data

Article

A Multimodal Dataset of Fact-Checked News from Chile's Constitutional Processes: Collection, Processing, and Analysis

Ignacio Molina [1], Brian Keith [1,*] and Mauricio Matus [2]

[1] Department of Computing and Systems Engineering, Universidad Católica del Norte, Antofagasta 1270709, Chile; ignacio.molina@alumnos.ucn.cl
[2] School of Journalism, Universidad Católica del Norte, Antofagasta 1270709, Chile; mmatus@ucn.cl
* Correspondence: brian.keith@ucn.cl

Abstract: This paper presents a multimodal dataset capturing fact-checked news coverage of Chile's constitutional processes from 2019–2023. The collection comprises 300 articles from three sources: *Fast Check*, *Fact Checking UC*, and *BioBioChile*, containing 242,687 words of text and visual content in 168 entries. The dataset implements advanced natural language processing through RoBERTa and computer vision techniques via EfficientNet, with unified multimodal analysis using the CLIP model. Technical validation through clustering analysis and expert review demonstrates the dataset's effectiveness in identifying narrative patterns within constitutional process coverage. The structured format includes verification metadata, precomputed embeddings, and documented relationships between textual and visual elements. This enables research into how misinformation propagates through multiple channels during significant political events. This paper details the dataset's composition, collection methodology, and validation while acknowledging specific limitations. This contribution addresses a gap in current research resources by providing verified multimodal content spanning two constitutional processes, supporting investigations in computational social science and misinformation studies.

Keywords: multimodal dataset; fact-checking; computational social science; misinformation; constitutional processes; natural language processing; computer vision; political discourse analysis; Chile; social media

check for updates

Academic Editor: Han Woo Park

Received: 9 January 2025
Revised: 22 January 2025
Accepted: 26 January 2025
Published: 28 January 2025

Citation: 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. https://doi.org/10.3390/data10020013

1. Introduction

Social networks have become a primary source of information [1], but they also facilitate the rapid spread of misinformation, including manipulated images [2,3]. This phenomenon has been particularly evident during Chile's recent constitutional processes [4], where the spread of false information has impacted public discourse and decision-making.

The propagation of disinformation and fake news on social networks represents a critical challenge in the digital era. Lazer et al. [5] and Vosoughi et al. [6] highlight the complexity of this phenomenon and the need for interdisciplinary approaches to address it. These studies are fundamental for understanding how misinformation spreads online and its social repercussions. Scholars have analyzed the impact of manipulated images on people's perceptions and beliefs, emphasizing the importance of visual disinformation [2,3]. However, these works focus on individual images and do not consider the broader narrative structures that can influence the dissemination of fake news [7,8].

To address these challenges, we present a multimodal dataset that captures fake news and fact-checked information from Chile's two recent constitutional processes. This dataset combines text and visual content from three primary sources: *Fast Check*, *Fact Checking UC*, and *BioBioChile*, covering the periods of both constitutional processes. The dataset includes 300 news items, 168 of which contain associated images.

The significance of our dataset lies in its focus on a crucial period in Chile's recent history, where the spread of misinformation played a substantial role in shaping public opinion. The multimodal nature of our collection, combining text analysis with image processing, provides researchers with a foundation to study how false information propagates across different media formats. This approach allows for a more nuanced understanding of how narratives develop and spread during significant political events.

The dataset has been processed using state-of-the-art natural language processing and computer vision techniques, including BERT [9], RoBERTa [10], and CLIP [11] models, enabling researchers to analyze both textual and visual aspects of misinformation. This processing facilitates various applications, from narrative analysis to pattern recognition in fake news dissemination.

This paper describes the dataset's collection methodology, structure, and potential applications. We also present validation results from both quantitative metrics and expert evaluation, demonstrating the dataset's reliability and utility for research in misinformation studies, computational social science, and related fields.

2. Materials and Methods

2.1. Data Sources

The dataset integrates content from three Chilean fact-checking and news organizations: *Fast Check*[1], *Fact Checking UC*[2], and *BioBioChile*[3]. The source selection process prioritized organizations that actively covered Chile's constitutional processes between 15 November 2019 and 17 December 2023, encompassing both constitutional referendums [12].

Fast Check operates as an independent fact-checking platform focusing on viral content and political statements. The organization implements a systematic verification methodology that categorizes claims into "True", "False", or additional nuanced classifications such as "Imprecise" or "Misleading". The platform provides detailed analysis with supporting evidence and maintains a public archive of verified claims [4].

Fact Checking UC functions as a university-based verification initiative from the *Pontificia Universidad Católica de Chile*. Their methodology incorporates academic rigor in the verification process, with fact-checks conducted by journalism students under faculty supervision. The platform utilizes a three-tier classification system for claims: "True", "False", and "Requires Context" [5].

BioBioChile represents a mainstream news organization that, while not primarily focused on fact-checking, provides coverage of both constitutional processes. Its inclusion enables analysis of how traditional media outlets report on disputed claims and fact-checks. The organization maintains standardized journalistic practices for news verification and source attribution.

The temporal scope of the dataset covers two significant phases of constitutional activity. The first constitutional process began on 15 November 2019, and concluded on 14 May 2022. This phase was succeeded by the second constitutional process, which commenced on 14 May 2022, and extended until 17 December 2023.

The dataset contains 300 unique entries distributed across sources:

- *Fast Check*: 36 fact-checks;
- *Fact Checking UC*: 139 fact-checks;
- *BioBioChile*: 125 news articles.

The variation in entry counts reflects both the organizations' differential focuses on constitutional issues and their distinct publication frequencies. *Fast Check*'s lower entry count correlates with its selective approach to fact-checking high-impact claims, while the higher count of *Fact Checking UC* reflects its systematic coverage of constitutional debates [13].

Each source presents information through standardized web interfaces, enabling systematic data extraction. *Fast Check* and *Fact Checking UC* provide structured fact-check formats including claim, verdict, and evidence. *BioBioChile* follows standard news article formatting with headlines, body text, and associated media.

2.2. Data Collection Methodology

The data collection process employed Python-based web scraping techniques using the BeautifulSoup library, which was selected for its reliable HTML parsing capabilities and extensive documentation[4]. Figure 1 shows an overview of the data collection methodology.

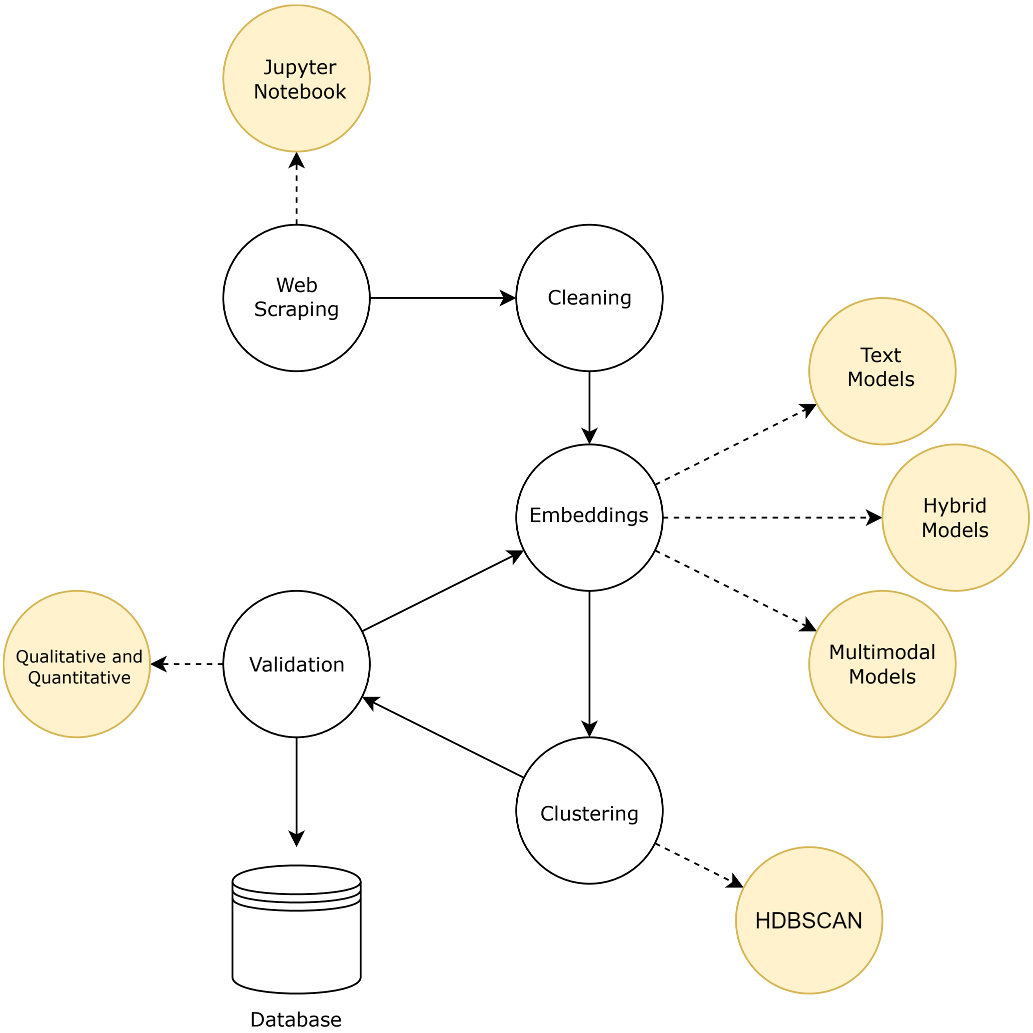

Figure 1. Data collection and processing methodology.

2.2.1. Web Scraping Implementation

The scraping architecture processes HTML content through BeautifulSoup's parser, creating a parse tree structure that enables systematic data extraction. This approach preserves the semantic relationships between different content elements while maintaining the integrity of the data. The implementation handles rate limiting through controlled request intervals to prevent server overload.

2.2.2. Source-Specific Extraction Processes

For *Fast Check*, the extraction process targets articles through Google search results using the query "Constitución Chile *Fast Check*" within the specified constitutional periods. The process extracts the text of the article, verification status, publication date, associated links, and embedded images. Verification status classifications include "True", "False", and additional nuanced categories [4].

Fact Checking UC content extraction utilizes the platform's internal search functionality, processing articles through their standardized HTML structure. The extraction captures the verification verdict, the text of the article, the supporting evidence, and the multimedia content. Due to platform structure limitations, publication dates require manual verification.

BioBioChile requires three distinct extraction approaches due to varying content structures across the website sections:

- Standard news section (/noticias);
- Special coverage section (/especial);
- Constitutional process section (/especial/nuevo-proceso-constituyente).

We show the web scraping methodology in Figure 2.

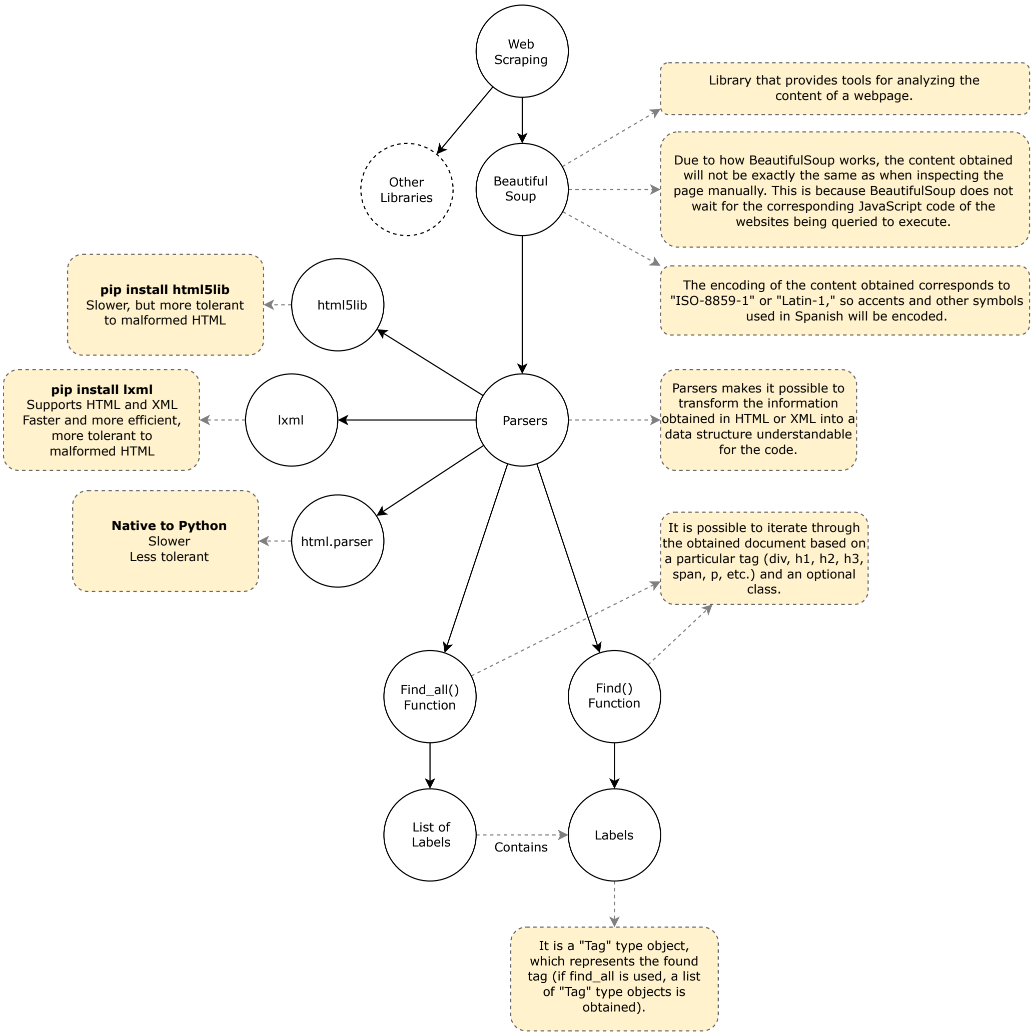
Web Scraping
Other Libraries
Beautiful Soup
Library that provides tools for analyzing the content of a webpage.
Due to how BeautifulSoup works, the content obtained will not be exactly the same as when inspecting the page manually. This is because BeautifulSoup does not wait for the corresponding JavaScript code of the websites being queried to execute.
The encoding of the content obtained corresponds to "ISO-8859-1" or "Latin-1," so accents and other symbols used in Spanish will be encoded.
pip install html5lib
Slower, but more tolerant to malformed HTML
html5lib
pip install lxml
Supports HTML and XML
Faster and more efficient, more tolerant to malformed HTML
lxml
Parsers
Parsers makes it possible to transform the information obtained in HTML or XML into a data structure understandable for the code.
Native to Python
Slower
Less tolerant
html.parser
It is possible to iterate through the obtained document based on a particular tag (div, h1, h2, h3, span, p, etc.) and an optional class.
Find_all() Function
Find() Function
List of Labels
Contains
Labels
It is a "Tag" type object, which represents the found tag (if find_all is used, a list of "Tag" type objects is obtained).

Figure 2. Web scraping methodology.

2.2.3. Quality Control Procedures

The quality control procedures for the dataset involved a series of validation steps to ensure the integrity and reliability of the data. First, consistency verification was used to maintain a uniform data format across all extracted content. This step ensured that all information was in accordance with standardized structural and formatting requirements. Then, redundancy elimination was performed by identifying and removing duplicate entries by comparing URLs, thereby streamlining the dataset and avoiding over-representation of data points. Finally, content coherence was evaluated by verifying the logical relationship between the article text and its associated media, ensuring that all content elements were contextually aligned.

During the quality control process, several challenges were encountered and systematically addressed. One issue involved non-executable JavaScript content, which required manual confirmation to ensure that the webpages were extracted properly. In particular, we found empirically that BeautifulSoup retained sufficient data for our purposes, at least in the scraped websites.

Additionally, platform-specific character encoding presented compatibility challenges that were resolved by implementing tailored encoding solutions. Finally, the diversity in image storage formats across different platforms necessitated adaptations in the processing pipeline to accommodate these variations effectively.

2.2.4. Ethical Considerations

The data collection adheres to the following ethical guidelines:

- Compliance with *robots.txt* specifications.
- Implementation of rate limiting to prevent server stress.
- Extraction of publicly available content only.
- Attribution maintenance for all collected content.

The scraping scripts are modular and documented, enabling adaptation for different source structures while maintaining consistent output formats. All extracted data undergo version control, with processing steps logged for reproducibility. The implementation includes error handling for network issues, malformed HTML, and unexpected content structures [5].

2.3. Dataset Structure and Organization

The dataset maintains a structured format optimized for multimodal analysis, with standardized fields across all sources. Data storage utilizes JSON format to preserve hierarchical relationships between text and associated media.

2.3.1. Data Fields

Each entry in the dataset contains the following fields:

```
{
    "newscast": "Source platform identifier",
    "title": "Article headline",
    "description": "Brief article summary",
    "date": "Publication date (YYYY-MM-DD)",
    "link": "Source URL",
    "author": "Content creator or platform",
    "text": "Full article content",
    "veracity": "Fact-check classification",
    "images": ["Array of image URLs"],
    "links": ["Array of referenced URLs"]
}
```

The veracity field contains standardized classifications (*True*, *False*, and *Other*).

2.3.2. Multimodal Components

The dataset comprises 300 total entries, with 168 entries containing one or more associated images. The text content includes approximately 242,687 words, with an additional 6745 words extracted from the text present in the images.

2.3.3. Preprocessing Implementation

Text preprocessing maintains original formatting and case to preserve semantic context. The preprocessing pipeline generates embedding representations using the following:

- **Text**: RoBERTa model with 560 million parameters.
- **Images**: EfficientNet implementation [14].
- **Multimodal**: CLIP model (openai/clip-vit-base-patch32) for combined analysis.

All preprocessing steps are documented and reproducible through the publicly available implementation code.

2.4. Processing Pipeline Details

The transformation of raw data into analyzable formats required a structured processing pipeline. Our first phase focused on data preparation, where we processed text content by removing HTML artifacts and standardizing formats. This included the extraction of text embedded within images through optical character recognition. We processed images by converting them to standard resolutions and formats, maintaining explicit linkages with their source articles through unique identifiers. The processed content was stored in JSON format to preserve the relationships between text, images, and metadata.

The second phase prepared the data for analysis through three parallel paths. For text content, we generated embeddings using BERT and RoBERTa models, converting the 300 textual entries into vector representations. For image content, we used EfficientNet to generate embeddings for the 300 images in our dataset. The multimodal analysis path used CLIP to process 168 entries containing both text and images, generating combined representations that captured cross-modal relationships.

Each processing path followed the same subsequent steps: dimensionality reduction using t-SNE or UMAP, followed by clustering with HDBSCAN. This unified approach to processing enabled the comparative analysis presented in Section 4, while maintaining consistent treatment across different content types.

The resulting processed dataset maintains the original content structure while adding computed representations, allowing for both individual and combined analysis of text and visual elements. This processing pipeline ensures reproducibility and enables the systematic evaluation of potential misinformation patterns across different modalities.

2.5. Model and Clustering Hyperparameters

The analysis pipeline employs three core models with specific configurations (Table 1). RoBERTa utilizes the xlm-roberta-large architecture with standard truncation and padding. EfficientNet implements the B0 variant with ImageNet weights and global average pooling. CLIP uses the base patch32 architecture with batch processing for multimodal analysis.

The dimensionality reduction and clustering configurations varies across analysis types (Table 2). The hybrid approach achieved a silhouette score of 0.734589 using UMAP with random state 70 and HDBSCAN with epsilon 2.5. The multimodal analysis produced the highest silhouette score of 0.849641 with UMAP random state 20 and HDBSCAN epsilon 1.5. Text analysis employed additional UMAP parameters including neighbor count and minimum distance constraints. These configurations were selected through systematic parameter search to optimize cluster cohesion while maintaining interpretable groupings.

Table 1. Model configuration parameters.

Model	Parameter	Value
RoBERTa	Base Model Truncation Padding	xlm-roberta-large True True
EfficientNet	Base Model Weights Include Top Pooling	EfficientNetB0 ImageNet False Average
CLIP	Base Model Batch Size	clip-vit-base-patch32 8

Table 2. UMAP and HDBSCAN parameters across different analyses.

Analysis	Method	Parameter	Value
Hybrid	UMAP	Components Random State	2 70
	HDBSCAN	Min Cluster Size Min Samples Selection Epsilon Silhouette Score	4 5 2.5 0.734589
Multimodal	UMAP	Components Random State	2 20
	HDBSCAN	Min Cluster Size Min Samples Selection Epsilon Silhouette Score	9 5 1.5 0.849641
Text	UMAP	Components Random State N Neighbors Min Distance Init	2 12 3 0.0 random
	HDBSCAN	Min Cluster Size Min Samples Selection Epsilon	5 3 2.2

Each parameter combination reflects trade-offs between cluster granularity and stability. Higher epsilon values in HDBSCAN produce fewer but more stable clusters, while lower values capture finer-grained relationships at the cost of potential noise inclusion. UMAP configurations balance local and global structure preservation based on the specific requirements of each analysis type.

3. Data Description

3.1. Dataset Composition

The dataset comprises entries collected from three Chilean media sources that cover constitutional processes between 2019 and 2023. Table 3 presents the distribution of entries across sources.

Table 3. Distribution of entries across sources.

Source	Number of Entries	Percentage
Fact Checking UC	139	46.3%
BioBioChile	125	41.7%
Fast Check	36	12.0%
Total	300	100%

The verification status distribution varies significantly between fact-checking platforms, as shown in Table 4.

Table 4. Verification status distribution by platform.

Platform	True	False	Other Classification
Fast Check	5	22	9
Fact Checking UC	16	16	107
BioBioChile	Not Applicable		

The temporal distribution corresponds to two distinct constitutional processes:

- **First Process** (15 November 2019 to 14 May 2022): Constitutional Convention Formation and Initial Draft Development.
- **Second Process** (14 May 2022 to 17 December 2023): Expert Council Deliberations and Final Proposal.

Content analysis reveals four primary thematic categories:

1. **Constitutional Process Mechanics**: Procedural frameworks and voting mechanisms;
2. **Political Actor Statements**: Claims verification from public figures;
3. **Social Impact Claims**: Analysis of projected social policy changes;
4. **Economic Implications**: Assessment of financial and property rights modifications.

The multimodal composition of the dataset includes visual elements in 168 entries (56% of total). The textual corpus contains 242,687 words, supplemented by 6745 words extracted from embedded image text through optical character recognition processes. The distribution of multimodal content is shown in Figure 3.

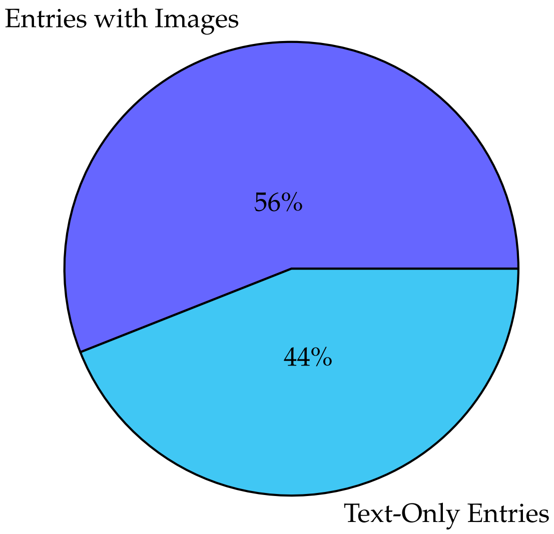

Figure 3. Distribution of multimodal content.

This composition enables analysis of information dissemination patterns during both constitutional processes, with particular emphasis on verification methodologies and multimodal content interaction.

3.2. Data Validation and Quality Assessment

The validation process implemented three key verification mechanisms. First, the extraction process verified the consistency of the format across all entries, which is particularly important given the varied structure of the source websites. Second, a redundancy detection system identified and eliminated duplicate entries through URL comparison. Third, content coherence validation ensured proper relationships between articles and their associated media elements.

The initial dataset contained approximately 400 images, which was reduced to 300 after removing decorative images and elements of the website interface that did not contribute to the news content. These 300 images were associated with 168 articles from the dataset. This reduction improved dataset coherence by ensuring all visual elements directly related to news content.

The completeness assessment revealed varying levels of data availability across sources. *Fast Check* and *Fact Checking UC* provided structured fact-checking information including claim verification status, supporting evidence, publication dates, and source attribution.

However, several limitations and potential biases warrant consideration:

1. **Temporal Coverage**: The dataset's temporal scope is constrained to the constitutional processes, potentially excluding relevant content from adjacent periods.
2. **Source Limitations**: *Fast Check*'s representation (36 entries) is notably smaller than other sources, reflecting their selective fact-checking approach.
3. **Format Constraints**: BeautifulSoup's inability to execute JavaScript resulted in accessing simplified versions of some web pages.
4. **Image Association**: In cases where news items contained multiple images, the relationship between specific text segments and images required manual verification.

The extraction process addressed technical challenges through specific error handling, including the implementation of rate limiting to prevent server timeouts, handling different character encoding through standardization across sources, logging errors for failed extraction attempts, and using version control for all extraction scripts.

4. Technical Validation

We implemented a three-stage validation process to evaluate our dataset and processing methods. First, we independently validated each modality using quantitative metrics. Second, we evaluated multimodal representations through both concatenation and specialized models. Finally, we performed expert validation to verify the semantic coherence of the identified patterns.

Our validation framework employs three key metrics across all analyses. The silhouette score measures the cohesion of the clusters, with values ranging from -1 to 1, where higher scores indicate better-defined groups. Average distance measures the spacing between data points, providing insight into embedding quality. Maximum and minimum distances help identify potential outliers and tight clusters.

In particular, for clustering evaluation, we used HDBSCAN with consistent hyperparameters across all analyses [15]. We note that the usage of HDBSCAN requires parameter optimization for optimal performance. The implementation tested multiple combinations of hyperparameters, including minimum cluster size, minimum samples, and cluster selection epsilon values. We show an overview of the hyperparameters in Figure 4.

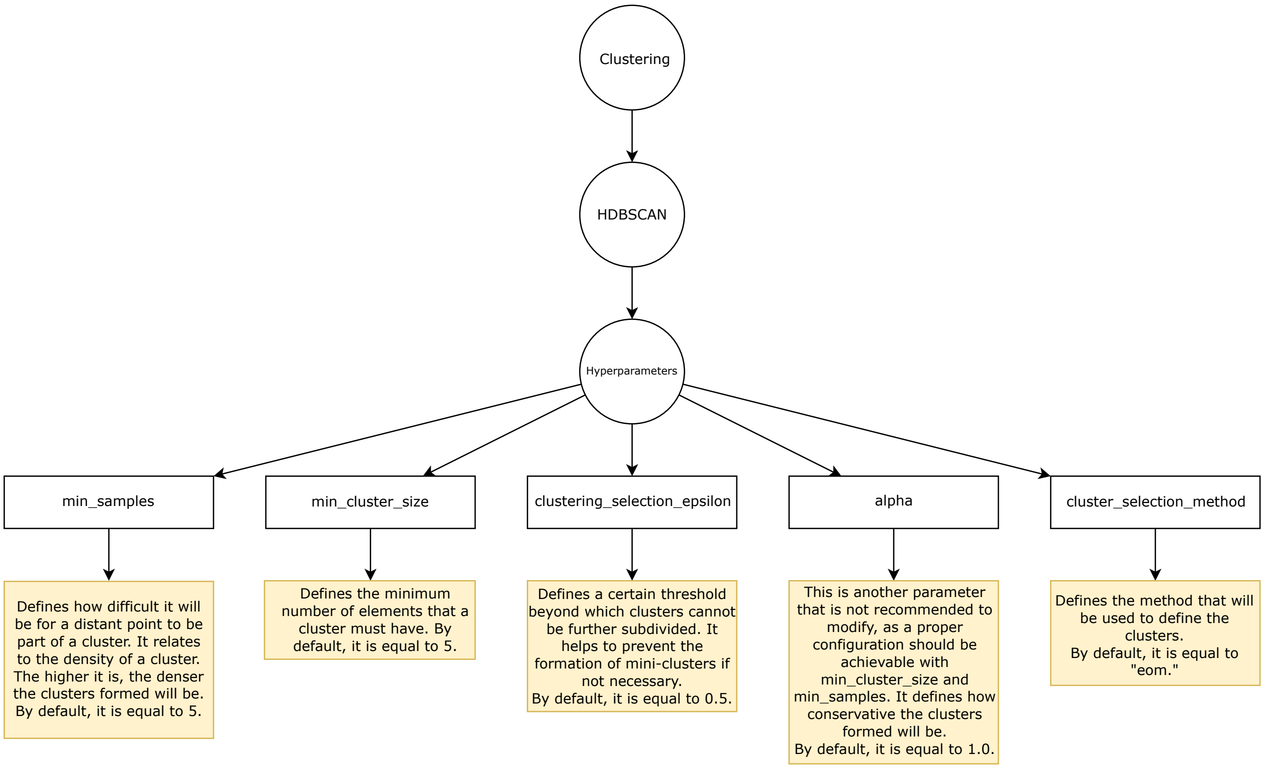
Clustering
HDBSCAN
Hyperparameters
min_samples
min_cluster_size
clustering_selection_epsilon
alpha
cluster_selection_method
Defines how difficult it will be for a distant point to be part of a cluster. It relates to the density of a cluster. The higher it is, the denser the clusters formed will be. By default, it is equal to 5.
Defines the minimum number of elements that a cluster must have. By default, it is equal to 5.
Defines a certain threshold beyond which clusters cannot be further subdivided. It helps to prevent the formation of mini-clusters if not necessary. By default, it is equal to 0.5.
This is another parameter that is not recommended to modify, as a proper configuration should be achievable with min_cluster_size and min_samples. It defines how conservative the clusters formed will be. By default, it is equal to 1.0.
Defines the method that will be used to define the clusters. By default, it is equal to "eom."

Figure 4. Hyperparameters considered for HDBSCAN.

The final clustering configuration balanced cluster coherence with meaningful group sizes, producing interpretable clusters that facilitated expert analysis of narrative patterns within the constitutional process coverage.

4.1. Text Analysis

The implementation of the text analysis utilized two primary models: BERT and RoBERTa. The RoBERTa model, which contains 560 million parameters, demonstrated superior performance in capturing semantic relationships within the content of the constitutional process. We showcase the text-based validation methodology in Figure 5.

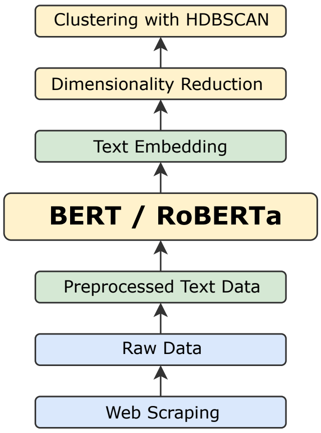

Figure 5. Text-based validation and analysis of the extracted data.

4.1.1. Text Analysis Results

Initial text processing generated embeddings through both models, and was followed by dimensionality reduction using t-SNE [16] and UMAP [17] techniques. The UMAP implementation produced more concentrated clusters compared to t-SNE, while maintaining meaningful distance relationships between data points.

Quantitative evaluation of the clustering results revealed significant differences between the models. RoBERTa achieved superior performance across key metrics, as shown in Tables 5 and 6.

Table 5. Comparative model performance with t-SNE.

Metric	BERT	RoBERTa
Average Distance	88.4007	83.5239
Minimum Distance	0.4746	0.3482
Maximum Distance	276.6690	264.5120
Silhouette Score	0.6445	0.6999

Table 6. Comparative model performance with UMAP.

Metric	BERT	RoBERTa
Average Distance	7.3546	5.9415
Minimum Distance	0.0007	2.4616
Maximum Distance	25.3451	23.6373
Silhouette Score	0.6593	0.6728

In particular, with t-SNE, RoBERTa achieved a silhouette score of 0.6999 compared to BERT's 0.6445, indicating more defined clusters. RoBERTa also produced lower distance measurements, with an average distance of 83.5239 versus BERT's 88.4007, suggesting more compact groupings.

Under UMAP, both models maintained cluster quality but with reduced distances. RoBERTa's silhouette score of 0.6728 exceeded BERT's 0.6593, while maintaining consistent minimum distances above 2.4. These metrics support RoBERTa's selection for subsequent analysis tasks. The distance metrics indicate that UMAP produced more concentrated clusters than t-SNE while preserving data relationships.

4.1.2. Expert-Based Validation

A journalism and communications expert (co-author of this publication) analyzed cluster contents to verify the computational findings.

The expert examined Cluster 5 (see Figure 6), which contains 11 articles about economic and political issues during the constitutional process. Through systematic content analysis, the expert identified two distinct narrative patterns within the RoBERTa results:

1. **Political Polarization Group** (entries 130, 97, 135, 18, 132): Demonstrated clear political polarization patterns and coordinated discrediting campaigns. In particular, the expert identified the following:
    - Consistent use of negative framing in headlines.
    - Repeated references to specific political figures.
    - Similar rhetorical devices across articles.
    - Temporal clustering of publication dates.
2. **Economic Impact Group** (entries 146, 153, 61, 166, 158, 144): Focused on citizen concerns regarding economic implications, particularly related to pension funds and financial stability. In particular, the expert identified the following:
    - Focus on economic impact predictions.
    - Citations of similar expert sources.
    - Common metaphors about financial stability.
    - Parallel structure in presenting opposing viewpoints.

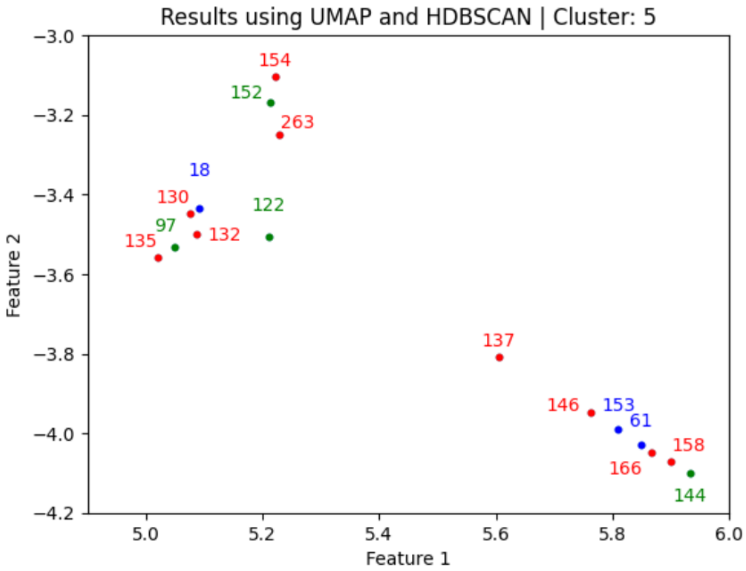

Figure 6. Plot of Cluster 5 for expert-based validation. The colors were added to make it easier to match the document IDs with their corresponding points.

These expert-identified patterns align with our quantitative metrics. The high silhouette score (0.6999) of Cluster 5 corresponds to the expert's observation of clear thematic separation. The average intra-group distance (5.9415) reflects the expert-noted similarity in narrative construction within each group.

Thus, the expert validation confirmed that the clustering results captured meaningful relationships between articles, aligning with established research on polarization in electoral campaigns [18]. This validation supports the effectiveness of the RoBERTa model in identifying coherent narrative structures within the constitutional process coverage.

4.1.3. Insights from Text-Based Validation

The complementary relationship between our quantitative and qualitative validation provides three key insights:

- Clustering metrics identify potential narrative groups, which expert analysis can then verify through close reading.
- Expert-identified patterns help explain why certain articles cluster together, moving beyond purely statistical relationships.
- The alignment between computational and expert analysis suggests our embeddings capture meaningful semantic relationships in how information spreads during political events.

This multi-method validation approach demonstrates that our dataset and analysis methods can identify both statistical and narratively meaningful patterns in constitutional process coverage.

4.2. Image Analysis

The image analysis utilized EfficientNet for the visual content processing, which was selected specifically for its speed and representation quality compared to other models. We note that to ensure that all images were properly associated with the corresponding image we had to generate "duplicate" embeddings, as each news article could have multiple images associated with it. We show this approach in Figure 7.

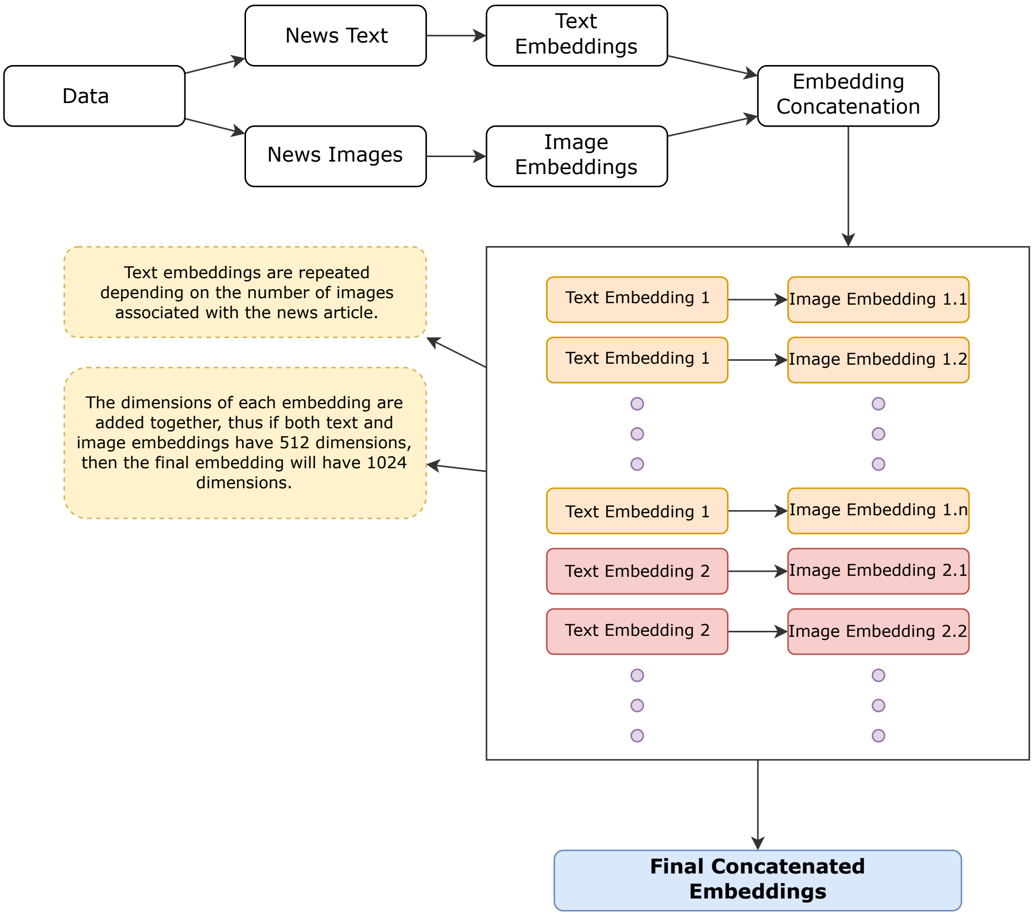
Data
News Text
News Images
Text Embeddings
Image Embeddings
Embedding Concatenation
Text embeddings are repeated depending on the number of images associated with the news article.
The dimensions of each embedding are added together, thus if both text and image embeddings have 512 dimensions, then the final embedding will have 1024 dimensions.
Text Embedding 1
Image Embedding 1.1
Text Embedding 1
Image Embedding 1.2
Text Embedding 1
Image Embedding 1.n
Text Embedding 2
Image Embedding 2.1
Text Embedding 2
Image Embedding 2.2
Final Concatenated Embeddings

Figure 7. Concatenation of text and image embeddings approach in cases with multiple images per news article.

The embeddings generated with EfficientNet were then used in two distinct approaches:

1. **Hybrid Modeling**: Image embeddings were concatenated with textual embeddings from RoBERTa, enabling analysis of text–image relationships within the news articles. We showcase the hybrid approach in Figure 8.
2. **Specialized Modeling**: The CLIP model (openai/clip-vit-base-patch32) processed both images and text in a unified framework, providing an alternative approach to multimodal analysis. We showcase the fully multimodal approach in Figure 9.

4.3. Multimodal Analysis

The multimodal analysis implemented two distinct approaches to examine text–image relationships in the coverage of the constitutional process. The first approach concatenated embeddings from RoBERTa (text) and EfficientNet (images), while the second used the CLIP model for unified multimodal processing.

The concatenation approach linked each text entry with its associated images, resulting in multiple data points for articles containing several images. This methodology led to repeated text embeddings according to the number of associated images in each article, creating a mixed representation of text–image relationships. We show the overview of this approach in Figure 8.

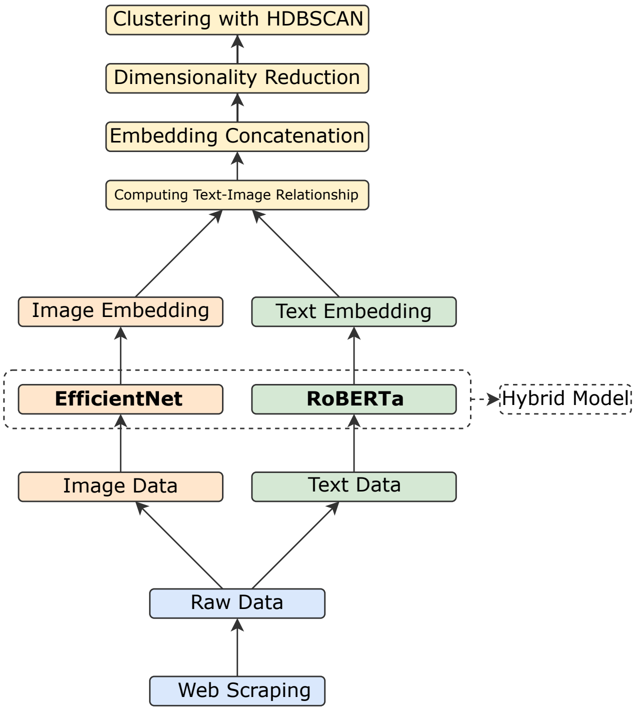

Figure 8. Hybrid-based validation and analysis of the extracted data.

The implementation of the CLIP model (openai/clip-vit-base-patch32) processed both modalities within a unified framework. Analysis of the resulting embeddings, following dimensionality reduction with UMAP, revealed more concentrated clusters compared to the concatenation approach, though both methods identified clear groupings within the data. We show the overview of this approach in Figure 9.

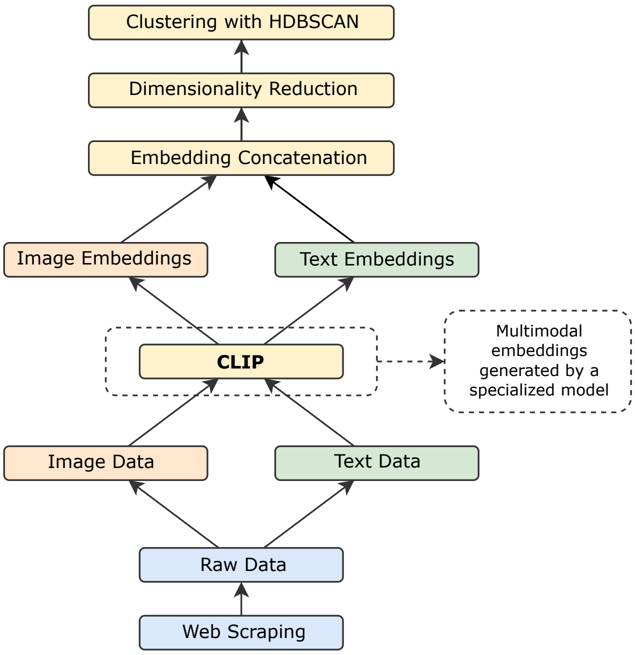

Figure 9. Multimodal-based validation and analysis of the extracted data.

The cluster analysis, implemented through HDBSCAN, identified three distinct groups in the multimodal representations with the dimensionality-reduced feature space. The configuration achieved a silhouette score of 0.8496, indicating robust cluster separation.

Multimodal Models Validation

In general, CLIP-based embeddings demonstrated particular effectiveness in maintaining the proximity between related content, with images associated with the same text typically appearing close to each other within the reduced-dimensional space.

Furthermore, the cluster analysis produced a dominant group that contains approximately 40% of the data points, a secondary group with 30%, and a small group of outlier points.

The distribution of points in the feature space demonstrates clear boundaries between clusters with minimal overlap, suggesting that CLIP successfully captured consistent patterns in the relationships between textual and visual elements.

This structural clarity in the embedding space provides a foundation for future detailed content analysis by domain experts to interpret the semantic significance of these groupings within the context of constitutional process coverage.

5. Discussion

This dataset provides a structured foundation for analyzing misinformation patterns during Chile's constitutional processes. The combination of fact-checked content with multimodal analysis capabilities enables several research applications while acknowledging specific limitations.

5.1. Example Usage Scenarios

The dataset structure supports multiple research applications in computational social science and misinformation studies. Researchers can utilize multimodal embeddings to investigate how visual and textual elements combine to spread false information. The inclusion of fact-checking verdicts from established platforms enables the study of verification methodologies and their effectiveness in constitutional discourse.

The temporal coverage that spans both constitutional processes provides opportunities for comparative analysis. Researchers can examine how misinformation patterns evolved between the two processes, potentially identifying changes in narrative strategies or the prevalence of themes.

Concrete Data Example

The structured format of the dataset facilitates integration with existing natural language processing and computer vision frameworks. The standardized JSON structure and precomputed embeddings enable direct application of analysis tools without extensive preprocessing requirements. We show a concrete example in Figure 10.

For instance, considering the previous example, the dataset shows how routine procedural decisions, such as referring property rights proposals to appropriate committees, generated widespread misinformation. Researchers can examine how verification status, source platform, and temporal patterns correlate with spread dynamics.

In general, the dataset structure supports research applications through its multimodal content. Researchers could trace how a misinterpreted Commission vote transforms into viral misinformation. For example, the JSON format allows examination of how the original claim text correlates with social media amplification patterns. Meanwhile, the linked images enable analysis of visual elements that accompany false narratives.

```
{
    "newscast": "www.fastcheck.cl",
    "title": "Constitutional Convention eliminates private property rights: #False",
    "date": "10/02/2022",
    "description": "Social media claims about Constitutional Convention
                   eliminating private property rights",
    "veracity": "False",
    "text": "The fact-check revealed the Commission of Knowledge Systems
            rejected a proposal because it fell outside their jurisdiction,
            not due to its content. The˜proposal belonged in the
            Fundamental Rights Commission.",
    "images": [
        {
            "image": "path/to/image.jpg",
            "text": "Tweet from Constitutional Convention account"
        }
    ]
}
```

Figure 10. Example entry from the dataset showing the JSON structure used to store fact-checking information and associated media content.

5.2. Topical Diversity

The dataset captures a diverse range of topics related to the constitutional processes of Chile. There are documents related to constitutional process mechanics that focus on electoral procedures, convention protocols, and the interpretations of the underlying legal framework. There are documents related to statements from political actors, which encompass fact-checked claims from public officials, political parties, and commentaries from experts. Furthermore, the are documents related to social issues, such as the proposed changes to healthcare, education, indigenous rights, and environmental protections. Finally, there are documents focusing on the economic implications that discuss property rights, pension reforms, natural resource management, and fiscal policy proposals.

5.3. Broader Impact Considerations

This dataset contributes to the understanding of misinformation dynamics in significant political processes. The findings from expert validation demonstrate how coordinated narrative patterns emerge in constitutional debates, particularly regarding economic implications and political polarization.

The multimodal approach addresses a gap in current research methodologies. Although previous studies have focused on the textual or visual aspects of misinformation, this dataset enables investigation of their interaction. This capability proves particularly relevant given the increasing prevalence of visual elements in social media disinformation.

The inclusion of multiple verification sources provides insight into how different fact-checking methodologies approach similar content. This aspect supports research on fact-checking practices and their role in public discourse during constitutional processes.

5.4. Ethical Usage of the Dataset

This dataset contains misinformation from Chile's constitutional processes and thus requires protocols for responsible use. The inclusion of false claims serves research purposes, but requires precautions against propagation. Users must maintain source attributions and fact-check classifications to prevent decontextualization of the content. We implement safeguards through dataset structure and documentation. Each entry preserves the verification status and platform metadata.

5.5. Limitations

5.5.1. Dataset Issues

The dataset composition presents opportunities for future expansion. In particular, the composition of the dataset reflects the inherent constraints in the media landscape of Chile. First, the source distribution presents an inherent limitation, with *Fast Check* providing substantially fewer entries (36) compared to *Fact Checking UC* (139) and *BioBioChile* (125).

The distribution of the sources matches the availability of fact-checking resources for constitutional processes in Chile that could be scraped with relative ease. Furthermore, this imbalance reflects the platforms' different approaches to content verification but may impact comparative analyses. Future work could incorporate additional sources to broaden the coverage of fact-checking approaches.

The temporal scope of data collection focused on constitutional process periods to maintain specificity in our analysis. Including content from transition periods in future research could enrich our understanding of narrative development across extended timeframes.

5.5.2. Validation Issues

Our validation methodology demonstrated the effectiveness of computational analysis for both textual and visual content. Expert validation of text-based clustering confirmed the detection of meaningful patterns. Future work could extend this expert validation to multimodal representations, providing insight into cross-modal narrative structures.

5.5.3. Web Scraping Implementation

The web scraping implementation successfully collected content from static webpage elements. Future technical developments could incorporate dynamic content processing to expand data collection capabilities. The current process for associating multiple images with text segments could benefit from automation to increase processing efficiency.

5.6. Future Work

There are several promising directions for future work. Extended expert validation and automated processing methods would enhance the analysis of narrative patterns. Integration of additional temporal periods and technical capabilities would deepen our understanding of information flow during constitutional processes.

Future research could also expand the model comparison beyond BERT and RoBERTa to include more advanced models such as XLM, GPT, and other transformer architectures. For multimodal analysis, comparisons between CLIP and alternatives like DALL-E or Stable Diffusion could reveal performance differences across embedding techniques.

We note that the limitations of this work do not diminish the dataset's utility for research purposes, but should inform methodological decisions in its application. Future work might address these constraints through expanded data collection or refined multimodal analysis techniques.

6. Conclusions

This research presents a multimodal dataset capturing fact-checked news content from Chile's constitutional processes. Through the implementation of advanced natural language processing and computer vision techniques, we have demonstrated the feasibility of analyzing complex political narratives using a combination of textual and visual data.

The primary contribution of the dataset lies in its structured approach to capture multimodal misinformation patterns. The integration of 300 fact-checked entries, including 168 with associated images, provides researchers with verified content for studying how false information propagates during significant political events. The inclusion of content from multiple verification platforms enables a comparative analysis of fact-checking methodologies and their effectiveness.

Our analysis revealed the effectiveness of transformer-based approaches in capturing textual patterns through measurable improvements in cluster cohesion. RoBERTa embeddings achieved higher silhouette scores than BERT, indicating better capture of semantic relationships. The expert validation of text-based clusters confirmed that computational grouping identified meaningful narrative patterns, particularly in coverage of economic and political issues. While our dataset spans constitutional processes through text and images, our validation focused primarily on textual content. The multimodal analysis shows technical promise through clustering metrics, but requires further expert evaluation to verify semantic coherence.

The dataset provided insights into information patterns during Chile's constitutional processes, though with constraints in source distribution. While we collected 300 articles across three sources, the uneven distribution between Fast Check and Fact Checking UC affects the breadth of our analysis.

This dataset serves as a foundation for understanding how misinformation manifests in multimodal forms during crucial political processes. The methodologies developed for their creation and analysis provide a framework for similar efforts in other contexts, contributing to the broader field of computational social science and misinformation studies.

Author Contributions: The authors contributed to this research as follows: Conceptualization, B.K.; methodology, B.K.; software, I.M.; validation, I.M., B.K. and M.M.; formal analysis, I.M.; investigation, I.M.; resources, B.K.; data curation, I.M.; writing—original draft preparation, I.M. and B.K.; writing—review and editing, I.M., B.K. and M.M.; visualization, I.M.; supervision, B.K.; project administration, B.K. Expert validation and knowledge contribution was provided by M.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Project 202311010033-VRIDT-UCN.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are publicly available at the following repository https://github.com/MolodyGs/CapstoneProject (accessed on 11 December 2024).

Acknowledgments: The authors would like to thank Claudio Meneses, Diego Urrutia, and Eduardo Aguilar for their feedback and discussion throughout the project.

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

Notes

[1] https://www.fastcheck.cl/ (last accessed on 11 December 2024).

[2] https://factchecking.cl/ (last accessed on 11 December 2024).

[3] https://www.biobiochile.cl/ (last accessed on 11 December 2024).

[4] The code used to extract all data is publicly available at https://github.com/MolodyGs/CapstoneProject (accessed on 11 December 2024).

References

1. Porter, E.; Wood, T.J. Political misinformation and factual corrections on the Facebook news feed: Experimental evidence. *J. Politics* **2022**, *84*, 1812–1817. [CrossRef]
2. Seo, D.H. Much Ado about Disinformation: A Critical Approach to Coping with Information Manipulation in a Post-Truth World. *Sci. Mil. S. Afr. J. Mil. Stud.* **2024**, *52*, 105–120. [CrossRef]
3. Shen, C.; Kasra, M.; Pan, W.; Bassett, G.A.; Malloch, Y.; O'Brien, J.F. Fake images: The effects of source, intermediary, and digital media literacy on contextual assessment of image credibility online. *New Media Soc.* **2019**, *21*, 438–463. [CrossRef]

4. 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. [CrossRef]
5. 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. [CrossRef] [PubMed]
6. Vosoughi, S.; Roy, D.; Aral, S. The spread of true and false news online. *Science* **2018**, *359*, 1146–1151. [CrossRef] [PubMed]
7. Keith Norambuena, B.F.; Mitra, T.; North, C. A survey on event-based news narrative extraction. *ACM Comput. Surv.* **2023**, *55*, 1–39. [CrossRef]
8. 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. [CrossRef]
9. Devlin, J. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
10. Liu, Y. Roberta: A robustly optimized bert pretraining approach. *arXiv* **2019**, arXiv:1907.11692.
11. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning, Virtual, 18–24 July 2021 ; PMLR; pp. 8748–8763.
12. Delucchi, A.S.; Ugarte, V.R. The Chilean constitutional process narrated through a spiral. *Stud. Soc. Justice* **2024**, *18*, 969–991. [CrossRef]
13. Shao, C.; Ciampaglia, G.L.; Varol, O.; Yang, K.C.; Flammini, A.; Menczer, F. The spread of low-credibility content by social bots. *Nat. Commun.* **2018**, *9*, 4787. [CrossRef] [PubMed]
14. Tan, M.; Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; PMLR; pp. 6105–6114.
15. McInnes, L.; Healy, J.; Astels, S. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.* **2017**, *2*, 205. [CrossRef]
16. Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
17. McInnes, L.; Healy, J.; Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv* **2018**, arXiv:1802.03426.
18. Lindh, J.; Fabrega, J.; Gonzalez, J. The Fragility of Consensus: Ideological Polarization in Post-Pinochet Chile. *Rev. Cienc. Política* **2019**, *39*, 99–127.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.