracies and misinformation being propagated by bots and users, and the spread of pandemic visualizations by coronavirus skeptics (Ahmed et al. 2020, Ferrara 2020, Lee et al. 2021, Lopez et al. 2020, Yang et al. 2020). There has even been a collaborative effort at an international scale to collate and create a COVID-19 Twitter chatter dataset, demonstrating a sizable energy in regard to Twitter-based analysis within the research community (Banda et al. 2020).

Research on digital news and "mainstream news" and their relation to COVID-19 coverage, however, has not been as common. There are some studies that analyze this, for example, Hart, Chinn and Soroka (2020) take a computer-assisted content analytic approach to analyzing newspaper and network news coverage, finding a high degree of politicization and polarization of COVID-19 news coverage from March to May 2020. Another similar study conducted by Motta, Stecula and Farhart (2020) uses digital news stories gathered via MediaCloud to analyze the ways in which right-wing media organizations spread misinformation about COVID-19 during the early stages of the pandemic. Both of these studies, however, were published in the early stages of the pandemic and focused on particular major news sources, which leaves room for further inquiry, particularly with the development of COVID throughout the summer and the rest of 2020.

One study that seems to be similar in scope is the work done by Yejin (2020). Their project, which leveraged the BYU-BNC corpora and its recently added Coronavirus Corpus, looked to analyze the ways in which COVID-19 was conceptualized in contemporary media using methods such as word collocation analysis (Yejin 2020). This Coronavirus Corpus is fairly large in its scope in terms of its content and timespan, especially since it seems to be consistently updated. Currently, it has 115,986 pieces of texts from online magazines and newspapers from 20 different English-speaking countries from January 1st, 2020 to the current time. This project, however, works to expand upon the dataset used by Yejin (2020) by constructing a new and potentially more holistic corpus of digital news.

In terms of analytical techniques, this project uses an analytical pipeline that is modeled after prior works that leveraged quantitative modeling for the evaluation of large textual corpora. In many ways, this work was inspired by Martin and McCrain (2019), which attempted to consolidate a large-scale corpus of broadcast transcripts to show changes in news content using a local versus national politics distinction. In that work, they primarily leveraged topic modeling analysis to determine whether an individual transcript was covering local or national news content, and the changes were estimated using a difference-in-difference estimator. This work, similarly, seeks to construct a comprehensive corpus of digital news articles and utilize unsupervised topic models and supervised learning classification methods to show changes in content along a political lean dimension.

This study also improves upon the previous work by analyzing changes through a contextual lens. This is primarily done through creating word embedding models of different slices of the corpus and using a semantic shift algorithm to verify the contextual changes. This process of detecting semantic shift is well-documented, notably by Hamilton, Leskovec and Jurafsky (2016) and Azarbonyad et al. (2017). Guo, Xypolopoulos and Vazirgiannis (2021) has also applied the semantic shift detection algorithm to a COVID-19 Twitter dataset. Within the

COVID-19 literature base, however, there has not been a systematic study to our knowledge that uses these different methods in tandem.

## Data and Methods

### General Overview

To analyze the effects of media polarization on the coverage of COVID-19, a unique corpus that encompasses a wide variety of news sources from across the political spectrum and from different levels of news coverage was required. Thus, this study has set out to start the collation of these news sources, with the hope that continued work may also be done on this front similar to the work started by Banda et al. (2020) in the COVID Twitter research space. The corpus that was built for this study is a novel collection of digital news articles from various news sources across the political spectrum. It includes articles from both national news sources such as CNN, NBC news, FOX news etc. as well as local/regional news sources such as the Arizona Republic, the Atlanta Journal Constitution, the Minneapolis Star Tribune etc. These articles were collected via an API called NewsAPI, which finds articles in real-time and indexes them internally for up to 3 years ("Documentation" 2018). The API was queried using a keyword-based approach for any article from January 1, 2020 to December 30, 2020 that contained the words "COVID" or "coronavirus", and the results were processed and stored in Google Cloud storage. This long timespan allows for a comprehensive coverage of the COVID-19 pandemic from its beginnings as an isolated story at the start of 2020 to the global phenomenon it becomes three months later, where it has endured as a long-term pandemic, affecting life at all levels regardless of background.

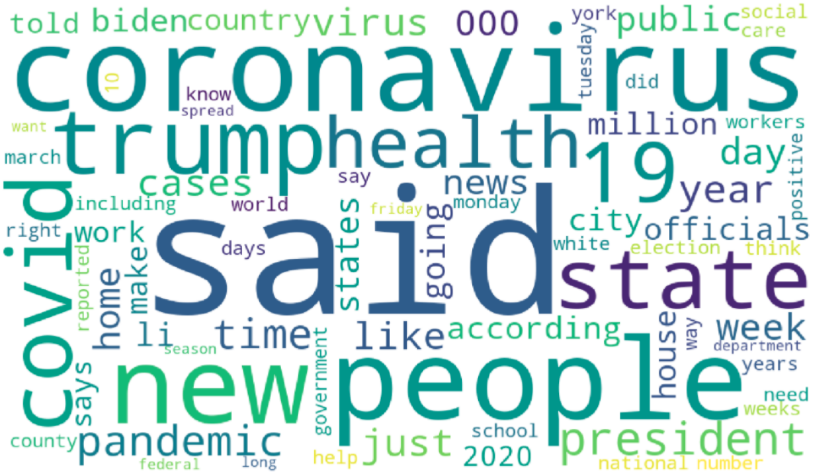
While there are a variety of news sources that have been collected, this paper focuses on a subset of the news sources (n =88), which are believed to be a representative sample of the U.S. media landscape. This subset has 839,713 unique tokens (not including stopwords) and approximately 209 million total words, with an average of 713 words per document. In terms of pre-processing, residual html tags from the API and stopwords were removed as much as possible prior to running any analysis.

In comparison to the Coronavirus Corpus noted earlier, this corpus has a couple differences. First, it is relatively more expansive, as the subsample has more than twice the number of documents (291,810 articles vs. 115,986). Second, this corpus contains minimal transformations and edits on the original full text, which the Coronavirus Corpus was required to do to avoid copyright/legal issues.

Some caveats to point out with the data at this stage are as follows. First, URLs that provided a video link were skipped in the collection process because there was no way to extract the text information. Thus, discussion about this corpus can only focus on the text-based news. The second caveat is that some news sources had empty content features within the API. This was resolved somewhat by manually scraping the missing content, however that is also limited by a website’s restrictions on web scraping tools.

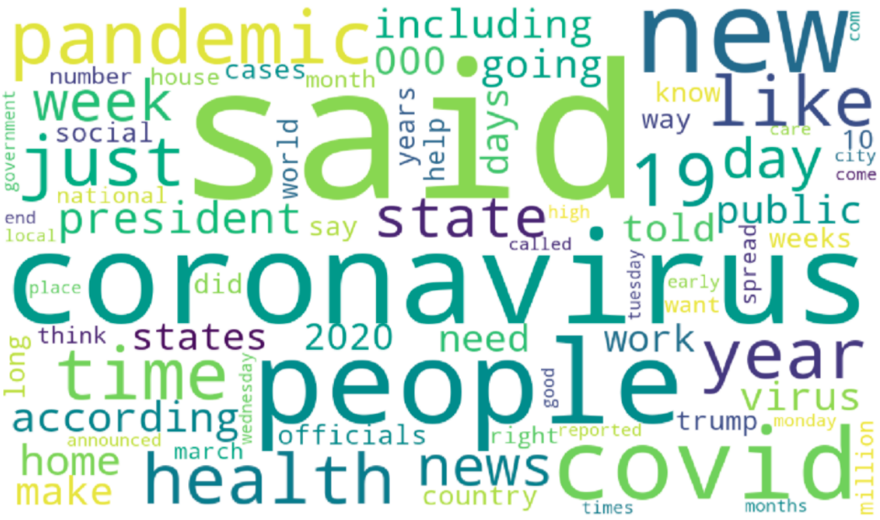
Figure 1 shows a simple visualization of the overall corpus after removing common English stop words using sci-kit learn’s built-in stopwords list. As expected, words associated with the crisis such as “coronavirus”, “covid”, and “pandemic” appear alongside political terms like “Trump”, “officials”, and “government”. There also seems to be words related to the impact of the pandemic i.e. words such as “million”, “workers”, “cases”, and “care”.

Figure 1: Full Corpus WordCloud (Font Size varies with Count Frequency)



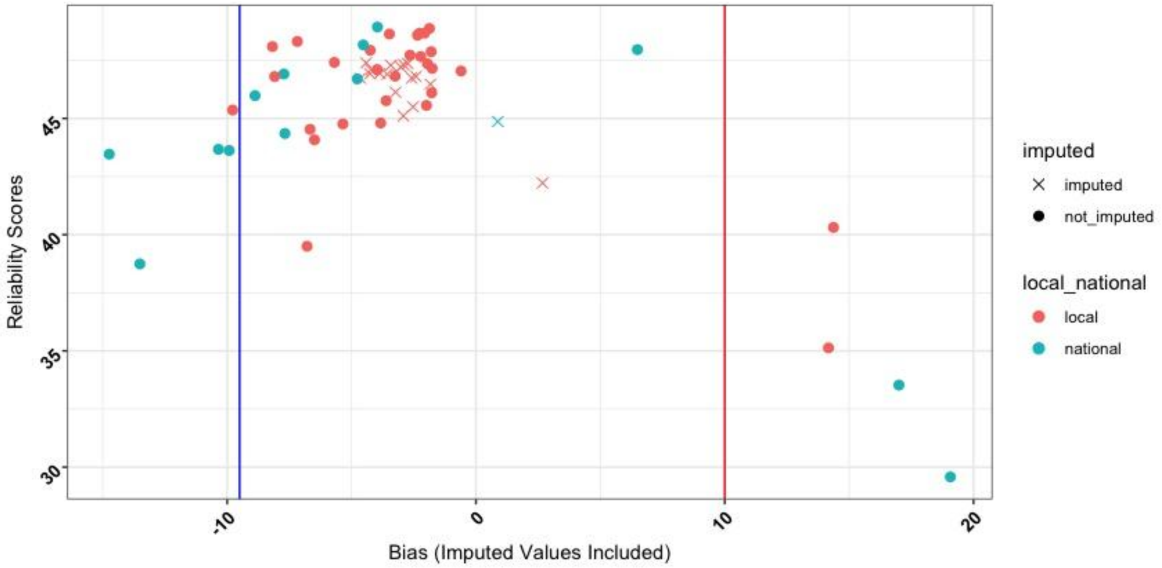
A tf-idf representation, which also uses counts but also takes into account the unique significance of the word in each document, shows a similar distribution of the words (Figure 2). Slight differences between this and the simple count-based word cloud are the different weightings, for example the word "Trump" seems to be weighted less in the tf-idf version while the words "pandemic" and "virus" are weighted more.

Figure 2: Full Corpus WordCloud (Font Size varies inversely with TF-IDF Weight)



Additional data that is leveraged include political lean and reliability of the news source, which was collected from Ad Fontes Media bias database. While not a perfect representation of political bias, their qualitative, multi-analyst approach to rating based on criteria such as veracity, language, political position etc. seems to suggest a holistic approach to political bias rating (Otero 2019). For the sources used, they were filtered down to 60 sources that had more than 100 articles over the course of the timespan. This was done to prevent noisy topic probability estimates from affecting the imputation. Of the 60 sources, there were 18 sources that had missing political bias scores. These scores were imputed with topic probabilities and reliability scores as covariates using a random forest model, specifically the missForest package developed by Stekhoven and Bühlmann (2012). Using a random forest model, in theory, places minimal assumptions on the functional form of the model (Stekhoven and Bühlmann 2012). This procedure returned a normalized root-mean-square out-of-bag error of 0.2, which suggests a reasonable imputation for our needs. Figure 3 shows the distribution of the news sources in terms of their political lean and reliability scores. The local/national distinction was determined by whether the newspaper was supposedly tied to a locality ex. Los Angeles Times would be considered a local newspaper despite occasionally covering national news.

Figure 3: Bias Vs. Reliability (60 Sources)



For these political lean groupings, separate tf-idf counts were calculated with the top-15 from each segment shown in table 1. The differences between the tf-idf counts are minimal for the most part other than some weighting differences. For example, the word "president" is weighted slightly more in the left-leaning bucket than in the right-leaning bucket (the word is just outside the top-15 range). The tf-idf counts for the local/national buckets are roughly the same as well. The major differences along that dimension is the higher weighting of the word "president" and "trump" in the national bucket while words such as "state" and "cases" are weighted slightly more in the local bucket.

Table 1: TF-IDF Ranking of Words by Political Lean

| **Left-Leaning** | **Center** | **Right-Leaning** |
|---|---|---|
| said (1.209) | said (1.201) | coronavirus (1.138) |
| coronavirus(1.257) | coronavirus (1.246) | said (1.207) |
| people (1.400) | people (1.464) | people (1.581) |
| new (1.461) | 19 (1.467) | new (1.618) |
| pandemic (1.483) | covid (1.470) | pandemic (1.676) |
| covid (1.499) | new (1.476) | news (1.691) |
| 19 (1.509) | pandemic (1.497) | covid (1.712) |
| health (1.676) | time (1.568) | 19 (1.743) |
| just (1.707) | health (1.704) | time (1.770) |
| like (1.716) | year (1.730) | just (1.882) |
| year (1.757) | state (1.730) | year (1.949) |
| week (1.820) | just (1.740) | health (1.950) |
| president (1.827) | week (1.766) | told (1.969) |
| day (1.832) | like (1.768) | like (1.989) |
| state (1.846) | day (1.807) | according (2.002) |

### Topic Model Analysis

To analyze the content and context of the news sources, a variety of methods were employed. For content analysis, two different topic modeling packages were used: Gensim and MALLET (Machine Learning for Language Toolkit). Both were used for different reasons. Gensim, which uses a Variational Bayes sampling method that is less precise compared to MALLET's Gibbs sampling, allows for an output to be quickly generated (Hoffman, Bach and Blei 2010). The toolkit also comes with an internal visualization tool using pyLDAvis, which creates a PCA-decomposed graph of the topics for interpretation. MALLET, on the other hand, is not only more precise, it retains as an output the topic probabilities, which can be used later to slice and examine the among-source and across-time trends for an individual topic (Yao, Mimno and McCallum 2009). For both of these implementations, the full corpus of articles was used. Because of their bag-of-words approach, both Gensim and MALLET performed better on the whole corpus, and subsequent among-source and across-time analyses were performed by re-aggregating the topic probabilities.

Figures 4-6 show the gensim PCA-decomposed visualization of the topic model that was run on the overall corpus. The number of topics was chosen by calculating the u_mass coherence score for a range of topics, with a lower u_mass score generally indicating better topic coherence. According to figure 7, the best number of topics is around 17-20 topics, and 20 topics were chosen for both gensim and later mallet. As shown in the visualizations, there are fairly distinct topics occurring within the corpus. Figure 4 highlights a sizeable topic (11.2% of tokens) that seem to include words related to federal/state public health actions such as "state", "federal", "official", "public", "plan" etc. Interestingly, this topic is separated out from topic 17 (Figure 5), which features explicit mentions of social distancing, masks and gatherings. Other notable topics include mentions of unemployment relief and social aid (Figure 6), and topics of vaccines (20), testing (14), case counts (4) and patient care (12), which are all clustered together in the lower-right quadrant.

Figure 4: Federal/State Public Health Topic

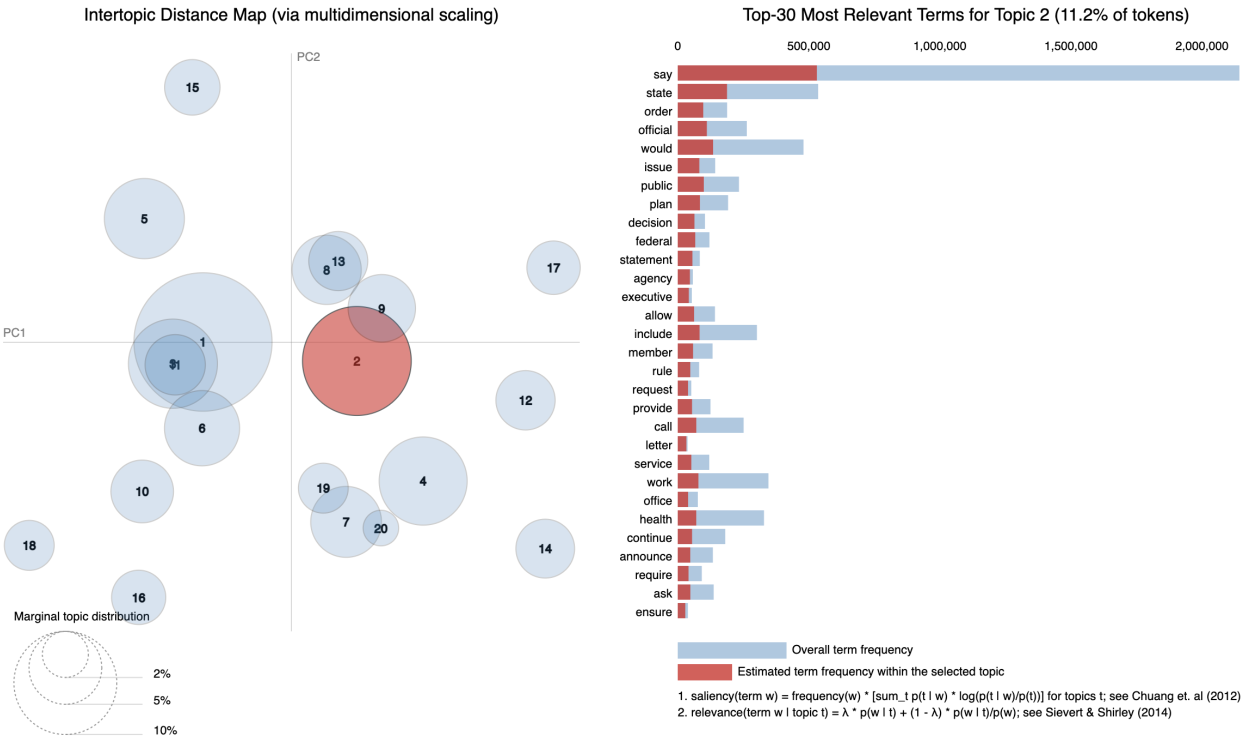


Figure 5: Public Health Measures Topic

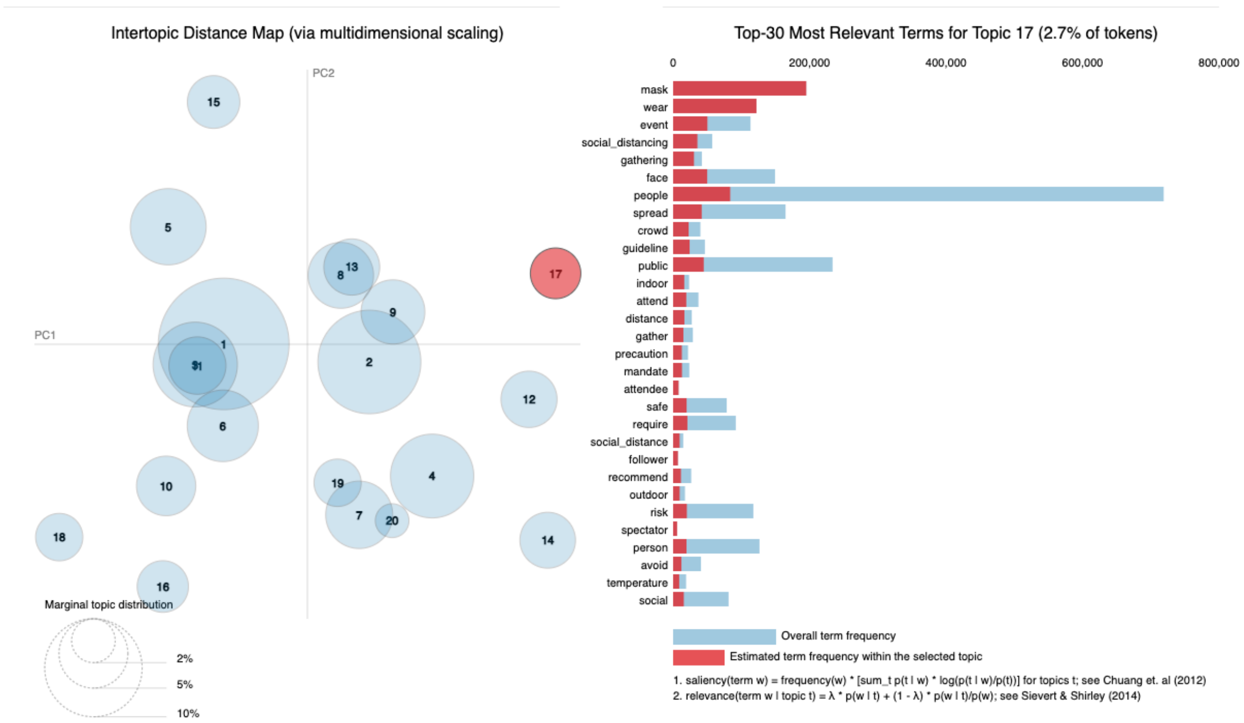


Figure 6: Unemployment Relief and Social Aid Topic

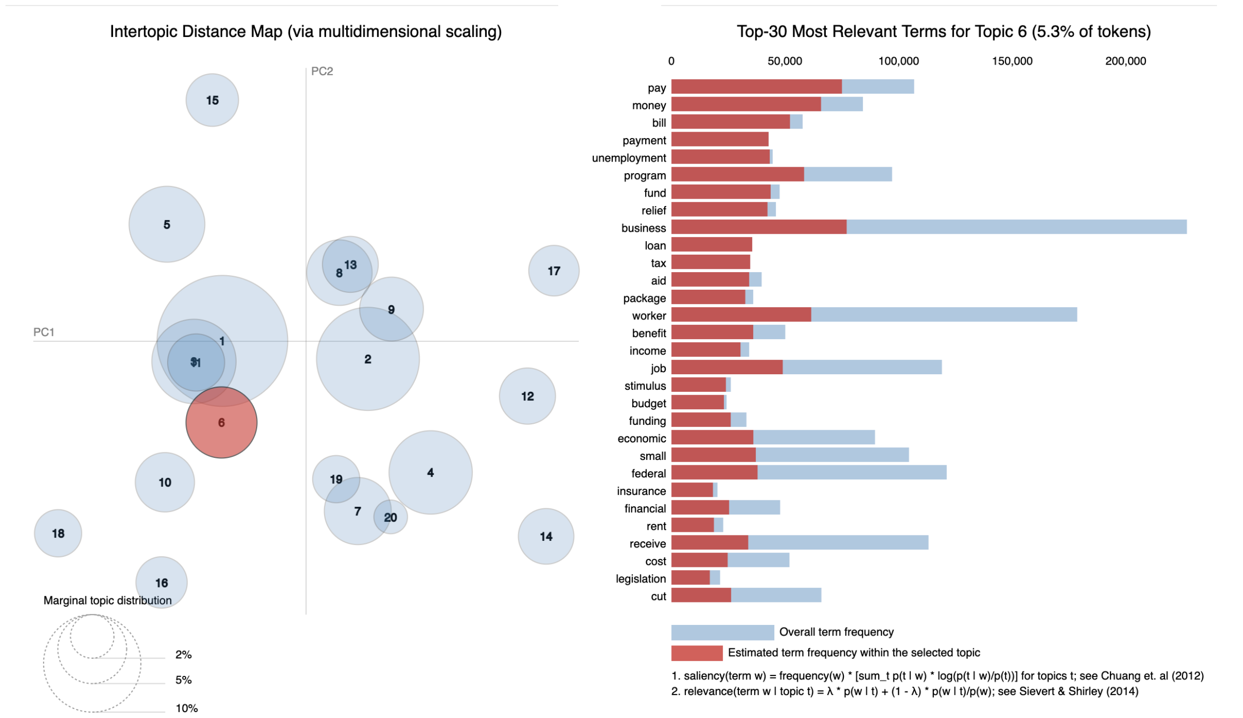
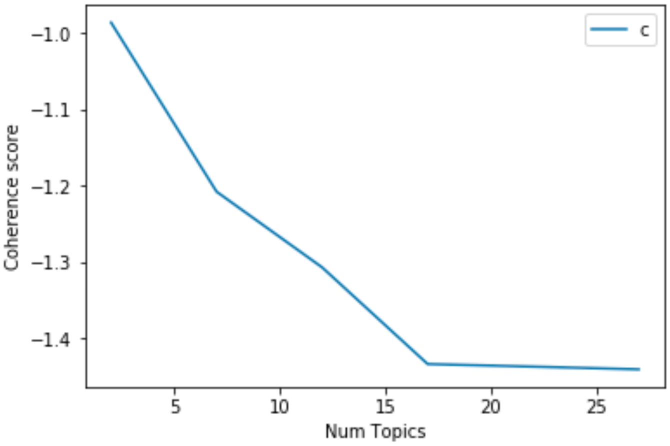
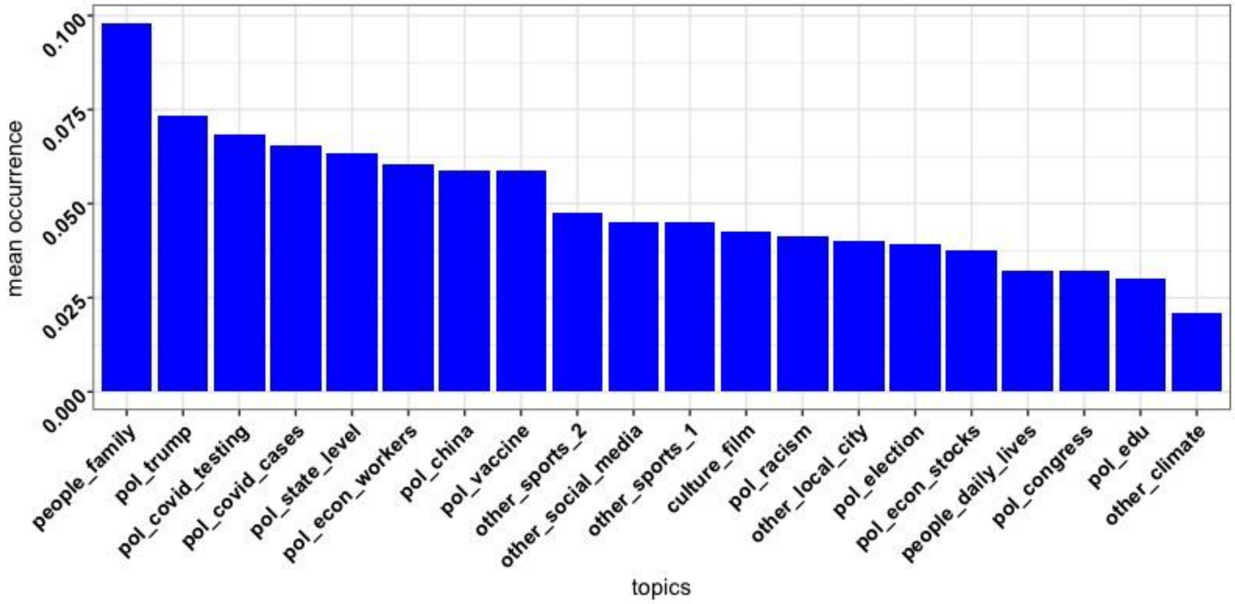


Figure 7: Mallet Coherence Score by Number of Topics



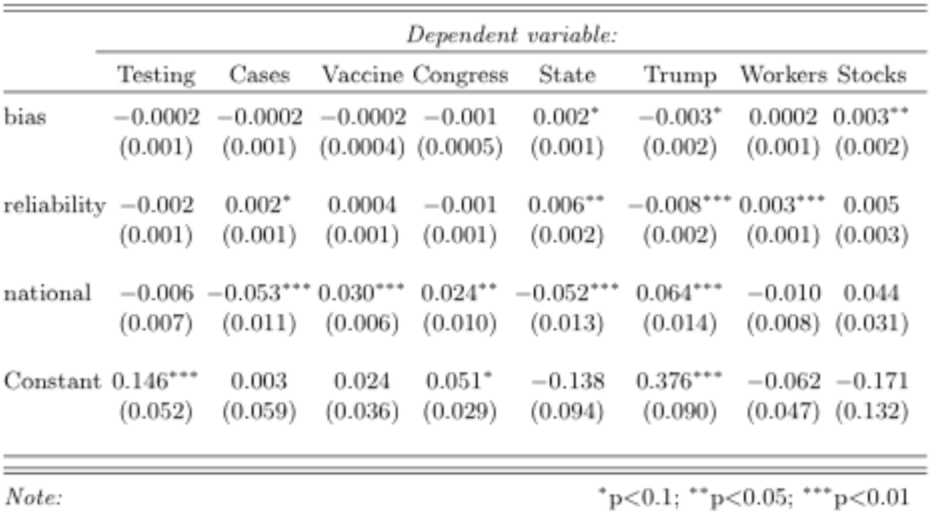
The MALLET version of the topic model provided similar results, with some slight variation in topic percentages. Looking at Figure 8, the overall trend still holds with distinct topics related to covid (testing, cases, vaccines etc.). This visualization seems to suggest that the mask/social distancing topic found by the gensim implementation was subsumed by the larger covid testing/cases topics.

Figure 8: Mallet Mean Occurrences of Topics in Corpus



Using the topic proportions from MALLET, a few regressions were tested using both the political lean scale and the lean buckets. These results, which use robust standard errors to account for heteroskedasticity in the bias data, are shown in the following tables. From the results, it appears that the political lean of the news source plays a smaller role than previously expected on the topic frequencies. The possible exceptions are the topics surrounding states and Trump, which seems to align somewhat with the idea that right-leaning news sources would focus less on Trump's actions related to the pandemic and instead deflect by covering more state and local level politics. That result, however, does not seem to be robust, as the bucketed version of this regression seems to indicate minimal effect of political lean on the state and Trump topics. In fact, whether a news source is locally or nationally-based seems to have a more significant correlation with the battery of topics, even through the robustness checks.

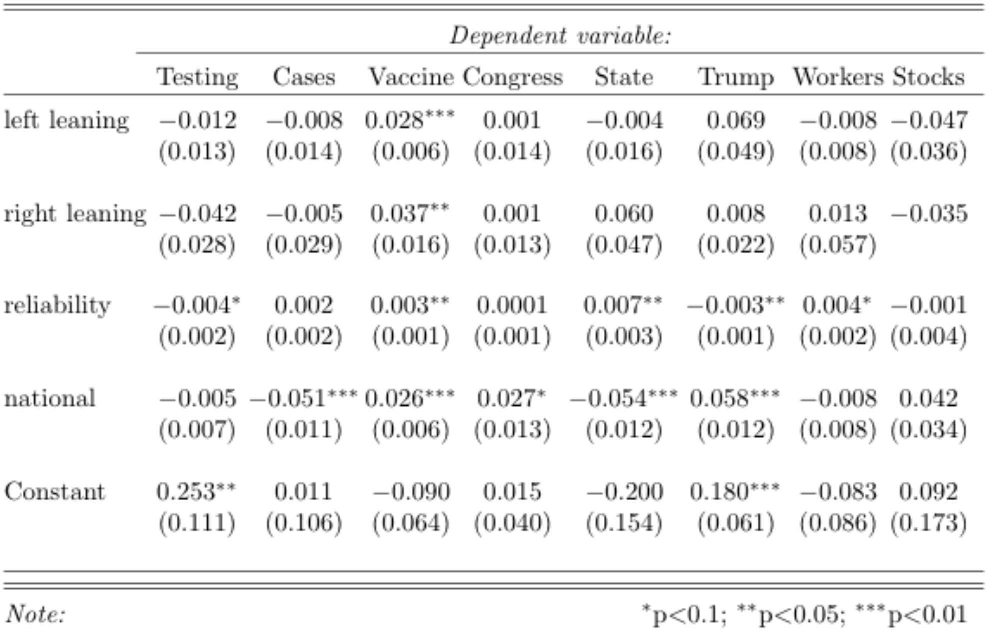
Table 2: Topic Probability Averages Regressed on Political Lean Score



| | *Dependent variable:* | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Testing | Cases | Vaccine | Congress | State | Trump | Workers | Stocks |
| bias | −0.0002 | −0.0002 | −0.0002 | −0.001 | 0.002* | −0.003* | 0.0002 | 0.003** |
| | (0.001) | (0.001) | (0.0004) | (0.0005) | (0.001) | (0.002) | (0.001) | (0.002) |
| reliability | −0.002 | 0.002* | 0.0004 | −0.001 | 0.006** | −0.008*** | 0.003*** | 0.005 |
| | (0.001) | (0.001) | (0.001) | (0.001) | (0.002) | (0.002) | (0.001) | (0.003) |
| national | −0.006 | −0.053*** | 0.030*** | 0.024** | −0.052*** | 0.064*** | −0.010 | 0.044 |
| | (0.007) | (0.011) | (0.006) | (0.010) | (0.013) | (0.014) | (0.008) | (0.031) |
| Constant | 0.146*** | 0.003 | 0.024 | 0.051* | −0.138 | 0.376*** | −0.062 | −0.171 |
| | (0.052) | (0.059) | (0.036) | (0.029) | (0.094) | (0.090) | (0.047) | (0.132) |

*Note:* *p<0.1; **p<0.05; ***p<0.01

Table 3: Topic Probability Averages Regressed on Political Lean Buckets

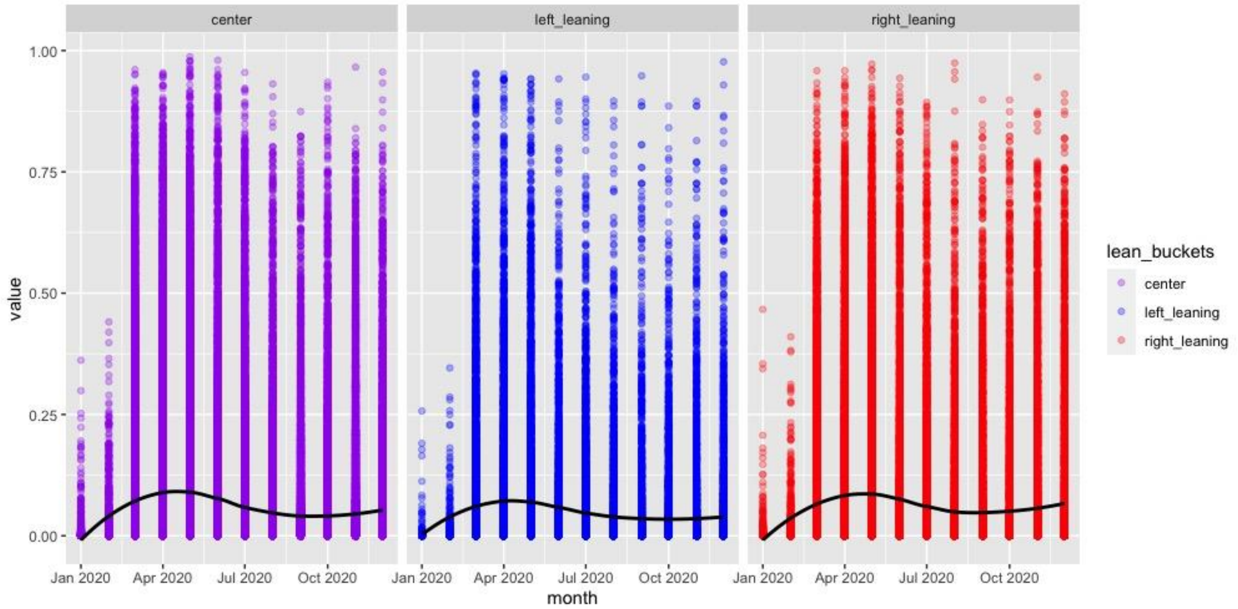


| | *Dependent variable:* | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Testing | Cases | Vaccine | Congress | State | Trump | Workers | Stocks |
| left leaning | −0.012 | −0.008 | 0.028*** | 0.001 | −0.004 | 0.069 | −0.008 | −0.047 |
| | (0.013) | (0.014) | (0.006) | (0.014) | (0.016) | (0.049) | (0.008) | (0.036) |
| right leaning | −0.042 | −0.005 | 0.037** | 0.001 | 0.060 | 0.008 | 0.013 | −0.035 |
| | (0.028) | (0.029) | (0.016) | (0.013) | (0.047) | (0.022) | (0.057) | |
| reliability | −0.004* | 0.002 | 0.003** | 0.0001 | 0.007** | −0.003** | 0.004* | −0.001 |
| | (0.002) | (0.002) | (0.001) | (0.001) | (0.003) | (0.001) | (0.002) | (0.004) |
| national | −0.005 | −0.051*** | 0.026*** | 0.027* | −0.054*** | 0.058*** | −0.008 | 0.042 |
| | (0.007) | (0.011) | (0.006) | (0.013) | (0.012) | (0.012) | (0.008) | (0.034) |
| Constant | 0.253** | 0.011 | −0.090 | 0.015 | −0.200 | 0.180*** | −0.083 | 0.092 |
| | (0.111) | (0.106) | (0.064) | (0.040) | (0.154) | (0.061) | (0.086) | (0.173) |

*Note:* *p<0.1; **p<0.05; ***p<0.01

Figure 9 shows the state topic over time, split by the lean buckets. There does in fact seem to be a slight increase in the state topic from the right-leaning bucket, especially when compared to the left-leaning bucket. The spike is most prominent during the April-May months (likely a coronavirus spike) and the October-November months (as a function of election season).

Figure 9: State Level Topic Proportions over Time



**Classifier Section**

The next point of analysis centers on the likelihood of articles to center their story on micro-political action or macro-political actions to combat COVID. To answer this question, a supervised-learning text classifier was employed to determine an article's orientation. A randomized subsample of the news sources (n = 1532) was extracted and labeled by the author in order to train a text classification model. During the labeling process, one of four labels were given to each text piece, the counts of which are shown in table 4. The categories are as follows:

**Neither Micro nor Macro-political actions (1)**: This category consists of content that is not included in the following three categories. Examples include (but are not limited) articles about the NFL/NBA season, international news stories, and various "lifestyle" stories. Articles about government or politics unrelated to COVID ex. campaigns, George Floyd/race protests etc. are also included in this class. Essentially, if an article did not fit the next few categories, then it was coded as "1".

**Primarily Micro-political Actions (2)**: This category largely corresponds to on-the-ground actions that would be considered a part of the ongoing effort to mitigate effects of COVID. This would include public health measures if the article content centered on the individual ex. the process of social distancing, mask wearing etc. This can also include nongovernmental organizations and their implementation of these public health measures ex. discussions of NFL, prisons, etc. and their testing/quarantine practices. This category also encompasses news stories that focus on frontline workers both medical and non-medical. If an article's content was roughly 50% or more of these aspects, then the article was coded as "2".

**Primarily Macro-political Actions (3)**: This category primarily includes governmental actions such as congressional legislations that dealt with COVID relief. Those legislations/policies include but are not limited to unemployment aid, small-business loans, bailouts, vaccine development etc. State level policymaking was included as well if it was primarily shown as a top-down approach, for example articles that indicated the governor was shutting down schools or indoor dining venues. If an article was found to be primarily talking about governmental actions or any other macro-political actions, then the article was coded as "3".

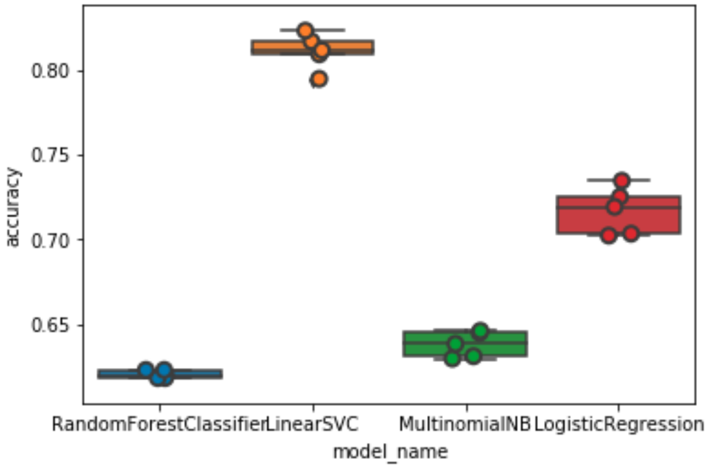
**Mixture of Micro and Macro-political actions (4)**: This category is a mixture of the prior two categories. Considering the nature of the pandemic, it would be expected for articles to mention both micro-political and macro-political actors in their content. This could include governors urging citizens to quarantine and social distance or vaccine distribution plans that include community actions of information spreading or volunteering. This category also includes individualized experiences related to COVID for government officials, for example if a senator contracted COVID and was forced to quarantine. If the article was found to be a mixture, then it was coded as “4”.

Table 4: Counts of Labels used for Classifier

| Label | Meaning | Count |
|---|---|---|
| 1 | No mention of any micro or macro-political actions to combat COVID | 811 |
| 2 | Primarily of micro-political/individual measures to combat COVID | 223 |
| 3 | Primarily of macro-political measures to combat COVID | 273 |
| 4 | Some mention of both micro-political and macro-political actions | 225 |

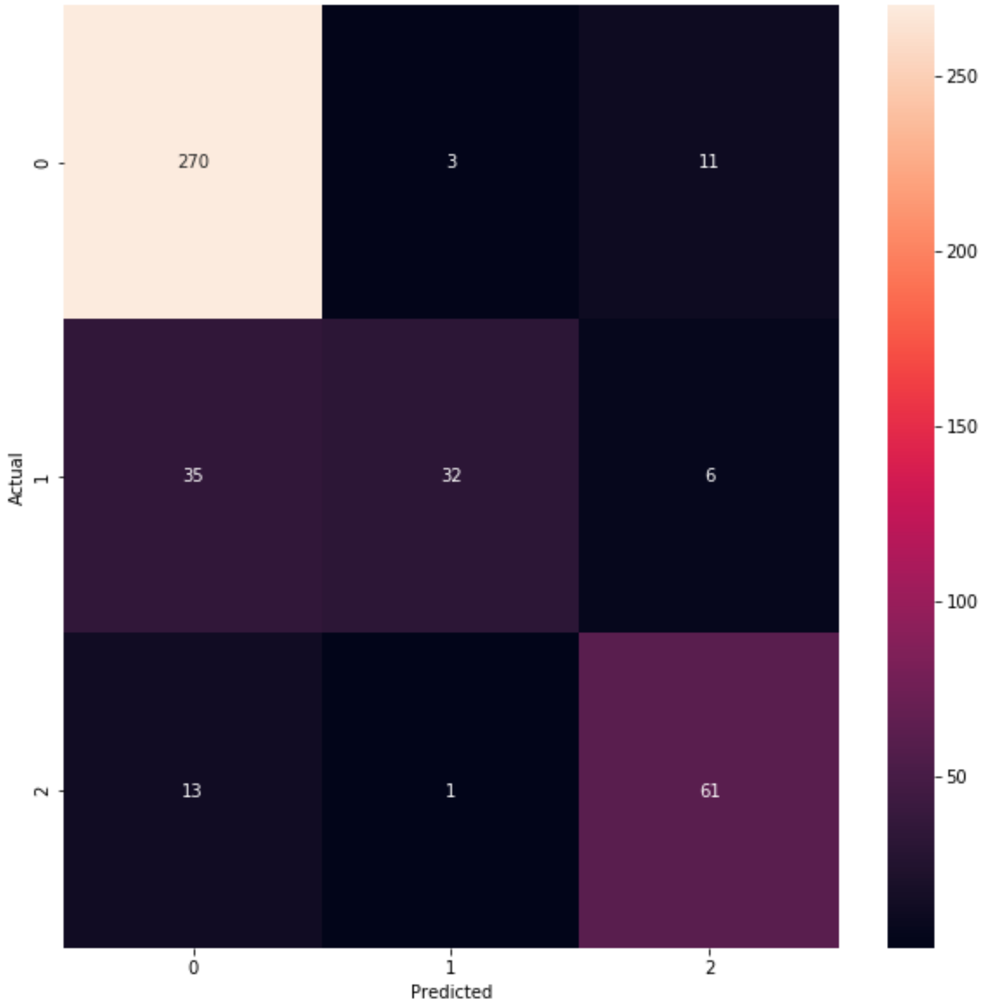
Because the primary concern is obtaining the most accurate predictions, the data was trained on a few different models that are a part of the scikit-learn package, specifically a random forest classifier, a linear support vector classifier (linearSVC), a multinomial Naïve-Bayes model, and a logistic regression model. In addition to these scikit-learn approaches, a long short-term memory recurrent neural network was also trained using the data. Each model was run through a 5-fold cross validation to determine accuracy, with the highest reported accuracy being the LinearSVC model at 73%. This classification model, while not perfect, seems to be the best that can be made with the limited sample size, especially since there is such a large variance in the types of articles that mention COVID-19. As a test, the mixed class (label 4) was removed from the model, with the reasoning that this mixed class would have too many similar elements as class 2 and 3. With the limited sample size, this heterogeneous class would likely make the model more inaccurate. Indeed, once the mixed class was removed, the test accuracy increased to approximately 80% as shown in the following figure.

Figure 10: 5-Fold Cross-Validated Accuracies



The confusion matrix of the LinearSVC's predictions using a 33% test set indicates that the model has trouble identifying differences between labels 1 and 2 i.e. articles that mention micro-political actions versus articles that have no mentions of either micro-political or macro-political actions. This could be because there are legitimate similarities between the classes. The most obvious example would be the sports-related articles. In some of those articles, the content primarily focuses on the season and games themselves, only tangentially mentioning COVID. Those articles were clearly labeled as "1". At the same time, other sports-related articles actually talked about the quarantine practices and individual players abiding by the public health measures, which were interpreted as micro-political actions and thus labeled as "2". The other likely contributor to this is the relative difference between the frequency of the micro-political class and the other labels, with the micro-political class comprising only 14.6% of the total labels.

Figure 11: LinearSVC Test Set (33%) Confusion Matrix



Given the relatively low accuracy and the evidence from the confusion matrix that shows the bad performance of the model with labelling micro-political action, there was some hesitancy with using this biased model on the full corpus and using those predictions to determine correlations between these classes and political lean. Because of this, the focus of this section of analysis shifted to doing a close reading of the texts used for labelling, particularly of the false positives and negatives given by the test set predictions. In addition to the issues related to sports stories, there are a couple more examples of micro-political actions being mislabeled as class

"1". One of them is a story published by Breitbart in May 2020 that included mentions of the Chinese government:

> "*A man in Long Island, New York, whose father died of complications from the coronavirus in April is joining others in a lawsuit against China….Guasto developed severe respiratory issues and was taken to a local hospital, where he tested positive for the Chinese coronavirus, said his son, Richard Jr. On April 15, Guasto died of complications from the disease….Guasto is now one of 15,000 plaintiffs suing the Chinese government….*"

As for the cases of micro-political labels being misclassified as macro-political (19 cases in the test set), the general trends are a little less clear. Some of them are universities closures such as an article written by CBS news in August 2020:

> "*The University of North Carolina at Chapel Hill is abruptly ending in-person instruction for undergraduates after a cluster of coronavirus cases emerged. The outbreak is a worrying development for education officials and parents….The decision was made after over 130 COVID-19 infections were reported, with clusters confirmed at four residences. More than 300 students were quarantined as of Monday, and others scrambled to make plans to return home….*"

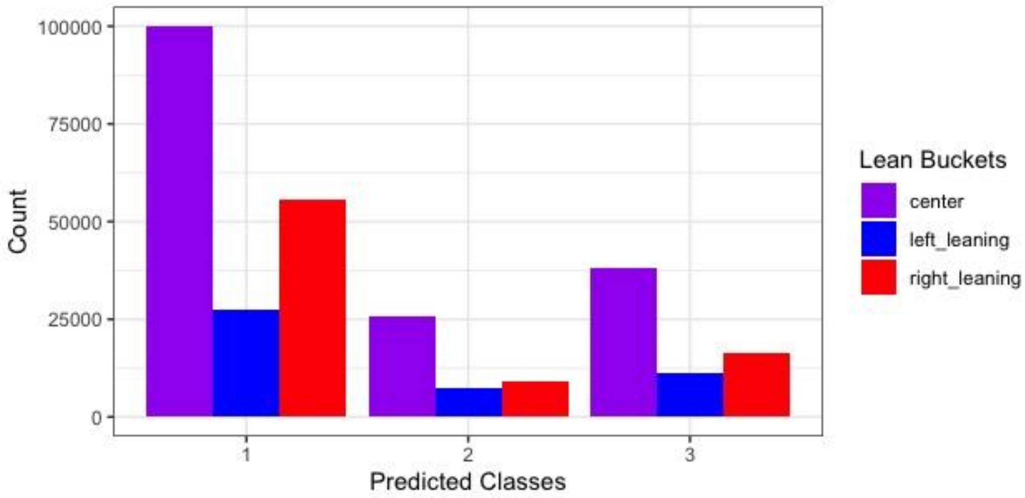
Another example is a story by NBC news in April 2020 about the cruise ships stranded in mid-trip due to coronavirus restrictions. This one could have reasonably been labelled as a mixed class (4). Regardless, it seems that the model, in its current state, is extremely sensitive to any mentions of governmental bodies, which is why this story is being classified as a macro-political class despite primarily being about nongovernmental quarantining.

> "*When Andrea Anderson and her husband boarded the MS Zaandam cruise ship….they didn't know that their trip of a lifetime would disastrously coincide with a global pandemic that would leave them shut out and stranded at sea. Unable to find a port willing to accept them, their ship has been stuck in a holding pattern for nearly two weeks as it desperately goes from country to country….officials, including Gov. Ron DeSantis, say the state simply does not have the resources to take on an extra burden amid a growing health crisis….While the governor has expressed staunch disapproval with the ship disembarking, the final say lies in the hands of the Broward County Commission who was not able to come to a decision on Tuesday….*"

These results suggest that there is a sizable group of articles where the model over-predicts based on mentions of governmental bodies or officials. This seems to suggest more that this is a sample-size issue, with more labels of class "2" being necessary. The other issue, with the over-prediction on class "1" might also be fixed by this.

Despite the potentially untrustworthy nature of the model as shown, it is still used for predictions on the rest of the corpus. Figure 12 shows the distribution of the predicted classes by the political lean buckets that were constructed. It is important to note that the frequency of class 1 and 2 are suspect given the confusion matrix results.

Figure 12: Counts of the Predicted Class by Lean Bucket



These classes are manipulated to create a micro-political class versus other dummy and a macro-political class versus other dummy. These dummies are then regressed on lean buckets using a logistic regression, the results of which are shown in the following table.

Table 5: Class Dummies Regressed on Political Lean Buckets

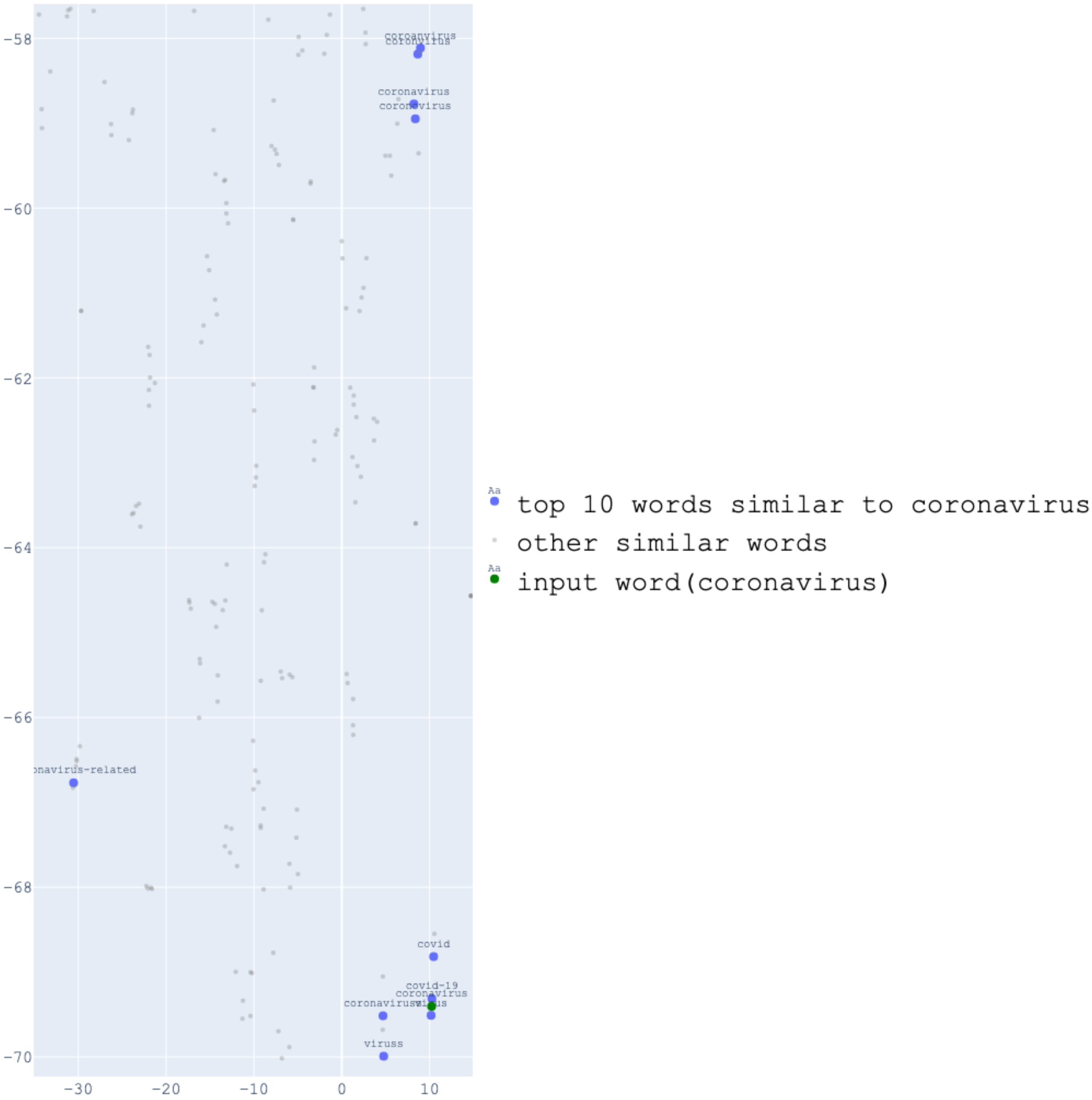
| | *Dependent variable:* | |
|---|---|---|
| | Micro-Political Class | Macro-Political Class |
| left leaning | $0.191^{***}$<br>(0.017) | $0.146^{***}$<br>(0.014) |
| right leaning | $0.172^{***}$<br>(0.037) | $0.243^{***}$<br>(0.031) |
| reliability | $0.043^{***}$<br>(0.003) | $0.034^{***}$<br>(0.002) |
| national | $-0.266^{***}$<br>(0.012) | $0.230^{***}$<br>(0.010) |
| Constant | $-3.585^{***}$<br>(0.134) | $-2.920^{***}$<br>(0.111) |
| Observations | 291,014 | 291,014 |
| Log Likelihood | −119,420.200 | −154,947.700 |
| Akaike Inf. Crit. | 238,850.500 | 309,905.500 |
| *Note:* | | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

### Word Embedding Analysis

For examining the contextual changes of certain words throughout the corpus, word2vec embeddings were created for the corpus. Word2vec embeddings are vectorized representations of words found within the corpus. These vectors are trained using a shallow neural network to learn the linguistic context of words in the corpus, for example, in a continuous bag-of-words architecture of the word2vec model, the current word is predicted from a window of surrounding context words (Mikolov et al. 2013). A window size of 5 and a dimensionality of 100 for the word vectors was chosen as the parameters for the model. With these vector embeddings, cosine similarities can be calculated between words, with larger cosine similarities corresponding to more closely associated words.

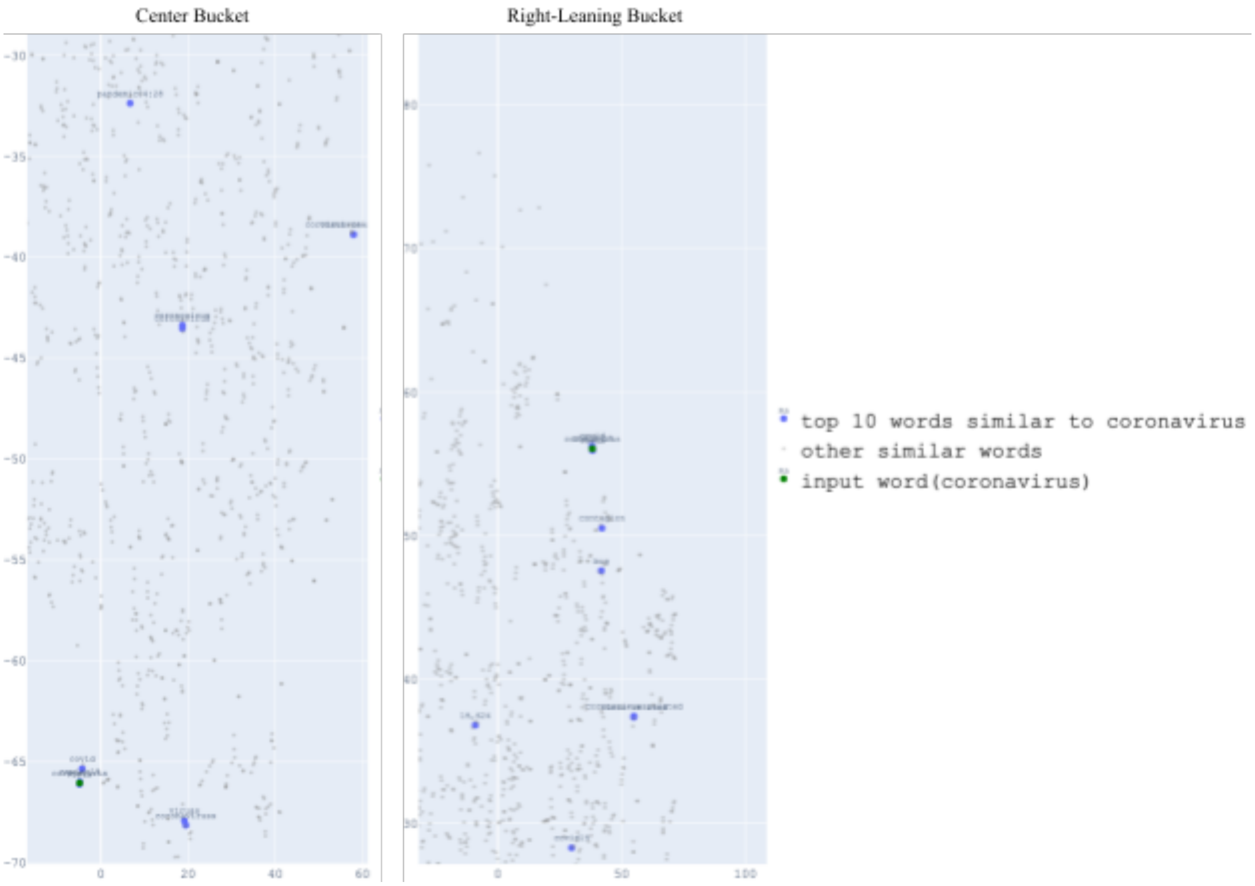
These word embeddings are also analyzed across the different political leanings of news sources. Because of the training process, however, the word2vec model on the whole corpus is unable to be disaggregated once the model is created; thus, separate "mini-models" based off the source lean (grouped by the political lean scores used during the topic modeling section) were made for comparison. T-distributed stochastic neighbor (t-SNE) plots, which try to project high-dimensional data down to a two-dimensional plane, were used to visualize and interpret the word embedding models (Chen, Tao and Lin 2018). Specifically, words such as coronavirus, masks, trump, and vaccine were chosen based on the close reading done in the classification section and given to the model, returning the top-10 closest words, which were then mapped via t-SNE to scatterplots. Figure 13 shows an example of this using a word2vec model of the full corpus. The semantic shift algorithm used by Hamilton, Leskovec and Jurafsky (2016) was also utilized to look for and verify contextual changes in the same words used for the t-SNE plots.

Figure 13: Top-10 Most Similar Words to Coronavirus (Full Corpus)



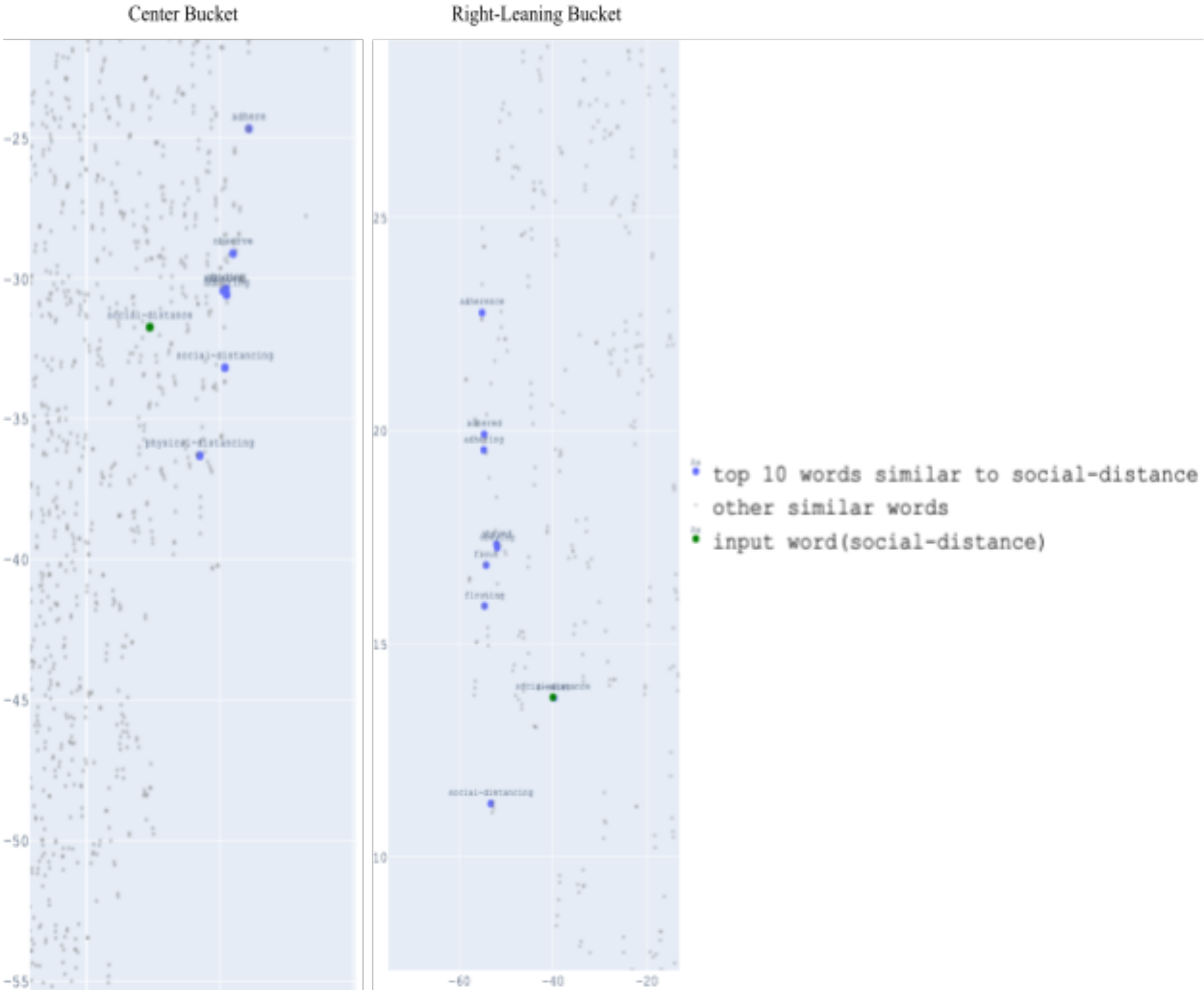
Looking at the sliced versions, noticeable differences in context appear. For example, "coronavirus" appears alongside words such as "bug" and "contagion" in the right-leaning news bucket. The center news bucket, however, mostly contains variations of the word "coronavirus", with minimal usage of similar sensational words. The semantic drift measure returned a cosine difference score of about 0.12, which suggests some semantic differences between the two. This trend holds for other variations, such as covid-19, which has a semantic drift score of 0.13.

Figure 14: Top-10 Most Similar Words to Coronavirus



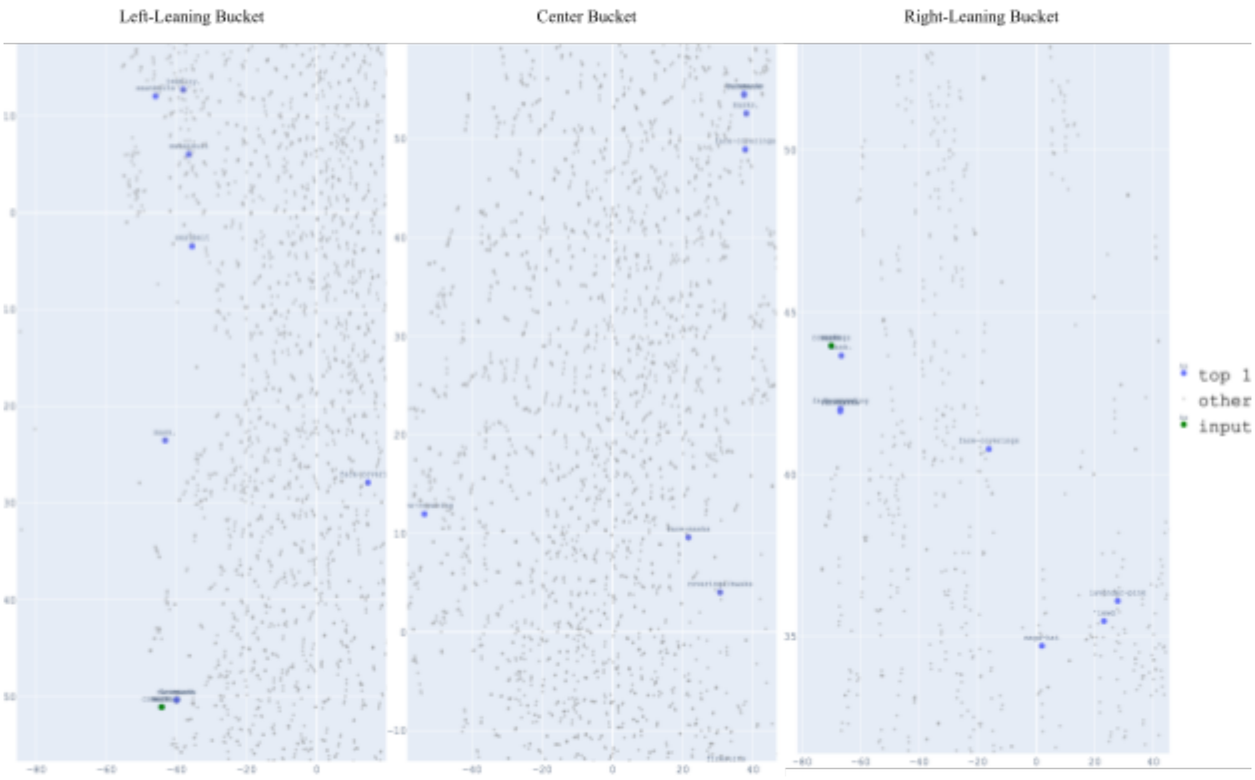
As another example, the most similar words of “social-distance” within the center-leaning bucket are words such as “adhering”, “abiding”, and “observe”. The right-leaning bucket, on the other hand, contains words such as “obeyed” and interestingly “flout”, which are antonyms of each other. Other interesting neighboring words not shown by the visualization are words such as “unscientific” and “anti-socializing” in the right-leaning bucket, whereas the center bucket maintains fairly consistent words such as “cdc-recommended” and “respecting”. Overall though, the semantic shift of the two models is about 0.06, which is not that large. This is likely due to the most close words being similar even though the fringe words are somewhat different.

Figure 15: Top-10 Most Similar Words to Social-distance



In a similar vein, the word "mask", has mostly similar words throughout the three bucketed models except for the right-leaning bucket. The right-leaning bucket contains the word "maga-hat", which by itself suggests some level of polarization. The semantic shift measure seems to corroborate this difference to some degree, with the score being a 0.13 and 0.07 difference between the left-leaning/center buckets and the right-leaning bucket respectively.

Figure 16: Top-10 Most Similar Words to Mask



Other words that were tested, such as “government”, “relief”, “ppe”, “vaccine”, etc. yielded minimal changes in neighboring words and semantic shift scores.

## Discussion

The information conveyed by news stories is commonly assumed to be interpolated and “spun” by their source based on partisan lean. This study, however, has so far found only minimal signals of this effect. From the topic modeling analysis, most of the topics (ex. testing, vaccines, working/unemployment conditions etc.) had non-significant correlations with the political lean scores and buckets after taking into account robust standard errors. The couple of topics that seemed to be significant was the topic related to state-level politics. With an effect size of 0.002, this would suggest that a right-leaning news source (ex. FOX News) would have an average state topic proportion that is 4% higher than a left-leaning news source (ex. MSNBC), at least for their digital text articles.

The classifier also shows some interesting results, though with the caveat that coefficients are likely biased due to the accuracy issues. The logistic regression suggests that the probability of an individual article being about micro-political action increases more for left-leaning articles than for right-leaning articles. The substantive value, however, is fairly small, with only a .2% difference (15.4% left-leaning to 15.2% right-leaning). Conversely, focus on macro-political action increases slightly more for right-leaning articles than left-leaning articles by about 2 percentage points (26.9% left-leaning to 28.9% right-leaning). These substantive values assume the average reliability score and national news sources for interpretation.

One of the problems with these content-focused approaches, however, is the inability to determine directionality of the content. For example, it is possible that left-leaning and right-leaning sources do cover the COVID preventative measures such as social-distancing and mask-wearing at the same proportions but with one side espousing the benefits while the other disparages the measures. This contextual effect could be picked up by the word embedding models that were built, and for some words there were signs of a signal. The words that were most likely to have large semantic differences are the public health measures initiated during the pandemic, including words such as “social-distance” and “mask”. In particular, the right-leaning news sources seem to flout and use words that de-emphasize the scientific and safety aspects of these public health measures. For example, Breitbart, wrote in a piece around September 2020:

> *“Sen. Dianne Feinstein (D-CA) was busted strolling through an airport Friday after demanding the Federal Aviation Administration (FAA) implement mandatory masks for*

> *travelers….[T]hey tell us masks are absolutely necessary to save lives, and pass all these stifling laws forcing perfectly healthy people into masks, and relentlessly shame anyone who pushes back or even questions the mask dogma….Why are the very people telling us that masks save lives, not wearing masks? If you were 80 or 87-years-old and truly believed a mask would save you from catching the coronavirus, which is especially deadly to the elderly, would you not wear a mask in, of all places, an airport?"*

While clearly an opinion piece, there is a reasonable possibility that a person reading this would believe that masks are not actually mandatory. At the very least, it would serve to politicize and polarize mask wearing, suggesting to conservative viewers some form of "liberal hypocrisy" in regard to these public health measures. A more hard-news example comes from the New York Post in June 2020:

> "*An Arizona town will push forward with its summer festivities but leave the coronavirus mask requirement behind. Mayor Bryce Hamblin of Eagar, an east Arizona city with a population around 5,000, said in a recent statement that he has no plans to cancel a roster of events for the summer season…[the] response from the onset of COVID-19 pandemic has been that we will err on the side of freedom….*"

While the article later goes on to give information about the case counts within Arizona, there is little to no mention about the national guidance on mask wearing. This is all in comparison to left-leaning and center news sources that place emphasis on the scientific aspects of mask-wearing, social-distancing etc as well as the long-term large-scale effects of the coronavirus. For example, the Washington Post wrote in May 2020:

> "*University Town Center is emblematic of retail outlets, restaurants, hair salons and other businesses around the country that have started to reopen, even though a deadly*

> *virus with no cure or vaccine is circulating widely. More than half of all states have eased some restrictions on businesses or movements put in place because of the coronavirus over the past two weeks...Nationwide, coronavirus has killed more than 75,000 Americans. More than 1.2 million people have tested positive for the coronavirus. Epidemiologists and disease modelers warn that social distancing is the only real weapon against the coronavirus...*"

These findings largely make sense, since these words like masks and social distance have taken on new meaning in the context of the coronavirus pandemic and have largely been polarized. This would also help explain why partisan news exposure has such a noticeable effect on an individual's likelihood to follow public health measures placed by various levels of government as noted by various prior studies (Burstzyn et al. 2020, Jamieson and Albaraccin 2020, Simonov et al. 2020). These results also seem to corroborate in part the results found by Motta, Stecula, and Farhart (2020), although not in the explicit conspiratorial language but rather in the more subtle way of associating the public health measures with non-science. It is important to note that this signal persisted even after the initial months of the pandemic.

Unfortunately, other words picked out through the close reading, particularly those that relate to macro-political elements such as "unemployment", "relief", "PPE" did not show as much semantic shifts between sources. This could be due sample size issues as word embeddings are data-hungry algorithms[1]. An alternative explanation would be that these words are not polarized to the extent that word embedding vectors would be able to detect them. If that is the case, this brings new questions of why the polarization of some public health measures and macro-political actions were able to be detected via word embedding analysis while others were

[1] Some state-of-the-art embedding models such as word2vec and GLoVe are trained on billions of words (Bahdanau et al. 2017)

not. Ultimately, even though this study leveraged big-data methodologies to draw out latent features of the corpus, simple measures such as the counts and content analytic approaches are still informative of the characteristics of these types of corpora (Wicke and Bolognesi 2020).

Moving forward, it is important to continue expanding inquiry into the digital news landscape. Like the prominent Twitter-based analyses, this study was able to show the discussion of low-credibility information, this time by large-scale media conglomerates rather than individuals or networks of individuals. This was able to be detected using algorithms similar to ones implemented by Guo, Xypolopoulos and Vazirgiannis (2021) on Twitter despite digital news not having the same level of virality that Twitter has. Importantly, this low-credibility information is not the same as the explicit conspiracy theories found by Ahmed et al. (2020), which were not picked up by either the close reading or the corpus-level data analysis. These more subtle forms of misinformation are comparatively harder to notice and push back against even though it can have the same macro-political implications of making people skeptical of public health measures that are instrumental to resolving the pandemic.

## Conclusion

This study has constructed and processed a unique, novel corpus of COVID-19 digital news articles and utilized that corpus for preliminary analyses on how different news sources have been discussing COVID-19, particularly with respect to the sources' political lean. Spanning 12 months and covering multiple news sources across the political spectrum, this corpus is likely the most comprehensive dataset of COVID-19-related digital news articles so far, and it still has room to grow. Creating a corpus like this is uniquely important in the context of the COVID-19 pandemic, especially since analysis of this corpus can reveal how information about public health emergencies is being disseminated into the general populace. With most biomedical sources indicating that the next pandemic is not a question of “if” but “when”, it is paramount that analysis continues on the COVID-19 corpus to determine the differences in news coverage and respond accordingly (Brownson et al. 2020).

In terms of analysis, topic modeling and regression analysis on the topic proportions were used to examine the content of this corpus, with some indication that right-leaning news covered state and local responses to COVID more than left-leaning news in their text articles. More contextual analysis was done using word2vec word embedding models, which showed that public health measures such as social distancing and mask wearing were often portrayed with less gravitas by right-leaning media compared to left-leaning and center sources. These signals, despite being faint, show that political lean has a fairly subtle, yet impactful correlation with how news sources talk about COVID-19 in their text-based articles, particularly through the contextual analysis lens. It is likely that these trends can and likely will re-emerge in future crisis situations considering the persistence and endurance of these effects during the current coronavirus pandemic.

As for next steps, the clearest one would be to include the full corpus of articles from the data collected stage. These text analysis algorithms become much more powerful as the sample size becomes larger. In a similar vein, the text classifier could use more labels to achieve its original goal of classifying micro-political versus macro-political actions throughout the COVID crisis. Another iteration of this study would likely also utilize word collocation and ngrams analysis. These would get at the contextual questions more by examining words that are in close proximity to each other. The ngrams would also be a useful modification for the word embedding models to include groupings of words in the vector representations. Finally, perhaps a way to get at the directionality issue of the topic model would be to use some version of a sentiment score or other measure to split topics.

## References

Ahmed W, López Seguí F, Vidal-Alaball J, Katz MS. “COVID-19 and the ‘Film Your Hospital’ Conspiracy Theory: Social Network Analysis of Twitter Data” *J Med Internet Res* 22 no. 10 (2020). https://doi.org/10.2196/22374.

Ali, Safia Samee. “Cruise ship passengers desperately plead with Florida to allow them in”. *NBC News*, April 1, 2020. https://www.nbcnews.com/news/us-news/cruise-ship-passengers-desperately-plead-florida-allow-them-n1173576

Azarbonyad, Hosein, Mostafa Dehghani, Kaspar Beelen, Alexandra Arkut, Maarten Marx, and Jaap Kamps. “Words are malleable: Computing semantic shifts in political and media discourse.” *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (November 2017): 1509–1518. https://doi.org/10.1145/3132847.3132878

Bahdanau, Dzmitry, Tom Bosc, Stanisław Jastrzębski, Edward Grefenstette, Pascal Vincent, Yoshua Bengio. “Learning to Compute Word Embeddings On the Fly” (June 2017). *arXiv*:1706.00286

Banda, Juan M., Ramya Tekumalla, Guanyu Wang, Jingyuan Yu, Tuo Liu, Yuning Ding, Katya Artemova, Elena Tutubalina, Gerardo Chowell. “A large-scale COVID-19 Twitter chatter dataset for open scientific research -- an international collaboration”. (November 2020). *arXiv*:2004.03688.

Brownson, Ross C., Thomas A. Burke, Graham A. Colditz, and Jonathan M. Samet. “Reimagining Public Health in the Aftermath of a Pandemic” *American Journal of Public Health* 110 (October 2020): 1605-1610. https://doi.org/10.2105/AJPH.2020.305861

Bursztyn Leonardo, Aakaash Rao, Christopher P. Roth and David H. Yanagizawa-Drott.

"Misinformation during a Pandemic". *National Bureau of Economic Research*, Working Paper 27417. (Sept 2020). https://www.nber.org/papers/w27417

Carter, Shawn M. "Arizona mayor says he won't require masks for town's summer events". *New York Post*, June 30, 2020. https://nypost.com/2020/06/30/arizona-mayor-says-he-wont-require-masks-for-towns-summer-events/

Chen, Juntian, Yubo Tao and Hai Lin. "Visual exploration and comparison of word embeddings" *Journal of Visual Languages & Computing.* 48 (August 2018): 178-186. https://doi.org/10.1016/j.jvlc.2018.08.008

"Documentation". (2018). https://newsapi.org/docs.

Elkind, Elizabeth. "Doctor says UNC Chapel Hill "should not have been opened" after school backtracks on in-person classes". *CBS News*, August 18, 2020. https://www.cbsnews.com/news/covid-19-unc-chapel-hill-in-person-classes-doctor-dyan-hes/

Ferrara, Emilio. "What types of COVID-19 conspiracies are populated by Twitter bots?" *First Monday*, 25 no. 6 (June 2020). https://doi.org/10.5210/fm.v25i6.10633

Furr, Amy. "Man Whose Dad Died of Coronavirus Joins 15K in Lawsuit Against China". *Breitbart News*, May 16, 2020. https://www.breitbart.com/health/2020/05/16/man-whose-dad-died-coronavirus-joins-15k-lawsuit-china/

Guo, Yanzhu, Christos Xypolopoulos and Michalis Vazirgiannis. "How COVID-19 Is Changing Our Language : Detecting Semantic Shift in Twitter Word Embeddings". (February 2020). *arXiv*:2102.07836.

Hamilton, William L., Jure Leskovec and Dan Jurafsky. "Cultural Shift or Linguistic

Drift? Comparing Two Computational Measures of Semantic Change”. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, (September 2016): 2116–2121. *arXiv*:1606.02821

Hart, P. Sol, Sedona Chinn , and Stuart Soroka. “Politicization and Polarization in COVID-19 News Coverage.” *Science Communication,* 42 No. 5 (August 2020): 679-697. https://doi.org/10.1177/1075547020950735

Hoffman, Matthew, Francis R. Bach, and David M. Blei. "Online learning for latent dirichlet allocation." *In neural information processing systems*, (2010)

Ingraham, Christopher. “New research explores how conservative media misinformation may have intensified the severity of the pandemic” *The Washington Post,* June 25, 2020.

Jamieson, Kathleen and Dolores Albarracin. “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*, Vol. 1. (April 2020). https://doi.org/10.37016/mr-2020-012

Johns Hopkins University. “COVID-19 Dashboard.” Accessed February 28, 2021. https://coronavirus.jhu.edu/map.html

Kim, Yejin. “On media coverage of the COVID-19 outbreak: A corpus-based collocation study.” *SNU Working Papers in English Linguistics and Language* 17, (August 2020): 47-73. http://hdl.handle.net/10371/168776

Lee, Crystal, Tanya Yang, Gabrielle Inchoco, Graham M. Jones, Arvind Satyanarayan. “Viral Visualizations: How Coronavirus Skeptics Use Orthodox Data Practices to Promote Unorthodox Science Online.” (January 2021). *arXiv*:2101.07993.

Lopez, Christian, Malolan Vasu and Caleb Gallemore. “Understanding the perception of

COVID-19 policies by mining a multilanguage Twitter dataset”. (March 2020). *arXiv*:2003.10359

Martin, Gregory and Joshua McCrain. “Local News and National Politics.” *American Political Science Review* 113, no. 2 (February 2019): 372-384. https://doi:10.1017/S0003055418000965

Martin, Gregory and Ali Yurukoglu. “Bias in Cable News: Persuasion and Polarization.” *American Economic Review* 107, no.9 (September 2017): 2565–2599. https://doi:10.1257/aer.20160812

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. “Distributed Representations of Words and Phrases and their Compositionality” (October 2013). *arXiv:*1310.4546

Motta, Matthew, Dominik Stecula, and Christina E. Farhart. “How Right-Leaning Media Coverage of COVID-19 Facilitated the Spread of Misinformation in the Early Stages of the Pandemic.” *Canadian Journal of Political Science* (May 2020): 335–342. https://doi:10.1017/S0008423920000396

Nolte, John. “Nolte: Dianne Feinstein Busted Not Wearing Mask After Demanding Mask Mandate”. *Breitbart News*, September 29, 2020. https://www.breitbart.com/politics/2020/09/29/nolte-dianne-feinstein-busted-not-wearing-mask-after-demanding-mask-mandate/

Otero, Vanessa. “Ad Fontes Media’s First Multi-Analyst Content Analysis Ratings Project.” *Ad Fontes Media White paper,* (August 2019).

Simonov, Andrey, Syzmon K. Sacher, Jean-Pierre H. Dubé, and Shirsho Biswas. “The

Persuasive Effect of Fox News: Non-Compliance with Social Distancing During the Covid-19 Pandemic". *National Bureau of Economic Research*, Working Paper 27237, (July 2020). https://doi:10.3386/w27237

Stekhoven, Daniel J, and Peter Bühlmann. "MissForest--non-parametric missing value imputation for mixed-type data." *Bioinformatics* 28, no. 1 (January 2012): 112-118. https://doi:10.1093/bioinformatics/btr597

Yao, Limin, David Mimno, and Andrew McCallum. "Efficient methods for topic model inference on streaming document collections." *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. (January 2009). https://doi:10.1145/1557019.1557121

Wicke, Philipp, and Marianna M Bolognesi. "Framing COVID-19: How we conceptualize and discuss the pandemic on Twitter." *PloS one* vol. 15, no. 9 (September 2020). https://doi.org/10.1371/journal.pone.0240010

Wootson, Cleve. "A mall in Florida has reopened — steps away from a coronavirus testing facility". *The Washington Post,* May 8, 2020. https://www.washingtonpost.com/politics/a-mall-in-florida-has-reopened--steps-away-from-a-coronavirus-testing-facility/2020/05/07/595483e2-8fca-11ea-a9c0-73b93422d691_story.html

Yang, Kai-Cheng, Christopher Torres-Lugo, and Filippo Menczer. "Prevalence of Low-Credibility Information on Twitter During the COVID-19 Outbreak". (June 2020). *arXiv*:2004.14484.