NewsFraud-SC: A Dataset of News about Corruption Extracted from SC Multiple Portals

Ana Clara Stupp de Souza[1], **Carina F. Dorneles**[1]

[1]Departamento de Informática e Estatística
Universidade Federal de Santa Catarina (UFSC) – Florianópolis – SC – Brazil

ana.stupp02@gmail.com, carina.dorneles@ufsc.br

Abstract. *This article presents NewsFraud-SC, a dataset composed of news articles about illicit activities in the public sector, originating from multiple news portals in the state of Santa Catarina. With a total volume of 70 MiB, the dataset was constructed using an automated collection pipeline that integrates web scraping techniques, content cleaning via readability algorithms, and Named Entity Recognition (NER). Each record in the dataset is structured in JSON format, containing detailed metadata such as source, URL, and date, as well as previously extracted entities, including people, organizations, locations, and monetary values involved in the reported incidents. The work describes the collection methodology, descriptive statistics of the corpus, highlighting the distribution of topics and the density of entities. With 10,913 extracted entities and well-defined metadata, this corpus allows researchers to bypass the traditional data engineering bottleneck typically associated with web scraping and manual content cleaning. As a scientific contribution, the dataset is made available to the academic community, aiming to foster research in the extraction of complex relationships and government transparency technologies.*

1. Introduction

Corruption and illicit activities in the public sector represent one of the greatest challenges to the transparency and integrity of democratic institutions, and are frequently exposed through journalistic investigations [Lyra et al. 2022]. Large volumes of news are generated daily, reporting cases of mismanagement, bid rigging, and misappropriation of funds, making the monitoring of these events an essential task for oversight bodies and civil society in monitoring public ethics [Weichselbraun et al. 2020].

Widespread visibility of news about fraud involving public funds is essential so that oversight bodies and civil society can monitor public ethics and ensure the transparency of democratic institutions. However, systematic access to this information is hampered by the fragmentation of data on the web. Each news portal has its own architecture, which imposes significant technical barriers to manual and structured data collection. The noise present on websites, the lack of standardization, and the volatility of links make building a reliable database a complex data engineering problem [Sizov et al. 2003]. In this scenario, the consolidation of a structured and persistent dataset becomes indispensable: it not only preserves the memory of investigative events that could be lost with the deletion of web pages, but also serves as the necessary foundation for training AI models capable of identifying corruption patterns and correlations between different cases, transforming isolated texts into actionable intelligence for public transparency.

In this context, this article proposes NewsFraud-SC, a structured dataset composed of news from portals in Santa Catarina about illicit activities in the public sector. The paper describes an automated collection methodology that uses web scraping techniques, readability algorithms for content cleaning, and Named Entity Recognition (NER) to extract detailed metadata. With 70 MiB of data organized in JSON format, 10,913 extracted entities and a good set of metadata, the resource is presented as a scientific contribution to the development of intelligent systems, aiming to strengthen research in government transparency and technologies applied to the legal domain[1].

We believe that making the NewsFraud-SC dataset available lies in the need for structured, high-quality data to bridge the gap between journalistic reporting and investigative intelligence. Public sector corruption and mismanagement are frequently exposed through media coverage, yet this information remains fragmented across disparate portal architectures with no standardized format. By providing a curated corpus with 10,913 extracted entities and high metadata completeness, this dataset enables researchers to overcome the data engineering bottleneck inherent in web scraping and content cleaning. Scientifically, NewsFraud-SC serves as a benchmark for evaluating Named Entity Resolution solutions within the legal and investigative domains, fostering the development of AI models capable of identifying complex corruption patterns and ensuring long-term digital preservation of critical public ethics records.

2. NewsFraudSC

The NewsFraud-SC is a structured dataset composed of news articles from multiple portals in Santa Catarina, specifically 43 unique newspapers portals, concerning illicit activities in the public sector. Before to characterize this resource, we present the data collection methodology, the data dictionary, and some statistics description of the dataset.

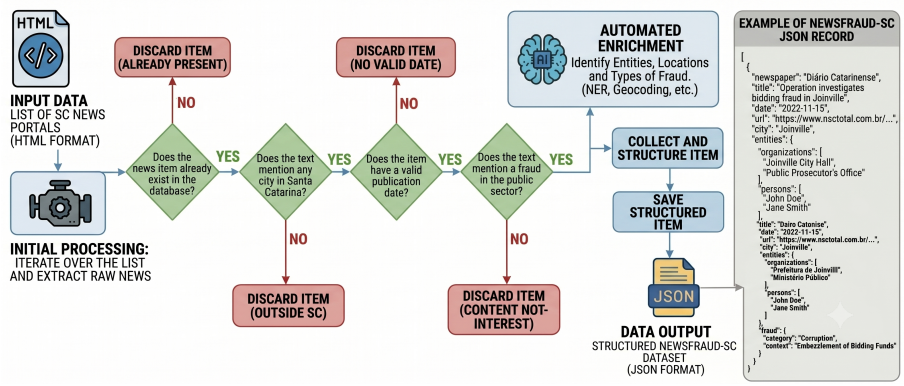

Figure 1. Data Collection Methodology

2.1. Data Collection Methodology

Figure 1 illustrates the steps of the data collection methodology. The process begins with URL analysis: since titles are contained within web addresses, the system performs keyword matching to identify the topic before accessing the page. If the news item is relevant and not yet in the capture history (*processed_urls.json*), the tool analyzes the

[1]Code, data and everything else is at https://github.com/dornelescfd/cono

HTML, extracting only the main text (without menus and ads). Next, a geographic filter is applied based on a list of cities in Santa Catarina. If the content does not have a local link, it is ignored to ensure that only cases within the state are processed. The resulting structured data is stored in *corruption_articles.json* and integrated into the application's frontend. Methodology details have been published in [Souza and Dorneles 2025].

2.2. Dataset and Dictionary

The structural composition of the NewsFraud-SC dataset is detailed in Table 1, which provides the data dictionary and the corresponding JSON schema for each record. The schema encompasses primary metadata, such as `source`, `url`, and publication `date`, alongside the raw and sanitized versions of the news content, `text` and `content` respectively. Furthermore, it specifies the structured attributes resulting from the NER and extraction pipeline, including arrays for geographical locations thought `cities`, `people`, `organizations`, and monetary `values`. This standardized format ensures interoperability for downstream analytical tasks and provides a clear mapping for the identified corruption-related categories and legal actors involved in the reported incidents.

Table 1. Data Dictionary and JSON Schema for the NewsFraud-SC Dataset.

JSON Field	JSON Data Type	Description
newspaper	string	Name of the news source or media outlet.
title	string	Article headline (may contain the URL as a placeholder in some entries).
url	string	Direct permalink to the original news article.
text	string	Raw HTML source code collected from the page.
date	string	Publication date as identified in the document.
content	string	Sanitized and normalized main body text (structure-free).
cities	array	List of identified municipalities from Santa Catarina.
people	array	Named entities identified as "Person" via NER.
organizations	array	Named entities identified as "Organization" via NER.
fraud	array	Detected corruption-related keywords or fraud categories.
value	array	Monetary values and financial amounts mentioned in the text.
prosecutor	string/null	Name of the Prosecutor or Attorney identified in the case.

2.3. Dataset Statistics

The dataset analysis comprises a total of 377 records sourced from 43 unique news outlets across Santa Catarina. A significant concentration of data is observed in the distribution of occurrences, with **ND Mais**[2] emerging as the primary source, accounting for 320 of the

[2]https://ndmais.com.br/

analyzed entries. The remaining records are distributed across a diverse range of 42 local and regional platforms, such as Portal Menina, SC Todo Dia, and Tudo Sobre Floripa, most of which contribute between one and three occurrences each. This distribution highlights a centralized reliance on major state-wide digital portals while maintaining a broad, multifaceted reach into smaller regional journalistic sources, ensuring a comprehensive geographical coverage of the reported irregularities.

Table 2 presents a quantitative breakdown of named entities, comprising People, Organizations, and Fraud types, identified across the top 11 Newspaper portals within the NewsFraud-SC dataset. A significant disparity in data volume is observed, with ND Mais emerging as the dominant source, contributing 4,409 mentions of people and 4,213 of organizations. This outlier status reflects the portal's extensive coverage of public sector irregularities in Santa Catarina compared to regional outlets. While portals such as A Notícia and Jornal AW show a more balanced distribution of entities, others like SC em Pauta and Portal Menina exhibit a higher density of organizational mentions relative to individual names, highlighting variations in reporting styles and the focus on institutional involvement in corruption cases.

Table 2. Distribution of Named Entities (People, Organizations, and Fraud Types) by Newspaper Portal in the NewsFraud-SC Dataset.

Newspaper	People	Organizations	Fraud Types
ND Mais	4,409	4,213	342
A Notícia	40	41	3
Jornal AW	38	66	5
SC em Pauta	31	77	4
Linha Popular	29	20	2
Portal Minutta	21	18	3
Jornal Folha Videira	18	20	2
Tudo Sobre Floripa	16	20	4
Portal Menina	14	65	6
4oito	13	14	1
Portal GCD	13	45	6

Table 3 shows high information density and significant lexical diversity. On a global scale, the dataset achieves an average of 23.43 entities per document, with an overall NER Density of 6.17 entities per 1,000 characters. At the source level, Linha Popular presents the highest mean NER Density (10.22), followed closely by ND Mais (7.74), which also displays a higher standard deviation (3.75) due to its larger and more heterogeneous volume of reports. Furthermore, the Average Vocabulary Richness (TTR) of 0.5617 suggests that the dataset is composed of diverse journalistic narratives rather than repetitive templates. The variance in density and TTR across the top 11 newspapers highlights the complexity of the corpus.

3. Characterization

The characterization of the NewsFraud-SC dataset is conducted through a comprehensive analysis that evaluates its structural, thematic, and economic dimen-

Table 3. Global and Detailed Metadata Statistics for the NewsFraud-SC Dataset

Global Dataset Metrics			
Average Entities per Document		23.43	
Overall NER Density (Mean)		6.17	
Overall Vocabulary Richness (TTR)		0.5617	
Detailed Metrics by Top Newspapers			
Newspaper	**NER Density (Mean)**	**NER Density (Std Dev)**	**Avg TTR**
Linha Popular	10.22	0.71	0.60
ND Mais	7.74	3.75	0.58
Informe Floripa	7.33	—	0.72
Destaque Sul	7.22	—	0.68
Notícias Governo SC	6.27	—	0.63
Portal Aconteceu	5.69	—	0.64
Jornal Metas	4.74	—	0.68
Jornal Folha Videira	4.47	6.24	0.41
Diário Catarinense	4.27	—	0.68
J. Vale do Itapocu	4.26	1.22	0.66
Tribuna de Itapoá	4.23	—	0.64

sions. To ensure the corpus's reliability for downstream tasks such as Retrieval-Augmented Generation (RAG), we employ feature completeness and NER density metrics [Jurafsky and Martin 2026], measuring the concentration of actionable facts per record. The geographical and institutional landscape is mapped via spatial heatmaps and entity incidence analysis, while the dataset's diversity is quantified through Shannon Entropy and Simpson Diversity [Jost 2006] indices to detect potential source biases. Furthermore, the economic impact is assessed using logarithmic distributions of monetary values, and thematic consistency is explored through Pareto-style categorical sparsity and document size granularity analysis. Together, these metrics provide a multidimensional validation of the dataset's representative capacity and its utility for investigative intelligence

3.1. Named Entities

Figure 2 provides a quantitative overview of the named entities, People and Organization, with the highest incidence in the analyzed dataset, revealing a landscape heavily shaped by law enforcement and regulatory bodies. On the institutional front (Top 10 Organizations), prominent roles are held by the MPSC (with 150 mentions) and the "Polícia Civil" (Civil Police), followed by specialized units such as GAECO. This distribution suggests an information flow centered on investigative and judicial activities. Simultaneously, the analysis of the most cited individuals (Top 10 People) highlights key figures of regional and national political relevance, such as Renê, Ed Pereira, and Raimundo Colombo. It is worth noting that related acronyms and names (e.g., "PF" and "Polícia Federal" (Federal Police)) have been consolidated in these visualizations to ensure data accuracy and reflect the true density of the institutional and political themes within the dataset. Deduplication was not the focus of this work.

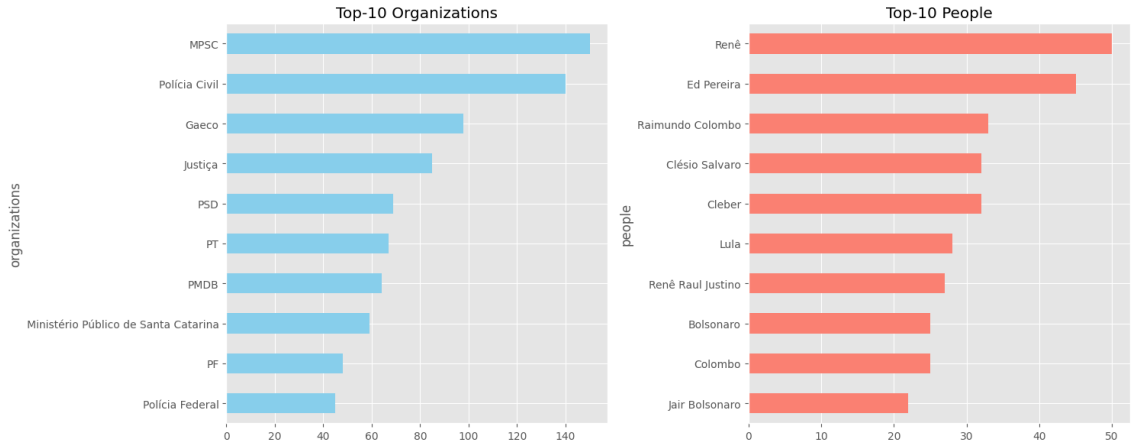

Figure 2. Top-10 Organizations and top-10 People

Figure 3 shows the Named Entity Recognition (NER) Density Report, which indicates a high concentration of actionable information across the dataset, with a total of 10,913 entities successfully extracted. The Mean Density of 6.37 entities per 1,000 characters and a Median Density of 6.26 entities per 1,000 characters demonstrate a consistent and robust presence of factual data, such as people, organizations, and cities—throughout the records. Furthermore, the observation of a Max Density reaching 22.30 entities per 1,000 characters highlights specific high-density fact centers that are particularly valuable for downstream tasks like Retrieval-Augmented Generation (RAG) systems, where information richness is critical for accuracy.

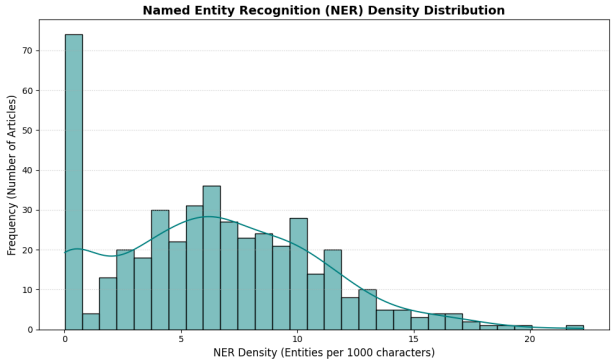

Figure 3. NER Density Distribution

Metric	Value
Mean Density	6.37
Median Density	6.26
Max Density Observed	22.30
Total Entities Extracted	10,913

Table 4. NER Density Summary

3.2. HeatMap

The geographic distribution of reported irregularities is synthesized in the heatmap presented in Figure 4. This visualization provides a multidimensional characterization of the dataset by correlating identified municipalities with specific categories of administrative misconduct. Beyond a simple frequency count, this graphic reveals regional patterns of corruption and media coverage density. For instance, the high concentration of specific fraud types in the state's capital versus the prevalence of procurement-related mentions in smaller municipalities highlights the dataset's capacity to reflect real-world socioeconomic dynamics. Furthermore, this analysis serves as a validation layer for the crawler's

geographic filtering logic, ensuring that the processed 70.67 MiB of data maintains strict territorial relevance to the state of Santa Catarina while providing a robust foundation for regionalized investigative intelligence

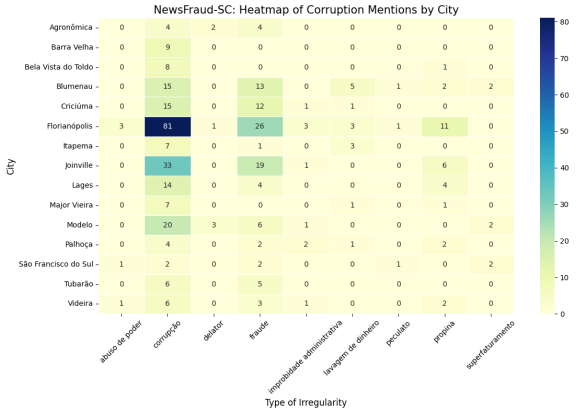

Figure 4. Mentions/City

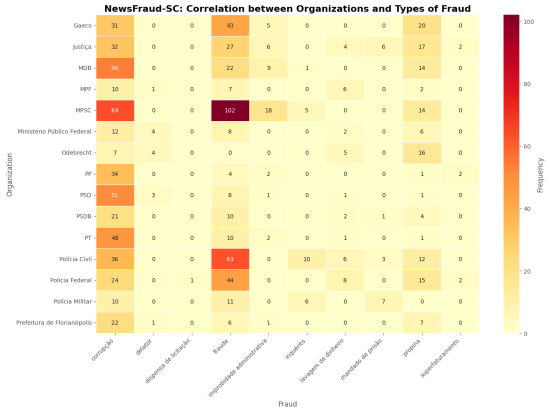

Figure 5. Mentions/Organizations

4. Document Distribution and Fields Completeness

The document length distribution, illustrated in Figure 6, exhibits a unimodal distribution with a positive skew, characterized by a prominent modal peak at approximately 300 words. This concentration suggests a consistent editorial pattern across the regional news portals, where information is typically condensed into concise reports. The presence of a long tail extending towards 5,000 words indicates a subset of extensive investigative pieces or detailed legal transcriptions. From a computational linguistics perspective, this granularity is ideal for Natural Language Processing (NLP) tasks, as the typical document size allows for the extraction of complex relationships and named entities without the risk of context fragmentation, ensuring that semantic dependencies between actors, locations, and monetary values are preserved within a single processing window.

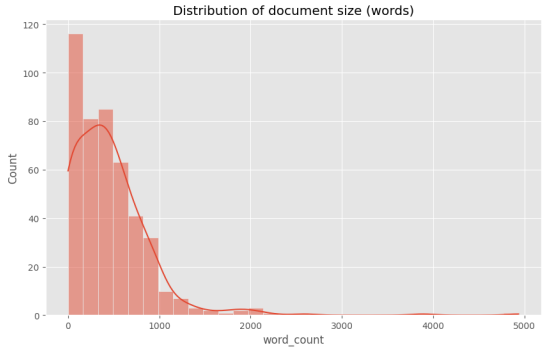

Figure 6. Doc. Distribution

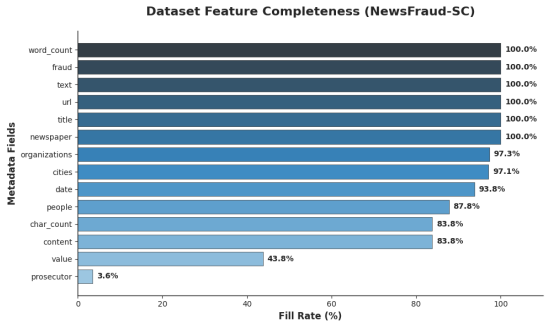

Figure 7. Fields Completeness

The robustness of the automated extraction pipeline is evaluated through the Dataset Feature Completeness analysis, as illustrated in Figure 7. The results demonstrate a 100% completion rate for primary metadata and structural fields, such as newspaper, title, url, text, date, and content, ensuring a baseline of high-quality raw data. In

contrast, attributes dependent on Named Entity Recognition (NER) and heuristic extraction exhibit variable density: value was identified in 43.8% of the records, while people reached 87.8%, and cities reached 97.1%. The prosecutor field currently shows a 3.6% detection rate, indicating either the absence of this specific actor in the source texts or the need for more specialized extraction patterns. This variance in feature density reflects the complexity of identifying specific legal and financial entities within unstructured regional news, providing a transparent measure of the current information coverage within the NewsFraud-SC dataset.

5. Word Cloud

Figures 8 and 9 provide a comparative analysis between the news textual content and the assigned fraud classifications (we kept the terms in Portuguese to remain faithful to the original terms in the dataset.). While the word cloud for the content attribute reveals operational and procedural terms, emphasizing key actors and actions such as ”empresa” (company), ”operação” (operation), ”polícia civil” (civil police), and ”investigação” (investigation), the cloud for the fraud attribute synthesizes the legal and criminal nature of the events. In the latter, the focus shifts toward the typology of the irregularities, with a predominance of the terms ”corrupção” (corruption), ”fraude” (fraud), ”improbidade administrativa” (administrative misconduct), and ”lavagem de dinheiro” (money laundering). This distinction demonstrates that while the news content describes the who and how of the proceedings, the fraud attribute isolates the essence of the reported illicit activity, validating the model's effectiveness in extracting specific categories of financial and administrative crimes.

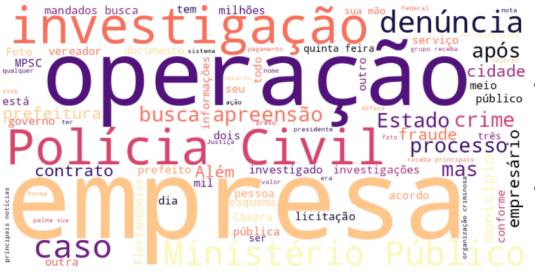

Figure 8. News content attribute

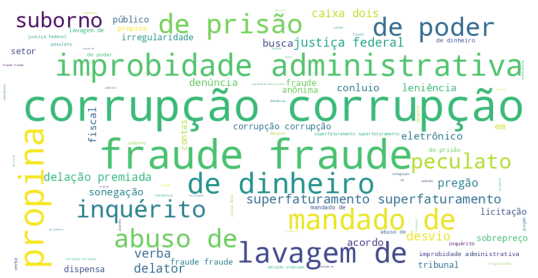

Figure 9. Fraud attribute

5.1. Degree Centrality Co-occurrence

Table 5 illustrates the Degree Centrality Co-occurrence within the NewsFraud-SC dataset, highlighting the organizations that act as structural ”hubs” by aggregating a high number of unique individuals in the same reporting context. With a Global Contextual Connectivity of 12.97, the dataset demonstrates a dense network of relationships, essential for constructing robust Knowledge Graphs. Entities such as Transparência Brasil and specialized task forces like Operação Fundo do Poço exhibit exceptionally high centrality, indicating their role as focal points in investigative narratives where multiple actors are interconnected. The presence of political parties (e.g., União Brasil) with a significant number of total connections (366) across multiple documents further validates the dataset's capacity to capture complex, multi-layered corruption networks, providing the necessary relational depth to evaluate context-aware chunking and prompt engineering strategies.

Table 5. Degree Centrality Co-occurrence: Top Organizations by Connectivity in NewsFraud-SC

Global Connectivity Metric			
Global Contextual Connectivity		12.97 people/org per doc	
Top Organizations by Connectivity (Hubs)			
Organization	**Avg. Degree Centrality**	**Doc. Frequency**	**Total Connections**
Transparência Brasil	80.00	2	160
Consórcio Médio Vale Itajaí	80.00	2	160
Água Azul	72.00	2	144
Operação Fundo do Poço	58.50	4	234
Secretaria de Obras	44.50	2	89
Grupo Serrana Engenharia	43.00	2	86
Prefeitura de Tubarão	41.50	2	83
Evento Olesc 2023	38.00	2	76
Gaeco de Lages	37.33	3	112
PSDB-PR	37.00	2	74
União Brasil	36.60	10	366

5.2. Distribution and Source Diversity

To analyze the economic impact and the scale of the irregularities within the NewsFraud-SC dataset, we utilize a Logarithm Boxplot and a Logarithmic Histogram to represent the distribution and density of the values involved and present in Figure 10. Given the high variance in the data, where the most frequent value is 1.0, while other entries reach significantly higher magnitudes, the logarithmic scale is essential to normalize the visualization and prevent extreme outliers from distorting the graphical representation. The Values Distribution (Boxplot) allows for a clear identification of the median and the dispersion of financial data across 191 unique values, while the Values Density (Histogram) illustrates the concentration of lower, value entries against the sparse but critical high-value occurrences, providing a comprehensive overview of the financial landscape of reported corruption and administrative misconduct in Santa Catarina.

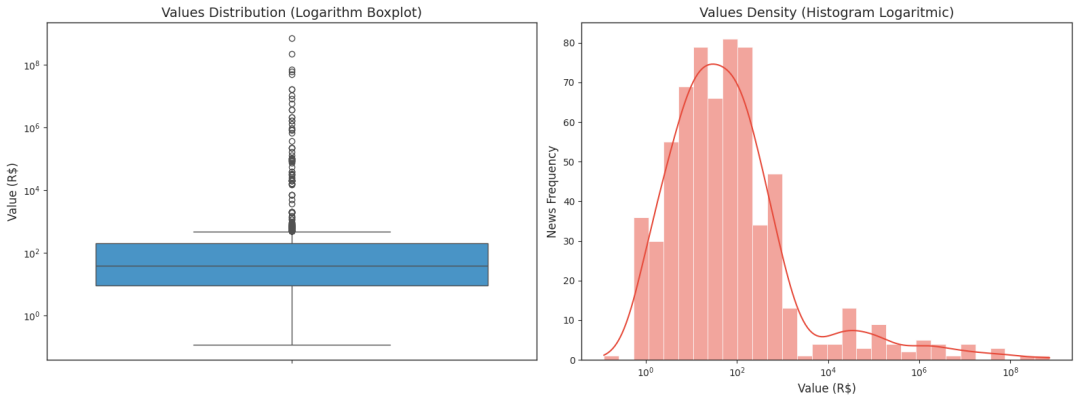

Figure 10. Lorenz Curve: Source Bias Visualization

The Source Diversity and Bias Report, presented in Figure 11, reveals a highly

balanced distribution of information across the analyzed news outlets. With a Shannon Entropy (H) of 2.1640 and a Simpson Diversity (1-D) of 0.8800, the dataset exhibits significant variety, suggesting that no single source disproportionately dominates the narrative. This is further confirmed by the Equitability (Evenness) score of 0.9849, a value near the theoretical maximum of 1.0, which indicates that the 9 unique news outlets contribute almost equally to the total volume of records. These metrics validate the structural reliability of the dataset, ensuring that the subsequent thematic and entity analysis is not heavily skewed by the editorial bias of a specific journalistic source. The Categorical Sparsity and Long Tail Analysis, presented in Figure 12, identifies a significant concentration within the dataset, as the top 20% of categories, specifically the ”fraude” label, account for 60.00% of all recorded occurrences. The remaining 40% of the distribution is spread across a long tail of diverse typologies, including ”propina”, ”corrupção”, ”desvio de verba”, and ”lavagem de dinheiro”, which contribute incrementally to the cumulative total. This Pareto-style distribution highlights a predominant focus on general fraud while maintaining a sparse but varied representation of other illicit activities, illustrating both the thematic core and the categorical sparsity of the reported incidents.

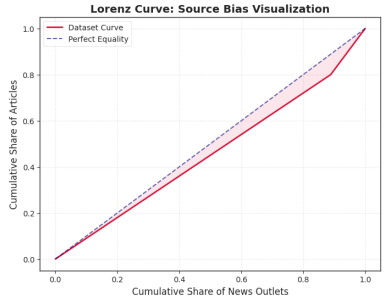

Figure 11. Values Distribution and Density

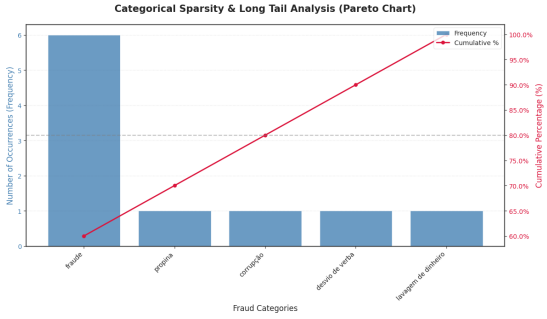

Figure 12. Categorical Sparsity and Long Tail Analysis

6. Download and Citation Request

The NewsFraud-SC is publicly available at `https://github.com/dornelescfd/cono`, containing the crawler code and the code used to generate all the analysis graphs described in this article, as well as other graphs not available here due to space limitations. If this dataset is used for any purpose, please cite this article.

7. Conclusion

This papers demonstrate the mapping of the dynamics of corruption within the state of Santa Catarina. We described the extracted entities, such as organizations and public figures, and correlate them with specific fraud typologies revealed an ecosystem of irregularities closely tied to public procurement and administrative misconduct. Beyond its technical contributions, NewsFraud-SC stands as a vital resource for investigative journalism and social oversight by bringing transparency to unstructured data. Future research may expand this analysis to other regions, improving automated entity resolution and integrating temporal analysis to track the evolution of influence networks over time

Artificial Intelligence Use: we have used the Gemini 3 Flash AI model to assist with grammatical corrections in some sections of this paper.

References

Jost, L. (2006). Entropy and diversity. *Oikos*, 113(2):363–375.

Jurafsky, D. and Martin, J. H. (2026). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, with Language Models*. 3rd edition. Online manuscript released January 6, 2026.

Lyra, M. S., Damásio, B., Pinheiro, F. L., and Bacao, F. (2022). Fraud, corruption, and collusion in public procurement activities, a systematic literature review on data-driven methods. *Applied Network Science*, 7(1):83.

Sizov, S., Graupmann, J., and Theobald, M. (2003). From focused crawling to expert information: An application framework for web exploration and portal generation. In *Proceedings VLDB Conference*, pages 1105–1108. Morgan Kaufmann, San Francisco.

Souza, A. and Dorneles, C. (2025). Cono: Um coletor automatizado de notícias sobre corrupção em santa catarina. In *Anais da XX Escola Regional de Banco de Dados*, pages 129–132, Porto Alegre, RS, Brasil. SBC.

Weichselbraun, A., Hörler, S., Hauser, C., and Havelka, A. (2020). Classifying news media coverage for corruption risks management with deep learning and web intelligence. WIMS 2020, page 54–62, New York, NY, USA. Association for Computing Machinery.