

Social Indicators from News-Based Gender Violence Events

Samuel R. S. Oliveira¹, Daniel Zitei¹, Ricardo M. Marcacini¹

¹Institute of Mathematics and Computer Sciences (ICMC), University of São Paulo (USP)

{samuelrubens,daniel.zitei,ricardo.marcacini}@usp.br

Abstract. *Official indicators are essential for monitoring gender violence, but they often arrive late and provide limited contextual detail about specific events. This paper explores online news as a complementary social sensing layer for producing low-latency indicators of gender violence. We propose a pipeline that collects news reports, retrieves target-related documents, removes duplicates, normalizes temporal and geographic information, and uses language models to classify gender violence events and extract structured metadata. Each event includes attributes such as violence type, location, victim–aggressor relationship, severity, outcome, and event dynamics. Using Brazilian news from 2020 to April 2026 as a case study, we show how structured news-based events can support temporal monitoring, spatial analysis, comparison with official data, and executive summaries of emerging peaks. The results indicate that news can anticipate signals, enrich interpretation, and support public managers with more timely and semantically detailed evidence.*

1. Introduction

The World Health Organization estimates that nearly one in three women worldwide has experienced physical or sexual violence by an intimate partner or sexual violence by a non-partner during her lifetime [WHO 2024 2024]. In Brazil, the problem is monitored through multiple institutional systems, including police records, judicial databases, healthcare notifications, reporting channels, and national statistical initiatives. These systems are essential for official accountability, policy evaluation, and long-term monitoring. Examples include the Brazilian Public Security Yearbook [FBSA 2025], the Atlas of Violence [IPEA 2025], and the National Map of Gender Violence, which integrates data from health, justice, public security, and survey-based sources [Senado Federal 2024].

Despite their relevance, official indicators are not sufficient to support all monitoring needs. Administrative data on violence against women require intersectoral coordination among justice, police, health, social services, statistical agencies, and other institutions [Vasconcelos et al. 2024]. Therefore, official statistics should be complemented by additional observational layers that can provide low-latency signals and richer contextual descriptions of violence-related events.

The web has become an important source of social signals about real-world phenomena [Gôlo et al. 2021]. Previous studies have explored digital traces as social sensors for monitoring epidemics, disasters, elections, and other social processes [Ginsberg et al. 2009, Sakaki et al. 2010, Leetaru and Schrodt 2013]. In this direction, WebSensors introduced the idea of learning sensors from web news events to monitor and understand real-world phenomena from large-scale collections of Portuguese news [Marcacini et al. 2017]. Later, the notion of *learning to sense* was formalized as the task

of learning representations and decision functions over event collections with semantic, temporal, and geographic dimensions [Gôlo et al. 2021]. This conceptual basis is especially relevant for gender violence monitoring because news reports often describe not only that an event occurred, but also where it happened, when it happened, who was involved, what type of violence was reported, and what contextual dynamics were present.

This paper investigates how online news can operate as a complementary social sensing infrastructure for gender violence monitoring. The central idea is that journalistic reports can provide timely and context-rich signals about violence-related events, especially when official statistics are delayed, aggregated, or limited in semantic detail. Instead of treating news only as textual evidence, we model news reports as observations of real-world events that can be collected, filtered, structured, and aggregated into indicators. Our work is grounded in the WebSensors perspective, which frames web news as a source of social signals for monitoring real-world phenomena [Marcacini et al. 2017]. In this view, a computational sensor is not a physical device, but a model or pipeline capable of observing digital traces and identifying events of interest. This perspective was later formalized through the notion of *learning to sense*, in which event collections are represented and analyzed across semantic, temporal, and geographic dimensions [Gôlo et al. 2021]. Gender violence monitoring is particularly aligned with this approach because news reports often describe not only that an event occurred, but also where it happened, when it happened, who was involved, what type of violence was reported, and what contextual dynamics were present.

To operationalize this approach, we use ReVeJo, the *Repositório de Veículos Jornalísticos*, as the news data infrastructure. ReVeJo is a Brazilian news intelligence platform designed to support large-scale social sensing from journalistic content by collecting news from multiple media outlets, normalizing and deduplicating articles, extracting metadata, entities, locations, temporal references, 5W1H attributes, thematic information, and causal relations, and transforming unstructured news into structured knowledge representations [ReVeJo 2025]. Unlike traditional news repositories focused mainly on storage and retrieval, ReVeJo provides an analytical infrastructure for monitoring events, comparing narratives, studying media coverage, and supporting public policy analysis and computational social science research. Building on ReVeJo and on the WebSensors perspective, we propose a pipeline that converts news reports into structured gender violence events and then aggregates these events into social indicators. This paper makes the following contributions:

- We propose a news-based social sensing pipeline for gender violence monitoring, connecting automated news collection, event retrieval, duplicate removal, temporal normalization, semantic classification, and indicator construction¹.
- We define a structured event representation for gender violence reports, including attributes such as violence type, territorial location, victim–aggressor relationship, severity, outcome, and event dynamics.
- We show how structured news events can generate low-latency social indicators for frequency analysis, temporal monitoring, territorial analysis, semantic exploration, and peak detection.

¹Project repository: <https://github.com/LABIC-ICMC-USP/GenderViolenceEvents>.

- We present a Brazilian case study that compares news-based indicators with official data, emphasizing their complementary role in identifying territorial concentrations, recurrent patterns, and emerging violence narratives.
- We discuss the limitations of news-based monitoring, including media visibility bias, source heterogeneity, duplicate reporting, incomplete temporal information, and the fact that news indicators represent public visibility rather than the true incidence of violence.

2. Related Work

The idea of using the web as a sensor for real-world phenomena has been explored in different domains, including public health, disaster response, political monitoring, and social event detection. Early studies showed that digital traces can provide useful signals about offline processes, such as influenza trends from search queries [Ginsberg et al. 2009] and earthquakes from social media posts [Sakaki et al. 2010]. These works helped establish the broader view that online information streams can complement traditional data collection systems, especially when official data are delayed, incomplete, or available only at high levels of aggregation.

News media are particularly relevant in this context because they report events with temporal, spatial, and semantic information. Unlike many social media posts, news articles often include structured narratives about what happened, where it happened, who was involved, and what consequences followed. This makes news an important source for event extraction, environmental scanning, and social monitoring. Wei and Lee [Wei and Lee 2004] investigated event detection from online news documents as a way to support environmental scanning, showing the relevance of news streams for identifying external situations that may affect decision-making. Large-scale initiatives such as GDELT also rely on global news sources to build event databases that connect actors, locations, themes, tone, and temporal information [Leetaru and Schrodtt 2013]. Digital data have also been studied as a resource for government, public administration, and policy-making. Ceron and Negri [Ceron and Negri 2016] discuss how online public opinion can support different stages of the policy cycle. Their work emphasizes that social media can reveal how citizens react to public issues and how these reactions are mobilized over time.

In the Brazilian context, WebSensors introduced a computational approach for learning sensors from web news events [Marcacini et al. 2017]. The key idea is to use machine learning methods to monitor real-world phenomena from collections of news events. This perspective is aligned with the notion that a sensor does not need to be a physical device, i.e., it can also be a computational model that observes digital traces and identifies events of interest. Later, Gôlo, Rossi, and Marcacini [Gôlo et al. 2021] formalized the task of *learning to sense* from events by combining textual, temporal, and geographic dimensions.

Our work follows this line of research but focuses on gender violence monitoring. While previous web sensing approaches have been applied to broader event collections, we investigate how news can be transformed into structured gender violence events and then aggregated into social indicators. This requires a domain-specific representation of violence-related events, including attributes such as violence type, victim–aggressor relationship, severity, outcome, territorial location, and event dynamics. Unlike GDELT,

which relies on automated event coding schemes designed for broad geopolitical conflict monitoring and primarily covers English-language sources globally, our approach combines a Portuguese-language-focused infrastructure with LLM-based extraction tailored to the semantic structure of gender violence events, enabling fine-grained attributes that are not available in general-purpose event databases.

3. Methodology

This work follows the idea of *learning to sense* (detecting real-world phenomena from digital traces), in which a computational model observes digital traces and transforms them into structured representations of real-world phenomena [Marcacini et al. 2017, Gôlo et al. 2021]. Here, the target phenomenon is gender violence reported in online news. The objective is to transform journalistic reports into structured events and aggregate them into social indicators that complement official statistics.

Let $\mathcal{D}$ be the corpus of collected news reports, as defined in Equation 1.

$$\mathcal{D} = \{d_i\}_{i=1}^N, \quad d_i = (h_i, a_i, u_i, p_i, t_i, s_i, x_i), \quad (1)$$

where h_i is the title, a_i is the description, u_i is the URL, p_i is the publisher, t_i is the publication date, s_i is the source domain, and x_i is the text analyzed by the model.

The first stage is target-oriented retrieval. Given a set of queries Q_T related to gender violence, the retrieval function selects a candidate subset of reports, as shown in Equation 2.

$$\mathcal{C}_T = R(\mathcal{D}, Q_T), \quad \mathcal{C}_T \subseteq \mathcal{D}. \quad (2)$$

The candidate reports are then normalized and deduplicated using repeated URLs, exact text hashes, and normalized textual content. Next, a large language model (LLM)-based sensing function receives each candidate report and a task prompt. Its output is defined in Equation 3.

$$f_\theta(x_i, p_T) = (\hat{y}_i, \hat{\rho}_i, \hat{\mathbf{z}}_i, \hat{\mathbf{q}}_i), \quad (3)$$

where $\hat{y}_i \in \{0, 1\}$ indicates whether the report describes a gender violence event, $\hat{\rho}_i \in [0, 1]$ is the relevance confidence, $\hat{\mathbf{z}}_i$ is the extracted event metadata, and $\hat{\mathbf{q}}_i$ stores quality scores. The metadata include 5W1H information, violence type, victim–aggressor relationship, severity, outcome, institutional response, event date, and geographic location.

A structured event is created when a report is classified as relevant, reaches a minimum confidence threshold, and is not considered duplicated. This rule is presented in Equation 4.

$$\mathcal{E}_T = \{e_i = \kappa(d_i, \hat{\mathbf{z}}_i, \hat{\mathbf{q}}_i) \mid \hat{y}_i = 1, \hat{\rho}_i \geq \lambda, \neg \text{dup}(d_i)\}, \quad (4)$$

where $\mathcal{E}_T$ is the final set of structured news-based gender violence events, κ is the event construction function, λ is the minimum relevance threshold, and $dup(d_i)$ indicates whether the report is duplicated.

Finally, the structured events are aggregated into social indicators. In plain terms, the indicator simply counts how many structured events match a given combination of violence type, location, and time period. For a violence type v , a territorial unit u , and a time interval Δ , the indicator is defined in Equation 5.

$$I(v, u, \Delta) = \sum_{e \in \mathcal{E}_T} w_e \mathbf{1}[type(e) = v] \mathbf{1}[loc(e) = u] \mathbf{1}[time(e) \in \Delta]. \quad (5)$$

In the simplest case, $w_e = 1$, meaning each structured news event contributes equally to the count. For example, $I(\text{femicide}, \text{São Paulo}, \text{Q1-2021}) = 14$ means that 14 femicide events were captured in news reports from the state of São Paulo during the first quarter of 2021. The same formulation supports indicators by violence type, time, territory, severity, outcome, and victim–aggressor relationship, and allows comparison with official statistics when both sources are aggregated using compatible temporal and geographic units.

4. Experimental Results

This section reports the first empirical results obtained from a prototype portal developed to test the integration between ReVeJo and WebSensors for monitoring gender violence through news-based indicators.

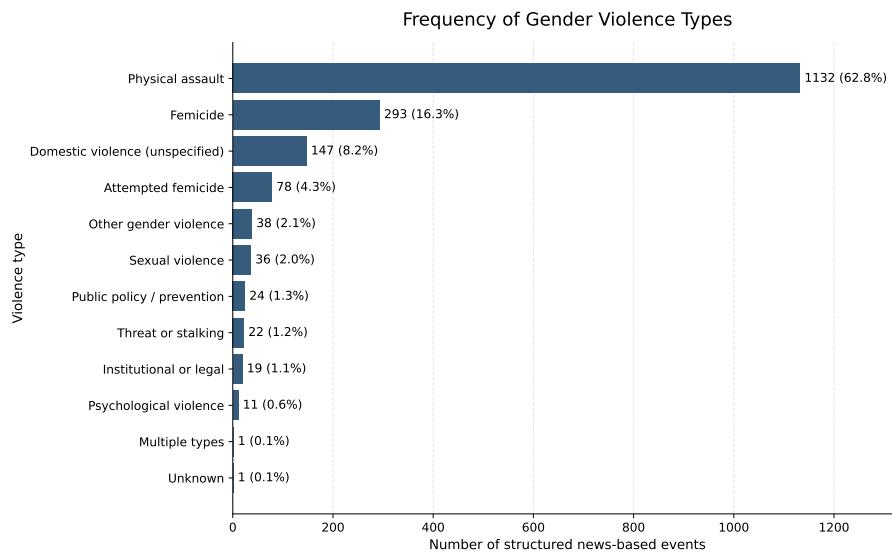


Figure 1. Frequency of structured news-based gender violence events by violence type, considering the analyzed period from 2020 to April 2026.

Figure 1 presents the frequency of structured news-based events by type of gender violence in the analyzed period, from 2020 to April 2026. The distribution is highly concentrated. Physical assault is the dominant category, with 1,132 events, corresponding to

62.8% of the structured news-based events. Femicide appears as the second most frequent category, with 293 events, or 16.3%. This result suggests that news-based monitoring captures not only fatal cases, but also a broader set of visible violent events, especially physical aggression and domestic violence dynamics. Figure 2 shows the quarterly evolution of the main violence categories as a stacked bar chart. The figure focuses on the most frequent categories, making it possible to compare the relative composition of the news-based events over time. This representation avoids interpreting changes in collection volume as direct changes in violence dynamics, and instead emphasizes the internal composition of the monitored events.

Across most quarters, physical assault remains the dominant category. However, the relative participation of femicide, domestic violence, attempted femicide, and other gender violence varies over time. This temporal variation indicates that news-based indicators can be used to monitor shifts in the public visibility of different forms of gender violence. For example, quarters with an increase in femicide or attempted femicide may indicate periods in which more severe events gained visibility in the news ecosystem.

To explore the complementarity between news-based indicators and official data, we compared the quarterly volume of events extracted from news with official victim records from the Brazilian public security data system. The official source used in this experiment was the *Ocorrências Criminais - Sinesp* dataset, made available by the Brazilian Ministry of Justice and Public Security.² In the version used in this study, the official victim-level spreadsheet covered data up to 2022, which illustrates that publication delays are a practical limitation for near-real-time monitoring.

Figure 3 compares the official series of female victims with the news-based indicator for the same period, from 2020 to 2022. Both series were independently normalized to the interval $[0, 1]$ and aligned using Dynamic Time Warping (DTW), a technique that compares the overall shape of two time series even when their peaks are slightly shifted in time. The objective was to observe whether the temporal shape of the news-based indicator could provide signals that are compatible with official dynamics, even when the absolute scales are different.

The short time span does not provide enough evidence for a robust statistical significance analysis. However, this combination can support different analytical questions: official data are better suited for validated measurement, whereas news-based events can provide earlier contextual signals about emerging situations, locations, and narratives. In practical terms, periods of higher news visibility tend to coincide with periods of higher official registrations, suggesting that the news-based indicator captures genuine fluctuation patterns rather than random noise.

Figure 4 compares the spatial distribution of the news-based indicator with official female homicide rates by state. The left panel shows the intensity of gender violence events captured by the news-based pipeline, including georeferenced localities. The right panel shows official female homicide rates per 100,000 inhabitants, based on recent data reported by the Atlas da Violência, released in May 2026 by Ipea and the Brazilian Forum on Public Security, which reports homicide statistics for 2024.

²<https://dados.mj.gov.br/dataset/sistema-nacional-de-estatisticas-de-seguranca-publica>

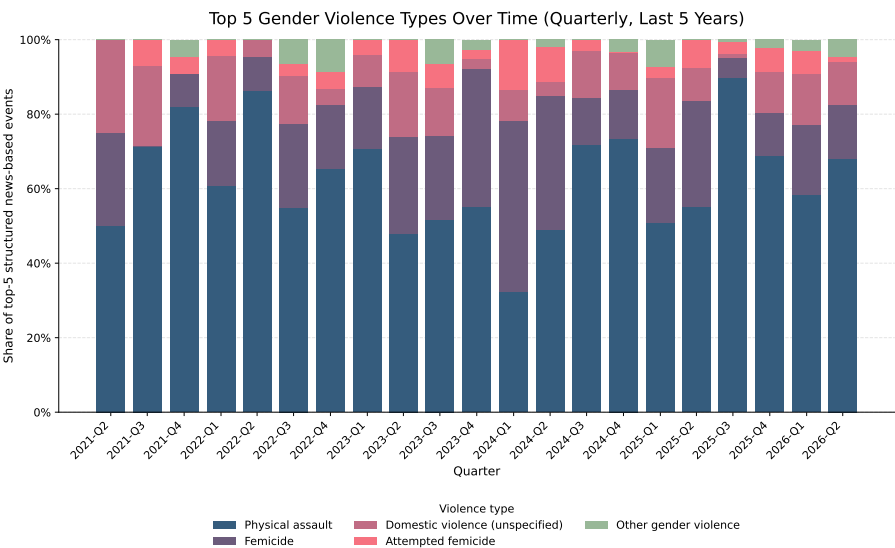


Figure 2. Quarterly evolution of the main gender violence categories in the news-based indicator.

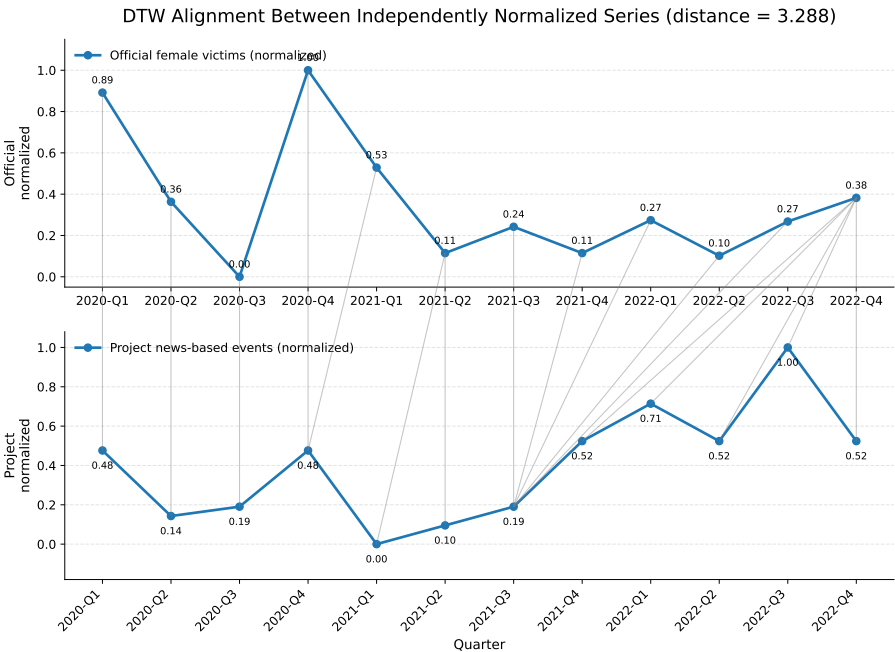


Figure 3. DTW alignment between the official series of female victims and the news-based gender violence indicator.

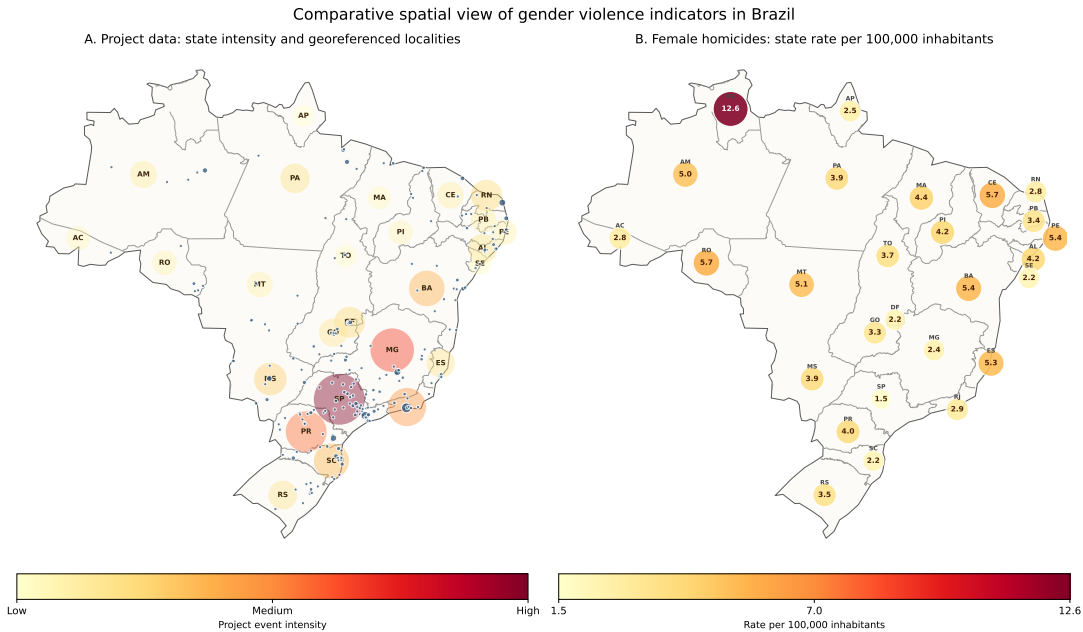


Figure 4. Comparison between the spatial coverage of news-based gender violence events and official female homicide rates by state. The left panel shows project event intensity and georeferenced localities, while the right panel shows official rates per 100,000 inhabitants.

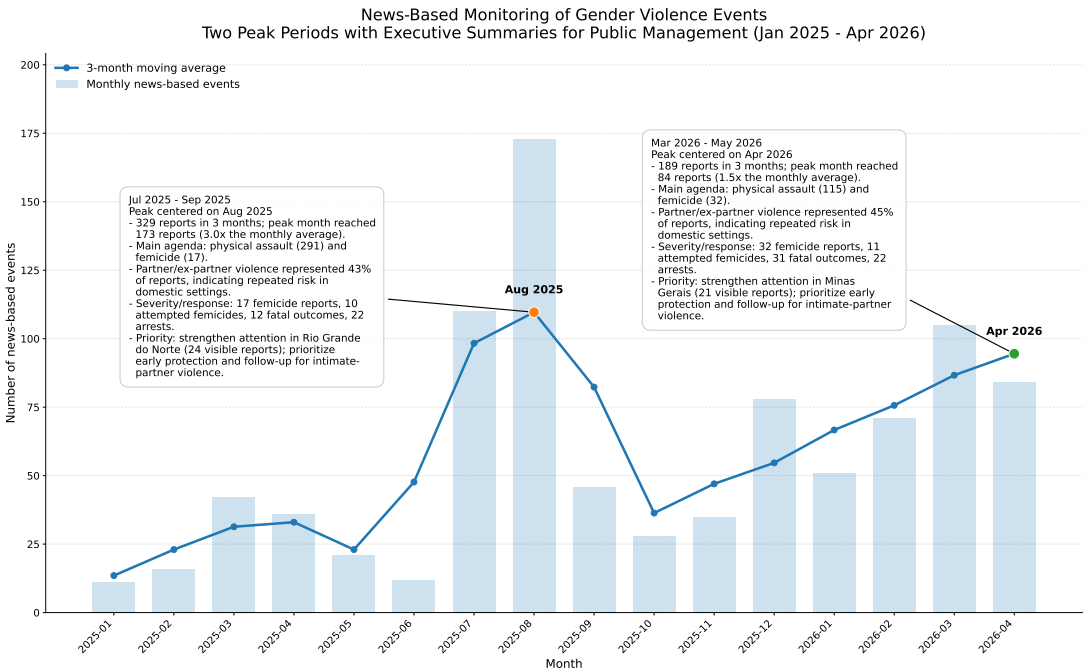


Figure 5. Recent news-based monitoring from January 2025 to April 2026. The figure highlights two peak periods and provides executive summaries generated from structured news metadata.

The comparison suggests a limitation of news-based monitoring, where the coverage of news reports is not spatially neutral. The news-based indicator tends to emphasize more populous states and regions where media production and digital visibility are stronger. In contrast, official homicide rates per 100,000 inhabitants highlight states with high relative risk, including less populous units that may not appear with the same intensity in the news-based collection.

This spatial mismatch has at least three possible explanations. First, the structure of the Brazilian digital media market concentrates news production in more populous and economically central states, so events in those regions are more likely to be captured by the pipeline regardless of actual incidence. Second, the collection strategy itself may introduce geographic bias depending on which outlets are indexed and how retrieval queries are defined. Third, the official reference normalizes by population, making sparsely populated but high-risk states more prominent relative to absolute event counts. Disentangling these sources of discrepancy requires further methodological investigation and is a priority for future work.

Finally, Figure 5 illustrates how recent news-based data can support public managers before consolidated official statistics become available. The figure focuses on monthly news-based events from January 2025 to April 2026 and applies a three-month moving average to identify periods of higher visibility. Two peak periods are annotated with executive summaries generated from the structured metadata extracted by the AI pipeline. These summaries show the potential value of combining news monitoring with AI-based event extraction, with concise analytical narratives about what happened, where the reports were concentrated, what types of violence were most visible, and which management priorities may require attention.

5. Concluding Remarks

Our results indicate that news-based gender violence events can complement official indicators in three main ways. First, they provide a low-latency observational layer that can be updated before consolidated official datasets are released. Second, they preserve semantic and contextual information that is often lost in aggregated administrative statistics, such as the reported relationship between victim and aggressor, event severity, institutional response, and territorial details. Third, they enable analytical products that are useful for public management, such as temporal alerts, spatial visibility maps, and executive summaries of peak periods.

However, the results also show important limitations. News-based indicators are affected by media visibility, source availability, duplication, uneven territorial coverage, and differences in journalistic attention. Therefore, they should not be interpreted as incidence estimates. Their main value is complementary: they can help identify emerging patterns, guide further investigation, and support situational awareness while official systems remain the reference for validated measurement. Beyond volume comparisons, other structured attributes extracted by the pipeline, such as victim–aggressor relationship, event severity, and institutional response, could in principle be compared with disaggregated administrative records from health and justice systems, providing richer cross-validation. This kind of comparison is contingent on official data availability at compatible levels of temporal and geographic granularity, and is left as an avenue for future work.

Funding declaration: This work was supported by CAPES, CNPq (Process No. 316507/2023-7), and the São Paulo Research Foundation (FAPESP), Brazil (Grants 2023/10100-4, 2024/15430-5, 2022/09091-8).

Use of Artificial Intelligence: Generative AI tools were used only for text improvement.

References

- Ceron, A. and Negri, F. (2016). The “social side” of public policy: Monitoring online public opinion and its mobilization during the policy cycle. *Policy & Internet*, 8(2):131–147.
- FBSA (2025). 19º anuário brasileiro de segurança pública. <https://forumseguranca.org.br/publicacoes/anuario-brasileiro-de-seguranca-publica/>. Accessed: 2026-05-26.
- Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., and Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014.
- Gôlo, M. P. S., Rossi, R. G., and Marcacini, R. M. (2021). Learning to sense from events via semantic variational autoencoder. *PLOS ONE*, 16(12):e0260701.
- IPEA (2025). Atlas da violência 2025. <https://www.ipea.gov.br/atlasviolencia/artigo/290/atlas-da-violencia-2025>. Accessed: 2026-05-26.
- Leetaru, K. and Schrodtt, P. A. (2013). Gdelt: Global data on events, location, and tone, 1979–2012. In *ISA Annual Convention*.
- Marcacini, R. M., Rossi, R. G., Nogueira, B. M., Martins, L. V., Cherman, E. A., and Rezende, S. O. (2017). Websensors analytics: Learning to sense the real world using web news events. In *XXIII Brazilian Symposium on Multimedia and the Web: Workshops and Posters*, pages 169–173.
- ReVeJo (2025). Repositório de veículos jornalísticos. <https://alfa.revejo.net.br/>. Accessed: 2026-05-26.
- Sakaki, T., Okazaki, M., and Matsuo, Y. (2010). Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*, pages 851–860.
- Senado Federal (2024). Mapa nacional da violência de gênero. <https://www12.senado.leg.br/institucional/omv/mapadaviolencia>. Accessed: 2026-05-26.
- Vasconcelos, N. M. d., Bernal, R. T. I., Souza, J. B. d., Bordoni, P. H. C., Stein, C., and Coll (2024). Underreporting of violence against women: an analysis of two data sources. *Ciência & Saúde Coletiva*, 29(10):e07732023.
- Wei, C.-P. and Lee, Y.-H. (2004). Event detection from online news documents for supporting environmental scanning. *Decision Support Systems*, 36(4):385–401.
- WHO 2024 (2024). Violence against women. <https://www.who.int/news-room/fact-sheets/detail/violence-against-women>. Accessed: 2026-05-26.