data

MDPI

Article

Web Scraping Chilean News Media: A Dataset for Analyzing Social Unrest Coverage (2019–2023)

Ignacio Molina [1], José Morales [2] and Brian Keith [1,*]

[1] Department of Systems and Computing Engineering, Universidad Católica del Norte, Antofagasta 1270398, Chile; ignacio.molina@alumnos.ucn.cl
[2] School of Journalism, Universidad Católica del Norte, Antofagasta 1270398, Chile; jmorales03@ucn.cl
* Correspondence: brian.keith@ucn.cl

Abstract

This paper presents a dataset of Chilean news media coverage during the social unrest and constitutional processes from 2019 to 2023. Using Python-based web scraping with BeautifulSoup and Selenium, we collected articles from 15 Chilean news outlets between 15 November 2019 and 17 December 2023. The initial collection of 1254 articles was filtered to 931 usable data points after removing non-relevant content, duplicates, and articles unrelated to the Chilean social outburst. Each news outlet required specific extraction approaches due to varying HTML structures, with some outlets inaccessible due to paywalls or anti-scraping mechanisms. The dataset is structured in JSON format with standardized fields including title, content, date, author, and source metadata. This resource supports research on media coverage during political events and provides data for Spanish-language processing tasks. The dataset and extraction code are publicly available on GitHub.

Keywords: web scraping; Chilean social outburst; news media dataset; data collection; estallido social

Academic Editor: Davide Martinenghi

Received: 10 September 2025
Revised: 28 October 2025
Accepted: 30 October 2025
Published: 31 October 2025

Citation: Molina, I.; Morales, J.; Keith, B. Web Scraping Chilean News Media: A Dataset for Analyzing Social Unrest Coverage (2019–2023). *Data* **2025**, *10*, 174. https://doi.org/10.3390/data10110174

1. Introduction

The Chilean *social outburst* beginning on 18 October 2019, known as the *estallido social*, triggered a period of intense unrest and political activity that lasted through multiple constitutional processes until 2023 [1,2]. Pleyers [3] analyzes this uprising within the global wave of social movements, while Brannen et al. [4] document an 11.5% annual increase in global protests from 2009 to 2019, positioning Chile within this trend.

This unrest period generated extensive media coverage across Chilean outlets, creating an opportunity to study information dynamics during political transformation [5]. The events spanned from the initial protests in October 2019 through the *Agreement for Peace and New Constitution* in November 2019, the 2020 constitutional plebiscite, the *Constitutional Convention* period (2021–2022), and subsequent referendum processes through 2023 [6].

The rapid spread of information—and misinformation—through digital channels during this period highlighted the need for datasets that capture the response of the media landscape to these events [7]. Although social media platforms have been extensively studied for their role in political mobilization [8–10], traditional news media remains an important source of authoritative information that shapes public discourse and influences policy debates [11–13].

This dataset was developed as part of ongoing research investigating news narratives with the purpose of exploring misinformation patterns. In general, our aim is to understand

how narratives unfold in different news media. In particular, we decided to focus on the Chilean constitutional processes, as they possess interesting qualities due to their highly politicized nature and the fact that there were two processes: one led by the left-wing politicians followed by another one led by right-wing politicians, with both failing at their respective plebiscites. We note that while our previous work focused on multimodal data from fact-checking platforms [14], this dataset focuses on traditional news outlets.

This paper presents a data resource that documents the construction of a standardized dataset from heterogeneous Chilean media sources. While the primary contribution is the dataset itself, the documentation of platform-specific extraction solutions provides methodological value for future web-scraping efforts in regional media contexts where sources employ diverse technical architectures and access restrictions. This work addresses a gap in the available resources for research on Chilean Spanish media coverage and political events.

There are several challenges that must be overcome for data collection in the Chilean media landscape. In particular, major outlets employ different content management systems, varying levels of access restrictions, and different technical architectures. We provide a standardized dataset from multiple news sources covering the period of social unrest and constitutional debates. However, we note that not all media outlets were accessible.

In more detail, this article documents the web-scraping methodology [15] used to collect 1254 news articles from 15 Chilean outlets. Following the initial collection procedure, the filtering process that yielded 931 validated articles, and the technical solutions developed for platform-specific challenges. We present the structure of the dataset and extraction results and make both the data and the code publicly available for research use.

2. Background and Related Work

Web scraping has become a relevant technique for social science research, enabling the systematic collection of digital trace data at scale [15]. Previous work has established frameworks for scraping news media [16], with recent advances in extraction accuracy. Ref. [17] achieved 97.69% F1-score using newspaper-specific parsers, while ref. [18] provides legal and ethical frameworks for web-scraping research. Beyond these cases, there have been applications of web scraping on social media platforms [19], and government websites [20]. However, these frameworks often focus on well-structured and English-language content, leaving gaps in methodologies for regional media ecosystems with diverse technical implementations.

Several notable datasets have been created to study political events and social movements. The GDELT Project [21] provides global event data extracted from news sources, while ACLED [22] focuses on conflict and protest events. However, news media datasets for Chile remain relatively scarce.

In the Chilean context, our previous work [14] introduced a fact-checking dataset from Chilean platforms, establishing a foundation for multimodal analysis of misinformation during constitutional processes.

In general, the development of non-English language processing resources has lagged behind English, particularly for regional variants [23]. The Chilean Spanish used in news media presents unique challenges, including local idioms, political terminology, and cultural references that standard NLP tools may not capture effectively [24].

Thus, for natural language processing research, the dataset provides a substantial corpus of Chilean Spanish text from professional sources. This addresses the scarcity noted by Cañete et al. [25,26] for regional Spanish-language resources. Recent advances like Gutierrez et al.'s MarIA models [27] and Pérez et al.'s `pysentimiento` toolkit could be directly applied to this dataset [28].

Furthermore, research on misinformation during political crises has highlighted the importance of the availability of media datasets [29]. Studies have shown that traditional media plays a central role in either amplifying or debunking false narratives [30]. The Chilean social outburst provides a unique case study, as it involved both grassroots mobilization and institutional responses, creating complex information dynamics [1].

3. Data Collection Methodology

3.1. Source Selection and Temporal Scope

The data collection targeted 15 major Chilean news outlets selected to represent diverse editorial perspectives and audience segments across the political spectrum. The selection included traditional newspapers (*El Mercurio*, *La Tercera*, *La Segunda*), tabloid publications (*La Cuarta*, *Las Últimas Noticias*), digital-native platforms (*EMOL*, *El Mostrador*, *Ex-Ante*), investigative journalism outlets (*Ciper Chile*, *The Clinic*), television news websites (*CNN Chile*), weekly magazines (*HoyxHoy*), international coverage (*El País* Chile edition), business press (*Diario Financiero*), and alternative media (*Gamba*).

This diversity captures multiple narrative frames [31] and editorial approaches to the social outburst period. However, the Chilean media landscape operates within a significant concentration of ownership [32]. *La Cuarta* and *Las Últimas Noticias*, historically characterized as sensationalist media targeting middle and lower socioeconomic strata, belong to major conglomerates: *La Cuarta* to Copesa (which also owns *La Tercera*) and *Las Últimas Noticias* to El Mercurio S.A.P. Studies have shown that these reference newspapers tend towards editorial homogeneity [33]. However, we note that regional and digital transformations continue to reshape journalistic practices [34,35]. Despite ownership concentration, these outlets maintain distinct editorial voices and target audiences that contribute to the representational breadth of the dataset.

The temporal scope spans from 15 November 2019 to 17 December 2023. The start date, roughly one month after the initial unrest, corresponds to the intensification of protests following the 18 October—the start of the *estallido social* itself—while the end date encompasses the conclusion of the second constitutional process attempt. This time frame was selected to analyze the information flow during the entire cycle of political mobilization and institutional response.

3.2. Web-Scraping Architecture

The extraction system was built using Python 3.13 with `BeautifulSoup` for HTML parsing [36] and `Selenium` WebDriver for dynamic content [37]. This approach follows best practices established by Donig et al. [38] for handling heterogeneous website structures with JavaScript-rendered content.

Each news outlet required individual treatment due to unique website structures, resulting in 15 separate Jupyter notebooks containing outlet-specific extraction logic. This approach allowed for iterative development and testing of the different data collection processes. The notebooks are stored in the repository's `src/extract_code/` directory, with each filename corresponding to its target outlet (e.g., `la_tercera.ipynb`, `emol.ipynb`).

The general extraction pipeline consisted of four stages. First, the URLs of the articles were discovered through search interfaces, archive pages, or sitemap crawling, depending on the outlet structure. Second, HTML content was retrieved using direct HTTP requests for static pages or `Selenium` for JavaScript-rendered content. Third, `BeautifulSoup` parsed the HTML to extract article components using outlet-specific CSS selectors. Finally, extracted data were validated for completeness and stored in JSON format with standardized field names across all sources.

For outlets with dynamic content loading, we employed `Selenium` with *undetected-chromedriver* to avoid detection as automated traffic. Rate limiting was implemented with configurable delays between requests (minimum 1 s) to respect server resources. The system handled various pagination methods, including page-based navigation, infinite scroll, and date-range queries. The error handling included retry logic for network failures and recording of extraction failures for manual review. Figure 1 shows the general content extraction method with `Selenium` and Figure 2 shows a specific example of how this code is applied with a single outlet.

```
def extract_data_from_page(
    get_author: Callable[[WebElement], str],
    get_description: Callable[[WebElement], str],
    get_content: Callable[[WebElement], str],
    input_file, output_file,
    limit_of_pages: int = 1):

    extracted_pages = {"pages": []}
    extracted_pages_with_content = {"pages": []}

    with open(input_file, 'r', encoding='utf-8') as file:
        extracted_pages = json.load(file)

    for page_index, page in enumerate(extracted_pages["pages"]):
        if page_index >= limit_of_pages:
            break
        try:
            options = uc.ChromeOptions()
            options.add_argument("--no-sandbox")
            options.add_argument("--disable-blink-features=
                AutomationControlled")
            driver = uc.Chrome(options=options)
            driver.set_page_load_timeout(20)
            driver.get(page['link'])
            body=WebDriverWait(driver, 20).until(
                EC.presence_of_element_located((By.TAG_NAME, 'body'))
            )

            author = get_author(body)
            description = get_description(body)
            content = get_content(body)

            page["content"] = content
            page["author"] = author
            page["description"] = description
            extracted_pages_with_content["pages"].append(page)

            with open(output_file, 'w', encoding='utf-8') as file:
                json.dump(extracted_pages_with_content, file,
                    ensure_ascii=False, indent=4)

            driver.quit()

        except Exception as e:
            print(f"[ERROR] An error occurred while processing page {
                page_index}: {e}")
```

Figure 1. Simplified content extraction code using `Selenium`.

```python
def get_content(body: WebElement):
    content = ""
    for element in body.find_elements(By.CLASS_NAME, "article-body__paragraph"):
        content += element.text.strip() + "\n"
    return content

def get_author(body: WebElement):
    try:
        return body.find_element(By.CLASS_NAME, "article-body__byline__author").text.strip()
    except:
        return "not found"

def get_description(body: WebElement):
    try:
        return body.find_element(By.CLASS_NAME, "article-head__subtitle").text.strip()
    except:
        return "not found"

extract_data_from_page(
    input_file="test_input.json",
    output_file="test_output.json",
    get_author=get_author,
    get_description=get_description,
    get_content=get_content
)
```

Figure 2. Example code usage of the extraction code using the webpage of *La Tercera*.

3.3. Comparison with Other Extraction Methods

Our extraction methodology draws insights from established efforts in automated protest event data collection, while adapting to the specific challenges of the Chilean media landscape. The MPEDS framework [39] demonstrated automated generation of protest event data from news sources, achieving 85% precision in event detection. Similarly, Dorff et al. [40] compared multiple protest datasets, identifying systematic biases in news-based extraction that we addressed through outlet diversification.

The Chilean context presents unique challenges compared to English-language extraction systems. Although GDELT [21] processes global news in multiple languages, its generic approach achieves a lower accuracy for regional Spanish variants and often misses local coverage. Although frameworks like Fundus [17] have demonstrated that newspaper-specific parsers can achieve 97.69% F1-score. However, such frameworks have limited coverage of Latin American media outlets.

Table 1 compares our dataset with existing news and event datasets, emphasizing structural differences in data collection and validation approaches. GDELT [21] provides event-level metadata extracted from global news but does not preserve the full text of the article, limiting its utility for narrative and discourse analysis.

ACLED [22] manually codes conflict events into structured formats (actor, action, location, date), but similarly lacks the full content of the article necessary for studying media framing or linguistic patterns. Fundus [17] achieves high extraction accuracy (97.69% F1-score) using manually crafted extractors for 39 publishers, but focuses exclusively on US and European media with no documented coverage of Latin American outlets—a gap documented in regional NLP research [23,24]. Our dataset contributes validated full-text Chilean articles with documented extraction quality, addressing the scarcity of Spanish-language news corpora from regional media sources. While our 74.2% validation rate reflects conservative manual filtering (removing non-articles, off-topic content, and extraction failures), the resulting corpus provides reliable full-text data for Chilean discourse analysis during the constitutional period.

Table 1. Comparison with existing datasets: methodological and coverage characteristics.

Dataset	Spatiotemporal Scope	Data Points	Granularity	Validation	Chilean Coverage	Ref.
GDELT	Global, 1979–2012	Millions of events	Event tuples (no full text)	Automated extraction, no reported validation	Generic (includes Chile in global aggregate)	[21]
ACLED	Global conflicts, 1997–2010	Event records	Coded events (actors, actions, locations)	Expert manual coding + automated assistance	Includes LATAM conflicts, not media-focused	[22]
Fundus	39 publishers (at publication), primarily US/EU	Full articles	Complete article text + structured metadata	97.69% F1 on gold standard (automated)	No LATAM coverage documented	[17]
Ours	10 Chilean outlets, November 2019–December 2023	931 articles (1254 collected)	Full text + metadata (title, date, author, source)	74.2% validation via manual review (relevance + completeness)	High (Chile-specific, constitutional process focus)	N/A

4. Technical Implementation and Challenges

4.1. Platform-Specific Obstacles

Table 2 summarizes the extraction results and obstacles encountered for each outlet. Of the 15 targeted sources, five were completely inaccessible, two provided only image-based content, and eight yielded extractable text data.

Table 2. Data extraction results and obstacles by news outlet (as of 3 September 2025).

Outlet	Access Status	Raw Data	Usable
El Mercurio	Login required	-	-
La Cuarta	Successfully extracted	183	153
The Clinic	Successfully extracted	89	75
CNN Chile	Successfully extracted	45	28
Diario Financiero	Anti-scraping protection	-	-
La Segunda	Login required	-	-
La Tercera	Successfully extracted	269	236
Las Últimas Noticias	Image-based format	-	-
HoyxHoy	Image-based format	-	-
Ciper Chile	Successfully extracted	17	16
EMOL	Successfully extracted	215	153
Ex-Ante	Successfully extracted	99	65
El Mostrador	Successfully extracted	99	86
El País	Successfully extracted	128	102
Gamba	Successfully extracted	110	17
Total		**1254**	**931**

El Mercurio and *La Segunda* implemented paywall systems requiring paid subscriptions for article access. While some preview content was visible, full articles remained inaccessible without authentication. *Diario Financiero* employed sophisticated anti-scraping techniques including disabled right-click functionality, obfuscated source code, blocked developer tools access, and request pattern detection. These protective measures, discussed in online forums, effectively prevented automated extraction. Las *Últimas Noticias* and *HoyxHoy* presented their content as scanned newspaper pages in image format, requiring OCR processing beyond the scope of this initial collection effort. The image-based presentation made text extraction infeasible with our current HTML parsing approach.

For accessible outlets, HTML structure variations required custom extraction logic for each platform. *CNN Chile*'s brief article format resulted in many pieces below our 100-character minimum threshold for valid content. *Gamba*'s alternative media format included significant non-news content (memes, user submissions) that reduced the usable article ratio to 15.5% of collected pages. Character encoding presented consistent challenges across Spanish-language content, with ISO-8859-1 and Latin-1 encodings requiring special handling for accented characters and Spanish-specific punctuation [41].

Table 3 documents the technical obstacles encountered during data collection, categorizing them by type and affected outlets. Five of the fifteen targeted outlets were completely inaccessible due to paywalls [42], anti-scraping protections [43], or image-based content presentation [44]. For accessible outlets, extraction challenges included character encoding issues across all Spanish-language sources [45] and dynamic JavaScript content requiring browser automation [46]. Validation rates varied significantly based on content models and technical implementations, ranging from 94.1% for *Ciper Chile* to 15.5% for *Gamba*. These obstacles represent practical limitations of automated content extraction from heterogeneous regional media sources.

Table 3. Technical obstacles encountered during data collection.

Obstacle Type	Affected Outlets	Specific Details	Ref.
Paywall systems	El Mercurio, La Segunda	Paid subscriptions required; preview content visible but full articles inaccessible without authentication	[42]
Anti-scraping protections	Diario Financiero	Disabled right-click, obfuscated source code, blocked developer tools, request pattern detection	[43]
Image-based format	Las Últimas Noticias, HoyxHoy	Content presented as scanned newspaper pages; text extraction would require OCR	[44]
Brief content	CNN Chile	Many articles below 100-character minimum threshold; resulted in 62% validation rate	This work
Mixed content	Gamba	Significant non-news content (memes, user submissions); resulted in 15.5% validation rate	This work
Character encoding	All Spanish outlets	Default BeautifulSoup encoding corrupted Spanish characters; required sequential UTF-8, Latin-1, ISO-8859-1 attempts	[45]
Dynamic content	Multiple outlets	JavaScript-rendered content required Selenium with –disable-blink-features = AutomationControlled flag	[46]

4.2. Content Extraction Methods

Different outlets required specific extraction strategies. Static HTML sites like *Ciper Chile* and *El Mostrador* could be scraped with simple HTTP requests and `BeautifulSoup` parsing. Dynamic sites loading content via JavaScript required use of Selenium with `undetected-chromedriver` to avoid detection as automated traffic. Our implementation used specific configuration flags including `-disable-blink-features` set to `AutomationControlled` to prevent bot detection, as shown in the code examples in Figures 1 and 2.

To respect server resources and avoid triggering anti-scraping protections, we implemented rate limiting with minimum one-second delays between requests, with random delays of 6-12 seconds between page loads (as documented in the repository code). The user-agent strings were configured to identify the purpose of the academic research. Some outlets embed article metadata in JSON structured data within script tags, providing a cleaner extraction path than HTML parsing. Search-based discovery worked for outlets with functional archive search, while others required crawling through paginated category pages or date-based archives.

Character encoding issues arose consistently due to Spanish-language content. The default `BeautifulSoup` encoding often corrupted Spanish characters, requiring explicit encoding detection and conversion. We implemented a multi-strategy approach attempting UTF-8, Latin-1, and ISO-8859-1 in sequence until successful character representation was achieved.

5. Dataset Description and Results

5.1. Data Filtering and Validation

From the initial collection of 1254 pages, the filtering process removed 323 items to produce 931 validated articles. We note that this process was performed through manual inspection of each retrieved article and did not follow an automated process (e.g., with a classifier) [47]. The manual filtering process followed criteria focused on three objective

factors: **(1)** content type—removing non-article pages like category indexes, author profiles, and multimedia galleries; **(2)** geographic relevance—excluding coverage of non-Chilean events, particularly other Latin American protests that appeared in search results; and **(3)** extraction quality—removing pages with insufficient extracted text (fewer than 100 characters). While manual review introduces human judgment, the criteria were primarily factual (is this about Chile? Is this an article?) rather than interpretive. We retained articles with even tangential mentions of Chilean social unrest to minimize exclusion bias. The complete dataset preserves metadata for all filtered items in the repository, allowing future researchers to review our decisions. While automated approaches using classifiers could improve reproducibility, our manual process ensured high-quality inclusion decisions for this initial dataset release.

85 pages were non-news content including category pages, author profiles, and multimedia galleries without text articles. Another 235 items contained information unrelated to the Chilean social outburst, either covering different topics entirely or referring to protests in other countries such as Colombia's 2019 demonstrations. Three additional articles were removed due to extraction failures that resulted in fewer than 100 characters of body text.

The filtering criteria focused on temporal relevance (within the November 2019 to December 2023 timeframe), geographic relevance (Chilean events and context), content type (news articles rather than other formats), and extraction quality (sufficient text content successfully extracted). Articles mentioning the Chilean social outburst only tangentially were retained if they provided context on the political situation. Opinion pieces and editorials were included when they directly addressed constitutional processes or protest events.

5.2. Dataset Structure and Statistics

The dataset is organized in JSON format with standardized fields across all sources, following Wilkinson et al.'s [48] FAIR principles for findable, accessible, interoperable, and reusable scientific data.

Each article contains nine fields: `newscast` (source outlet identifier), `title` (article headline), `description` (summary when available), `category` (outlet-provided classification), `date` (publication date), `image_link` (featured image URL), `author` (byline when available), `link` (original article URL), and `content` (full article text). Not all fields are populated for every article due to variations in outlet metadata availability.

Table 4 presents the key dataset metrics. The average article length of 2847 characters indicates substantial content depth, although this varies significantly by outlet. For example, *Ciper Chile*'s investigative pieces average over 4000 characters, while CNN Chile's news briefs average under 2000 characters. In terms of words, the average length is 1003. The validation rate represents the percentage of initially collected pages that passed filtering criteria, ranging from 94.1% for *Ciper Chile* to 15.5% for Gamba.

While formal precision/recall evaluation would require a manually-created ground truth dataset beyond this project's scope, the 74.2% overall validation rate provides an indirect quality indicator. This rate represents the percentage of collected pages that contained valid, relevant article content after removing non-article pages, off-topic content, and extraction failures. The validation process successfully identified incomplete extractions (insufficient text) and non-relevant content, suggesting reliable extraction for validated articles. The variation in validation rates between outlets (from 94.1% for Ciper Chile to 15.5% for Gamba) reflects differences in website structure clarity and content models rather than extraction quality issues.

Table 4. Dataset composition and quality metrics.

Metric	Value
Total pages collected	1254
Non-extractable content	85
Unrelated to Chilean context	235
Insufficient text content	3
Final validated articles	**931**
Accessible outlets	10
Inaccessible outlets	5
Average article length (words)	1003
Average article length (characters)	2847
Temporal span (days)	1493
Articles from La Tercera	236
Articles from EMOL	153
Articles from La Cuarta	153
Articles from El País	102
Articles from El Mostrador	86
Articles from The Clinic	75
Articles from Ex-Ante	65
Articles from CNN Chile	28
Articles from Gamba	17
Articles from Ciper Chile	16

5.3. Temporal and Content Distribution

Figure 3 illustrates the temporal distribution of the articles throughout the collection period. The frequency of the articles shows distinct peaks corresponding to major political events, with the highest concentration during the October 2020 plebiscite period [49,50] and around September 2022 for the constitutional referendum [51–53]. The distribution reveals how media attention fluctuated with the intensity of political activity [54–56] and how the unrest subsided as the unrest faded away from public consciousness [9,57].

The temporal distribution in Figure 3 reveals distinct patterns in media attention that correspond to the issue-attention cycle described by Downs (2016). The highest concentration of articles occurs during the October 2020 constitutional plebiscite period, when media coverage intensified around the decision to draft a new constitution. A second major peak appears around September 2022 during the first constitutional referendum, when Chileans voted on the proposed constitution. These peaks demonstrate how media attention fluctuates with the intensity of political activity, concentrating during decision points and diminishing during interim periods. The gradual decline in coverage toward 2023 reflects how the unrest faded from public consciousness as institutional processes concluded, consistent with patterns observed in other social movements.

Content analysis reveals variation in coverage focus across outlets. *La Tercera* and *EMOL* provided consistent daily coverage with a wide topical range. *Ciper Chile* produced fewer but longer investigative pieces that examined the underlying causes and systemic issues. *The Clinic* and *Gamba* offered alternative perspectives that were often critical of mainstream narratives. *El País* provided international context comparing Chilean events to regional movements. This diversity in approach and focus provides multiple lenses for analyzing the social unrest period.

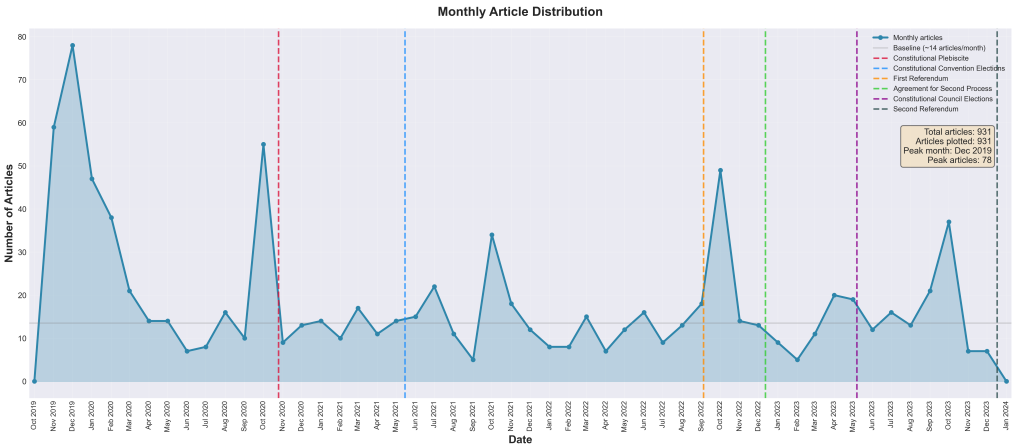

Figure 3. Temporal distribution of articles showing coverage intensity during key political events.

Figure 4 shows the relationship between the articles collected and validated for each outlet, revealing different patterns of content quality. Outlets with higher validation rates typically had cleaner HTML structures and clearer article boundaries. The low validation rate for *Gamba* (15.5%) reflects its mixed content model that includes user-generated content, memes, and brief commentary alongside traditional articles.

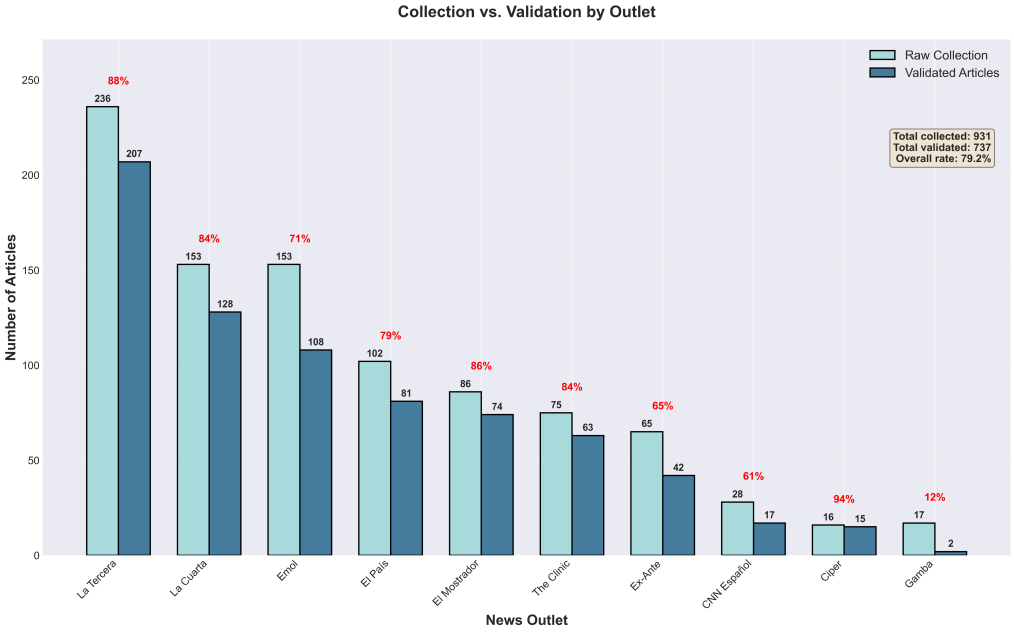

Figure 4. Data collection and validation results by news outlet.

The article length distributions in Figures 5 and 6 serve complementary analytical purposes. Figure 5 displays raw article length distributions using box plots, facilitating direct comparison of median values and quartile ranges across outlets. This visualization reveals that *Ciper Chile*'s investigative pieces average over 2000 words, while *CNN Chile*'s news briefs average about a 1000 words, demonstrating how different outlet types (investigative journalism, daily news, and alternative media) produce content of varying depth. Figure 6 employs a logarithmic scale with violin plots to better visualize the full distribution shapes and identify patterns obscured in the linear scale. The violin plots reveal whether outlets produce content of consistent length (unimodal distributions) or vary between brief updates and longer analyses (multimodal distributions). For example, *The Clinic* shows a bimodal pattern indicating both brief news items and longer feature articles, while *Ciper Chile* demonstrates a unimodal distribution concentrated at longer lengths, consistent with

its investigative journalism focus. Together, these figures reveal distinct editorial patterns in coverage approaches.

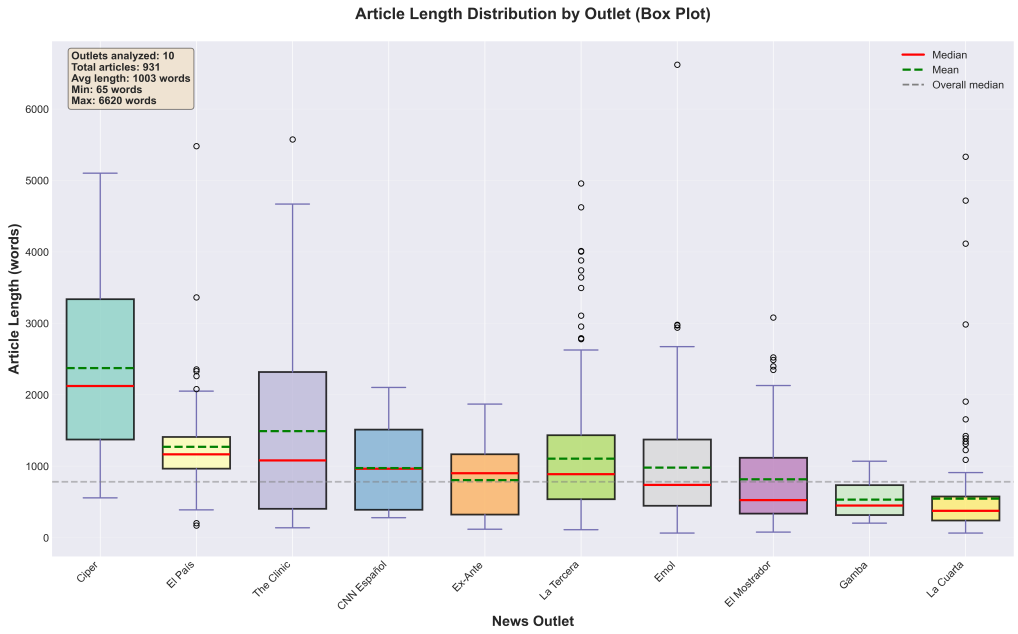

Figure 5. Article length distributions revealing editorial patterns across outlets.

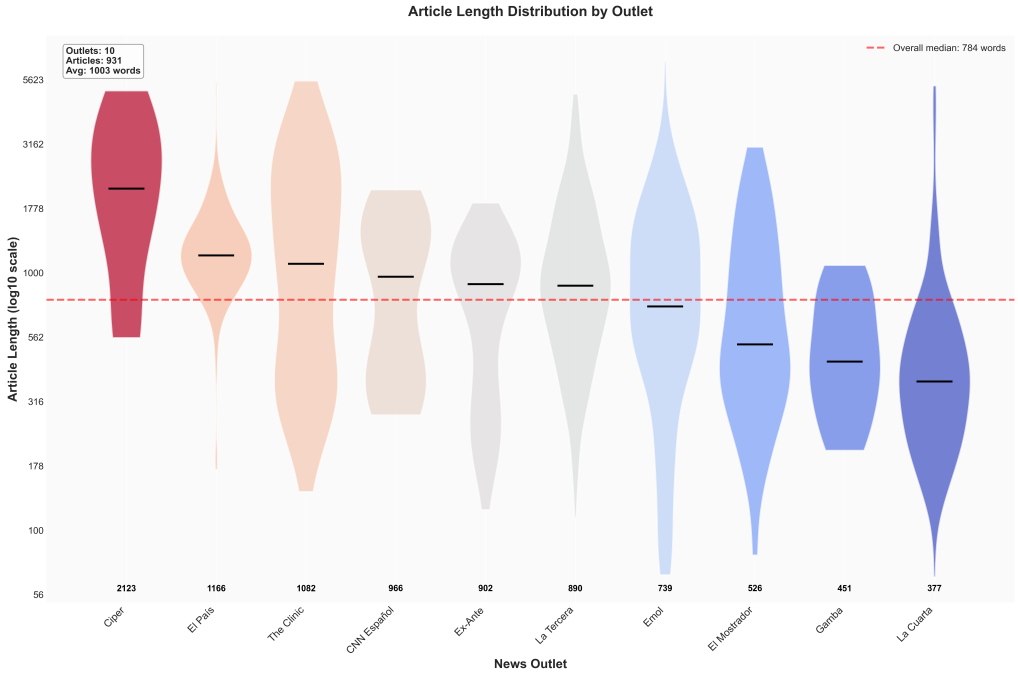

Figure 6. Article length in logarithmic scale to showcase the distribution differences between outlets.

5.4. Data Availability and Repository Structure

The complete dataset and extraction code are publicly available on GitHub at https://github.com/MolodyGs/CSON-Data (accessed on 10 September 2025). The repository follows a structured organization that facilitates both data access and code reproducibility. The `src/archive/` directory contains the consolidated dataset (`DATA.json`). The `src/archive/pages_extracted/` subdirectory organizes the final extracted data by outlet, with each outlet having its own folder containing JSON files. The `src/extract_code/`

directory houses the Jupyter notebooks used for extraction, named according to their target outlet. The data file is included as Supplementary Materials in this article.

The supporting documentation includes a `README` file describing the repository and extraction process. The Jupyter notebooks contain markdown comments that further explain the process. Moreover, the modular notebook structure allows researchers to understand and adapt the extraction logic for specific outlets or to extend the collection to additional sources. However, we note that as outlets evolve and their web page's inner workings change, the extraction code may eventually no longer work.

The dataset is provided as Supplementary Materials to this article to ensure permanent availability through the journal's archival system, independent of external repository changes. The extraction code is available on GitHub (https://github.com/MolodyGs/CSON-Data (accessed on 10 September 2025)) with version control for transparency and reproducibility. We acknowledge that web scraping code may become obsolete as websites evolve their structures and anti-scraping measures, a known limitation of web-scraping research. However, the dataset itself remains valuable independent of the extraction code, as it captures a specific historical period. Researchers can submit issues or pull requests to the GitHub repository to report errors, suggest improvements, or contribute extensions. While we cannot guarantee active maintenance of the extraction scripts as websites evolve, the dataset documentation and standardized JSON format ensure long-term usability for research purposes.

6. Discussion

6.1. Applications and Research Potential

This dataset enables specific research applications demonstrated in the existing literature. For media framing analysis, researchers can examine the media framing and context of social unrest in Chile [31,58,59] to address research questions such as: "How did framing of constitutional processes differ between outlets with different editorial orientations?" [8,60] and "What narrative patterns emerged during peak protest periods versus institutional process periods?" The temporal span allows tracking narrative evolution throughout the constitutional process period [55,61] and enables event-based narrative analysis and extraction [62] correlating media coverage with protest intensity, political announcements, and electoral outcomes.

For natural language processing research, the dataset provides a substantial corpus of Chilean Spanish text from professional sources. This addresses the scarcity of regional Spanish-language resources for training and evaluating Natural Language Processing models [63]. The variety of writing styles across outlets—from *Ciper*'s investigative journalism to *CNN*'s brief news updates—provides diverse linguistic patterns for model training [64].

Potential applications include sentiment analysis of political discourse [65–67] to address questions such as: "How does sentiment expressed in professional news coverage compare to social media discourse during the same events?" Researchers could also apply named entity recognition for Chilean political figures and organizations, and event extraction for building timelines of the social unrest [7,62,68]. For narrative extraction, researchers can extract the underlying narratives from the news information landscape [69–71]. In this context, the extracted narratives could be evaluated using large language models [72], allowing the development of an end-to-end extraction and evaluation approach based on the collected news data.

In the context of misinformation dynamics research researchers could use the data in conjunction with fact-checking information from the same period [14] to investigate specific questions: "What patterns emerge when comparing news coverage with fact-checked claims?" and "How do claims circulate between mainstream news media and

verification platforms during political crises?" This could enable studies of misinformation dynamics [52,73], source attribution patterns [74], and the relationship between news reporting and fact-checking activities [75].

6.2. Limitations

Several limitations affect the completeness of the dataset and potential applications. The exclusion of paywall-protected outlets (*El Mercurio*, *La Segunda*, *Diario Financiero*) removes important voices from the Chilean media landscape, particularly conservative-leaning perspectives that could provide balance to the political spectrum represented, potentially skewing the political spectrum representation toward center-left and alternative media voices that were more accessible. These outlets historically reach different demographic segments and provide editorial perspectives that would enhance the dataset's coverage. We attempted to mitigate this bias by including multiple outlets across the accessible political spectrum, but acknowledge that the dataset could underrepresent certain viewpoints present in Chilean media discourse.

Similarly, the inability to process image-based content from Las Últimas Noticias and HoyxHoy excludes popular tabloid coverage that reaches middle and lower socioeconomic strata through sensationalist presentation styles. These outlets, despite belonging to major conglomerates, maintain distinct editorial voices and target audiences that differ from digital-first platforms. The focus on national outlets also limits geographic diversity of perspectives from regional media. This decision was driven by technical constraints, as numerous small regional outlets would each require individual extraction solutions. Future work could expand to regional outlets through collaborative efforts or automated extraction framework adaptations.

The inability to process image-based content from *Las Últimas Noticias* and *HoyxHoy* excludes popular tabloid coverage that often reaches different demographic segments than digital-first outlets.

Technical constraints in the extraction process may have resulted in incomplete content capture for JavaScript-heavy sites despite Selenium usage. Dynamic content updates, such as article corrections or additions after initial publication, may not be reflected in the dataset because the extraction occurred at specific time points. However, we note that as this is a retrospective study from 2025, we expect that such changes should be very minor, if there are any changes at all due to the time that has passed since the events. Character encoding issues, while addressed through multiple conversion strategies, may persist for some special characters or unconventional formatting.

The filtering process, while necessary for data quality, introduces selection effects. The removal of 235 articles as "unrelated" involved judgment calls about relevance to Chilean social outburst. Articles with indirect connections or those that provide a broader Latin American context may have been inappropriately excluded. We note that the 100-character minimum threshold, while filtering extraction failures, may have removed legitimate brief news items or breaking news alerts.

6.3. Future Directions

Several avenues exist for expanding and improving the dataset. Implementing OCR technology [76] or more advanced Vision-Language Model [77] techniques could unlock the image-based content from *Las Últimas Noticias* and *HoyxHoy*, adding perspectives from the popular press to the corpus. Developing authentication methods for paywall-protected sites, with appropriate permissions, would provide access to influential outlets currently excluded. Including regional and local news sources would capture the geographic variation in coverage beyond the Santiago-centric national media.

There are also potential analytical extensions that could enhance the research value of the dataset. For example, automated topic modeling could provide subject classifications beyond the categories supplied in the outlet [78]. Sentiment analysis [67,79] could quantify the emotional tone in articles and time periods. Entity linking [80] could connect mentions across articles to build actor networks and use entity-based connections to extract narratives [81,82]. Integration with social media data could reveal cross-platform information flows during the period of social unrest [10,83]. Further data could be obtained by extracting the main event descriptors [84]—the fundamental answers to the 5W1H questions (Who, What, Where, When, Why, and How)—from news articles using automated approaches that reflect the human reasoning process [85], which can also be used as transparency cues to help audiences assess the quality of the information [86].

Moreover, considering that the dataset includes multiple journalistic genres—traditional news articles, opinion pieces, and editorials—future studies could benefit from genre-specific analysis. News articles and informative chronicles, written in the third person and based on reporting and investigation, represent distinct journalistic products that follow different verification standards than opinion-based content, which is particularly relevant when analyzing the media structure of social unrest [58,59]. Furthermore, opinion articles and editorials express the views of individual authors that may not necessarily represent the institutional position of the media outlet [69]. Furthermore, the relationship between news reporting and fact-checking activities [75] differs fundamentally from opinion content, where subjective interpretation is expected.

6.4. Ethical Considerations

The data collection process adhered to ethical web-scraping practices while recognizing the limitations and responsibilities involved [18]. Wenget et al. [87] provide empirical evidence on `robots.txt` compliance based on 3.9 million requests, informing our approach to respectful data collection. Thus, we followed the directives of the `robots.txt` files and implemented rate limiting to avoid server overload, with minimum one-second delays between requests. No attempts were made to circumvent authentication systems or access restricted content. The extraction identified itself with appropriate user agent strings that indicate an academic purpose.

All collected content was publicly available at the time of extraction, with no user comments or personal information collected. The dataset preserves attribution to original sources and maintains links to original articles. Although the content remains the intellectual property of the respective outlets, the compilation of the dataset for research purposes is aligned with the principles of academic fair use. Researchers using this dataset should cite both this work and original news sources, respect copyright in any republication, and consider the sensitivity of analyzing political discourse during social conflict.

While the dataset enables research on media dynamics during political crises, we acknowledge risks of potential misuse for political manipulation or targeted misinformation campaigns. The aggregation of news content creates both research opportunities and responsibilities, as noted by Lazer et al. [88] in discussing computational misinformation research. The dataset is provided specifically for academic research purposes to advance understanding of democratic discourse during periods of social transformation. Users should adhere to ethical guidelines when analyzing political discourse during social conflicts, considering the sensitivity of content related to protests, institutional legitimacy, and constitutional processes. We emphasize that the dataset captures professional journalism subject to editorial standards and verification processes, distinguishing it from social media content, and researchers should contextualize findings within the ownership structures and professional norms of the Chilean media landscape.

The aggregation of news content creates both opportunities and responsibilities. While allowing for research on media dynamics during political crises [29], the dataset could potentially be misused for political manipulation or targeted misinformation campaigns [88]. Users should approach the data with awareness of its limitations, potential biases, and the broader social context of Chile's social unrest [89]. The dataset should contribute to understanding and improving democratic discourse rather than undermining it.

7. Conclusions

This paper documents a dataset of 931 Chilean news articles collected through web scraping of 15 major outlets covering the social unrest period from November 2019 to December 2023. The data collection process successfully extracted content from 10 outlets while documenting access barriers that prevented complete coverage. The resulting dataset provides a structured resource for studying media narratives during one of Chile's most significant political periods in recent decades.

The modular extraction approach using outlet-specific Jupyter notebooks proved to be effective in handling the heterogeneous landscape of Chilean news websites. Each platform's unique HTML structure, content delivery method, and access restrictions required customized solutions, from simple `BeautifulSoup` parsing to Selenium-based dynamic content rendering. The documentation of these platform-specific challenges and solutions contributes methodological knowledge for future web-scraping efforts in similar contexts.

Despite limitations in coverage and technical implementation, the dataset offers substantial value for research. The 931 validated articles provide temporal coverage of key events, including constitutional plebiscites, convention proceedings, and referendum campaigns. The diversity of sources—from investigative journalism to alternative media—captures multiple perspectives on the social unrest. The standardized JSON format and the availability of public code ensure the reproducibility and extensibility of this work.

As Chile continues to grapple with social demands and potential unrest, understanding the dynamics of the media remains essential. This dataset provides a foundation for computational analysis of political discourse, contributing to broader efforts to understand information flow during periods of democratic transformation. Integration with fact-checking data and potential expansion to additional sources could further enhance its value for researchers studying Chilean politics, media systems, and social movements.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/data10110174/s1.

Author Contributions: Conceptualization, B.K. and I.M.; methodology, I.M. and B.K.; software, I.M.; validation, I.M., J.M. and B.K.; formal analysis, I.M. and J.M.; investigation, I.M. and J.M.; resources, B.K.; data curation, I.M.; writing-original draft preparation, I.M., J.M. and B.K.; writing—review and editing, I.M., J.M. and B.K.; visualization, I.M.; supervision, B.K.; project administration, B.K.; funding acquisition, B.K. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been supported by ANID (National Research and Development Agency of Chile) FONDECYT de Iniciación en Investigación 2025 Grant 11250039 Project "Interactive Narrative Analytics: Developing scalable knowledge-based narrative extraction models and visual analytics systems for sensemaking in complex information landscapes." This research was also supported by Project 202311010033-VRIDT-UCN.

Institutional Review Board Statement: Ethical review and approval were waived for this study by the Scientific Ethics Committee of Universidad Católica del Norte (protocol code CEC 076/2025) because the research involves the collection and analysis of publicly available news articles from web pages and does not involve participation with human subjects.

Informed Consent Statement: Not applicable. This study does not involve human subjects.

Data Availability Statement: The complete dataset and extraction code are publicly available on GitHub at https://github.com/MolodyGs/CSON-Data (accessed on 10 September 2025).

Acknowledgments: During the preparation of this manuscript, the authors used AI-based editing tools for the purposes of paraphrasing and improving English writing. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Somma, N.M.; Bargsted, M.; Disi Pavlic, R.; Medel, R.M. No water in the oasis: The Chilean Spring of 2019–2020. *Soc. Mov. Stud.* **2021**, *20*, 495–502.
2. Garcés, M. *Estallido Social y una Nueva Constitución para Chile*; LOM Ediciones: Santiago, Chile, 2020.
3. Pleyers, G. The Chilean awakening in a global decade of social movements. In *Citizenship Utopias in the Global South*; Routledge: London, UK, 2024; pp. 68–84.
4. Brannen, S.J.; Haig, C.S.; Schmidt, K. The age of mass protests: Understanding an escalating global trend. *Cent. Strateg. Int. Stud.* **2020**, *4*, 1–42.
5. Lo, W.H.; Lam, B.S.Y.; Cheung, M.M.F. The dynamics of political elections: A big data analysis of intermedia framing between social media and news media. *Soc. Sci. Comput. Rev.* **2021**, *39*, 627–647.
6. Delucchi, A.S.; Ugarte, V.R. The Chilean constitutional process narrated through a spiral. *Stud. Soc. Justice* **2024**, *18*, 969–991.
7. Mendoza, M.; Valenzuela, S.; Núñez-Mussa, E.; Padilla, F.; Providel, E.; Campos, S.; Bassi, R.; Riquelme, A.; Aldana, V.; López, C. A study on information disorders on social networks during the Chilean social outbreak and COVID-19 pandemic. *Appl. Sci.* **2023**, *13*, 5347.
8. Luna, J.P.; Toro, S.; Valenzuela, S. Amplifying counter-public spheres on social media: News sharing of alternative versus traditional media after the 2019 Chilean uprising. *Soc. Media+ Soc.* **2022**, *8*, 20563051221077308.
9. Scherman, A.; Rivera, S. Social media use and pathways to protest participation: Evidence from the 2019 Chilean social outburst. *Soc. Media+ Soc.* **2021**, *7*, 20563051211059704.
10. Keith Norambuena, B.F.; Mitra, T.; North, C. Characterizing social movement narratives in online communities: The 2021 Cuban Protests on Reddit. *arXiv* **2022**, arXiv:2208.04465.
11. Daud, R.S. The role of political communication in shaping public opinion: A comparative analysis of traditional and digital media. *J. Public Represent. Soc. Provis.* **2021**, *1*, 80–89.
12. Adelabu, O.T.; Owojuyigbe, M.J.; Adeyemo, F.D. New media versus traditional media: Power dynamics and the struggle for credibility. *Afr. J. Soc. Behav. Sci.* **2025**, *15*, 1418–1428.
13. Brown, D.K.; Harlow, S. Protests, media coverage, and a hierarchy of social struggle. *Int. J. Press.* **2019**, *24*, 508–530.
14. Molina, I.; Keith, B.; Matus, M. A Multimodal Dataset of Fact-Checked News from Chile's Constitutional Processes: Collection, Processing, and Analysis. *Data* **2025**, *10*, 13.
15. Munzert, S.; Rubba, C.; Meißner, P.; Nyhuis, D. *Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining*; John Wiley & Sons: Chichester, UK, 2014.
16. Krotov, V.; Johnson, L.; Silva, L. Legality and ethics of web scraping. In Proceedings of the Twenty-Fourth Americas Conference on Information Systems, New Orleans, LA, USA, 10–14 August 2020.
17. Dallabetta, M.; Dobberstein, C.; Breiding, A.; Akbik, A. Fundus: A simple-to-use news scraper optimized for high quality extractions. *arXiv* **2024**, arXiv:2403.15279.
18. Brown, M.A.; Gruen, A.; Maldoff, G.; Messing, S.; Sanderson, Z.; Zimmer, M. Web scraping for research: Legal, ethical, institutional, and scientific considerations. *arXiv* **2024**, arXiv:2410.23432.
19. Marres, N.; Weltevrede, E. Scraping the Social? Issues in live social research. *J. Cult. Econ.* **2013**, *6*, 313–335.
20. Neumann, M.; Linder, F.; Desmarais, B. Government websites as data: A methodological pipeline with application to the websites of municipalities in the United States. *J. Inf. Technol. Politics* **2022**, *19*, 411–422.
21. Leetaru, K.; Schrodt, P.A. GDELT: Global data on events, location, and tone, 1979–2012. In Proceedings of the ISA Annual Convention, San Francisco, CA, USA, 3–6 April 2013; Volume 2, pp. 1–49.
22. Raleigh, C.; Linke, A.; Hegre, H.; Karlsen, J. Introducing ACLED: An armed conflict location and event dataset. *J. Peace Res.* **2010**, *47*, 651–660.
23. Team, N.; Costa-Jussà, M.R.; Cross, J.; Çelebi, O.; Elbayad, M.; Heafield, K.; Heffernan, K.; Kalbassi, E.; Lam, J.; Licht, D.; et al. No language left behind: Scaling human-centered machine translation. *arXiv* **2022**, arXiv:2207.04672.
24. Aguilar, C.; Acosta, O. A critical review of the current state of natural language processing in Mexico and Chile. In *Natural Language Processing for Global and Local Business*; Business Science Reference: Hershey, PA, USA, 2021; pp. 365–389.

25. Cañete, J.; Chaperon, G.; Fuentes, R.; Ho, J.H.; Kang, H.; Pérez, J. Spanish pre-trained bert model and evaluation data. *arXiv* **2023**, arXiv:2308.02976.
26. Cañete, J.; Donoso, S.; Bravo-Marquez, F.; Carvallo, A.; Araujo, V. ALBETO and DistilBETO: Lightweight Spanish language models. *arXiv* **2022**, arXiv:2204.09145.
27. Gutiérrez-Fandiño, A.; Armengol-Estapé, J.; Pàmies, M.; Llop-Palao, J.; Silveira-Ocampo, J.; Carrino, C.P.; Gonzalez-Agirre, A.; Armentano-Oller, C.; Rodriguez-Penagos, C.; Villegas, M. Maria: Spanish language models. *arXiv* **2021**, arXiv:2107.07253.
28. Pérez, J.M.; Rajngewerc, M.; Giudici, J.C.; Furman, D.A.; Luque, F.; Alemany, L.A.; Martínez, M.V. pysentimiento: A python toolkit for opinion mining and social nlp tasks. *arXiv* **2021**, arXiv:2106.09462.
29. Tucker, J.A.; Guess, A.; Barberá, P.; Vaccari, C.; Siegel, A.; Sanovich, S.; Stukal, D.; Nyhan, B. *Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature*; Technical report; Hewlett Foundation: Menlo Park, CA, USA, 2018.
30. Tsfati, Y.; Boomgaarden, H.G.; Strömbäck, J.; Vliegenthart, R.; Damstra, A.; Lindgren, E. Causes and consequences of mainstream media dissemination of fake news. *Ann. Int. Commun. Assoc.* **2020**, *44*, 157–173.
31. Concha Macías, S.; Keith Norambuena, B. Evaluating the Ability of Computationally Extracted Narrative Maps to Encode Media Framing. In Proceedings of the Text2Story@ECIR, Glasgow, Scotland, UK, 24 March 2024; pp. 17–28.
32. Del Valle Rojas, C.; Mellado Ruiz, C.; Salinas Meruane, P.; González Rodríguez, G. La concentración de la propiedad de los medios de comunicación en Chile: De la propiedad al mercado de la publicidad: Los desafíos pendientes. *Sapiens* **2011**, *12*, 44–58.
33. Gronemeyer, M.E.; Porath, W. Tendencias de la posición editorial en diarios de referencia en Chile: El arte de dosificar la crítica frente a la actuación de los actores políticos. *Rev. Cienc. Política* **2017**, *37*, 177–202.
34. González-Trujillo, R.; Olate-Hidalgo, C.; Grassau, D. Impacto del entorno digital en los medios tradicionales chilenos: percepciones y actitudes predominantes de sus protagonistas. *Palabra Clave* **2022**, *25*. https://doi.org/10.5294/pacla.2022.25.4.7.
35. Saavedra, J. Comunicación, medios y movimientos sociales en Chile, balance de (un cuarto de) siglo. *Comun. Medios* **2023**, *32*, 12–23.
36. Richardson, L. Beautiful Soup Documentation. 2007. Available online: https://www.crummy.com/software/BeautifulSoup/ (accessed on 1 July 2025).
37. Selenium Contributors. Selenium WebDriver. 2022. Available online: https://www.selenium.dev/documentation/webdriver/ (accessed on 1 July 2025).
38. Donig, S.; Eckl, M.; Gassner, S.; Rehbein, M. Web archive analytics: Blind spots and silences in distant readings of the archived web. *Digit. Scholarsh. Humanit.* **2023**, *38*, 1033–1048.
39. Hanna, A. Mpeds: Automating the Generation of Protest Event Data. *Deposited at SocArXiv*. 2017. Available online: https://osf.io/preprints/socarxiv/xuqmv (accessed on 1 July 2025).
40. Dorff, C.; Gallop, M.; Minhas, S. Networks of violence: Predicting conflict in Nigeria. *J. Politics* **2020**, *82*, 476–493.
41. *ISO/IEC 8859-1:1998*; Information Technology—8-Bit Single-Byte Coded Graphic Character Sets—Part 1: Latin Alphabet No. 1. International Organization for Standardization: Geneva, Switzerland, 1998.
42. Dhillon, P.S.; Panda, A.; Hemphill, L. How digital paywalls shape news coverage. *PNAS Nexus* **2025**, *4*, 511.
43. Cuevas, A.; Miedema, F.; Soska, K.; Christin, N.; van Wegberg, R. Measurement by proxy: On the accuracy of online marketplace measurements. In Proceedings of the 31st USENIX Security Symposium (USENIX Security 22), Boston, MA, USA, 10–12 August 2022; pp. 2153–2170.
44. Strange, C.; McNamara, D.; Wodak, J.; Wood, I. Mining for the meanings of a murder: The impact of OCR quality on the use of digitized historical newspapers. *Digit. Humanit. Q.* **2014**, *8*, 1–17.
45. Khan, M.; Ullah, K.; Alharbi, Y.; Alferaidi, A.; Alharbi, T.S.; Yadav, K.; Alsharabi, N.; Ahmad, A. Understanding the research challenges in low-resource language and linking bilingual news articles in multilingual news archive. *Appl. Sci.* **2023**, *13*, 8566.
46. Fayzrakhmanov, R.R.; Sallinger, E.; Spencer, B.; Furche, T.; Gottlob, G. Browserless web data extraction: Challenges and opportunities. In Proceedings of the 2018 World Wide Web Conference, Lyon, France, 23–27 April 2018; pp. 1095–1104.
47. Lorenzini, J.; Kriesi, H.; Makarov, P.; Wüest, B. Protest event analysis: Developing a semiautomated NLP approach. *Am. Behav. Sci.* **2022**, *66*, 555–577.
48. Wilkinson, M.D.; Dumontier, M.; Aalbersberg, I.J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.W.; da Silva Santos, L.B.; Bourne, P.E.; et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **2016**, *3*, 160018.
49. Morales Quiroga, M. Chile's perfect storm: Social upheaval, COVID-19 and the constitutional referendum. *Contemp. Soc. Sci.* **2021**, *16*, 556–572.
50. Heiss, C.; Suárez-Cao, J. Constitution-making in the 21st century: Lessons from the Chilean process. *PS Political Sci. Politics* **2024**, *57*, 282–285.
51. Sazo, D. Chile 2022: De las grandes expectativas al creciente pesimismo. *Rev. Cienc. Política* **2023**, *43*, 193–222.
52. Charney, J.; Mayer, L.; Santander, P. Information Disorder in the Chilean Constitutional Process: When Disinformation Originates with the Political Authorities Themselves. *Eur. J. Crim. Policy Res.* **2025**, *31*, 477–499.

53. Fierro, P. Feeling the split: Territorial divide and political emotions in the Chilean constituent processes (2022–2023). *Environ. Plan. C Politics Space* **2023**. https://doi.org/10.1177/23996544251349295.
54. Downs, A. Up and down with ecology: The "issue-attention cycle". In *Agenda Setting*; Routledge: London, UK, 2016; pp. 27–33.
55. Boydstun, A.E.; Hardy, A.; Walgrave, S. Two faces of media attention: Media storm versus non-storm coverage. *Political Commun.* **2014**, *31*, 509–531.
56. Vasterman, P.L. Media-hype: Self-reinforcing news waves, journalistic standards and the construction of social problems. *Eur. J. Commun.* **2005**, *20*, 508–530.
57. Cox, L.; González, R.; Le Foulon, C. The 2019 Chilean social upheaval: A descriptive approach. *J. Politics Lat. Am.* **2024**, *16*, 68–89.
58. Riffo-Pavón, I.; Basulto, Ó.; Segovia, P. El Estallido Social chileno de 2019: Un estudio a partir de las representaciones e imaginarios sociales en la prensa. *Rev. Mex. Cienc. Políticas y Soc.* **2021**, *66*, 345–368.
59. Gallegos, O.F.B.; González, S.F.; Riquelme, C.M. Significación social de la violencia en narrativas de prensa escrita tradicional chilena: Un caso de estudio en el contexto del estallido social en Chile (18-O). *Prism. Soc. Rev. Investig. Soc.* **2023**, *2023*, 289–326.
60. Issacharoff, S.; Verdugo, S. Populismo constituyente, democracia y promesas incumplidas: El caso de la Convención Constitucional Chilena (2021–2022) Constituent populism, democracy, and failed promises: The case of the Chilean Constitutional Convention (2021–2022). *Int. J. Const. Law* **2023**, *21*, 1517–1548.
61. Waldherr, A. Emergence of news waves: A social simulation approach. *J. Commun.* **2014**, *64*, 852–873.
62. Keith Norambuena, B.F.; Mitra, T.; North, C. A survey on event-based news narrative extraction. *Acm Comput. Surv.* **2023**, *55*, 1–39.
63. Tellez, E.S.; Moctezuma, D.; Miranda, S.; Graff, M.; Ruiz, G. Regionalized models for Spanish language variations based on Twitter. *Lang. Resour. Eval.* **2023**, *57*, 1697–1727.
64. Acuña-Duarte, A.A.; Campos, P.G.; León, J.A.; Salazar, C.A. Tweeting to be a constitution-writer in Chile: Social media activity, public discourse, and electoral outcomes during pandemic times. *Technol. Soc.* **2024**, *79*, 102740.
65. Rozado, D.; Hughes, R.; Halberstadt, J. Longitudinal analysis of sentiment and emotion in news media headlines using automated labelling with Transformer language models. *PLoS ONE* **2022**, *17*, e0276367.
66. Ruz, G.A.; Henríquez, P.A.; Mascareño, A. Bayesian constitutionalization: Twitter sentiment analysis of the chilean constitutional process through bayesian network classifiers. *Mathematics* **2022**, *10*, 166.
67. Keith Norambuena, B.; Lettura, E.F.; Villegas, C.M. Sentiment analysis and opinion mining applied to scientific paper reviews. *Intell. Data Anal.* **2019**, *23*, 191–214.
68. Calvo Figueras, B.; Caselli, T.; Broersma, M. Finding Narratives in News Flows: The Temporal Dimension of News Stories. *DHQ Digit. Humanit. Q.* **2021**, *15*, 1.
69. Zhang, Z. In *From Media Hype to Twitter Storm: News Explosions and Their Impact on Issues, Crises and Public Opinion*; Taylor & Francis: London, UK, 2019.
70. Keith Norambuena, B.F.; Mitra, T. Narrative maps: An algorithmic approach to represent and extract information narratives. *Proc. Acm-Hum.-Comput. Interact.* **2021**, *4*, 1–33.
71. German, F.; Keith, B.; North, C. Narrative Trails: A Method for Coherent Storyline Extraction via Maximum Capacity Path Optimization. In Proceedings of the Text2Story 2025 Workshop@ECIR2025. CEUR-WS, Lucca, Italy, 10 April 2025; pp. 15–22.
72. Keith, B. LLM-as-a-Judge Approaches as Proxies for Mathematical Coherence in Narrative Extraction. *Electronics* **2025**, *14*, 2735.
73. Saldaña, M.; Orchard, X.; Rivera, S.; Bustamante-Pavez, G. "Your house won't be yours anymore!" Effects of Misinformation, News Use, and Media Trust on Chile's Constitutional Referendum. *Int. J. Press/Politics* **2024**, *26*, 2197–2209.
74. Cárcamo-Ulloa, L.; Cárdenas-Neira, C.; Scheihing-García, E.; Sáez-Trumper, D.; Vernier, M.; Blaña-Romero, C. On politics and pandemic: How do Chilean media talk about disinformation and fake news in their social networks? *Societies* **2023**, *13*, 25.
75. Moreno-Gil, V.; Ramon, X.; Rodríguez-Martínez, R. Fact-checking interventions as counteroffensives to disinformation growth: Standards, values, and practices in Latin America and Spain. *Media Commun.* **2021**, *9*, 251–263.
76. Santos, Y.; Silva, M.; Reis, J.C. Evaluation of optical character recognition (ocr) systems dealing with misinformation in portuguese. In Proceedings of the 2023 36th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), Rio Grande, Brazil, 6–9 November 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 223–228.
77. Zhang, J.; Huang, J.; Jin, S.; Lu, S. Vision-language models for vision tasks: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 5625–5644.
78. Barberá, P.; Boydstun, A.E.; Linn, S.; McMahon, R.; Nagler, J. Automated text classification of news articles: A practical guide. *Political Anal.* **2021**, *29*, 19–42.
79. Wankhade, M.; Rao, A.C.S.; Kulkarni, C. A survey on sentiment analysis methods, applications, and challenges. *Artif. Intell. Rev.* **2022**, *55*, 5731–5780.
80. Shanaz, A.L.F.; Ragel, R.G. Wikidata based person entity linking in news articles. In Proceedings of the 2021 10th International Conference on Information and Automation for Sustainability (ICIAfS), Negambo, Sri Lanka, 11–13 August 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 66–70.

81. Keith Norambuena, B.F.; Mitra, T.; North, C. Narrative sensemaking: Strategies for narrative maps construction. In Proceedings of the 2021 IEEE Visualization Conference (VIS), New Orleans, LA, USA, 24–29 October 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 181–185.
82. Keith Norambuena, B.F.; Mitra, T.; North, C. Design guidelines for narrative maps in sensemaking tasks. *Inf. Vis.* **2022**, *21*, 220–245.
83. Grill, G. Future protest made risky: Examining social media based civil unrest prediction research and products. *Comput. Support. Coop. Work. (CSCW)* **2021**, *30*, 811–839.
84. Keith, B.; Horning, M.; Mitra, T. Evaluating the inverted pyramid structure through automatic 5w1h extraction and summarization. In Proceedings of the Computational Journalism C+ J 2020, Boston, MA, USA, 20–21 March 2020; pp. 1–7.
85. Muñoz, C.; Mendoza, M.; Lobel, H.; Keith, B. Imitating Human Reasoning to Extract 5W1H in News. In Proceedings of the Companion Proceedings of the ACM on Web Conference 2025, Sydney, Australia, 28 April–2 May 2025; pp. 1199–1203.
86. Keith Norambuena, B.; Farina, K.R.; Horning, M.; Mitra, T. Watching the watchdogs: Using transparency cues to help news audiences assess information quality. *Media Commun.* **2023**, *11*, 274–285.
87. Kim, T.; Bock, K.; Luo, C.; Liswood, A.; Wenger, E. Scrapers selectively respect robots. txt directives: Evidence from a large-scale empirical study. *arXiv* **2025**, arXiv:2505.21733.
88. Lazer, D.M.; Baum, M.A.; Benkler, Y.; Berinsky, A.J.; Greenhill, K.M.; Menczer, F.; Metzger, M.J.; Nyhan, B.; Pennycook, G.; Rothschild, D.; et al. The science of fake news. *Science* **2018**, *359*, 1094–1096.
89. Santos, F.G. Social movements and the politics of care: Empathy, solidarity and eviction blockades. *Soc. Mov. Stud.* **2020**, *19*, 125–143.