

# ReCOVary: A Multimodal Repository for COVID-19 News Credibility Research

Xinyi Zhou  
zhouxinyi@data.syr.edu  
Data Lab, EECS Department  
Syracuse University

Emilio Ferrara  
emiliofe@usc.edu  
Information Sciences Institute,  
University of Southern California

Apurva Mulay  
asmulay@syr.edu  
Data Lab, EECS Department  
Syracuse University

Reza Zafarani  
reza@data.syr.edu  
Data Lab, EECS Department  
Syracuse University

## ABSTRACT

First identified in Wuhan, China, in December 2019, the outbreak of COVID-19 has been declared as a global emergency in January, and a pandemic in March 2020 by the World Health Organization (WHO). Along with this pandemic, we are also experiencing an “infodemic” of information with low credibility such as fake news and conspiracies. In this work, we present ReCOVary, a repository designed and constructed to facilitate research on combating such information regarding COVID-19. We first broadly search and investigate ~2,000 news publishers, from which 60 are identified with extreme [high or low] levels of credibility. By inheriting the credibility of the media on which they were published, a total of 2,029 news articles on coronavirus, published from January to May 2020, are collected in the repository, along with 140,820 tweets that reveal how these news articles have spread on the Twitter social network. The repository provides multimodal information of news articles on coronavirus, including textual, visual, temporal, and network information. The way that news credibility is obtained allows a trade-off between dataset scalability and label accuracy. Extensive experiments are conducted to present data statistics and distributions, as well as to provide baseline performances for predicting news credibility so that future methods can be compared. Our repository is available at <http://coronavirus-fakenews.com>.

## CCS CONCEPTS

• **Information systems** → **Collaborative and social computing systems and tools**; **Clustering and classification**; • **Security and privacy** → *Social aspects of security and privacy*.

## KEYWORDS

Repository; COVID-19; coronavirus; pandemic; infodemic; information credibility; fake news; multimodal; social media

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](https://permissions.acm.org).

CIKM '20, October 19–23, 2020, Virtual Event, Ireland

© 2020 Association for Computing Machinery.

ACM ISBN 978-1-4503-6859-9/20/10...\$15.00

<https://doi.org/10.1145/3340531.3412880>

## ACM Reference Format:

Xinyi Zhou, Apurva Mulay, Emilio Ferrara, and Reza Zafarani. 2020. ReCOVary: A Multimodal Repository for COVID-19 News Credibility Research. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM '20), October 19–23, 2020, Virtual Event, Ireland*. ACM, New York, NY, USA, 8 pages. <https://doi.org/10.1145/3340531.3412880>

## 1 INTRODUCTION

As of June 4<sup>th</sup>, the COVID-19 pandemic has resulted in over 21 million confirmed cases and 775,893 deaths globally.<sup>1</sup> Governments have enforced border shutdowns, travel restrictions, and quarantines to “flatten the curve”. The COVID-19 outbreak has had a detrimental impact on not only the healthcare sector but also every aspect of human life such as education and economic sectors [11]. For example, over 100 countries have imposed nationwide (even complete) closures of education facilities, which has led to over 900 million learners being affected.<sup>2</sup> Statistics indicate that 3.3 million Americans applied for unemployment benefits in the week ending on March 21<sup>st</sup> and the number doubled in the following week, before which the highest number of unemployment applications ever received in one week was 695,000 in 1982.<sup>3</sup>

Along with the COVID-19 pandemic, we are also experiencing an “infodemic” of information with low credibility regarding COVID-19.<sup>4</sup> Hundreds of news websites have contributed to publishing false coronavirus information.<sup>5</sup> Individuals who believe false news articles (e.g., claiming that eating boiled garlic or drinking chlorine dioxide, an industrial bleach, can cure or prevent coronavirus), might take ineffective or extremely dangerous actions to protect themselves from the virus.<sup>6</sup>

Given this background, research is motivated to combat this infodemic. Hence, we design and construct a multimodal repository, ReCOVary, to facilitate reliability assessment of news on COVID-19. We first broadly search and investigate ~2,000 news publishers,

<sup>1</sup><https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200604-covid-19-sitrep-136.pdf>

<sup>2</sup><https://en.unesco.org/covid19/educationresponse>

<sup>3</sup><https://www.npr.org/2020/03/26/821580191/unemployment-claims-expected-to-shatter-records>

<sup>4</sup><https://www.un.org/en/un-coronavirus-communications-team/un-tackling-%E2%80%98infodemic%E2%80%99-misinformation-and-cybercrime-covid-19>

<sup>5</sup><https://www.newsguardtech.com/coronavirus-misinformation-tracking-center/>

<sup>6</sup><https://www.factcheck.org/2020/02/fake-coronavirus-cures-part-1-mms-is-industrial-bleach/>

from which 60 with various political polarizations and from different countries are identified with extreme [high or low] credibility. As past literature has indicated, there is a close relationship between the credibility of news articles and their publication sources [20]. In total, 2,029 news articles on coronavirus are finally collected in the repository along with 140,820 tweets that reveal how these news articles are spread on the social network. The main contributions of this work are summarized as follows:

- (1) We construct a repository to support the research that investigates (i) how news with low credibility is created and spread in the COVID-19 pandemic and (ii) ways to predict such “fake” news. The manner in which the ground truth of news credibility is obtained allows a scalable repository, as annotators need not label each news article that is time-consuming and instead they can directly label the news site;
- (2) ReCOVeRY provides multimodal information on COVID-19 news articles. For each news article, we collect its news content and social information revealing how it spreads on social media, which covers textual, visual, temporal, and network information; and
- (3) We conduct extensive experiments using ReCOVeRY data, which includes data analyses (data statistics and distributions) and baseline performances for predicting news credibility. These baselines allow future methods to be compared to. Baselines are obtained using either single-modal or multimodal information of news articles and utilize either traditional statistical learning or deep learning.

The rest of this paper is organized as follows. We first review the related datasets in Section 2. Then, we detail how the data is collected in Section 3. The statistics and distributions of the data are presented and analyzed in Section 4. Experiments that use the data to predict news credibility are designed and conducted in Section 5, whose results can be used as benchmarks. We conclude in Section 6.

## 2 RELATED WORK

Related datasets can be generally grouped as (I) COVID-19 datasets and (II) “fake” news and rumor datasets.

*COVID-19 Datasets.* As a global emergency, the outbreak of COVID-19 has been labelled as a black swan event and likened to the economic scene of World War II [11]. With this background, a group of datasets have emerged, whose contributions range from real-time tracking of COVID-19 to help epidemiological forecasting (e.g., [5] and [18]) and collecting scholarly COVID-19 articles for literature-based discoveries (e.g., CORD-19<sup>7</sup>), to tracking the spreading of COVID-19 information on Twitter (e.g., [2]).

Specifically, researchers at Johns Hopkins University develop a Web-based dashboard<sup>8</sup> to visualize and track reported cases of COVID-19 in real-time. The dashboard is released on January 22<sup>nd</sup>, presenting the location and number of confirmed COVID-19 cases, deaths, and recoveries for all affected countries [5]. Another dataset shared publicly on March 24<sup>th</sup> is constructed to aid the analysis and tracking of the COVID-19 epidemic, which provides real-time individual-level data (e.g., symptoms; date of onset, admission, and confirmation; and travel history) from national, provincial, and

municipal health reports [18]. The Allen Institute for AI has contributed a free and dynamic database of more than 128,000 scholarly articles about COVID-19, named CORD-19, to the global research community.<sup>7</sup> The intention is to mobilize researchers to apply recent advances in Natural Language Processing (NLP) to generate new insights to support the fight against COVID-19. Furthermore, Chen et al. [2] release the first large-scale COVID-19 twitter dataset. The dataset, updated regularly, collects COVID-19 tweets that are posted from January 21<sup>st</sup> and across languages.

Though these datasets have been broadly investigated and have contributed to the research on coronavirus pandemic, they do not provide the ground truth on the credibility of information on coronavirus to help fight the coronavirus infodemic.

*“Fake” News and Rumor Datasets.* Existing “fake” news and rumor datasets are collected with various focuses. These datasets may (i) only contain news content that can be full articles (e.g., NELA-GT-2018 [12]) or short claims (e.g., FEVER [16]); (ii) only contain social media information (e.g., CREDBANK[10]), where news refers to user posts; or (iii) contain both content and social media information (e.g., LIAR [17], FakeNewsNet [14], and FakeHealth [4]).

Specifically, NELA-GT-2018 [12] is a large-scale dataset of around 713,000 news articles from February to November 2018. News articles are collected from 194 news medium with multiple labels directly obtained from NewsGuard, Pew Research Center, Wikipedia, OpenSources, MBFC, AllSides, BuzzFeed News, and PolitiFact. These labels refer to news credibility, transparency, political polarizations, and authenticity. FEVER dataset [16] consists of ~185,000 claims and is constructed following two steps: claim generation and annotation. First, the authors extract sentences from Wikipedia, and then the annotators manually generate a set of claims based on the extracted sentences. Then, the annotators label each claim as “supported”, “refuted”, or “not enough information” by comparing it with the original sentence from which it is developed. On the other hand, some datasets focus on user posts on social media, for example, CREDBANK [10] includes more than 60 million tweets grouped into 1049 real-world events, each of which is annotated by 30 human annotators, while some contain both news content and social media information. By collecting both claims and fact-check results (labels, i.e., “true”, “mostly true”, “half-true”, “mostly false”, and “pants on fire”) directly from PolitiFact, Wang establishes the LIAR dataset [17] containing around 12,800 verified statements made in public speeches and social medium. The aforementioned datasets only contain textual information valuable for NLP research with limited information on how “fake” news and rumors spread on social networks, which motivate the construction of FakeNewsNet and FakeHealth dataset [4, 14]. The FakeNewsNet dataset collects fact-checked (real or fake) full news articles from PolitiFact (#=1,056) and GossipCop (#=22,140), respectively and tracks news spreading on Twitter. The FakeHealth dataset collects verified (real or fake) news reviews from HealthNewsReview.org with detailed explanations and social engagements regarding news spreading on Twitter that includes a user-user social network. Note that FakeHealth concentrates on healthcare data, so does CoAID, a recently released dataset for COVID-19 misinformation research [3].

<sup>7</sup><https://www.semanticscholar.org/cord19>

<sup>8</sup><https://coronavirus.jhu.edu/map.html>

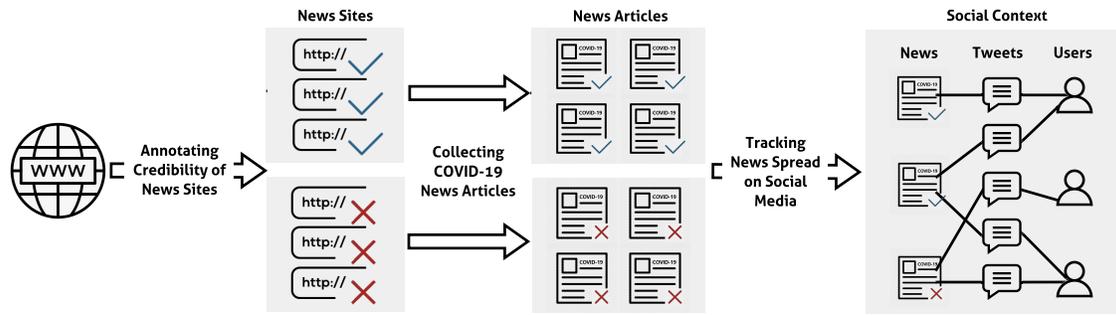


Figure 1: Data Collection Process for ReCOVery

In general, compared to datasets such as NELA-GT-2018, FEVER, and LIAR, our repository provides multimodal information and social engagements of news articles. Compared to CREDBANK and FakeNewsNet, ReCOVery aims to fight the coronavirus infodemic and presents a novel approach to collecting and annotating data, which allows the trade-off between data scalability and label accuracy. Compared to FakeHealth and CoAID, news articles in ReCOVery are from a mix of domains that include healthcare.

### 3 DATA COLLECTION

The overall process that we collect the data, including news content and social media information, is presented in Figure 1. To facilitate scalability, news credibility is assessed based on the credibility of the media (site) that publishes the news article. Based on the process outlined in Figure 1, we will further detail how the data is collected, answering the following three questions: (1) how to identify reliable (or unreliable) news sites mainly releasing real news (or fake news)? (which we address in Section 3.1); having determined such news sites, (2) how do we crawl COVID-19 news articles from these sites and what news components are valuable for collection? (Section 3.2); and given COVID-19 news articles, (3) how can we track their spread on social networks? (Section 3.3)

#### 3.1 Filtering News Sites

To determine a list of reliable and unreliable news sites, we primarily rely on two resources: *NewsGuard* and *Media Bias/Fact Check*.

*NewsGuard*.<sup>9</sup> NewsGuard is developed to review and rate news websites. Its reliability rating team is formed by trained journalists and experienced editors, whose credentials and backgrounds are all transparent and available on the site. The performance (credibility) of each news website is assessed based on the following nine journalistic criteria:

- (1) Does not repeatedly publish false content, (22 points)
- (2) Gathers and presents information responsibly, (18 points)
- (3) Regularly corrects or clarifies errors, (12.5 points)
- (4) Handles the difference between news and opinion responsibly, (12.5 points)
- (5) Avoids deceptive headlines, (10 points)
- (6) Website discloses ownership and financing, (7.5 points)
- (7) Clearly labels advertising, (7.5 points)

<sup>9</sup><https://www.newsguardtech.com/>

- (8) Reveals who’s in charge, including possible conflicts of interest, and (5 points)
- (9) The site provides the names of content creators, along with either contact or biographical information, (5 points)

where the overall score of a site is between 0 to 100; 0 indicates the lowest credibility, and 100 indicates the highest credibility. A news website with a NewsGuard score higher than 60 is often labeled as reliable; otherwise, it is unreliable. NewsGuard has provided ground truth for the construction of news datasets such as NELA-GT-2018 [12] for studying misinformation.

*Media Bias/Fact Check (MBFC)*.<sup>10</sup> MBFC is a website that rates factual accuracy and political bias of news medium. The fact-checking team consists of Dave Van Zandt, the primary editor and the website owner, and some journalists and researchers (more details can be found on its “About” page). MBFC labels each news media as one of six factual-accuracy levels based on the fact-checking results of the news articles it has published (more details can be found on its “Methodology” page): (i) very high, (ii) high, (iii) most factual, (iv) mixed, (v) low, and (vi) very low. Such information has been used as ground truth for automatic fact-checking studies. [1]

*What Are Our Criteria?* Referenced by NewsGuard and MBFC, our criteria for determining reliable and unreliable news sites are:

- ✓ **Reliable** A news site is reliable if its NewsGuard score is greater than 90, **and** its factual reporting on MBFC is very high or high.
- ✗ **Unreliable** A news site is unreliable if its NewsGuard score is less than 30, **and** its factual reporting on MBFC is below mixed.

Our search towards news medium with high credibility is conducted among news medium listed in MBFC (~2,000). To find news medium with low credibility, we search in MBFC and the newly released “Coronavirus Misinformation Tracking Center”<sup>5</sup> of NewsGuard, which provides a list of websites publishing false coronavirus information. Ultimately, we obtain a total of 60 news sites, from which 22 are the sources of reliable news articles (e.g., *National Public Radio*<sup>11</sup> and *Reuters*<sup>12</sup>) and the remaining 38 are sources to collect unreliable news articles (e.g., *Human Are Free*<sup>13</sup> and *Natural*

<sup>10</sup><https://mediabiasfactcheck.com/>

<sup>11</sup><https://www.npr.org>

<sup>12</sup><https://www.reuters.com>

<sup>13</sup><http://humansarefree.com/>

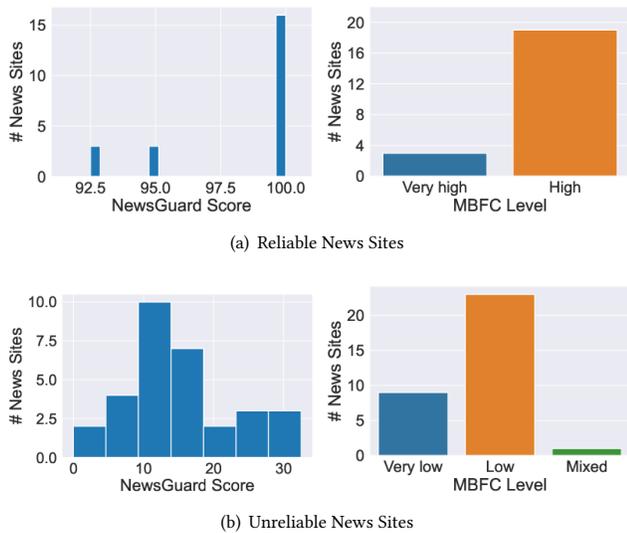


Figure 2: Credibility Distribution of Determined News Sites

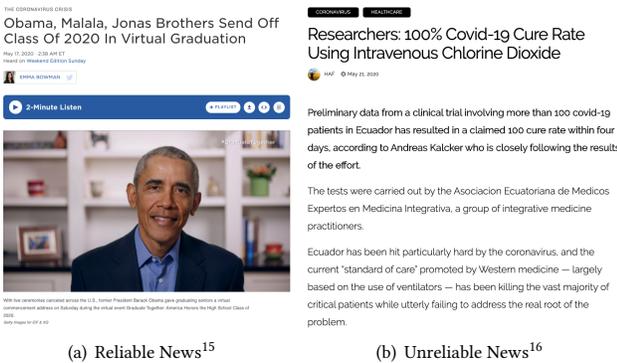


Figure 3: Examples of News Articles Collected

News<sup>14</sup>). The full list of sites considered in our repository is also available at <http://coronavirus-fakenews.com>. Note that several “fake” news medium are not included, such as *70 News*, *Conservative 101*, and *Denver Guardian*, since they no longer exist or their domains have been unavailable.

Also note that to achieve a good trade-off between dataset scalability and label accuracy, we utilize more extreme threshold scores (30 and 90) compared to the initial one provided by NewsGuard (60). In this way, the selected news sites exhibit an *extreme* reliability (or unreliability), which helps reduce the number of false positives and false negatives in news labels in our repository; ideally, each news article published on a reliable site is factual, and on an unreliable

<sup>14</sup><https://www.naturalnews.com>

<sup>15</sup><https://www.npr.org/sections/coronavirus-live-updates/2020/05/17/857512288/obama-malala-jonas-brothers-send-off-class-of-2020-in-virtual-graduation>

<sup>16</sup><https://humansarefree.com/2020/05/researchers-100-covid-19-cure-rate-using-intravenous-chlorine-dioxide.html>

site is false. Figure 2 illustrates the credibility distributions of reliable and unreliable news sites. It can be observed from the figure that for reliable news, most of them have a full mark on NewsGuard and are labeled as “high”ly factual by MBFC; “very high” is rare for all sites listed in MBFC. In contrast, unreliable news sites share an average NewsGuard score of ~15 and a low factual label by MBFC; similarly, “very low” is rarely given on MBFC.

### 3.2 Collecting COVID-19 News Content

To crawl COVID-19 news articles from selected news sites, we first determine whether the news article is about COVID-19; the process is detailed in Section 3.2.1. Next, we detail how the data is crawled and the news content components that are included in our repository in Section 3.2.2.

3.2.1 *News Topic Identification.* To identify news articles on COVID-19, we use a list of keywords:

- SARS-CoV-2,
- COVID-19, and
- Coronavirus.

News articles whose content contains any of the keywords (case-insensitive) are considered related to COVID-19. These three keywords are the official names announced by the WHO on February 11<sup>th</sup>, where SARS-CoV-2 (standing for Severe Acute Respiratory Syndrome CoronaVirus 2) is the virus name, and Coronavirus and COVID-19 are the name of the disease that the virus causes. Before the WHO announcement, COVID-19 was previously known as the “2019 novel coronavirus,”<sup>17</sup> which also includes the coronavirus keyword which we are considering. We merely consider official names as keywords to avoid potential biases or even discrimination in naming. Furthermore, a news media (article) that is credible, or pretends to be credible, often acts professionally and adopts the official name(s) of the disease/virus. Compared to those articles that use biased and/or inaccurate terms, false news pretending to be professional is more detrimental and challenging to detect, which has become the focus of current fake news studies. [24] Examples of such news articles are illustrated in Figure 3.

3.2.2 *Crawling News Content.* Content crawler relies on Newspaper Python library.<sup>18</sup> The content of each news article corresponds to twelve components:

- (C1) *News ID:* Each news article is assigned a unique id as the identity;
- (C2) *News URL:* The URL of the news article. The URL helps us verify the correctness of the collected data. It can also be used as the reference and source when repository users would like to extend the repository by fetching additional information;
- (C3) *Publisher:* The name of the news media (site) that publishes the news article;
- (C4) *Publication Date:* The date (in yyyy-mm-dd format) on which the news article was published on the site, which provides temporal information to support the investigation of, e.g., the relationship between the misinformation volume and the outbreak of COVID-19 over time;

<sup>17</sup>[https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-\(covid-2019\)-and-the-virus-that-causes-it](https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that-causes-it)

<sup>18</sup><https://github.com/codelucas/newspaper>

- (C5) *Author*: The author(s) of the news article, whose number can be none, one, or more than one. Note that some news articles might have fictional author names. Author information is valuable in evaluating news credibility by either investigating the collaboration network of authors [15] or exploring its relationships with news publishers and content [21];
- (C6-7) *News Title and Bodytext* as the main textual information;
- (C8) *News Image* as the main visual information, which is provided in the form of a link (URL). Note that most images within the news page are noise – they can be advertisements, images belonging to other news articles due to the recommender systems embedded in news sites, logos of news sites and/or social media icons, such as Twitter and Facebook logos for sharing. Hence, we particularly fetch the main/head/top image for each news article to reduce noise;
- (C9) *Country*: The name of country where the news is published;
- (C10) *Political bias*: Each news article is labeled as one of ‘extremely left’, ‘left’, ‘left-center’, ‘center’, ‘right-center’, ‘right’, and ‘extremely right’ that is equivalent to the political bias of its publisher. News political bias is verified by two resources, AllSides<sup>19</sup> and MFBC, both of which rely on domain experts to label media bias; and
- (C11-12) *NewsGuard score and MBFC factual reporting* as the original ground truth of news credibility, which has been detailed in Section 3.1.

### 3.3 Tracking News Spreading on Social Media

We first use Twitter Premium Search API<sup>20</sup> to track the spread of collected news articles on Twitter. Specifically, our search is based on the URL of each news article and looks for tweets posted after the date when the news article was published to the current date (for the current version of the dataset, this date is May 26<sup>th</sup>). Twitter Search API can return the corresponding tweets with detailed information such as their IDs, text, languages of text, times of being created, statistics on retweeted/replied/liked. Also, it returns the information of users who post these tweets, such as user IDs and their number of followers/friends. To comply with Twitter’s Terms of Service,<sup>21</sup> we only publicly release the IDs of the collected data for non-commercial research use, but provide the instructions for obtaining the tweets using the released IDs for user convenience. More details can be seen in <http://coronavirus-fakenews.com>.

## 4 DATA STATISTICS AND DISTRIBUTIONS

The general statistics on our dataset is presented in Table 1. The dataset contains 2,029 news articles, most of which have both textual and visual information for multimodal studies ( $\#=2,017$ ), [23] and have been shared on social media ( $\#=1,747$ ). The dataset is imbalanced in news class – the proportion of reliable versus unreliable news articles is around 2:1. The number of users who spread reliable news ( $\#=78,659$ ) plus that of users spreading unreliable news ( $\#=17,323$ ) is greater than the total number of users included in the dataset ( $\#=93,761$ ). This observation indicates that users can both engage in spreading reliable and unreliable news articles.

**Table 1: Data Statistics**

	Reliable	Unreliable	Total
<b>News articles</b>	1,364	665	2,029
w/ images	1,354	663	2,017
w/ social information	1,219	528	1,747
<b>Tweets</b>	114,402	26,418	140,820
<b>Users</b>	78,659	17,323	93,761

Next, we visualize the distributions of data features/attributes.

*Distribution of News Publishers.* Figure 4 shows the number of COVID-19 news articles published in each [extremely reliable or extremely unreliable] news site. There are five unreliable publishers with no news on COVID-19; hence, they are not presented in the figure. We keep these publishers in our repository as the data will be updated over time and these publishers may publish news articles on COVID-19 in the future.

*News Publication Dates.* The distribution of news publication dates is presented in Figure 5, where all articles are published in 2020. We point out that from January to May, the number of COVID-19 news articles published is significantly (exponentially) increased. The possible explanation for this phenomena is three-fold. First, from the time that the outbreak was first identified in Wuhan, China (December 2019) [7] to May 2020, the number of confirmed cases and deaths caused by SARS-CoV-2 have exponentially grown globally.<sup>1</sup> Meanwhile, the virus has become a world topic and has triggered more and more discussions on a world-wide scale. Second, some older news articles are no longer available, which has motivated us to timely update the dataset. Third, the keywords we have used to identify COVID-19 news articles are the official ones provided by the WHO in February.<sup>17</sup> Some news articles published in January are also collected, as before the WHO announcement COVID-19 was known as the “2019 novel coronavirus,” which also includes one of our keywords “coronavirus.” We have detailed the reasons behind our keyword selection in Section 3.2.1. Note that there are a small group of news articles whose publication dates are not accessible, which we denote as N/A in Figure 5.

*News Authors and Author Collaborations.* Figure 6 presents the distribution of the number of authors contributing to news articles, which is governed by a long-tail distribution: most articles are contributed by  $\leq 5$  authors. Instead of including the [real or virtual] names of the authors, some articles provide publisher names as authors. Considering such information has been available in the repository, we leave the author information of these news articles blank, i.e., their number of authors is zero. Furthermore, we construct the coauthorship network, shown in Figure 7. It can be observed from the network that node degrees also follow a power-law-like distribution: among 1,095 nodes (authors), over 90% of them have less than or equal to two collaborators.

*News Content Statistics.* Both Figures 8 and 9 reveal textual characteristics within news content (including news title and bodytext). It can be observed from Figure 8 that the number of words within news content follows a long-tail (power-law-like) distribution, with an average value of  $\sim 800$  and a median value of  $\sim 600$ . On the other

<sup>19</sup><https://www.all-sides.com/unbiased-balanced-news>

<sup>20</sup><https://developer.twitter.com/en/docs/tweets/search/overview/premium>

<sup>21</sup><https://developer.twitter.com/en/developer-terms/agreement-and-policy>

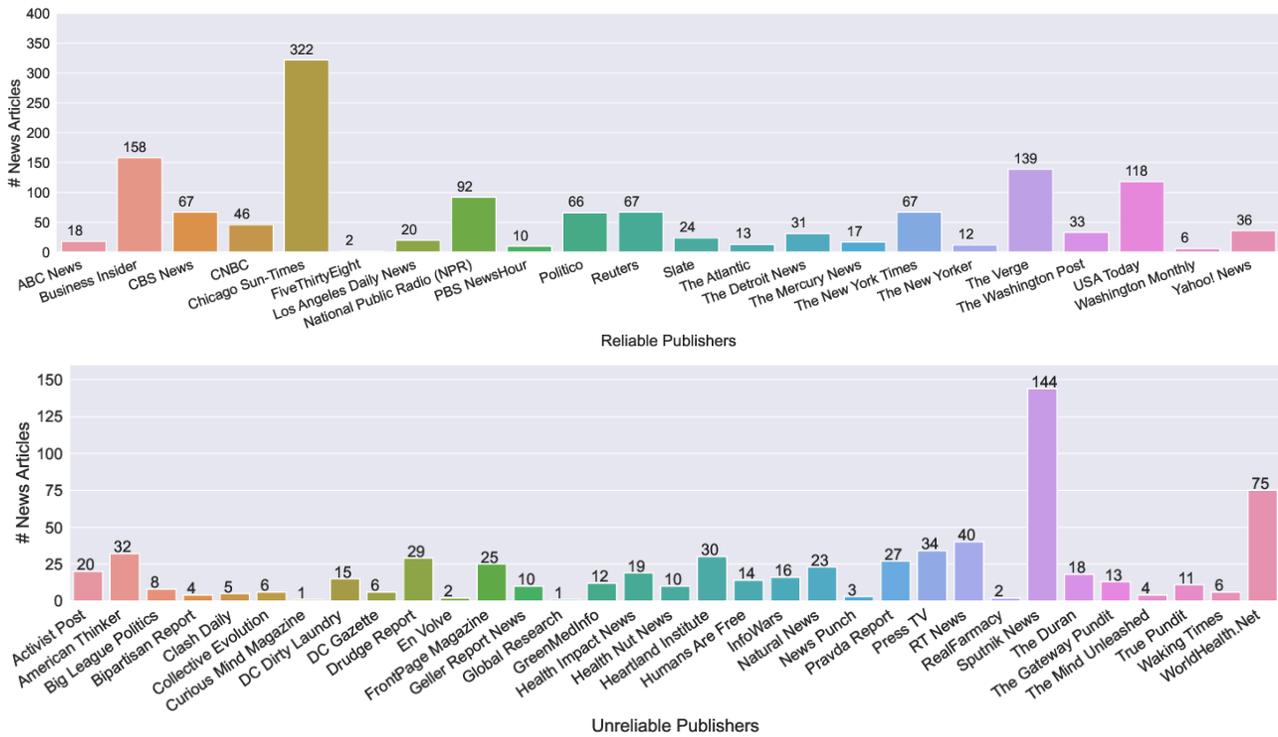


Figure 4: Distribution of News Publishers

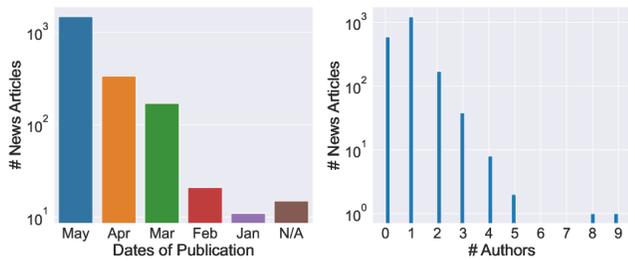


Figure 5: Publication Date

Figure 6: Author Count

hand, Figure 9 provides the word cloud for the entire repository. As the news articles collected share the same COVID-19 topic, some relevant topics and vocabularies have been naturally and frequently used by the news authors, such as “coronavirus” (#=6465), “COVID” (#=5413), “state” (#=4432), “test” (#=4274), “health” (#=3714), “pandemic” (#=3427), “virus” (#=2903), “home” (#=2871), “case” (#=2676), and “Trump” (#=2431) that are illustrated with word font size scaled to their frequencies.

**Country Distribution.** Figure 10 reveals the countries that news and news publishers belong to. It can be observed that in total six countries – United States (abbr. US), Russia (abbr. RU), United Kingdom (abbr. UK), Iran (abbr. IR), Cyprus (abbr. CY), and Canada (abbr. CA) – are covered, where US news and news publishers constitute the majority of the population.

**Political Bias.** Figure 11 provides the distribution of political bias of news and news medium (publishers). It can be observed from the figure that for both news and publishers, the distribution for those exhibiting a right bias (including extremely right (abbr. Ex. R), right (abbr. R), and right-center (abbr. R-C)) is more balanced compared to those exhibiting a left bias (including extremely left (abbr. Ex. L), left (abbr. L), and left-center (abbr. L-C)).

**News Spreading Frequencies.** Figure 12 shows the distribution of the number of tweets sharing each news article. The distribution exhibits a long tail – over 80% of news articles are spread less than 100 times while a few have been shared by thousands of tweets.

**News Spreaders.** The distribution of the number of spreaders for each news article is shown in Figure 13. It differs from the distribution in Figure 12 as one user can spread a news article multiple times. As for social connections of news spreaders, the distributions of their followers and friends are respectively presented in Figures 14 and 15, where the most popular spreader has over 40 million followers (or 600,000 friends).

## 5 FORMING BASELINES: USING ReCOVeRY TO PREDICT COVID-19 NEWS CREDIBILITY

In this section, several methods that often act as baselines are utilized and developed to predict COVID-19 news credibility using ReCOVeRY data, hoping to facilitate future studies. These methods (baselines) are first specified in Section 5.1. The implementation details of experiments are then provided in Section 5.2. Finally, we present the performance results for these methods in Section 5.3.

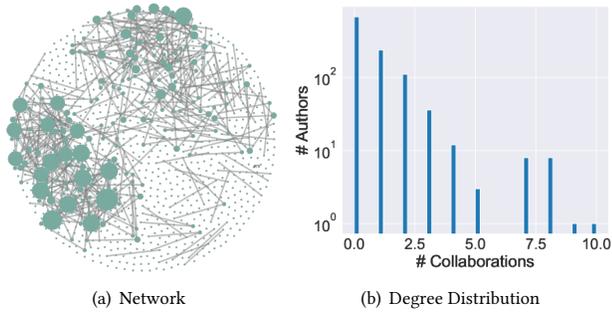


Figure 7: Author Collaborations

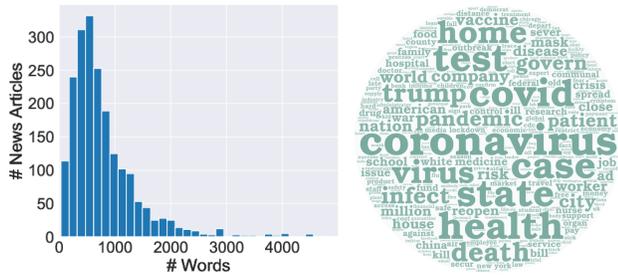


Figure 8: Word Count

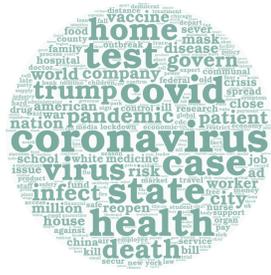


Figure 9: Word Cloud

## 5.1 Methods

We involve the following methods as our baselines. These methods can be grouped by their learning framework, which is either a traditional statistical learner such as SVM (e.g., LIWC) or a neural network (e.g., Text-CNN and SAFE). Baselines can also be grouped as single-modal methods (e.g., LIWC, RST, and Text-CNN) or multimodal methods (e.g., SAFE).

*LIWC* [13].<sup>22</sup> LIWC (Linguistic Inquiry and Word Count) is a widely-accepted psycholinguistic lexicon. Given a news story, LIWC can count the words in the text falling into one or more of 93 linguistic, psychological, and topical categories, based on which 93 features are extracted and often classified within a traditional statistical learning framework [22].

*RST*. RST (Rhetorical Structure Theory) organizes a piece of content as a tree that captures the rhetorical relation among its phrases and sentences. We use a pretrained RST parser [8]<sup>23</sup> to obtain the tree for each news article and count each rhetorical relation (in total, 45) within a tree, based on which 45 features are extracted and classified in a traditional statistical learning framework.

*Text-CNN* [9]. Text-CNN relies on a convolutional neural networks for text classification, which contains a convolutional layer and max pooling.

*SAFE* [23].<sup>24</sup> SAFE is a neural-network-based method that utilizes news multimodal information for fake news detection, where news representation is learned jointly by news textual and visual

<sup>22</sup><https://liwc.wpengine.com/>

<sup>23</sup><https://github.com/jiyfeng/DPLP>

<sup>24</sup><https://github.com/Jindi0/SAFE>

Table 2: Baselines Performance in Predicting COVID-19 News Credibility Using ReCOvery Data

Method	Reliable news			Unreliable news		
	Pre.	Rec.	$F_1$	Pre.	Rec.	$F_1$
LIWC+DT	0.779	0.771	0.775	0.540	0.552	0.545
RST+DT	0.721	0.705	0.712	0.421	0.441	0.430
Text-CNN	0.746	0.782	0.764	0.522	0.472	0.496
SAFE	0.836	0.829	0.833	0.667	0.677	0.672

information along with their relationship. SAFE facilitates recognizing the news falseness in its text, images, and/or the “irrelevance” between the text and images.

## 5.2 Implementation Details

The overall dataset is randomly divided into training and testing datasets with a proportion of 0.8:0.2. As the dataset has an unbalanced distribution between reliable and unreliable news articles ( $\approx 2:1$ ), we evaluate the prediction results in terms of precision, recall, and the  $F_1$  score. For methods relying on traditional statistical learners, multiple well-established classifiers are adopted in our experiments: Logistic Regression (LR), Naive Bayes (NB),  $k$ -Nearest Neighbor ( $k$ -NN), Random Forest (RF), Decision Tree (DT), and Support Vector Machines (SVM). We merely present the one performing best due to the space limitation. Codes are all available on <http://coronavirus-fakenews.com>.

## 5.3 Experimental Results

Prediction results are provided in Table 2. We observe that four baselines achieve an  $F_1$ -score (precision, recall) score of 71% (72%, 71%) to 83% (84%, 83%) in identifying reliable news and between 43% (42%, 44%) to 67% (67%, 68%) for unreliable news. Additionally, multimodal features are generally more representative than single-modal features in predicting news credibility. We point out that the four baselines are content-based methods; developing more advanced methods by mining social media [19] are encouraged.

## 6 CONCLUSION

To fight the coronavirus infodemic, we construct a multimodal repository for COVID-19 news credibility research, which provides textual, visual, temporal, and network information regarding news content and how news spreads on social media. The repository balances data scalability and label accuracy. To facilitate future studies, benchmarks are developed and their performances are presented on predicting news credibility using the data available in the repository. We point out that the data could be further enhanced (1) by including COVID-19 news articles in various languages such as Chinese, Russian, Spanish, and Italian, as well as the information on how these news articles spread on the popular local social media for those languages, e.g., Sina Weibo (China). Furthermore, (2) extending the dataset by introducing the ground truth of, for example, hate speech, clickbaits, and social bots [6] would help study the bias and discrimination bred by the virus, as well as the correlation among all information and accounts with low credibility. Both (1) and (2) will be our future work.

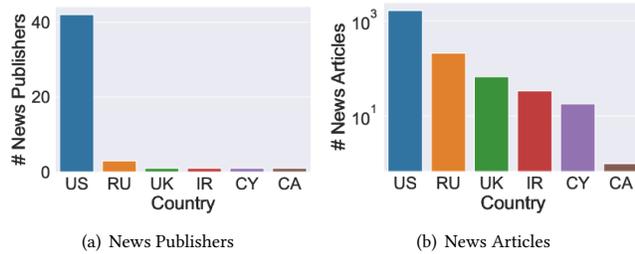


Figure 10: Country

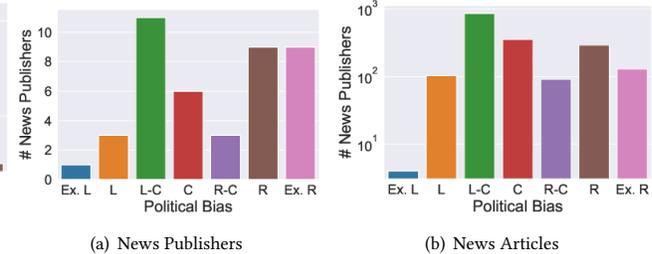


Figure 11: Political Bias

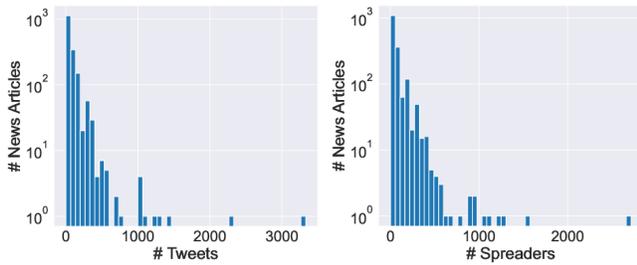


Figure 12: Spreading Frequency

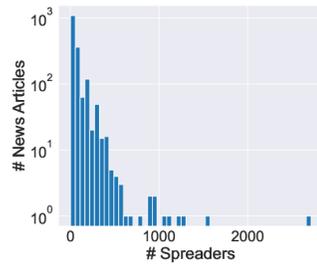


Figure 13: News Spreaders

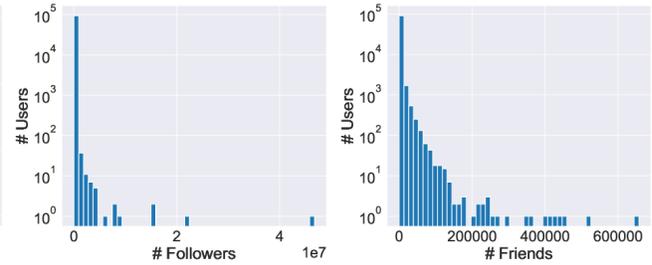


Figure 14: Follower Distribution

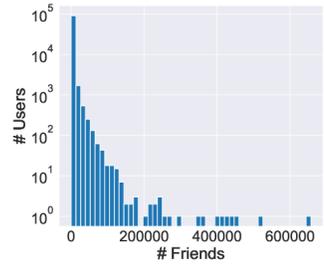


Figure 15: Friend Distribution

## ACKNOWLEDGMENTS

Emilio Ferrara is supported by the Defense Advanced Research Projects Agency (DARPA, grant number W911NF-17-C-0094).

## REFERENCES

- [1] Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. 2018. Predicting Factuality of Reporting and Bias of News Media Sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 3528–3539.
- [2] Emily Chen, Kristina Lerman, and Emilio Ferrara. 2020. Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set. *JMIR Public Health and Surveillance* 6, 2 (2020), e19273.
- [3] Limeng Cui and Dongwon Lee. 2020. CoAID: COVID-19 Healthcare Misinformation Dataset. *arXiv preprint arXiv:2006.00885* (2020).
- [4] Anyan Dai, Yiwei Sun, and Suhang Wang. 2020. Ginger Cannot Cure Cancer: Battling Fake Health News with a Comprehensive Data Repository. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 853–862.
- [5] Ensheng Dong, Hongru Du, and Lauren Gardner. 2020. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet infectious diseases* 20, 5 (2020), 533–534.
- [6] Emilio Ferrara. 2019. The history of digital spam. *Commun. ACM* 62, 8 (2019), 82–91.
- [7] Chaolin Huang, Yeming Wang, Xingwang Li, Lili Ren, Jianping Zhao, Yi Hu, Li Zhang, Guohui Fan, Jiuyang Xu, Xiaoying Gu, et al. 2020. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The lancet* 395, 10223 (2020), 497–506.
- [8] Yangfeng Ji and Jacob Eisenstein. 2014. Representation learning for text-level discourse parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, Vol. 1. 13–24.
- [9] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. 1746–1751.
- [10] Tanushree Mitra and Eric Gilbert. 2015. CREDBANK: A Large-scale Social Media Corpus with Associated Credibility Annotations. In *The Ninth International AAAI Conference on Web and Social Media*.
- [11] Maria Nicola, Zaid Alsafi, Catrin Sohrabi, Ahmed Kerwan, Ahmed Al-Jabir, Christos Iosifidis, Maliha Agha, and Riaz Agha. 2020. The socio-economic implications of the coronavirus and COVID-19 pandemic: A review. *International Journal of Surgery* (2020).

- [12] Jeppe Nørregaard, Benjamin D Horne, and Sibel Adah. 2019. NELA-GT-2018: A Large Multi-Labelled News Dataset for the Study of Misinformation in News Articles. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 13. 630–638.
- [13] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. *The development and psychometric properties of LIWC2015*. Technical Report.
- [14] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2018. FakeNewsNet: A Data Repository with News Content, Social Context and Dynamic Information for Studying Fake News on Social Media. *arXiv preprint arXiv:1809.01286* (2018).
- [15] Niraj Sitaula, Chilukuri K Mohan, Jennifer Grygiel, Xinyi Zhou, and Reza Zafarani. 2020. Credibility-based Fake News Detection. In *Disinformation, Misinformation and Fake News in Social Media: Emerging Research Challenges and Opportunities*. Springer.
- [16] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1. 809–819.
- [17] William Yang Wang. 2017. “Liar, Liar Pants on Fire”: A New Benchmark Dataset for Fake News Detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, Vol. 2. 422–426.
- [18] Bo Xu, Bernardo Gutierrez, Sumiko Mekaru, Kara Sewalk, Lauren Goodwin, Alyssa Loskill, Emily L Cohn, Yulin Hswen, Sarah C Hill, Maria M Cobo, et al. 2020. Epidemiological data from the COVID-19 outbreak, real-time case information. *Scientific Data* 7, 1 (2020), 1–6.
- [19] Reza Zafarani, Mohammad Ali Abbasi, and Huan Liu. 2014. *Social media mining: an introduction*. Cambridge University Press.
- [20] Reza Zafarani, Xinyi Zhou, Kai Shu, and Huan Liu. 2019. Fake News Research: Theories, Detection Strategies, and Open Problems. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 3207–3208.
- [21] Jiawei Zhang, Bowen Dong, and S Yu Philip. 2020. Fakedetector: Effective fake news detection with deep diffusive neural network. In *The 36th International Conference on Data Engineering*. IEEE, 1826–1829.
- [22] Xinyi Zhou, Atishay Jain, Vir V Phoha, and Reza Zafarani. 2020. Fake News Early Detection: A Theory-driven Model. *Digital Threats: Research and Practice* 1, 2 (2020), 1–25.
- [23] Xinyi Zhou, Jindi Wu, and Reza Zafarani. 2020. SAFE: Similarity-Aware Multi-Modal Fake News Detection. In *The 24th Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer.
- [24] Xinyi Zhou and Reza Zafarani. 2020. A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. *Comput. Surveys* (2020).