# Clustering Analysis of Website Usage on Twitter during the COVID-19 Pandemic

Iain J. Cruickshank[1 [0000−0002−4205−5806]] and Kathleen M. Carley[1 [0000−0002−6356−0238]]

Center for Computational Analysis of Social and Organizational Systems, Carnegie Mellon University, Pittsburgh PA, 15213 icruicks@andrew.cmu.edu, kathleen.carley@cs.cmu.edu

**Abstract.** In this study we analyzed patterns of external website usage on Twitter during the COVID-19 pandemic. We used a multi-view clustering technique, which is able to incorporate multiple views of the data, to cluster the websites' URLs based on their usage patterns and tweet text that occurs with the URLs. The results of the multi-view clustering of URLs used during the COVID-19 pandemic, from 29 January to 22 June 2020, revealed three, main clusters of URL usage. These three clusters differed significantly in terms of using information from different politically-biased, fake news, and conspiracy theory websites. Our results suggest that there are political biases in how information, to include misinformation, about the COVID-19 pandemic is used on Twitter.

**Keywords:** Clustering · COVID-19 · Multi-view Data

## 1 Introduction

The COVID-19 pandemic, which is caused by the SARS-CoV-2 virus, has caused immense societal and economic disruption across the world. The disruption caused by the pandemic has spread to nearly every facet of human social behavior to include how humans are interacting with information through mediums like social media [18], [1]. Thus far, much of the work with COVID-19 social media data has focused on the prevalence and spread of COVID-19 misinformation. There has been less work on understanding holistically what information social media users are interacting with and if there are any patterns in these interactions. In particular, a key source of information for social media users are external websites which often publish material that is shared through social media. This sharing of information from external websites, through the use of Uniform Resource Locators, or URLs, is often an important behavior not only for propagating useful information that individuals can use to help combat the pandemic, but also propagating conspiracy theories or fake news which can harm individuals and exacerbate the effects of the pandemic. Thus, it is important to understand if there are distinct patterns of usage of URLs on social media sites

to better understand what information is being propagated and give insight into what communities are using which information sources.

One means of discerning patterns from digital data is clustering. In this work, we propose the use of multi-view clustering to discern different patterns of URL usage on Twitter. Many naturally-occurring, social phenomena give rise to multiple types and views of data [9]. So, using a clustering method that can exploit information from all of those views results in a better clustering of the phenomena that gave rise to all of the data. To date and the best of the authors' knowledge, most clustering of social media data is not clustered by multi-view clustering. So, in this work, we use of multi-view clustering in order to better understand the usage patterns of URLs on social media during the COVID-19 pandemic. The main contributions of this work are summarized as follows:

- The first use of a multi-view clustering technique and approach to understand clusters of website usage on social media.
- Characterization of patterns of usage of all websites on Twitter, not just those related to misinformation, during the COVID-19 pandemic. The results show that politically-biased websites and those associated with fake news and conspiracy theories have different patterns of usage then those related to science or other content.

## 2 Related Research

Current studies into online social behavior during the COVID-19 pandemic have largely focused on how misinformation spreads during a pandemic. This is because good information is a key enabler to combat the effects of the pandemic whereas misinformation can exacerbate its effects [18], [14]. Recent studies into the prevalence and persistence of misinformation have shown that misinformation on the COVID-19 pandemic has been especially persistent and spreads through online social networks quickly [30], [1], [6]. The spread of COVID-19 misinformation has become so problematic and widespread that many many researchers are referring to it as an 'Infodemic' [14], [8], [1]. The Infodemic is characterized by a virus-like spread of misinformation across many different communication mediums, most notably online social networks. Additionally, other researchers have identified important mechanisms by which the misinformation propagates in social media. Recent research has identified the importance of bots in the spread of misinformation [11]. Other research has highlighted the role of alternative news sources and user characteristics like political beliefs in the spread of COVID-19 misinformation [6], [16]. Finally, some recent research has found that low-credibility information about the COVID-19 pandemic tends to be frequently used and have a high persistence on Twitter [30].

One of the common artifacts used for the determination of information veracity on social media sites are the URLs cited for that information. Recent research has shown social-media users use external websites for information relating to the COVID-19 pandemic [30], [6]. In fact, it is the use of these external websites which can allow for researchers to assess the spread of things like misinformation

during the pandemic [14], [8]. Despite the utility of external websites in assessing things like misinformation spread, there has not been a holistic look at the usage of websites on social media during the COVID-19 pandemic. It is unclear if there are different usage patterns or clusters of websites that exist during a pandemic, beyond those explicitly related to misinformation.

An increasingly used means of finding patterns or clusters within data is multi-view clustering. Multi-view clustering techniques are techniques designed to handle clustering of objects which can be described by more than one data source. Many different real-world, social phenomena give rise to 'views' of data which are often different types of data that can be used to describe the same set of actors. For example, social media users can post content, which could give rise to a text view, and have interactions with each other, which can give rise to network views. So, multi-view clustering aims to fuse the information from these different views of the data to produce one clustering of the objects that created the data [4], [32], [31], [3]. There has been a surge of new techniques developed in multi-view clustering for handling genetic data [34], [17], image data [31], [2], and more recently human, social-based data [9]. In particular, recent research with hashtags on Twitter during the COVID-19 pandemic has found multi-view clustering to be an effective means of characterizing topical discussion groups [10]. So, multi-view clustering can be used as a means of finding richer clusters from real-world data, than just clustering any particular view of then data by itself.

## 3 Methodology

Since there are different data modalities, or views, of how URLs are used on Twitter (i.e. how the URL is spread in Tweets, how the URL is described by the verbiage of the Tweets, etc.), we have adopted a multi-view clustering framework for clustering URLs in the COVID-19 Twitter data. By using a multi-view clustering, we can use all of the views of URL usage that previous research has identified as being important to their usage in one cohesive clustering. So, in this section we detail the methods used to obtain the multi-view clusters of the URLs. The first subsection describes the multi-view clustering technique used to produce the clusters. The second subsection describes the data statistics and processing done to obtain the views of the data used in the clustering.

### 3.1 Multi-view Clustering of URLs

In order to cluster the URLs, we adopted the multi-view clustering technique of Multi-view Modularity Clustering (MVMC). MVMC is a technique designed to work with multiple views, of any data type, of the same underlying social-based phenomena to produce one set of clusters [9], [10]. The technique works in two main steps. First, a graph is learned for every view of the data, then an iterative

procedure clusters all of the view graphs by optimizing the following modularity function:

$$\sum_{v=1}^{m} w^v \sum_{ij \in E^v} [A_{ij}^v - \gamma^v \frac{deg(i)^v \times deg(j)^v}{2 \sum A^v}] \delta(C_i, C_j) \quad (1)$$

where $v$ is a particular view, $m$ is the total number of views, $A^v$ is the adjacency matrix for the graph of the $v$th view, $deg(i)^v$ is the degree of the object $i$ in view, $v$, and $\delta(.,.)$ is the delta function which returns one if the two items are the same and 0 otherwise. The parameters that are used in the optimization are $w^v$ which is the weight assigned to the $v$th view, which controls how much impact upon the clustering solution the $v$th view should have, $\gamma^v$ which is the resolution parameter for the $v$th view, which controls for the resolution limit inherent in the modularity function [20], [13], [27], and $C_i$ which is the cluster assignment for object $i$. In order to optimize the cluster assignments as well as the weights and resolutions for each view, an iterative procedure is used. In the first step, the view graphs are clustered using a modularity optimization technique (i.e. Louvain [5] or Leiden [28]) with the current view weights and resolutions to produce provisional cluster assignments. Then, in the second step, the view weights and resolutions are updated by the provisional cluster assignments. This procedure is repeated until the view weights and resolutions no longer change. The pseudocode for the MVMC procedure is displayed in Algorithm 1 [1].

[1] A Python implementation of this algorithm is available on the lead author's GitHub page: https://github.com/ijcruic/Multi-view-Clustering-of-Social-Based-Data

**Algorithm 1** Multi-view Modularity Clustering (MVMC)

**input**:
- Adjacency for each view: $A^v$
- Max number of iterations: $max\_iter = 20$
- Starting resolutions: $\gamma_1^v = 1, \forall v \in m$
- Starting weights: $w_1^v = 1, \forall v \in m$
- Convergence tolerance: $tol = 0.01$

**output**: Cluster assignments
$clustering^* \leftarrow None$
$modularity^* \leftarrow -\infty$
**for** $i = 1 : max\_iter$ **do**
  $clustering_i \leftarrow cluster(A, w_i, \gamma_i)$
  $modularity_i \leftarrow RBmodularity(A, clustering_i, w_i, \gamma_i)$
  $\theta_{in}, \theta_{out} \leftarrow calculate\_thetas(A, clustering_i)$
  $\gamma_{i+1}^v \leftarrow \frac{\theta_{in}^v - \theta_{out}^v}{\log\theta_{in}^v - \log\theta_{out}^v}, \forall v \in m$
  $w_{i+1}^v \leftarrow \frac{\log\theta_{in}^v - \log\theta_{out}^v}{<\log\theta_{in}^v - \log\theta_{out}^v>_v}, \forall v \in m$
  **if** $abs(\gamma_{i+1} - \gamma_i) < tol$ AND $abs(weights_{i+1} - weights_i) < tol$ **then**
    $clustering^* \leftarrow clustering_i$
    $modularity^* \leftarrow modularity_i$
    BREAK
  **end if**
  **if** $iter >= max\_iter$ **then**
    $best\_iteration \leftarrow argmax(modularity)$
    $clustering^* \leftarrow clustering[best\_iteration]$
    $modularity^* \leftarrow modularity[best\_iteration]$
  **end if**
**end for**
**return** $clustering^*$

The algorithm begins by initializing all the resolution parameters, $\gamma_1^v$, and weight parameters, $w_1^v$ to one (or whatever the user may specify). The algorithm then goes on to cluster the view graphs, $A^v$, by a modularity maximization technique (i.e. Louvain, Leiden), $cluster()$, with the current resolution and weight settings. The output of this is then used to determine the propensities for internal edge formation $\theta_{in}^v$, and external edge formation, $\theta_{out}^v$ for each view. These values are then used to update the resolution, $\gamma^v$, and weight parameters, $w^v$, for each of the views. If the new weight and resolution parameters are the same as the previous ones (within tolerance), the algorithm then exits and returns the final clustering. If the algorithm fails to converge to stable resolution and weight parameters, within the maximum number of iterations allowed, then the algorithm returns whichever clustering produced the highest modularity.

One of the important elements in the aforementioned algorithm, Algorithm 1, is the computation of the edge propensities, $\theta$. In order to calculate these edge propensities, we follow the guidance outlined in previous works and assume edges form by a degree-corrected model [24], [25]. The following pseudocode, Algorithm 2, details the procedure for calculating these edge propensities:

**Algorithm 2** Calculation of Edge Propensities

**input**:
- Adjacency for each view: $A^v$
- clustering: $C$

**output**: Internal and external edge propensities $(\theta_{in}, \theta_out)$
**for** v = 1:m **do**
  $e_{in} = 0$
  $\kappa^2 = []$
  **for** c = 1:—C— **do**
    $e_c = \sum E_c^v$
    $e_{in} += e_c$
    $\kappa^2.append((\sum_{i \in V_c^v} deg(i))^2)$
  **end for**
  **if** $e_{in} = 0$ **then**
    $\theta_{in}^v \leftarrow \frac{1}{|E^v|}$
  **else**
    $\theta_{in}^v \leftarrow \frac{e_{in}}{\sum \frac{\kappa^2}{4 \sum E^v}}$
  **end if**
  **if** $e_{in} == \sum E^v$ **then**
    $\theta_{out}^v \leftarrow \frac{1}{|E^v|}$
  **else**
    $\theta_{out}^v \leftarrow \frac{\sum E^v - e_{in}}{\sum E^v - \sum \frac{\kappa^2}{4 \sum E^v}}$
  **end if**
**end for**
**return** $\theta_{in}, \theta_{out}$

The algorithm goes through each graph to calculate the propensities for each graph separately. For each graph, the algorithm begins by calculating the number of internal edges and the degree-corrected, null-model terms (i.e. $\kappa^2$) for each of the clusters [24]. Then, the algorithm checks as to whether the graph is directed or undirected and whether there are no internal or external edges and then calculates the final propensities for that view graph: $\theta_{in}^v$ and $\theta_{out}^v$. Once the propensities have been calculated for all of the view graphs, these are then returned.

### 3.2 Data Processing and View Graph Creation

The data for this analysis comes from Twitter's streaming API [2]. The data was collected using a list of keywords including "coronavirus", "wuhan virus", "wuhanvirus", "2019nCoV", "NCoV", "NCoV2019" [16]. The collected data spans the time period from 29 January 2020 to 22 June 2020 and consists of just over 500 million tweets that have, on average, 120,000 unique URLs per day. The data was first processed by grouping all tweets into daily collections, and then filtering for only those tweets with an English-language tag. Then, each of the URLs was processed in order to remove any query terms, so that all that is left is the base URL itself. Finally, for the clustering, only the top 50,000 most tweeted URLs across the entire time period were used.

The following figure, Figure 1 depicts the daily statistics concerning the use of URLs within the data set.

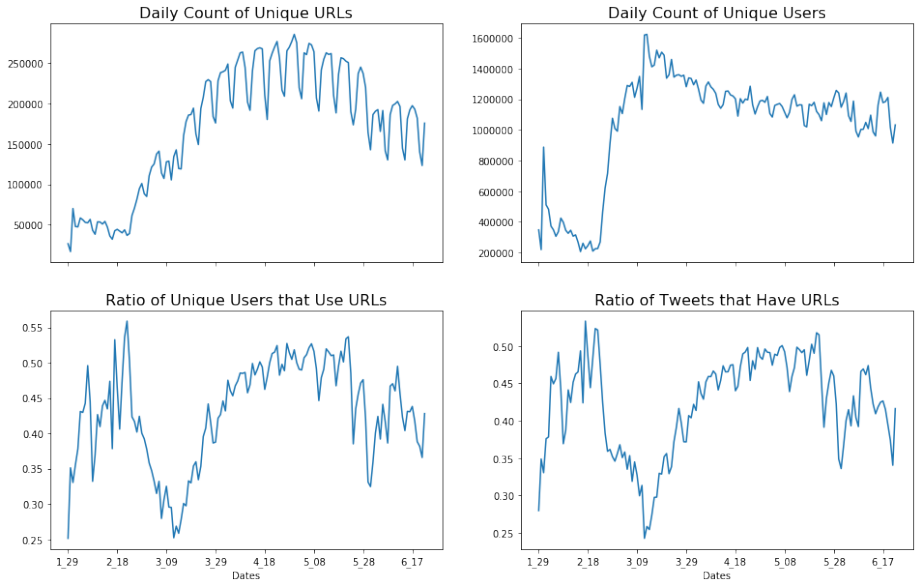


**Fig. 1.** Daily Statistics of the COVID-19 Twitter Data from 29 January 2020 to 22 June 2020. Use of URLs remains high both within tweets and by users, but does see a precipitous low-period of usage in the beginning of March, when many of the shut-downs were going into effect.

There are some distinct temporal regions in URL usage in the data. In the first place, both the number of daily unique URLs along with the daily count

[2] https://developer.Twitter.com/en/docs/tweets/filter-realtime/guides/basic-stream-parameters

of unique users starts small in the end of January and then surges around the end of February and beginning of March with the onset of lock-downs in many countries [19]. From there, the number of unique users and URLs slowly declines. In terms of the usage rates for URLs, they are both generally high for users and within tweets, with 47% of users using at least one URL on any given day and 42% of tweets containing a one URL. The usage ratios do, however, vary over the time-span of the data with there being low periods in URL usage at the end of January and again at the beginning of March. So, there are distinct temporal shifts in URL usage which generally mirrors patterns surrounding Twitter usage during the pandemic more broadly [7], [9].

In order to multi-view cluster the URLs, three views were created to describe the URLs. The first view was the text view, which consist of all of the text that co-occurs with a URL in all of the tweets that mention a URL. Tweet text has been commonly used to cluster social media devices, like hashtags [29], [12]. The next two views derive from the users that tweet the URLs. Since it has been noted in other works that retweeting behavior can differ from other tweeting behavior on Twitter [23], we broke the users into two views: those who retweet the URL versus those that tweet the URL.

Now, the first step of the MVMC procedure is to create graphs of each of the views. For each view, a similarity graph was created. So, for each view graph, an edge represents how similar two objects are with respect to that view. For the users view, the similarity graphs were created by multiplying the URLs-by-users matrices by their transpose (i.e. $A = XX^T$) to produce a URLs-to-URLs shared users graph. To produce the text view graph, the text was broken down into terms by a bag-of-words model and Term Frequency-Inverse Document Frequency (tf-idf) was applied to the URL-by-term matrix. We then used a symmetric k-Nearest Neighbor Graph (k-NN) with the number of nearest neighbors as $k = \sqrt{n}$, where $n$ is the number of objects being clustered, and cosine similarity was used to measure the similarity between the different URLs [22], [21]. To symmetrize the k-NN, the average strategy, $A' = \frac{1}{2}A + A^T$, which is common in spectral clustering methods [26], [33], was used to produce the final view graph. With that, we then had three different views of URL usage that were transformed to view graphs for clustering by MVMC.

## 4 Results

In this section, we describe the results of the multi-view clustering. In the first subsection we describe the cluster statistics. In the second subsection we compare the clusters to some additional labels from the domains that the URLs originate from. And, finally, in the third section, we analyze the content present within the most important of the clusters.

For clustering the COVID-19 URLs data, the following parameters for MVMC were used based on the parameter settings used in previous analyses using COVID-19 Twitter data [10]: The initial weights and resolutions were all set to one. The convergence tolerance for the resolutions was set to 0.01 and for

the weights to 0.01, and the procedure was allowed to run for a maximum of 20 iterations.

### 4.1 Multi-view Clustering Statistics

We begin the analysis of the multi-view clustering by looking at the clustering statistics. The procedure produced 71 clusters with an average membership of $704.7 \pm 3,361.8$ URLs. In terms of the cluster sizes, there are three main clusters of URLs, and several smaller clusters. 98.8% of the URLs fall within the first three clusters which are of sizes 18,987, 15,859, and 14,636 URLs respectively. Of the smaller clusters, they had URLs that were generally from common news sources such as the *New York Times* or *Washington Post*. The formation of these small clusters was generally a result of a lack of overlap in user bases in both the non-retweet and retweet user views, which is likely a result of the data being collected from the streaming API, meaning it is only a sample of the data available. So, the clustering produced a collection of small, outlier clusters, and three main URL usage clusters.

From the MVMC algorithm we also obtained some insight into the nature of the clusters. First, the algorithm converged in three iterations. Typically a faster convergence (i.e. less than 20 iterations) is linked to a stronger cluster structure being present in the data [9]. Also, the learned weights of the different views, $w^v$, were 1.036, 1.27, and 0.698 for the retweet users, non-retweet users, and tweet text views respectively. These weights indicate that the non-retweet users which tweet a URL were the most important view to the cluster structure followed by those users that retweet a tweet containing a URL. So, from the performance of the MVMC algorithm, there is a string cluster structure present in the data and the users who tweet a given URL form the most important view of that cluster structure.

### 4.2 Comparison of Clusters to Domain Labels

In order to understand the clusters produced by the URLs' different usage, we analyzed the domains of the URLs present in each of the three clusters. We compiled three different sets of labels for some of the commonly used domains of the URLs in the data. the first set of domain labels, which were compiled from various previously published articles on fake news and conspiracy theory domains, are color-based labels that relate to a domain's propensity to produce fake news articles [16], [15], [30]. For example, The 'black' domains contain websites which published exclusively fabricated stories while the 'red' list is a set of websites spreading falsehoods with a flawed editorial process, and the 'green' list are websites that follow full editorial processes in their news publications and are not known to produce fabricated content. We also compiled a set of domain labels based on biases of the domain and factual ratings of the domain which were compiled using several fact checking and bias checking websites [3].

[3] These bias and fact checking websites are: https://mediabiasfactcheck.com/, http://www.fakenewscodex.com/, and https://www.snopes.com/. For transparency,

The following figure, Figure 2, displays the breakdown of the various domain labels across the three main clusters of URLs.

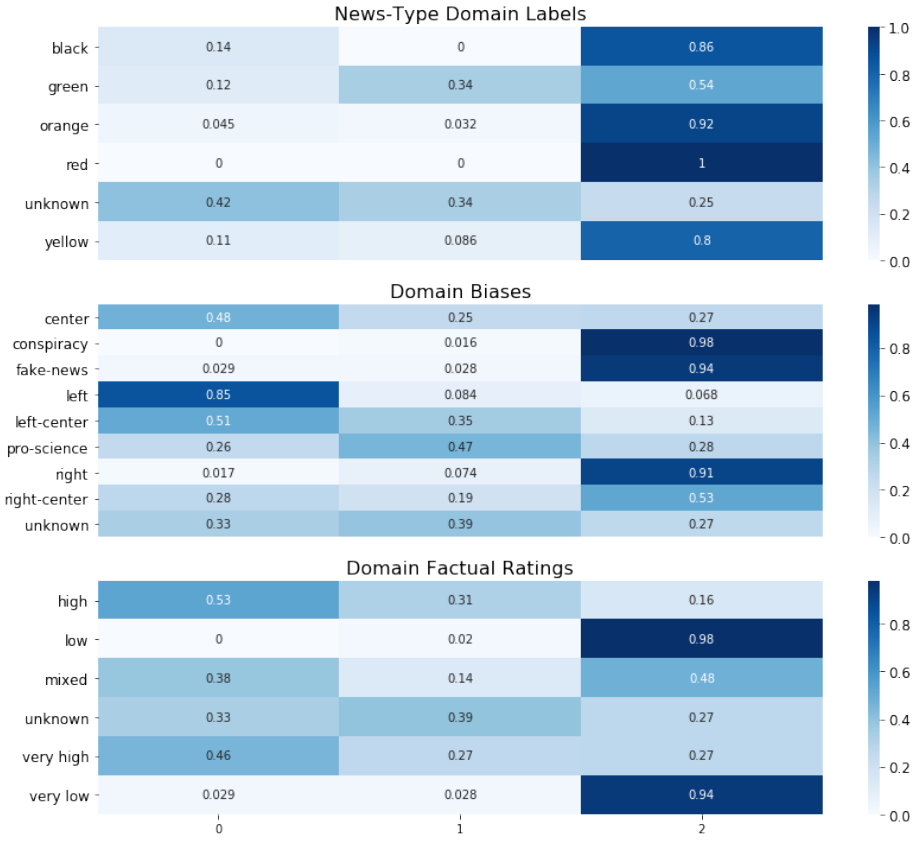


**Fig. 2.** Confusion matrices of three primary clusters of URLs and domain labels. The third cluster contains most of the URLs that come from domain labels that have been identified with producing fake news (i.e. 'black', 'red') as well as those domains with a US, politically right bias.

The domain labels are not evenly distributed across the three main clusters of URLs. The first cluster contains the predominance of US, politically-left leaning domains as well as those that are evaluated to produce information that is rated as being high or very high in factual content. The second cluster tends to have most of the domains associated with a pro-science bias in their information.

these labels along with the associated URLs are available in a public repository: https://figshare.com/articles/conference_contribution/Clustering_Analysis_of_Website_Usage_on_Twitter_during_the_COVID-19_Pandemic/13079657

The third cluster, in contrast to the other two, contains most of the domains associated with various types of news websites, especially those that produce fabricated content, websites that promulgate conspiracy theories, as well as those websites that tend to have a US politically right or right-center bias. So, the clusters of usage have different political and factual biases of the URLs present within the clusters. In particular, websites with high factual ratings and left-leaning biases tend to have different usage patterns than those websites with low-factual ratings and a right-leaning bias. It is also interesting to observe that the various color-based labels for news sites tend to congregate in the same cluster indicating that usage patterns of fake news can be similar to usage patterns of more legitimate news on Twitter.

### 4.3 Content of Clusters of Interest

In order to get a better sense of the differences between the three main clusters found by multi-view clustering we then analyzed the clusters' content. In particular, we looked at the commonly occurring verbiage that occurs with tweets that contain the various URLs in each cluster and any hashtags that tend to co-occur with the URLs. In order to analyze the verbiage that occurs in tweets with the URLs, we first removed any common English stop words along with commonly used terms for this data like 'COVID-19', emojis, URLs, hashtags, and Twitter specific text (i.e. 'RT' for retweet) from all of the tweets. The text was then combined across all URLs for each cluster. The following figure, Figure 3, displays a word map of the commonly used verbiage with URLs from the first cluster.

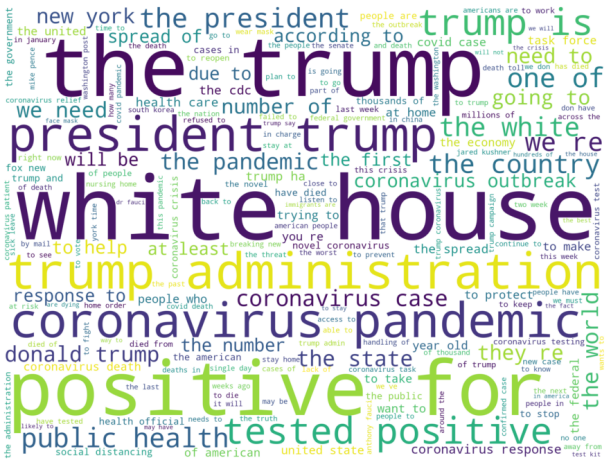


**Fig. 3.** Commonly used terms and phrases from the first cluster of URLs (label 0).

Much of the verbiage co-occurring in tweets with URLs from the first cluster centers around the US White House and President along with positive tests for the Coronavirus itself. Some of the hashtags which commonly occur with these URLs and their counts are: '#Trump':62,102, '#TrumpVirus':35,109, and '#MOG':29,061. As was noted in the previous analysis section, this cluster tends to have the most domains related to a left-leaning bias. This bias is again seen somewhat in the hashtags that co-occur with URLs in this cluster. And, from the verbiage that occur with this cluster, much of the URL usage centers on President Trump and his administration. So, the first cluster of URLs seems to be URLs used as means of criticizing President Trump and his administration's response to the COVID-19 pandemic.

The next figure, Figure 4, displays a word map of the commonly occurring verbiage with the URLs from the second cluster.

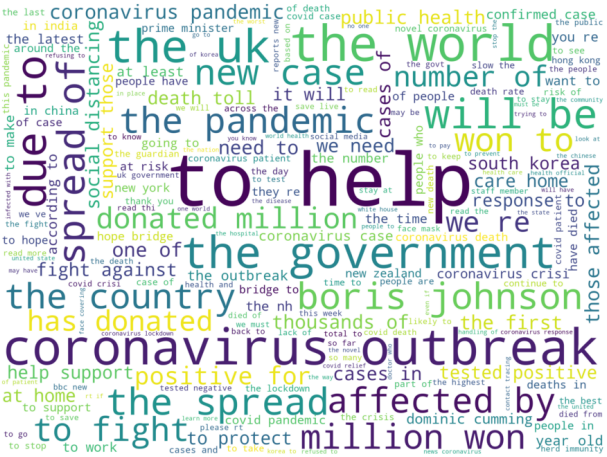


**Fig. 4.** Commonly used terms and phrases from the second cluster of URLs (label 1).

The verbiage co-occurring with the second cluster's URLs does not have as prominent of a dominant theme as the other clusters. There is verbiage related to the UK government as well as efforts related to fighting the spread of the COVID-19 pandemic, such as social-distancing. Some of the hashtags which commonly occur with these URLs and their counts are: '#IndiaFightsCorona':127,747, '#TogetherAtHome':82,962, '#lockdown':43,151, and '#StayHome':42,899. From the previous section, this particular cluster was not particularly high in any of the domain labels except for having the most URLs from pro-science biased domains. So, it would seem based on the verbiage and commonly occurring hashtags within this cluster that the URLs usage is generally meant to inform about actions taken to fight the COVID-19 pandemic.

Finally, the next figure, Figure 5, displays a word map of the commonly occurring verbiage with the URLs from the third cluster.

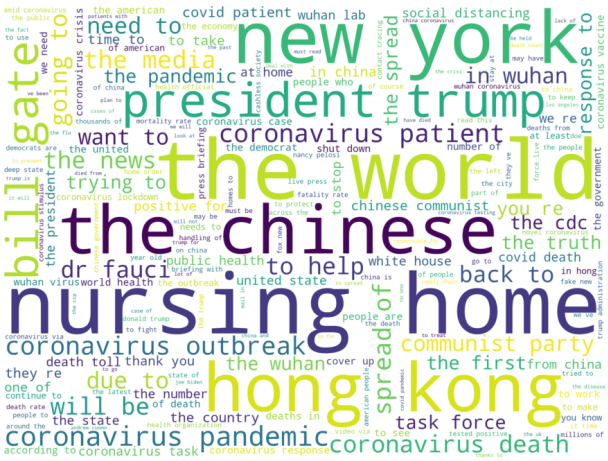


**Fig. 5.** Commonly used terms and phrases from the third cluster of URLs (label 2).

The verbiage from the third cluster centers around a couple of different topics. The prominent topics include verbiage around restrictions or impacts to nursing homes by the Coronavirus, criticisms of certain locations and their governments (i.e. China, Hong Kong, New York), and discussion around certain prominent individuals like Bill Gates, Dr. Anthony Fauci, or President Donald Trump. Some of the hashtags which commonly occur with these URLs and their counts are: '#FoxNews':66,955, '#Wuhan':62,760, '#QAnon':47,237, and '#MAGA':42,610. Much of the verbiage along with some of the hashtags relate to conspiracy theories surrounding the Pandemic as well as politically right leaning news websites. Thus, this cluster seems to contain both politically right biased news along with things like conspiracy theories, which would indicate both of these types of websites see similar usage patterns on Twitter.

## 5 Discussion

In this work we analyzed the usage patterns of websites on Twitter during the COVID-19 pandemic. We used multi-view clustering to do a clustering analysis on the 50,000 most used URLs on Twitter during the COVID-19 pandemic. The novel use of multi-view clustering allowed for then incorporation of multiple different views which could be used to describe how URLs are used in Twitter. In order to perform the multi-view clustering, we used three different views of usage of URLs on Twitter: the text the co-occurs in the tweets with the URLs, the users who retweet tweets of the URLs and the users who normally tweet the

URLs. From the performance of the multi-view clustering algorithm, MVMC, we observed that the users who tweet a URL are the most important view for finding cluster structures in the usage patterns of URLs in the data. From the multi-view clustering of the URLs, there were three main clusters of URL usage.

These clusters differed in composition largely along political biases and by their having domains associated with fake news and conspiracy theories. In particular, one of the clusters, which contained most of the politically right-biased domains also contained nearly all of the domains that have been associated with fake news, conspiracy theories and other low-reliability information sources. This is to say that the usage patterns in terms of the text that occurs with particular URLs and the users that tweet and retweet particular URLs tend to be the same for politically right biased news sites as it does for conspiracy theories or fake news. This similar pattern in usage of the different URLs may indicate that more politically conservative Twitter users may be more apt to share and interact with conspiracy theories and fake news than more politically liberal leaning users. Furthermore, these ideological splits in the clusters also implies that users who share content from one political bias tend not to share, either through tweeting or retweeting, content of other political biases. So, for a major world event which is not inherently political in nature, like the COVID-19 pandemic, Twitter users still engage in discussion about the world event with political overtones, especially with regards to the external content that they share. So the findings of this study could have implications in terms of how best to craft important medical information about the COVID-19 pandemic in order to reach a populace through social media. In order to reach a wider audience with good health information through a social-media site like Twitter, it may be necessary to craft multiple messages along political lines in order reach a large number of users.

There are some limitations and directions for future research presented by this study. First, the Twitter data used in this study was collected through Twitter's streaming API and so it is not all of the data that occurred on Twitter during the time period of investigation. So, while we used a large sample of the available Twitter data and the Tweets follow patterns observed by other studies, it is still important to note that the data used in this study was a sample of all of the available data. Additionally, it is also important to note that the domain bias and factual rating labels come from professional bias and domain rating websites, which can themselves be subject to biases when labeling domains. This is in many ways an unavoidable problem, but we have used labellings from industry-standard bias and fact-checking sources as the best means to mitigate the possible bias within the labels. Also, given the prevalence in bot activity, especially on Twitter, its not clear how much the tweeting activity of COIVD-19 information along political lines is manufactured by bots or genuine, real-life user behavior. Finally, it is unclear if the same usage patterns of URLs — namely similar usage in terms of content and interactions for fake news and conspiracy theories as right-leaning news media — holds on other social media platforms. So, for future research, we intend to investigate website usage on other social

media sites to see if there are similar patterns of website usage as was observed in this study. We also intend to conduct a more focused clustering analysis of those websites associated with conspiracy theories to see if there are different patterns of usage of different conspiracy theories during the COIVD-19 Pandemic.

## Acknowledgements

This work is supported in part by the Office of Naval Research under the Multidisciplinary University Research Initiatives (MURI) Program award number N000141712675, Near Real Time Assessment of Emergent Complex Systems of Confederates, the Minerva program under grant number N000141512797, Dynamic Statistical Network Informatics, a National Science Foundation Graduate Research Fellowship (DGE 1745016), and by the center for Computational Analysis of Social and Organizational Systems (CASOS). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the ONR or the U.S. government.

## References

1. Article19: Viral lies: Misinformation and the coronavirus. Tech. rep. (3 2020), https://www.article19.org/wp-content/uploads/2020/03/Coronavirus-briefing.pdf
2. Bai, S., Sun, S., Bai, X., Zhang, Z., Tian, Q.: Improving context-sensitive similarity via smooth neighborhood for object retrieval. Pattern Recognition **83**, 353 – 364 (2018). https://doi.org/https://doi.org/10.1016/j.patcog.2018.06.001, http://www.sciencedirect.com/science/article/pii/S0031320318302115
3. Baltrusaitis, T., Ahuja, C., Morency, L.: Multimodal machine learning: A survey and taxonomy. CoRR **abs/1705.09406** (2017), http://arxiv.org/abs/1705.09406
4. Baltrušaitis, T., Ahuja, C., Morency, L.: Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence **41**(2), 423–443 (2 2019). https://doi.org/10.1109/TPAMI.2018.2798607
5. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment **2008**(10), 10008 (10 2008). https://doi.org/10.1088/1742-5468/2008/10/P10008
6. Boberg, S., Quandt, T., Schatto-Eckrodt, T., Frischlich, L.: Pandemic Populism: Facebook Pages of Alternative News Media and the Corona Crisis – A Computational Content Analysis. arXiv e-prints arXiv:2004.02566 (4 2020)
7. Chen, E., Lerman, K., Ferrara, E.: COVID-19: The First Public Coronavirus Twitter Dataset. arXiv e-prints arXiv:2003.07372 (3 2020)
8. Cinelli, M., Quattrociocchi, W., Galeazzi, A.r., Valensise, C.M., Brugnoli, E., Schmidt, A.L., Zola, P., Zollo, F., Scala, A.: The COVID-19 Social Media Infodemic. arXiv e-prints arXiv:2003.05004 (3 2020)
9. Cruickshank, I.J.: Multi-view Clustering of Social-based Data. Ph.D. thesis, Carnegie Mellon University (7 2020)

10. Cruickshank, I.J., Carley, K.M.: Characterizing communities of hashtag usage on twitter during the 2020 covid-19 pandemic by multi-view clustering. Applied Network Science **5**(66) (9 2020). https://doi.org/10.1007/s41109-020-00317-8, https://appliednetsci.springeropen.com/articles/10.1007/s41109-020-00317-8
11. Ferrara, E.: #COVID-19 on Twitter: Bots, Conspiracies, and Social Media Activism. arXiv e-prints arXiv:2004.09531 (4 2020)
12. Figueiredo, F., Jorge, A.: Identifying topic relevant hashtags in twitter streams. Information Sciences **505**, 65 – 83 (2019). https://doi.org/https://doi.org/10.1016/j.ins.2019.07.062, http://www.sciencedirect.com/science/article/pii/S0020025519306668
13. Fortunato, S., Barthelemy, M.: Resolution limit in community detection. Proceedings of the National Academy of Science **104**(1), 36–41 (1 2007). https://doi.org/10.1073/pnas.0605965104
14. Gallotti, R., Valle, F., Castaldo, N., Sacco, P., De Domenico, M.: Assessing the risks of "infodemics" in response to COVID-19 epidemics. arXiv e-prints arXiv:2004.03997 (4 2020)
15. Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., Lazer, D.: Fake news on twitter during the 2016 u.s. presidential election. Science **363** (2019). https://doi.org/10.1126/science.aau2706, https://pubmed.ncbi.nlm.nih.gov/30679368/
16. Huang, B.: Learning User Latent Attributes on Social Media. Ph.D. thesis, Carnegie Mellon University (5 2020)
17. Huang, S., Chaudhary, K., Garmire, L.X.: More is better: Recent progress in multi-omics data integration methods. Frontiers in Genetics **8**, 84 (2017). https://doi.org/10.3389/fgene.2017.00084, https://www.frontiersin.org/article/10.3389/fgene.2017.00084
18. Hussain, W.: Role of social media in covid-19 pandemic **4** (4 2020). https://doi.org/10.37978/tijfs.v4i2.144, http://publie.frontiersscienceassociates.com/index.php/tijfs/article/view/144
19. Kantis, C., Kiernan, S., Bardi, J.: Timeline of the coronavirus: Think global health, https://www.thinkglobalhealth.org/article/updated-timeline-coronavirus
20. Lancichinetti, A., Fortunato, S.: Limits of modularity maximization in community detection **84**, 066122 (12 2011). https://doi.org/10.1103/PhysRevE.84.066122
21. Maier, M., Hein, M., von Luxburg, U.: Optimal construction of k-nearest neighbor graphs for identifying noisy clusters. arXiv e-prints arXiv:0912.3408 (12 2009)
22. Maier, M., von Luxburg, U., Hein, M.: How the result of graph clustering methods depends on the construction of the graph. arXiv e-prints arXiv:1102.2075 (2 2011)
23. Majmundar, A., Allem, J.P., Boley Cruz, T., Unger, J.B.: The why we retweet scale. PLOS ONE **13**(10), 1–12 (10 2018). https://doi.org/10.1371/journal.pone.0206076, https://doi.org/10.1371/journal.pone.0206076
24. Newman, M.E.J.: Community detection in networks: Modularity optimization and maximum likelihood are equivalent. arXiv e-prints arXiv:1606.02319 (6 2016)
25. Pamfil, A.R., Howison, S.D., Lambiotte, R., Porter, M.A.: Relating modularity maximization and stochastic block models in multilayer networks. CoRR **abs/1804.01964** (2018), http://arxiv.org/abs/1804.01964
26. Qiao, L., Zhang, L., Chen, S., Shen, D.: Data-driven graph construction and graph learning: A review. Neurocomputing **312**, 336 – 351 (2018). https://doi.org/https://doi.org/10.1016/j.neucom.2018.05.084, http://www.sciencedirect.com/science/article/pii/S0925231218306696

27. Reichardt, J., Bornholdt, S.: Statistical mechanics of community detection. Phys. Rev. E **74**, 016110 (7 2006). https://doi.org/10.1103/PhysRevE.74.016110, https://link.aps.org/doi/10.1103/PhysRevE.74.016110
28. Traag, V.A., Waltman, L., van Eck, N.J.: From louvain to leiden: guaranteeing well-connected communities. Nature: Scientific Reports **9** (2019). https://doi.org/https://doi.org/10.1038/s41598-019-41695-z, https://www.nature.com/articles/s41598-019-41695-z
29. Vicient, C., Moreno, A.: Unsupervised topic discovery in micro-blogging networks. Expert Systems with Applications **42**(17), 6472 – 6485 (2015). https://doi.org/10.1016/j.eswa.2015.04.014, http://www.sciencedirect.com/science/article/pii/S0957417415002444
30. Yang, K.C., Torres-Lugo, C., Menczer, F.: Prevalence of Low-Credibility Information on Twitter During the COVID-19 Outbreak. arXiv e-prints arXiv:2004.14484 (4 2020)
31. Yang, Y., Wang, H.: Multi-view clustering: A survey. Big Data Mining and Analytics **1**(2), 83–107 (6 2018)
32. Ye, F., Chen, Z., Qian, H., Li, R., Chen, C., Zheng, Z.: New approaches in multi-view clustering. Recent Applications in Data Clustering (2018). https://doi.org/10.5772/intechopen.75598, https://www.intechopen.com/books/recent-applications-in-data-clustering/new-approaches-in-multi-view-clustering
33. Zhu, X., Loy, C.C., Gong, S.: Constructing robust affinity graphs for spectral clustering. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1450–1457 (6 2014). https://doi.org/10.1109/CVPR.2014.188
34. Zitnik, M., Nguyen, F., Wang, B., Leskovec, J., Goldenberg, A., Hoffman, M.M.: Machine Learning for Integrating Data in Biology and Medicine: Principles, Practice, and Opportunities. arXiv e-prints arXiv:1807.00123 (6 2018)