3DLNews: A Three decade Dataset of US Local News Articles

Gangani Ariyarathne
William & Mary
Williamsburg, Virginia, USA
gchewababarand@wm.edu

Alexander C. Nwala
William & Mary
Williamsburg, Virginia, USA
acnwala@wm.edu

Abstract

We present *3DLNews*, a novel dataset with local news articles from the United States spanning the period from 1996 to 2024. It contains almost 1 million URLs (with HTML text) from over 14,000 local newspapers, TV, and radio stations across all 50 states, and provides a broad snapshot of the US local news landscape. The dataset was collected by scraping Google and Twitter search results. We employed a multi-step filtering process to remove non-news article links and enriched the dataset with metadata such as the names and geo-coordinates of the source news media organizations, article publication dates, etc. Furthermore, we demonstrated the utility of 3DLNews by outlining four applications.

CCS Concepts

• **Information systems** → **Web mining**.

Keywords

local news, US news dataset, news, news media

ACM Reference Format:
Gangani Ariyarathne and Alexander C. Nwala. 2024. 3DLNews: A Three-decade Dataset of US Local News Articles. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM '24), October 21–25, 2024, Boise, ID, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3627673.3679165

1 Introduction

With over 329 million Americans across 3,143 counties, the national media alone cannot provide news coverage for every community. Thus, local media plays a critical role by focusing on local issues, including government waste, corruption, the effectiveness of public schools, etc. In fact, more than half of original news content is produced by local media [11]. Local media was responsible for leading reports for many important stories, including exploring how the *Opioid epidemic* destroyed many lives in McDowell, West Virginia [15], chronicling the *Flint water crisis* before it received the national spotlight [16], and more recently reporting on the multi-faceted experiences of various communities experiencing the COVID-19 pandemic [18]. Given the significance of local media, local news datasets are crucial for studying the US and investigating various conditions experienced by residents of small towns and cities, surrounding issues of health, democracy, economy, etc.

CIKM '24, October 21–25, 2024, Boise, ID, USA

ACM ISBN 979-8-4007-0436-9/24/10
https://doi.org/10.1145/3627673.3679165

Existing news article datasets [9, 12, 13, 17, 19] either focus on global or national news, are paywalled, or are limited in scope; covering only specific geographical areas, timeframes, or topics. In response to this gap, we present *3DLNews*, the first-ever collection of US local news articles published from 1996 to 2024. 3DLNews was collected by scraping Google and Twitter (now X) search results. It contains about 1 million links of US local news articles from more than 14,000 websites of local newspapers, TV, and radio stations across all 50 states. The extracted URLs were carefully filtered to remove non-news article links. Furthermore, we enriched the dataset by including attributes such as, the names and geo-coordinates of the source news media, article publication dates, HTML text, etc. We published both the filtered and original dataset [2]. We demonstrated the usefulness of 3DLNews by outlining four use cases including, exploring the nationalization of local news, media bias and local news desert analyses, and community understanding.

2 Related Datasets

In contrast with existing news datasets, 3DLNews is focused on US local news, free, covers all 50 US states from 1996 to 2024, and features news articles across a wide variety of topics. Table 1 summarizes the similarities and differences between 3DLNews and the existing news datasets discussed next.

Media Cloud [17] (www.mediacloud.org), is an open-source web platform with news articles from thousands of global and national news outlets, blogs, etc. In addition to the news dataset, it offers web services such as, analytical tools for topic, media bias, network analysis, and trend tracking.

LexisNexis [19] (www.lexisnexis.com), is a commercial database with news articles, legal documents, business information, etc. Even though LexisNexis has a global focus, we do not know the quantity of US local news articles in the dataset since it is not free to access.

Nela-GT [6–8, 13], the News Ecosystem Learning Agent-Ground Truth (Nela-GT) dataset, is a labeled global news dataset for studying misinformation in news articles.

GDELT [12] (www.gdeltproject.org), the Global Database of Events, Language, and Tone (GDELT) dataset, is as a database that tracks broadcast, print, and web news in over 100 languages. It includes semantic labels for entities such as names of people, organizations, locations, events, etc. In addition to data, GDELT offers an online service for analyzing global societal trends.

NELA-Local [9] is a collection of over 1.4 million US local news articles collected from 313 local news outlets within 20 months (April 4, 2020 – December 31, 2021). Each article includes metadata associated with the community served such as, county demographics, politics, etc. Even though NELA-Local is most similar to 3DLNews since they both focus on US local news articles across multiple topics, there are significant differences. First, 3DLNews

covers a 28-year period (vs. the 20 months). Second, 3DLNews includes news articles from 14,086 news outlets (vs. 313). Third, unlike 3DLNews, NELA-Local is a longitudinal dataset which was created by extracting news articles daily from RSS feeds.

Table 1: Comparison of existing news datasets and 3DLNews.

Dataset	Time range	# Articles	Free/Paid
Media Cloud [17]	2008 – Present	~1.7 billion	Free
Lexis Nexis [19]	1980 – Present	83 billion	Paid
Nela-GT [13]	2018 – 2022	~7 million	Free
GDELT [12]	1979 – Present	~6 million	Free
NELA-Local [9]	2020 – 2021	1.4 million	Free
3DLNews [2]	**1996 – 2024**	**~1 million**	**Free**

3 Building 3DLNews

Here we explain our steps for creating 3DLNews.

3.1 Local news media dataset

We used an extended version of the Local Memory Project's [14] (LMP) US local news dataset as seeds for our data extraction. LMP's dataset consists of the websites of 5,993 local newspapers, 2,539 TV stations, and 1,061 radio stations, primarily extracted from thepaperboy.com in 2016. We extended it by crawling and scraping thepaperboy.com (again), web.archive.org/web/20221203031956/http://www.usnpl.com/, 50states.com (similar to Nela-Local [9]), and einpresswire.com/world-media-directory/3/united-states. This resulted in the publicly released local news dataset [2] outlined in Table 2. The "broadcast" type refers to either TV or radio stations, because we could not accurately distinguish them during scraping.

3.2 Data extraction and filtering

Step 1: We created Google search queries for each website in the local news media dataset (Table 2). For a single media website, e.g., timesstar.com, we constructed the following Google query:

```
https://www.google.com/search?tbs=cdr:1,
cd_min:1/1/1996,cd_max:12/31/1996
&q=news 20site:http://www.timesstar.com/.
```

This instructs Google to return "news" webpages exclusively from timesstar.com and published in 1996. We changed the Google date directives (`cd min`/`cd min`) to retrieve webpages published in a different year. Next, similar to Google, for each website, e.g., timesstar.com in the local news media dataset, we created Twitter search queries as follows:

```
'timesstar.com' until:2006-12-31 since:2006-01-01
```

This instructs Twitter to return tweets posted in 2006 and linked to timesstar.com. We changed the Twitter date directives (`until`/`since`) to retrieve webpages published in a different year.

The goal of the 3DLNews dataset is to provide a representative sample of US local news stories from 1996 to 2024, rather than capturing every published article. Therefore, our Google scraping focuses solely on the first page of results, and our Twitter scraping includes only the top 20 tweets per query.

Table 2: US local news media dataset.

Media Type	Number of websites
Newspapers	9,441
Radio	2,449
Broadcast	1,310
TV	886
Total	**14,086**

Table 3: 3DLNews: Counts of URLs collected (All URLs) and News article URLs after filtering applied (News URLs), and the counts of HTML files downloaded for All URLs (All HTML) and News URLs (News HTML).

Media	All URLs	All HTML	News URLs	News HTML
		Google		
Newspaper	853,543	636,967	502,530	436,031
Radio	140,401	113,383	52,925	52,117
TV	99,001	88,620	62,727	59,609
Broadcast	164,028	155,445	110,494	107,439
		Twitter		
Newspaper	199,996	155,083	54,295	30,981
Radio	102,494	41,917	3,794	2,637
TV	66,880	54,632	13,213	5,895
Broadcast	100,119	44,909	10,497	7,921
Total	**1,727,462**	**1,290,956**	**810,475**	**702,630**

Step 2: We issued Google and Twitter search queries to their respective search engines and scraped their links. For Google, we created queries from 1996 – 2024, for Twitter, 2006 – 2024. Table 3 presents the number of links scraped from Google and Twitter for each media type. It includes both news and non-news article links (e.g., homepages, menu pages) and corresponding counts for successfully downloaded HTML files. We removed non-news links by applying a filtering process outlined in Step 3. Table 3 presents the number of news articles (and HTML files) after filtering.

Next, we outline our filtering process for removing non-news article URLs from 3DLNews (Table 3), since there is no standard URL format for news articles. We provided access to the raw data, enabling researchers to apply their own filtering. This process was informed by an experiment in which we created a gold-standard dataset of news article URLs to understand two properties: path depth and word-boundary. The path depth of a URL is the number of hierarchies in the URL path property. For example, https://example.com/ has a path depth of zero while https://example.com/foo and https://example.com/foo/bar have path depths of one and two, respectively. A word-boundary is simply a symbol that separates words in a URL. For example, the word-boundary for the URL, https://example.com/this-is-a-page is '-'.

Step 3: First, we dereferenced all URLs to resolve redirects and retrieved final URLs that returned HTTP 200 codes. Second, we removed links with domains not present in our local news media

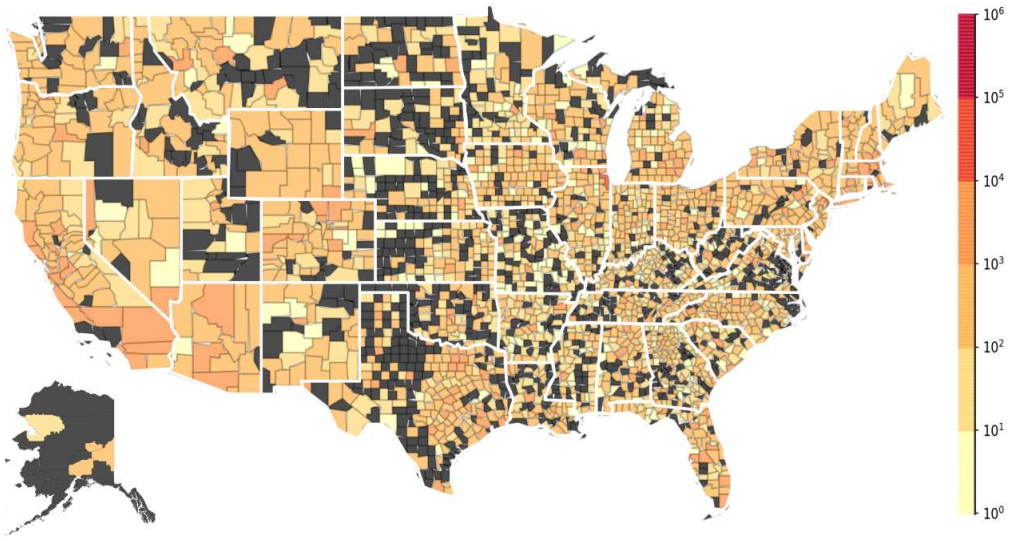

Figure 1: Local news articles in 3DLNews per US county. Black-colored counties indicate areas without news articles in 3DLNews.

dataset. Third, we lowercased all URLs, discarded trailing slashes, and removed duplicate URLs. Fourth, URLs with a path depth of zero, typically representing homepages, were removed since we only care about news article URLs which occur at a deeper path depth (e.g., 3). On some occasions however, we observed that news URLs occurred at lower path depths (e.g., 3). Therefore, we kept such news article URLs only if they included popular word-boundary separators such as '-', ' ', or '.' (e.g., http://kwgs.org/post/funeral-set-ou-quarterback-killed-crash). We kept all URLs with path depth 3. Our 3DLNews dataset [2] includes both raw and filtered versions with HTML text whenever available (Table 3).

3.3 Data Enrichment and format

We enhanced the usefulness of the news article URLs in 3DLNews by adding attributes to each URL. See Table 4 for the complete list of attributes. Next, we highlight a few.

link represents the news article URL. The *html filename* attribute points to the file containing the HTML text of the news article, while the *publication date* refers to the article publication date which was extracted using htmldate [3]. The *location* property of each URL includes the US state, city, and latitude/longitude of the source news media organization. *media metadata* contains information (e.g., newspaper or TV or radio name) about the news media website where the article was published. *source metadata* includes information (e.g., search query link) about the source (Twitter or Google) from which the article was scraped.

Each news article URL, along with its attributes (Table 4) is encapsulated in a JSON object within a single line in a file in 3DLNews.

4 Data Coverage and Descriptive Statistics

Here we summarize the composition of 3DLNews along the geographic and temporal dimensions.

4.1 Location Based Analysis

Figure 1 is a choropleth map illustrating the distribution of the counts of news articles per US county. Accordingly, 3DLNews covers 100% (50/50) of US states and about 68% (2,146/3,143) of all US counties. Cook county, Illinois, the second-most-populous county in the US, had the most news articles (13,006). In contrast, there were 59 counties with a single news article. The wide coverage of 3DLNews allows for robust analysis of regional news trends, community-specific issues, and the overall US local news ecosystem. Figure 1 also reveals the presence of news deserts (black-colored counties); areas with no local news coverage. This could be attributed to the absence of local news media organizations in these counties or blind spots in 3DLNews. Therefore, further research is needed to determine the actual cause.

Table 4: Properties of news article URLs in 3DLNews.

Property	Description
link	The URL of the local news article.
html filename	Filename with HTML content of the article.
publication date	Article publication date.
title	Title of the article.
media name	Name of local media organization.
media type	Type of media source (*Newspaper* or *TV* or *Radio station* or *Broadcast*). "Broadcast" refers to either TV or radio stations.
location	Location of the media organization. This includes: US state, city, & latitude/longitude.
media metadata	More information about the news media.
source	Platform (Google or Twitter) where the news article was extracted from.
source metadata	More information about the platform.
response code	Response code from issuing GET on link.
expanded url	Final target URL for links that redirect.

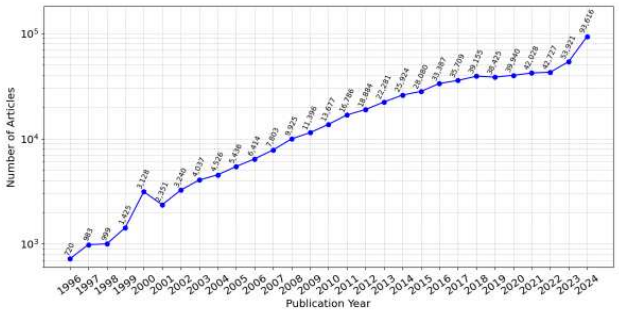

Figure 2: Counts of news articles in 3DLNews per year.

4.2 Time Based Analysis

Figure 2 depicts the distribution of the counts of local news articles from 1996 to 2024. Overall, the number of news articles gradually increases over time. This upward trend highlights the growing volume of online local news content over the years. In the earlier years of our dataset, fewer articles were available, which is indicative of the developing stage of digital news and the limited web presence of local news outlets during that period.

5 Use Cases

Here we demonstrate the utility of 3DLNews by outlining four possible applications that we intend to implement in future research.

5.1 Exploring the Nationalization of Local News

Local news outlets are meant to prioritize local stories that are relevant to their communities. This includes coverage of local politics, economy, culture, etc. A concerning contributor to the decline of local media in the US has been a steady increase in the nationalization of local news [4, 10] — the prioritization of national news (especially politics) over local news. 3DLNews presents a unique opportunity to quantify the degree of nationalization of local news across the US. This would involve using various natural language processing techniques to compare the topics of the news articles in 3DLNews with national news articles published in the same period.

5.2 Media Bias Analysis

Similar to the problem of the nationalization of local news, media bias in local news is another concerning issue that warrants further investigation, especially since bias could undermine the trust that local media enjoys over national media [5]. 3DLNews provides a comprehensive snapshot of the US local media landscape for studying bias in news coverage.

5.3 Studying US Local News Deserts

For the last two decades, thousands of local news stations have closed resulting in "news deserts" — communities without a local news outlet — spread across the US. In fact, since 2005, the US has lost almost 2,900 local newspapers and they continue to vanish at an average rate of more than two a week [1]. 3DLNews may be used to study this phenomenon through the analyses of the density of local news articles relative to various geographic regions.

5.4 Community Understanding

Given the broad geographical scope of 3DLNews, through content analysis, researchers can study the US through the lens of local media to gain deeper insights into living conditions in various communities or community attitudes surrounding various political, health, or economic issues.

The three-decade span of 3DLNews also provides a broad temporal scope for researchers to examine trends in local news coverage, to identify emerging topics, monitor changing public concerns, and to anticipate where the news agendas may be heading in the future. The insights gained from trend analysis and prediction have practical applications for journalists and possibly policymakers.

6 Discussion

Despite the broad geographic and temporal scope, and potential applications of 3DLNews, it has some limitations.

First, while our filtering attempts to minimize the likelihood of including non-news article URLs, since there is no standard URL format of news articles, we expect to have included some small proportion of non-news article URLs in 3DLNews. To address this and other limitations of 3DLNews, we provided access to the raw data (URLs, HTML text, etc), enabling researchers to apply their own filtering and/or research-specific analyses.

Second, 3DLNews excludes archived URLs (if they exist) for unavailable news articles. We plan to address this issue in future updates. Third, we were constrained by web scraping which limited the number of URLs we could collect.

Fourth, it is likely that 3DLNews includes articles from closed news organizations, since many news organizations have shut down between 1996 and 2024. Also, it is possible that 3DLNews does not include articles from closed news organizations whose information have been purged from search engine indexes. In a future research effort, we will quantify the proportion of articles for either cases and assess their impact. Furthermore, we intend to utilize the Internet Archive to understand how well local news articles are preserved.

Fifth, in our location-based analysis of news articles, we relied on the locations of the news media organizations, which might not always reflect the actual geographic areas covered by news stories. To remedy this, for future work, we will identify the specific geographic regions referenced in news stories.

7 Conclusion

We presented 3DLNews, a novel dataset with local news articles from the United States spanning the period from 1996 to 2024. 3DLNews is a significant contribution in the study of US local news coverage due to its broad geographical (covering 50 US states) and temporal scope (three-decade span). The dataset contains nearly 1 million URLs, enriched HTML text, and additional metadata for over 14,000 local news outlets across all 50 states and 68% of US counties. We believe that 3DLNews provides a valuable opportunity for researchers to study the US and/or US local media ecosystem.

Acknowledgments

We thank the NSF for funding this work (award no. 2245508). The NSF had no role in designing our study, data collection and analysis, decision to publish, or preparation of the manuscript.

References

[1] Penelope M. Abernathy. 2023. The State of Local News. https://localnewsinitiative.northwestern.edu/projects/state-of-local-news/2023/report/. Accessed: 2024-06-01.

[2] Gangani Ariyarathne and Alexander C. Nwala. 2024. A Three-decade Dataset of US Local News Articles. https://github.com/wm-newslab/3DLNews. Accessed: 2024-06-01.

[3] Adrien Barbaresi. 2020. htmldate: A Python package to extract publication dates from web pages. *Journal of Open Source Software* 5, 51 (2020), 2439. https://doi.org/10.21105/joss.02439

[4] Aisha Bradshaw. 2019. The nationalization of news. *Nature Human Behaviour* 3, 5 (2019), 421–421.

[5] Knight Foundation. 2023. American Views 2022: Part 2, Trust Media and Democracy. https://knightfoundation.org/reports/american-views-2023-part-2/. Accessed: 2024-06-01.

[6] Maurício Gruppi, Benjamin D Horne, and Sibel Adalı. 2020. *NELA-GT-2019: A Large Multi-Labelled News Dataset for The Study of Misinformation in News Articles.* Preprint 2003.08444. arXiv.

[7] Maurício Gruppi, Benjamin D Horne, and Sibel Adalı. 2021. *NELA-GT-2020: A large multi-labelled news dataset for the study of misinformation in news articles.* Preprint 2102.04567. arXiv.

[8] Maurício Gruppi, Benjamin D Horne, and Sibel Adalı. 2023. *NELA-GT-2022: A Large Multi-Labelled News Dataset for The Study of Misinformation in News Articles.* Preprint 2203.05659. arXiv.

[9] Benjamin D Horne, Maurício Gruppi, Kenneth Joseph, Jon Green, John P Wihbey, and Sibel Adalı. 2022. NELA-Local: A dataset of US local news articles for the study of county-level news ecosystems. In *Proceedings of the international AAAI conference on web and social media*, Vol. 16. 1275–1284.

[10] Kokil Jaidka, Sean Fischer, Yphtach Lelkes, and Yifei Wang. 2023. News nationalization in a digital age: An examination of how local protests are covered and curated online. *The ANNALS of the American Academy of Political and Social Science* 707, 1 (2023), 189–207.

[11] Mary E. Klas. 2019. Less Local News Means Less Democracy. https://niemanreports.org/articles/less-local-news-means-less-democracy/. Accessed: 2024-06-01.

[12] Kalev Leetaru and Philip A Schrodt. 2013. Gdelt: Global data on events, location, and tone, 1979–2012. In *ISA annual convention*, Vol. 2. Citeseer, 1–49.

[13] Jeppe Nørregaard, Benjamin D Horne, and Sibel Adalı. 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.

[14] Alexander C. Nwala, Michele C. Weigle, Adam B. Ziegler, Anastasia Aizman, and Michael L. Nelson. 2017. Local memory project: providing tools to build collections of stories for local events from local sources. In *Proceedings of ACM/IEEE Joint Conference on Digital Libraries (JCDL)* (Toronto, Ontario, Canada) *(JCDL '17)*. ACM, New York, NY, USA, 219––228. https://doi.org/10.1109/JCDL.2017.7991576

[15] Samantha Perry. 2016. Dangerous fentanyl, heroin becoming greater risks of fatal overdose across southern W.Va. https://archive.ph/QYSmN. Accessed: 2024-06-01.

[16] Denise Robbins. 2016. ANALYSIS: How Michigan And National Reporters Covered The Flint Water Crisis. https://mediamatters.org/research/2016/02/02/analysis-how-michigan-and-national-reporters-co/208290. Accessed: 2024-06-01.

[17] Hal Roberts, Rahul Bhargava, Linas Valiukas, Dennis Jen, Momin M Malik, Cindy Sherman Bishop, Emily B Ndulue, Aashka Dave, Justin Clark, Bruce Etling, et al. 2021. Media cloud: Massive open source collection of global news on the open web. In *Proceedings of the International AAAI Conference on Web and Social Media*. 1034–1045.

[18] Elisa Shearer. 2020. Local news is playing an important role for Americans during COVID-19 outbreak. https://www.pewresearch.org/short-reads/2020/07/02/local-news-is-playing-an-important-role-for-americans-during-covid-19-outbreak/. Accessed: 2024-06-01.

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