NELA-Local: A Dataset of U.S. Local News Articles for the Study of County-Level News Ecosystems

Benjamin D. Horne,[1] **Maurício Gruppi,**[2] **Kenneth Joseph,**[3] **Jon Green,**[4] **John P. Wihbey,**[5] **and Sibel Adalı**[2]

[1] School of Information Sciences, University of Tennessee Knoxville, Knoxville, TN, USA
[2] Computer Science, Rensselaer Polytechnic Institute, Troy, NY USA
[3] Computer Science and Engineering, University at Buffalo, Buffalo, NY USA
[4] Network Science Institute, Northeastern University, Boston, MA USA
[5] School of Journalism and Media Innovation, Northeastern University, Boston, MA USA
bhorne6@utk.edu, gouvem@rpi.edu, kjoseph@buffalo.edu, {jo.green, j.wihbey}@northeastern.edu, adalis@rpi.edu

Abstract

In this paper, we present a dataset of over 1.4M online news articles from 313 local U.S. news outlets published over 20 months (between April 4th, 2020 and December 31st, 2021). These outlets cover a geographically diverse set of communities across the United States. In order to estimate characteristics of the local audience, included with this news article data is a wide range of county-level metadata, including demographics, 2020 Presidential Election vote shares, and community resilience estimates from the U.S. Census Bureau. The `NELA-Local` dataset can be found at: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/GFE66K.

1 Introduction

Local news, news that primarily serves a specific geographic region (Abernathy 2018), is a fundamental component of the larger news ecosystem. Local news organizations can boost civic engagement, investigate wrong-doing, and inform decision making during crisis events, all at a community-specific level that national news outlets cannot fulfil (Hendrickson 2019; Gollust et al. 2017; Chauhan and Hughes 2017). This importance has been recently highlighted by the COVID-19 pandemic, in which, despite being a global event, local conditions are needed in decision making for both community members and public health experts (Gollust, Fowler, and Niederdeppe 2019; Branswell 2018).

Most existing work has focused on newsrooms adapting to the increased role of digital news consumption (Jenkins and Jeronimo 2021), and the alarming number of regions that no longer have local news outlets, often termed *news deserts* (Abernathy 2018). There is also a rich area of work focused on ownership. According to Abernathy (2016), for example, "since 2004 more than a third of the [U.S.] newspapers have changed ownership." Many of the outlets serving small to mid-sized communities have been bought and are operated by investment groups, often eroding the connection between the newspapers and local issues. With this shift in ownership comes diminished staffing and a diminished investment in operations in the name of profitability

(Abernathy 2016). Even if these cuts do not impact the sheer volume of news produced by local outlets, they can impact how labor-intensive those stories are, which in turn can reduce community knowledge of local events and the amount of substantive content produced[1]. In other words, local news produced may not actually meet the information needs of the local community.

Perhaps just as important, ownership values can not only change the volume and quality of news produced, but also change the news coverage itself. For example, there is evidence that ownership change can lead to local news coverage increasing around national politics, rather than the local area being served (Martin and McCrain 2019). Another example of this phenomenon has been termed the *Sinclair Effect* after the wide ownership of local cable news by the Sinclair Broadcast Group. One study found that outlets owned by Sinclair "produced more stories with dramatic elements, commentary, and partisan sources" (Hedding et al. 2019).

All of these studies demonstrate that local news is worthy, even crucial, for researchers to study. However, despite local media's known and well documented importance, local news article data is rare, particularly at a large scale over time. Hence, there are very few large scale studies of local news environments. For example, many of the current studies utilize journalist interviews, small, study specific collections that are not made publicly available, or large data about ownership that does not capture article content. While these data sources are certainly useful, many research questions cannot be sufficiently explored using them. For example: How do local outlets differ in the coverage of major events? What sub-topics of coverage are emphasized? How much content is shared across local outlets? What proportions of coverage are dedicated to local issues, rather than national events? To explore these types of large-scale, content-based questions, researchers need stable collections of news article data.

In this paper, we present the `NELA-Local` dataset, to fill this gap. The `NELA-Local` dataset contains over 1.4 million online articles from 313 local news outlets in the United States. Uniquely, this data contains nearly every article published by these 313 outlets between April 4th, 2020 and December 31st, 2021. Notably, this timeline covers several his-

[1]https://www.fcc.gov/general/information-needs-communities

torical events, in which local coverage of can be studied, including the 2020 U.S. Presidential Election, the January 6th U.S. capitol riots, and the COVID-19 pandemic. Furthermore, to aid studies using this dataset, we have mapped several open county-level datasets to each outlet, allowing researchers to estimate characteristics about each outlet's local audience. This county-level metadata includes demographics, political leanings, and community resilience estimates. Our hope is that this data can not only fuel novel, large-scale research on local news environments, but also benefit current lines of local media research.

2   Collection and Mapping Methods

There are two major parts of the `NELA-Local` dataset: 1. local news article data and 2. county-level metadata mapped to each outlet based on the county in which the outlet is headquartered. Below we describe the collection and mapping process for each part.

2.1   Local News Article Data

The primary, and most unique, part of our dataset is the local news article data. To collect the article data, we take the approach used in (Nørregaard, Horne, and Adalı 2019). Namely, we utilize RSS feeds from news websites to collect article URLs, and then follow those URLs to scrape full-text articles. The advantage of this method is that since data is collected as articles are published, nearly all articles published from each outlet can be collected. The challenge of this method is that it requires a predefined list of RSS feed URLs, which can be difficult to obtain without manual effort.

To this end, we start with a publicly available list of U.S. local news outlets[2]. This list comes from a social studies classroom resources website (www.50states.com). From our understanding, this list has been regularly expanded between 1996 and until at least 2018.

Using this list, we search for those that have active RSS feeds. This search is done using a mix of automatic and manual methods. First, we write a script that iterates over the website URLs and checks if common RSS feed paths exist within the website. For example, `www.examplenewswebsite.com/rss` or `www.examplenewswebsite.com/feed`. Then the script checks for paths on commonly used web feed management services like Feed Burner. Second, after the script is complete, two authors manually checked the found RSS feeds to ensure they appeared to be up-to-date and checked the websites without found RSS feeds for unusual RSS feed paths or services.

From this search, we found that 568 of the 3,300 outlets on the site had active RSS feeds. Next, one of the authors, who is an expert in local news, manually checked these outlets to ensure they were legitimate local news outlets. This manual filter removed 12 more outlets from the 568. Outlets that were removed that did not match the definition of local news (Abernathy 2016). Specifically, that the website

was a dedicated news site that primarily targets a geographically local audience. Hence, the outlets removed were either specifically for university students, explicitly national outlets, or were not news-oriented.

Using this filtered list of 556 outlets, we scrape each outlets' RSS feed twice a day, everyday between April 4th, 2020 and December 31st, 2021. Importantly, as the RSS feeds are crawled, we scrape full article text by following the URLs to the web pages that have the full article text, rather than scraping snippets of articles that may appear in the RSS feeds. This process is shown in Figure 1a.

We then perform several secondary robustness checks on the scraped data from the 556 outlets, reducing the final set of outlets to 313. First, we removed any outlets that had less than 50 articles during the time frame, indicating an outlet that publishes infrequently or does not maintain their RSS feed regularly. Second, we removed outlets that publish only in Spanish. Last, we removed sources that were later found to maintain multiple RSS feeds per news topic, but only one of the feeds were scraped (for example, one feed for Covid-19 news, another for general news). In some cases, these split-off RSS feeds were created after our data collection started. Importantly, this last step ensures that the data is of high enough quality to do analysis of topical coverage over time without artificially over-representing a topic (see Section 5 for use case discussion).

The final dataset has 313 outlets and 1,445,509 articles.

2.2   County-level Metadata

The second part of our dataset is county-level metadata. Specifically, we map each of the final 313 outlets in our dataset to the county in which it is headquartered, allowing us to utilize several open datasets. We obtain the Federal Information Processing Standard code (FIPS) for each county, and use the FIPS code to map outlets to the county-level datasets. The goal of this mapping is gain an approximate understanding of each outlets audience, similar to the process done in dataset presented in (Abernathy 2016). Note, there are fewer counties than there are outlets (255 counties and 313 outlets) as counties can have more than one outlet. Below we describe each of the county-level datasets.

Demographics   First, we map demographic data for each outlet using MIT Election Data and Science Lab's Codebook for 2018 Election Analysis Dataset[3]. Broadly, data we include from this dataset are: population, race, age, education, and economic information, including median household income and unemployment. Each of these data are described in detail in Table 5 in the Appendix.

Note, this dataset does not include Alaska. However, since we collected rich, complete news article data from three outlets in Alaska, we choose to keep them in the dataset. Instead of removing the data all together, we place 'Null' values for these counties in the demographics table (see Figure 1b) and 'None' values in the demographic columns of the CSV file (see Section 3.2).

---

[2]www.50states.com/news/

[3]https://github.com/MEDSL/2018-elections-unoffical/blob/master/election-context-2018.md

Politics   Second, we map county-level political leanings to each outlet. To do this, we link each outlet to the vote share of the last three U.S. Presidential elections: Biden vs. Trump in 2020, Clinton vs. Trump in 2016, and Obama vs. Romney in 2012. This data comes from two publicly available datasets. For the 2016 and 2012 election data, we again use MIT Election Data and Science Lab's Codebook for 2018 Election Analysis Dataset[4]. For the 2020 election data, we use Kieran Healy's 2020 Election Results data[5].

To ensure each counties vote share is comparable and standardized, we store each as the log-odds of a person in the county voting for the Republican candidate in each election. Hence, for example, higher log-odds for Biden vs. Trump in 2020 means that the county leaned towards Trump, or higher log-odds for Obama vs. Romney in 2012 means that the county leaned towards Romney.

Note, just like the demographic data, counties in Alaska are not included in these two datasets. We again place 'Null' values for these counties in the politics table (see Figure 1b) and 'None' values in the politics columns of the CSV file (see Section 3.2).

Each of these data are described in detail in Table 3 in the Appendix.

Risks   Third, to provide an assessment of *risks* per outlet audience, we map our data to the U.S. Census Bureau's *Community Resilience Estimates*[6]. According to the U.S. Census Bureau, these Community Resilience Estimates "provide an easily understood metric for how at-risk every neighborhood in the United States is to the impacts of disasters, including COVID-19 ()." More specifically, the estimates provided fall into three groups: the percentage of individuals had 0 risk factors, 1-2 risk factors, or 3+ risk factors, where a higher number of individuals in lower risk factors indicates a community that is more resilient to disasters. These estimates are determined by examining demographic, socioeconomic, and housing characteristics in the American Community Survey (ACS) microdata [7]. These characteristics included a variety of factors such as: household Income-to-Poverty Ratio (IPR), the number of caregivers per household, Unit-level crowding per household, health insurance coverage, and vehicle access.

All 313 outlets in our dataset are mapped to the risks data. Each of these data are described in detail in Table 4 in the Appendix.

3   Data Formats

In order to accommodate the widest audience possible, we provide several data formats.

---

[4]https://github.com/MEDSL/2018-elections-unoffical/blob/master/election-context-2018.md

[5]https://github.com/kjhealy/us_elections_2020_csv

[6]https://www.census.gov/programs-surveys/community-resilience-estimates.html

[7]https://www.census.gov/programs-surveys/community-resilience-estimates/technical-documentation/methodology.html

3.1   Normalized SQLite3 Database

The first format is a normalized, SQLite3 database with five tables: articles, outlets, demographics, politics, and risks. The schema for this database can be found in Figure 1b. The primary table in the database is the `articles` table, which includes the title, content, date, and URL for each article. Additionally, the `articles` table includes a identifier for the outlet of the article (called sourcedomain_id), which is the foreign key for the `outlets` table. The `outlets` table includes data about the location of the outlet, including the FIPS code. The FIPS code in the outlet table maps to all three of the county-level metadata tables: `demographics`, `politics`, and `risks`.

In addition to the database, we provide a short example Python script to use the database.

3.2   CSV

The second format we provide the dataset in is Comma-Separated Value (CSV) files. Specifically, we provide five CSV files, one for each table in the database: articles, outlets, demographics, politics, and risks. The columns in each CSV file are the same as the columns in each corresponding SQLite3 database table.

3.3   FAIR Principles

The `NELA-Local` dataset follows FAIR principles[8]. Namely, the data *Findable*, as it is persistently stored on Harvard Dataverse and is described with rich metadata. The data is *Accessible* and *Interoperable*, as it is retrievable through Harvard Dataverse's GUI and is stored in two widely-used, standard formats (SQLite3 and CSV), including sample Python scripts for extraction. Given the formats, the data can be parsed by both machine and human annotators. Additionally, this paper describes metadata and points users to other datasets that can complement and augment this dataset (see Section 6). Lastly, the data is *Re-usable*, as it is open for free use, ready to use out-of-the-box in a wide range of local news studies, and contains original article URLs and metadata links/documentation to maintain provenance.

4   Data Quality and Descriptive Statistics

To demonstrate the quality of our dataset, we provide several sets of descriptive statistics. There are two core traits that demonstrate the quality of this dataset: 1. The data comes from a geographically and demographically diverse set of counties. 2. The data contains nearly every article published by each outlet over time.

Geographical and demographic diversity   In Figure 2, we show a map of the counties in which our dataset contains at least 1 local news outlet. In total, our dataset contains outlets in 255 counties from 46 states (States not covered are: Delaware, Idaho, Maryland, and Wyoming). On average, the dataset has 1.23 outlets per county and 6.80 outlets per state (see KDE plots in Figure 4h and 4i). The max number of outlets in a single county is 11 (Middlesex County, MA) and the max number of outlets in a single state is 33

---

[8]https://www.force11.org/group/fairgroup/fairprinciples

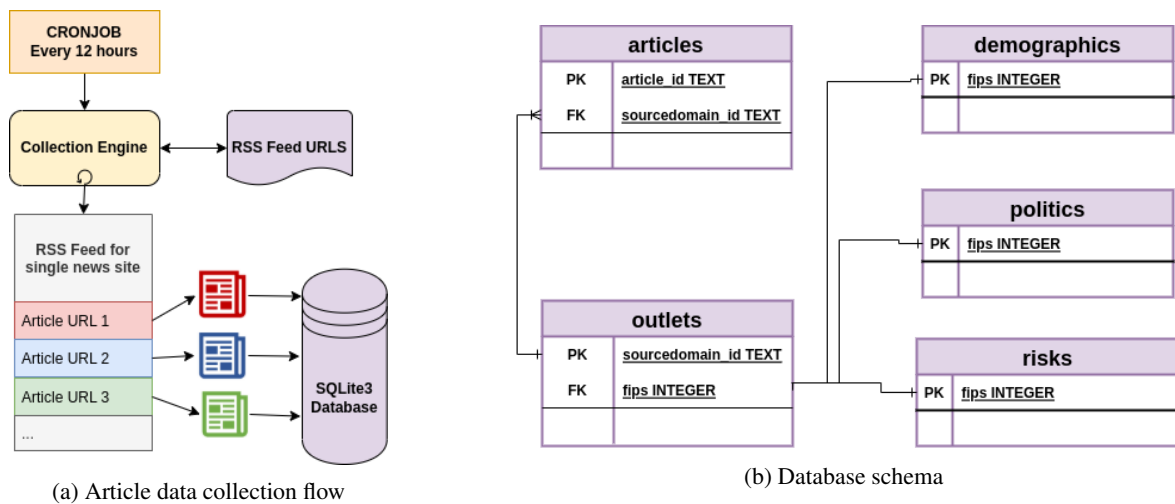(a) Article data collection flow

(b) Database schema

Figure 1: In (a) we show the flow of our news article collection. The general flow is as follows: 1. Every 12 hours start the collection engine. 2. Collection engine ingest and loops through a predetermined list of RSS feed links. 3. Each RSS feed is parsed for new article URLs, those URLs are followed to scrape the full article text from the outlets website. 4. New articles are stored in the SQLite3 database. In (b) we show the SQLite3 database schema. The main table is the `articles` table, which is mapped to a table for outlets and their location. Each outlet in the `outlet` table is then mapped to three different county-level metadata tables: `demographics`, `politics`, and `risks`
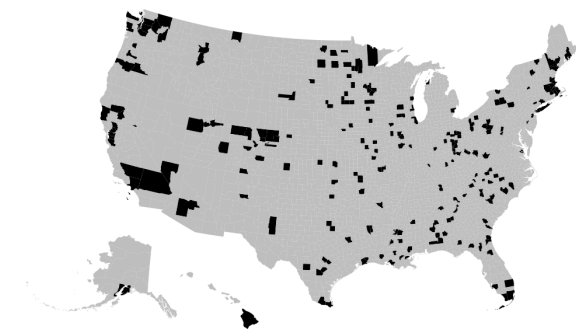
.

Figure 2: Map of counties in dataset, where black indicates an area where our dataset has at least 1 outlet.

(Massachusetts). Two notes about the skew towards Massachusetts: First, 20 of the outlets in Massachusetts are all run by the same parent company, WickedLocal, and could be combined or removed depending on the analysis being done. However, these 20 sites do produce different content. Second, Middlesex County is the most populous county in both Massachusetts and New England[9].

In addition to geography, we see diversity in various demographic factors. On average, the populations of counties in our dataset are 37.6% rural, with the minimum percent rural being 0% and the maximum being 100% (See KDE plot in Figure 4d). To add context, an example of a county with 0% rural population in our dataset is Denver County, Colorado and a county with 100% is Lewis County, Kentucky.

On average, counties contain a population of 355,998.82, with a minimum population of 4347 and maximum population of 10,057,155. Furthermore, the average median household income per county is $52,077.22, with a minimum of $28,136 and maximum of $115,224 (see KDE plot in Figure 4e). Again to add context, according to the Census Bureau, the U.S. median household income in 2018 was $63,179[10].

Similarly, the political leanings of counties in our dataset are diverse. Namely, on average, counties had a 0.25 log-odds of voting for Trump in 2020, with a minimum log-odds of -1.75 (left leaning) and a maximum log-odds of 1.83 (right leaning). We see very similar vote shares in 2016 and 2012. On average counties had a 0.29 log-odds of voting for Trump in 2016 and a 0.12 log-odds of voting for Romney in 2012. For the full distribution, see KDE plot in Figure 4a.

Hence, the `NELA-Local` dataset covers a wide range of outlet audiences, covering a diverse set of rural and urban settings. Overall, the audiences covered lean slightly right politically and have a slightly lower median household income than the U.S. as a whole.

Completeness over time In Figure 3, we show the volume of articles across time. Specifically, we show the number of articles per day (a), per week (b), and per month (c). There is one known data collection outage in May 2021 due to a cyber attack where our collection server was located[11]. However, we were able to recover most of the data missing during that time, but not all, hence the significant dip dur-

---

[9]https://data.census.gov/cedsci/profile?g=0500000US25017

[10]https://www.census.gov/library/stories/2019/09/us-median-household-income-not-significantly-different-from-2017.html

[11]https://www.news10.com/top-stories/rpi-restores-network-access-following-cyber-attack/

ing that time. To the best of our knowledge, the rest of the timeline follows the ebbs and flows of publishing patterns per outlet. Some outlets publish daily, while others publish weekly or bi-weekly.

5 Use Cases

Below we provide a set of three studies that we believe would benefit from our dataset. While these studies emphasize the large-scale temporal nature of the dataset, it is also true that the data can be used for smaller, in-depth qualitative studies (or mixed-methods approaches) as well. Nor, of course, is this discussion of use cases exhaustive.

5.1 Local Media Coverage of Disasters and Events

One critical use of the presented dataset could be to examine local media coverage of disasters and events. A long-standing literature on agenda setting and framing demonstrates that both the presentation and frequency of coverage of such events influences the importance that the public assigns to them (Entman 1993; Scheufele and Tewksbury 2007). Media coverage of disasters may also shed light on some aspects of the event, while leaving other details out of the coverage, therefore influencing what audiences believe about the event (Harbert 2010). Disasters are inherently localized, making local coverage of those disasters important in community sense-making, particularly during the uncertainty of a crisis event (Gollust, Fowler, and Niederdeppe 2019; Krafft et al. 2017).

As an example of this use of these data, Joseph et al. (2021) use an early subset of the NELA-Local dataset to examine the relationship between local news coverage of COVID-19 and local conditions. By mapping a subset of the presented dataset to county and state level COVID-19 case counts, deaths, and politics, the authors are able to provide new insights into factors associated with the degree of local COVID-19 coverage overtime and demonstrate how pandemic-related subtopics vary across local areas.

Given the major national and localized events that occurred in the U.S. during the time frame of our dataset (e.g. 2020 Presidential Election, U.S. capitol riots, local elections), the county-level metadata provided, and the fact that collection of these data continue over the present, similar studies on the relationship among coverage and local audience can be done "out-of-the-box."

5.2 Measuring the Nationalization of Local News

By pairing this dataset with national news datasets from similar timeframes (see discussion of Media Cloud and NELA-GT in Section 6), one can measure the similarity in coverage between national outlets and local outlets. This type of study could be done quantitatively, using text analysis techniques such as those in (Starbird et al. 2018) and (Horne, Nørregaard, and Adalı 2019). Or it could be done qualitatively by extracting subsets of stories from each dataset for thematic analysis. While current literature has provided evidence of local media increasingly reporting on national

events, particularly in politics, demonstrating this trend over time, topic, and outlets has not been explored.

Similarly, studies examining the impact of local media ownership on coverage can be done by mapping the NELA-Local dataset to ownership information (see discussion of the UNC database in Section 6). Finally, along this line, the extent to which different subsets of local news media mirror each other, and the potential causes of this (e.g. county-level demographics) could also be explored with the datasets presented here.

5.3 Characterizing Hybrid Local Media

This dataset can add novel contributions to the literature on hybrid media systems (Chadwick 2017), which are systems where old and new media logics, such as traditional news media and social media, are mixed together. Studies on this topic have examined the phenomenon of Twitter content being used in news articles, but these studies have been on small data sets of national news (Broersma and Graham 2013; Oschatz, Stier, and Maier 2021), not local news. Because the NELA-Local dataset contains the original URLs to all articles, embedded social media data can be scraped using a similar method to that used in (Gruppi, Horne, and Adalı 2021). If social media content is quoted or used as the source of the article, this will be captured in the article text already in the dataset.

6 Related Datasets

There are several related, but notably different, datasets to the NELA-Local dataset.

Media Cloud is an open source platform that is used for "collect[ing] data for studying the media ecosystem on the open web" (Roberts et al. 2021). The platform includes several web-based tools that operate on a stored set of media data. The focus of Media Cloud's stored data is distinctly different than NELA-Local. Namely, Media Cloud is focused on capturing global news coverage using a "combination of automated search ... [and] identified lists of influential sources.[12]" Whereas, our dataset is focused on U.S. local news outlets that serve distinct geographical regions. Hence, Media Cloud contains some overlap in outlets with our dataset, specifically those local outlets headquartered in large population centers, but does not contain outlets located in small, rural population areas.

LexisNexis is a commercial news database that has been widely used in academic studies (Deacon 2007; Weaver and Bimber 2008). It being expensive and proprietary is its main downfall. If a university has a subscription to it, stories can be downloaded through an API or the web interface. Because of lack of documentation and its commercial use, it is difficult to assess outlet overlap between LexisNexis data and our dataset. However, LexisNexis is focused on large, formal news sources, such as the Associated Press, which makes it unlikely to track small rural area news outlets.

There are event-based collections (Wang et al. 2016), such as **GDELT** (Leetaru and Schrodt 2013) and **Event Registry** (Rupnik et al. 2016). While the end-goal of these databases

---

[12]https://sources.mediacloud.org/#/home

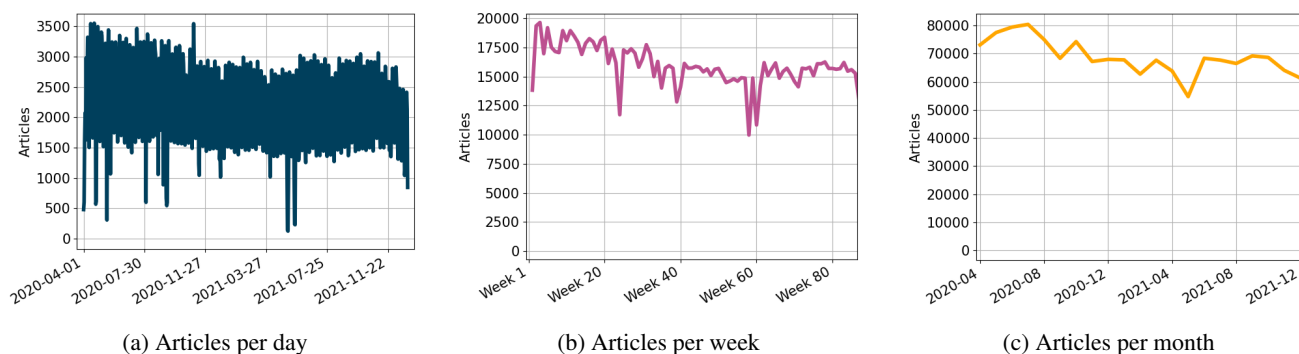| (a) Articles per day | (b) Articles per week | (c) Articles per month |

Figure 3: Articles over Time: (a) Per Day, (b) Per Week, (c) Per Month. There is one known data collection outage during May 2021 due to a cyber attack where our collection server was located, however, we were able to recover most of this missed data.

are quite different than our dataset, focused on storing events rather than news articles (news articles are used to find the events), they are occasionally used in news and media studies. In particular, GDELT stores full-text data in some cases and is open for academic use. However, GDELT has received criticism for incomplete documentation and lacking coverage of important U.S. sources (Kwak and An 2014; Weller, McCubbins, and a Glance Blog 2014). Event Registry overlaps in outlet coverage with GDELT, but is now a commercial entity (Kwak and An 2016; Roberts et al. 2021). Both event-based do not cover the U.S. Local outlets contained in `NELA-Local`.

The **NELA-GT** datasets are static, full-text news article datasets released on a yearly basis (Nørregaard, Horne, and Adalı 2019; Gruppi, Horne, and Adalı 2020, 2021). The primary goal of the NELA-GT datasets are to provide labeled data for machine learning tasks, therefore they include outlet-level veracity labels. Similar to Media Cloud, the NELA-GT datasets contain a large variety of news outlets around the globe and includes low-veracity, disinformation-peddling outlets. While similar in collection method, `NELA-Local` is not focused on providing data for veracity tasks, but rather providing data to examine local news environments. Furthermore, NELA-GT and `NELA-Local` contain mutually exclusive sets of outlets. Depending on the scope of study, we believe data from `NELA-Local`, NELA-GT, and Media Cloud can complement each other and could be used together (although carefully, as they each primarily serve different purposes).

All of these related data collections discussed above serve a broader, different purpose than the dataset presented in this paper. Our goal is to aid large-scale study of U.S. local news environments, while many of these data collections serve more general purpose, global studies of news media. **Yin Leon's Local News Dataset**[13] may be the closest related dataset in terms of scope. However, the dataset's focus is on TV stations, while ours is on online news articles. The dataset also covers a different time frame. None of the datasets mentioned, including Yin Leon's, contain county-level metadata like `NELA-Local`. However, Yin Leon's

dataset does contain ownership information.

Another dataset that is close in scope and focus to `NELA-Local` is the **UNC Database** (Abernathy 2016). The UNC Database contains information on about publication frequency, circulation statistics, and ownership of 7,927 newspapers. Like the data presented in this paper, the UNC Database contains county-level metadata such as Rural-Urban Continuum Codes. Unlike the data presented in this paper, the UNC Database does not contain article text data. The two datasets also cover different timeframes. This is another dataset that we believe could be used in conjunction with `NELA-Local` depending on the study.

7 Conclusion

In this paper, we presented a novel dataset of 1.4 million U.S. local news articles mapped to county demographics, politics, and risks data. We argued that the research community lacks large-scale, reliable local news datasets, particularly those containing full-text content for the analysis of topical coverage. By filling this gap, researchers can better understand what types of information local communities are receiving and what types of information those communities are lacking. Futhermore, we provided an extensive discussion of related datasets, some of which can be used to augment the presented dataset for a variety of studies.

The `NELA-Local` dataset, sample code, and further documentation can be found at: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/GFE66K.

---

[13]https://github.com/yinleon/LocalNewsDataset

A   Data Column Descriptions

In this appendix, we provide detailed descriptions of each data column in the `NELA-Local` dataset. Below are tables for each table in the database (articles, outlets, politics, risks, and demographics).
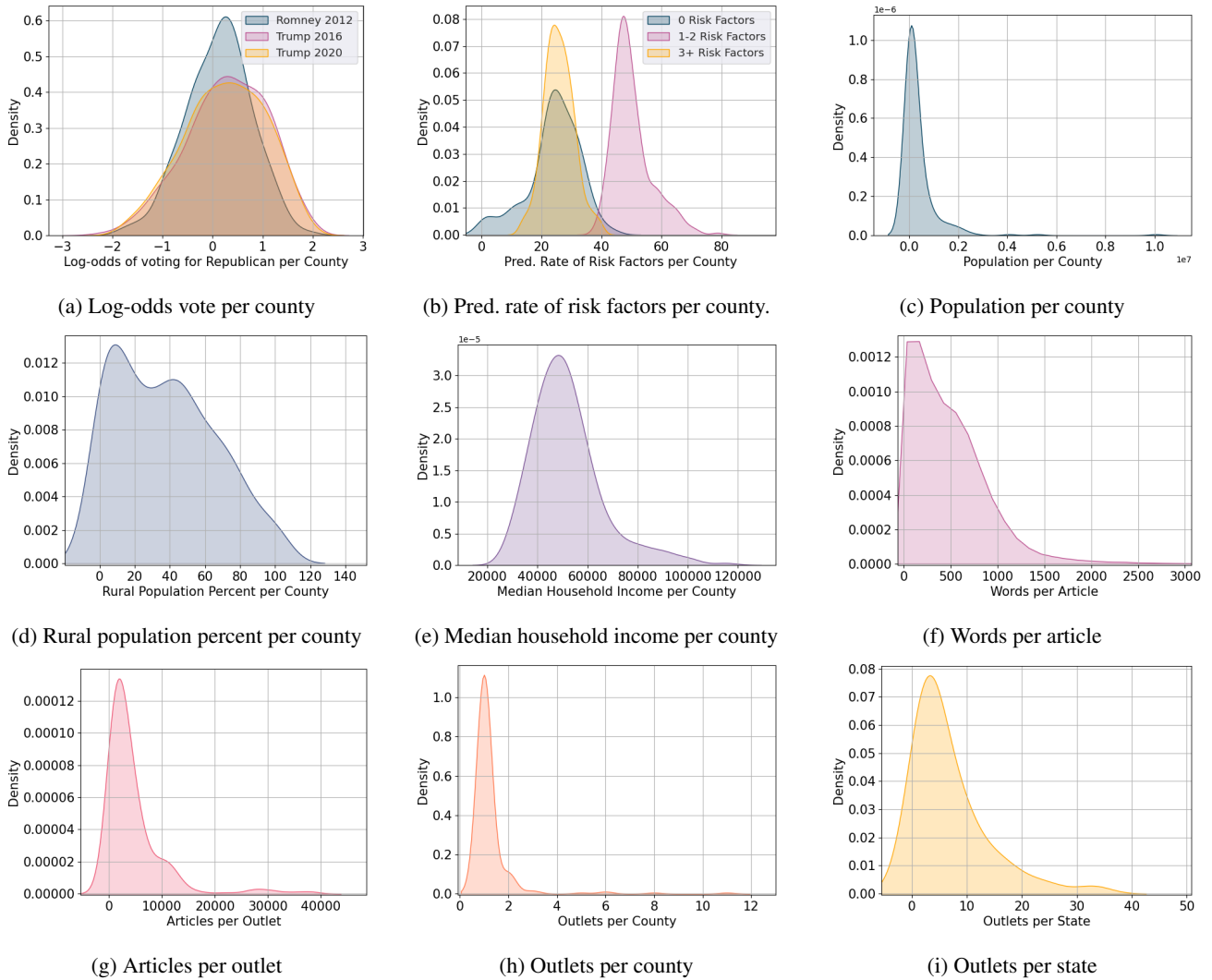

Figure 4: Kernel Density Estimate plots for (a) Log-odds of a person in a county in which a outlet is situated voting for the Republican Presidential Candidate in 2012, 2016, and 2020, (b) Estimated percentage of the population with 0, 1 to 2, or 3+ disaster/health related risks per county in the dataset, (c) Population of counties in the dataset, (d) Percent of population that is rural per county in the dataset, (e) Median household income per county in the dataset, (f) Words per article in the dataset, (g) Articles per outlet in the dataset, (h) Outlets per county in the dataset, and (i) Outlets per state in the dataset.

Column Name	Description
article_id	Unique id for article, made up of date, outlet name, and first 100 characters of the article title
sourcedomain_id	Unique outlet identifier, made up of outlet name and website domain
date	Date the article was published in format MM/DD/YYYY
title	Title of article
content	Full text content from article
url	URL of the article when the article was scraped

Table 1: Articles data description

Column Name	Description
sourcedomain_id	Unique identifier, made up of outlet name and website domain. Note, some outlets share websites if they have the same parent company.
fips	Federal Information Processing Standards code
source	Outlet name
description	Description of outlet from website, if available
onlinesince	How long the outlet has been online, if available
rank	Amazon Alexa's ranking of web domains for the outlet website, if available
state	Name of state outlet is headquartered in
city	Name of city outlet is headquartered in
lon	Longitude of county where outlet is headquartered
lat	Latitude of county where outlet is headquartered
county	Name of county outlet is headquartered in

Table 2: Outlets data description

Column Name	Description
fips	Federal Information Processing Standards code
logodds_Trump20	Log-odds of a person in the county in which the outlet is situated voting for Trump in 2020
logodds_Trump16	Log-odds of a person in the county in which the outlet is situated voting for Trump in 2016
logodds_Romney12	Log-odds of a person in the county in which the outlet is situated voting for Romney in 2012

Table 3: Politics data description. Note, the vote shares data to form the logodds comes from https://github.com/MEDSL/2018-elections-unoffical/blob/master/election-context-2018.md and https://github.com/kjhealy/us_elections_2020_csv.

Column Name	Description
fips	Federal Information Processing Standards code
predrt_0	Estimated percentage of the population with 0 disaster/health related risks
predrt_12	Estimated percentage of the population with 1 to 2 disaster/health related risks
predrt_3	Estimated percentage of the population with 3+ disaster/health related risks

Table 4: Risks data description. Note, this data comes from https://www.census.gov/programs-surveys/community-resilience-estimates.html.

Column Name	Description
fips	Federal Information Processing Standards code
total_population	Population of county, 2012-2016 ACS 5-Year Estimates
white_pct	Non-Hispanic whites as a percentage of total population
black_pct	Non-Hispanic blacks as a percentage of total population
hispanic_pct	Hispanics or Latinos as a percentage of total population
nonwhite_pct	Non-whites as a percentage of total population
foreignborn_pct	Foreign-born population as a percentage of total population
female_pct	Females as a percentage of total population
age29andunder_pct	Population 29 years or under as a percentage of total population
age65andolder_pct	Population 65 years or older as a percentage of total population
median_hh_inc	Median household income in the past 12 months (in 2016 inflation-adjusted dollars).
clf_unemploy_pct	Unemployed population in labor force as a percentage of total population in civilian labor force
lesshs_pct	Population with an education of less than a regular high school diploma as a percentage of total population
lesscollege_pct	Population with an education of less than a bachelor's degree as a percentage of total population
lesshs_whites_pct	White population with an education of less than a regular high school diploma as a percentage of total population
lesscollege_whites_pct	White population with an education of less than a bachelor's degree as a percentage of total population
rural_pct	Rural population as a percentage of total population in 2010
ruralurban_cc	Rural-urban continuum code from USDA Economic Research Service in 2013. Note, Table 6 contains the continuum code descriptions.

Table 5: Demographics data description. Note this data comes from https://github.com/MEDSL/2018-elections-unoffical/blob/master/election-context-2018.md

Column Name	Description
1	Counties in metro areas of 1 million population or more
2	Counties in metro areas of 250,000 to 1 million population
3	Counties in metro areas of fewer than 250,000 population
4	Urban population of 20,000 or more, adjacent to a metro area
5	Urban population of 20,000 or more, not adjacent to a metro area
6	Urban population of 2,500 to 19,999, adjacent to a metro area
7	Urban population of 2,500 to 19,999, not adjacent to a metro area
8	Completely rural or less than 2,500 urban population, adjacent to a metro area
9	Completely rural or less than 2,500 urban population, adjacent to a metro area

Table 6: Rural-urban continuum codes from USDA Economic Research Service in 2013 (https://www.ers.usda.gov/data-products/rural-urban-continuum-codes/) used in demographics table.

References

Abernathy, P. M. 2016. *The rise of a new media baron and the emerging threat of news deserts*. Center for Innovation and Sustainability in Local Media, University of North Carolina Chapel Hill.

Abernathy, P. M. 2018. *The expanding news desert*. Center for Innovation and Sustainability in Local Media.

Branswell, H. 2018. When Towns Lose Their Newspapers, Disease Detectives Are Left Flying Blind. *STAT*.

Broersma, M.; and Graham, T. 2013. Twitter as a news source: How Dutch and British newspapers used tweets in their news coverage, 2007–2011. *Journalism practice*, 7(4): 446–464.

Chadwick, A. 2017. *The hybrid media system: Politics and power*. Oxford University Press.

Chauhan, A.; and Hughes, A. L. 2017. Providing online crisis information: An analysis of official sources during the 2014 carlton complex wildfire. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 3151–3162.

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

Entman, R. M. 1993. Framing: Towards clarification of a fractured paradigm. *McQuail's reader in mass communication theory*, 390–397.

Gollust, S. E.; Baum, L. M.; Niederdeppe, J.; Barry, C. L.; and Fowler, E. F. 2017. Local Television News Coverage of the Affordable Care Act: Emphasizing Politics over Consumer Information. *American journal of public health*, 107(5): 687–693.

Gollust, S. E.; Fowler, E. F.; and Niederdeppe, J. 2019. Television News Coverage of Public Health Issues and Implications for Public Health Policy and Practice. *Annual Review of Public Health*, 40: 167–185.

Gruppi, M.; Horne, B. D.; and Adalı, S. 2020. NELA-GT-2019: A Large Multi-Labelled News Dataset for The Study of Misinformation in News Articles. *arXiv preprint arXiv:2003.08444*.

Gruppi, M.; Horne, B. D.; and Adalı, S. 2021. NELA-GT-2020: A Large Multi-Labelled News Dataset for The Study of Misinformation in News Articles. *arXiv preprint arXiv:2102.04567*.

Harbert, S. 2010. *Agenda setting and framing in hurricane ike news*. Gonzaga University.

Hedding, K. J.; Miller, K. C.; Abdenour, J.; and Blankenship, J. C. 2019. The Sinclair effect: Comparing ownership influences on bias in local TV news content. *Journal of Broadcasting & Electronic Media*, 63(3): 474–493.

Hendrickson, C. 2019. Local journalism in crisis: Why America must revive its local newsrooms. *Brookings Institution*. Available at https://apo.org.au/node/267591.

Horne, B. D.; Nørregaard, J.; and Adalı, S. 2019. Different spirals of sameness: A study of content sharing in mainstream and alternative media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, 257–266.

Jenkins, J.; and Jeronimo, P. 2021. Changing the Beat? Local Online Newsmaking in Finland, France, Germany, Portugal, and the UK. *Journalism Practice*, 1–18.

Joseph, K.; Horne, B. D.; Green, J.; and Wihbey, J. P. 2021. Local News Online and COVID in the US: Relationships among Coverage, Cases, Deaths, and Audience. *arXiv preprint arXiv:2111.08515*.

Krafft, P.; Zhou, K.; Edwards, I.; Starbird, K.; and Spiro, E. S. 2017. Centralized, parallel, and distributed information processing during collective sensemaking. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 2976–2987.

Kwak, H.; and An, J. 2014. Understanding news geography and major determinants of global news coverage of disasters. *arXiv preprint arXiv:1410.3710*.

Kwak, H.; and An, J. 2016. Two tales of the world: Comparison of widely used world news datasets gdelt and eventregistry. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 10.

Leetaru, K.; and Schrodt, P. A. 2013. Gdelt: Global data on events, location, and tone, 1979–2012. In *ISA annual convention*, volume 2, 1–49. Citeseer.

Martin, G. J.; and McCrain, J. 2019. Local news and national politics. *American Political Science Review*, 113(2): 372–384.

Nørregaard, J.; Horne, B. D.; and Adalı, S. 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*, volume 13, 630–638.

Oschatz, C.; Stier, S.; and Maier, J. 2021. Twitter in the News: An Analysis of Embedded Tweets in Political News Coverage. *Digital Journalism*, 1–20.

Roberts, H.; Bhargava, R.; Valiukas, L.; Jen, D.; Malik, M. M.; Bishop, C.; Ndulue, E.; Dave, A.; Clark, J.; Etling, B.; et al. 2021. Media Cloud: Massive Open Source Collection of Global News on the Open Web. *arXiv preprint arXiv:2104.03702*.

Rupnik, J.; Muhic, A.; Leban, G.; Skraba, P.; Fortuna, B.; and Grobelnik, M. 2016. News across languages-cross-lingual document similarity and event tracking. *Journal of Artificial Intelligence Research*, 55: 283–316.

Scheufele, D. A.; and Tewksbury, D. 2007. Framing, agenda setting, and priming: The evolution of three media effects models. *Journal of communication*, 57(1): 9–20.

Starbird, K.; Arif, A.; Wilson, T.; Van Koevering, K.; Yefimova, K.; and Scarnecchia, D. 2018. Ecosystem or echo-system? exploring content sharing across alternative media domains. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

Wang, W.; Kennedy, R.; Lazer, D.; and Ramakrishnan, N. 2016. Growing pains for global monitoring of societal events. *Science*, 353(6307): 1502–1503.

Weaver, D. A.; and Bimber, B. 2008. Finding news stories: a comparison of searches using LexisNexis and Google News. *Journalism & Mass Communication Quarterly*, 85(3): 515–530.

Weller, N.; McCubbins, K.; and a Glance Blog, P. V. 2014. Raining on the parade: Some cautions regarding the global database of events, language and tone dataset. *Political Violence@ a Glance Blog (20 de febrero). En: http://politicalviolenceataglance. org/2014/02/20/raining-on-the-parade-some-cautions-regarding-the-global-database-of-events-language-and-tone-dataset*.