Evaluating FrameNet-Based Semantic Modeling for Gender-Based Violence Detection in Clinical Records

Lívia Dutra[1,2], Arthur Lorenzi[1,3], Frederico Belcavello[1], Ely Matos[1]
Marcelo Viridiano[1], Lorena Larré[1], Olívia Guaranha[3], Erik Santos[3]
Sofia Reinach[3], Pedro de Paula[3], Tiago Torrent[1,4]
[1]Federal University of Juiz de Fora (FrameNet Brasil)
[2]University of Gothenburg, [3]Vital Strategies Brasil
[4]Brazilian National Council for Scientific and Technological Development (CNPq)
livia.vicente.dutra@svenska.gu.se, {marcelo.viridiano, lorena.tasca}@estudante.ufjf.br
{alorenzi, oguaranha, esantos, pcbpaula, sreinach}@vitalstrategies.org
{fred.belcavello, ely.matos, tiago.torrent}@ufjf.br

Abstract

Gender-based violence (GBV) is a major public health issue, with the World Health Organization estimating that one in three women experiences physical or sexual violence by an intimate partner during her lifetime. In Brazil, although healthcare professionals are legally required to report such cases, underreporting remains significant due to difficulties in identifying abuse and limited integration between public information systems. This study investigates whether FrameNet-based semantic annotation of open-text fields in electronic medical records can support the identification of patterns of GBV. We compare the performance of an SVM classifier for GBV cases trained on (1) frame-annotated text, (2) annotated text combined with parameterized data, and (3) parameterized data alone. Quantitative and qualitative analyses show that models incorporating semantic annotation outperform categorical models, achieving over 0.3 improvement in F1 score and demonstrating that domain-specific semantic representations provide meaningful signals beyond structured demographic data. The findings support the hypothesis that semantic analysis of clinical narratives can enhance early identification strategies and support more informed public health interventions.

1 Introduction

The World Health Organization estimates that one in three women experiences physical or sexual violence by an intimate partner at some point in her life (WHO, 2024). Gender-based violence (GBV) is therefore not only a social issue, but a major public health concern (Garcia-Moreno and Watts, 2011; Sweet, 2014; Öhman et al., 2020). In Brazil, healthcare professionals are legally required to report cases of violence. Yet underreporting remains widespread. Either because victims are unable or unwilling to report their experiences or because signs of violence go unrecognized within routine medical encounters. Research suggests that many professionals struggle to identify signs of abuse, lack appropriate support tools, and work within fragmented information systems that do not communicate effectively (Kind et al., 2013; Garbin et al., 2015).

Brazilian public health systems collect large amounts of data on hospitalizations, mortality, medical records, and violence notifications. However, these systems are not fully integrated and lack a shared individual identifier (Guaranha et al., 2025). As a result, it is difficult to follow trajectories of risk over time or across institutions. Most of this information is stored in parameterized fields, which facilitate statistical analysis but capture only structured aspects of clinical encounters. Electronic medical records, however, also include open-text fields where healthcare professionals describe symptoms, circumstances, and patient histories in more detail. These narrative records often contain rich descriptions of situations that may signal risk of violence, but they are rarely consider for analysis due to its complexity.

The research reported in this paper explores whether semantic analysis of clinical narratives can help identify potential cases of GBV earlier and more reliably. The underlying assumption is that linguistic patterns embedded in medical records may reveal indicators of violence that are not captured by structured data alone. In particular, we investigate the contribution of FrameNet-based annotation to identifying possible patterns of violence within routine health data.[1] To achieve that, three experimental setups are compared using a SVM classifier : (1) a model trained on manually and automatically frame-annotated open-text data; (2)

[1]This study is based on the first authors' master's thesis and has been presented as part of a book chapter to lustrate the social applicability of FrameNet (Gamonal et al.).

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 2, pages 49–57

a model trained on annotated open-text combined with parameterized data; and (3) a model trained exclusively on parameterized information. Model performance is assessed using precision, recall, and F1-score, along with qualitative analysis of possible semantic patterns.

The results show that models incorporating FrameNet-based semantic information outperform those relying solely on structured data, achieving an F1-score of 0.772, compared to 0.461 for the model trained exclusively on parameterized data. This result, backed by the qualitative analysis of the findings, suggests that semantic analysis of clinical narratives can provide meaningful support for the identification of gender-based violence in primary healthcare settings.

2 Frame-Based Models of Linguistic Cognition

FrameNet is a corpus-based computational lexical database grounded in Frame Semantics, a theory within Cognitive Linguistics proposed by Fillmore (1982). In Frame Semantics, word meaning is not treated as self-contained. Instead, meaning is understood through mental representations, called *frames*, which capture shared knowledge about recurrent situations, the participants involved in them, and the relations between those participants. A *frame* can therefore be seen as a structured background against which individual words are interpreted. Understanding a lexical item presupposes familiarity with this wider conceptual structure, since the meaning of any single element depends on how it fits into the scenario as a whole.

For instance, consider the lexical unit *diagnose.v*, which evokes the `Diagnosing` frame — Figure 1. The verb does not simply refer to an action. Rather, it presupposes a healthcare context in which several roles are necessarily present. At a minimum, there must be a healthcare professional who makes the diagnosis and a patient whose condition is being evaluated, defined as Frame Elements in the theory. Without these participants, the situation would be difficult to interpret as a diagnosis. The frame also allows for additional elements, such as the method in which the diagnosis was performed or the time and place that it happened. Thus, when the lexical unit *diagnose.v* appears in a clinical record, it activates a rich scenario of healthcare assessment.

FrameNet was then created in 1997 to implement this theoretical framework in a systematic,

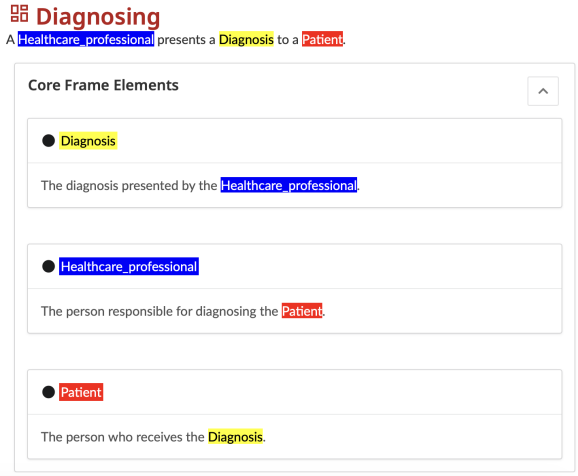

Figure 1: The `Diagnosing` frame.

corpus-based resource for English. The model has since been extended to several languages, including Brazilian Portuguese through FrameNet Brasil (FN-Br) (Torrent et al., 2022). Using its annotation tool (Torrent et al., 2024), FN-Br is able to link lexical units to the frames they evoke and annotate the semantic roles associated with those frames using authentic language data. In doing so, it is possible to gather insights and capture semantic representations that are both cognitively motivated and computationally interpretable for broader and more specific domains, as is the case of the current study, which focuses on semantic representations of healthcare and violence narratives. By modeling these domains explicitly and annotating relevant corpora, it becomes possible to identify recurring semantic configurations and relational patterns that may not be immediately visible in surface-level, as it is discussed next.

3 Corpora and Methods

This section presents the corpora and methods used in this study. The methodology involved modeling specific domains, manual and automatic annotation of data, and the design of GBV identification models and their quantitative and qualitative evaluation, as described next.

3.1 Corpora

The corpora used in this study were collected from health records produced in Recife, the capital of Pernambuco, Brazil, in collaboration with the Municipal Health Department. Data come primarily from two public health information systems: SINAN (Notifiable Diseases Information System)

and e-SUS AB (Primary Healthcare e-Medical Records System). In addition, definitions from the International Classification of Diseases (ICD) were incorporated to interpret diagnostic codes appearing in medical records. Also, causes of death were extracted from the Mortality Information System (SIM) as an indicator of true positive cases of GBV, in cases of violent deaths. The dataset is extensive, comprising more than three million records in e-SUS and more than 13,000 records in SINAN.

Regarding the systems structure, SINAN contains both parameterized fields and one open-text field ("observation") describing episodes of violence. For this study, only two components were used: (i) the parameterized indicator of a positive violence case and (ii) the corresponding open-text description. The e-SUS AB system records primary healthcare data through a combination of parameterized fields (e.g., age group, race, gender identity) and open-text fields following the SOAP model (Subjective, Objective, Assessment, Plan), along with additional narrative fields such as reason for referral, complements, and observations. Because diagnostic codes often appear without textual definitions in open-text fields, ICD descriptions were linked to the e-SUS records to enrich their semantic interpretability.

Given the presence of highly sensitive personal information in the corpora, strict ethical and legal safeguards were implemented as well as the anonymization of the data. The anonymization process combined automatic (NER models (Souza et al., 2019; Guillou, 2021; Cunha and Ramalho, 2022), fuzzy search of local place names, and regular expressions), semi-automatic (frequency-based detection of potential names), and manual verification methods to ensure the removal of Personally Identifiable Information (PII). Once this process was completed, anonymized samples of the data were used to model specific domains and for frame-based annotation.

3.2 Domain-Specific Modeling

As a means to fully capture the narratives in the open-fields of the public information systems, the specific semantic domains of Healthcare and Violence were modeled in FN-Br. This process entails the structuring of a cognitive representation that connects essential concepts of a given topic. This process consists of a twelve-step methodology, involving not only corpora collection and anonymization, but also lexical expansion, the modeling of new frames and relations between frames and lexical units – named Ternary Qualia Relations(Torrent et al., 2024) — and the lexicographic annotation of the corpora (Ruppenhofer et al., 2016), as described in Dutra et al. (2023) and Larré and Torrent (2024). This process was carried out by a group of ten researchers and resulted in 35 frames and 2,776 lexical units for the Healthcare domain along with 48 frames and 1,774 lexical units for the Violence domain.

3.3 Annotation

The annotation process was carried out in two phases: human and automatic. Human annotation followed the FrameNet methodology (Ruppenhofer et al., 2016) and was conducted using a sample of anonymized corpora with two aims. First, as the final stage of domain-specific modeling, annotation served the purpose of validating the domains; second, it was used to compose a dataset to train an automatic semantic labeler to be used in the entirety of the corpora. A total of seven trained annotators were part of the human annotation effort, which was carried out in a mirrored version of the FN-Br annotation tool (Torrent et al., 2024) with gated access to the data – so as to add one more layer of protection to the data being handled. The process focused on semantic annotation and consisted of selecting the frame evoked by each lexical unit in the sample sentences, which, then, generated an Annotation Set that allowed for the tagging of the frame elements represented in that narrative. The final number of annotated sentences was 2,352, resulting in over 14,600 Annotation Sets.

Automatic semantic labeling was performed using a newly trained version of LOME (Xia et al., 2021), a multilingual information extraction system that integrates a FrameNet parser within a pipeline based on XLM-RoBERTa, a BIO tagger and a Typer. For this study, LOME was trained in FrameNet Brasil's annotated data for English and Portuguese, including data from the Violence and Healthcare domains, expanding the training data used by Xia et al. (2021). The newly trained model had a micro-F1 of 50.68, slightly less than the original implementation (F1 = 56.34). Because this new instance was trained to deal with more challenging data, given its specificity, the performance is satisfactory. After training, the model was used to automatically annotate frames and frame frame elements in open-text sentences from e-SUS AB,

SINAN, and ICD records.

3.4 GBV Identification Models

As stated previously, this study focuses on evaluating the use of FrameNet-based semantic annotation to identify GBV cases and patterns in open-text medical records. In this sense, a model was developed to integrate FrameNet-annotated data and enable an assessment of feature importance for distinguishing violence from non-violence in e-SUS records.

Thus, as a means of accomplishing that, three experimental setups were conducted. All experiments used a linear Support Vector Machine (SVM). This choice was motivated by its interpretability, suitability for high-dimensional data, and consistency with the original project design. The three experimental setups used the same subset of e-SUS records, originally categorized based on ICD codes and links to SINAN notifications and SIM records. There were four labels:

- Violence: records with an ICD code for aggression or within two days of a SINAN notification or SIM record with the same code;
- Non-violence: ICD codes that have a small probability of being associated with violence, e.g. COVID-19 and some congenital malformations;
- Likely Violence: any record within 30 days of a notification of violence that does not have an ICD code for aggression;
- Unknown: any record that does not fall into one of the previous categories.

For this study, only violence and non-violence records were used, resulting in 801 cases (634 non-violence; 167 violence). Non-violence cases were undersampled to reduce class imbalance. Additionally, two specialists reviewed 100 "likely violence" cases, who reclassified them as 17 violence and 83 non-violence, increasing the complexity of the dataset by including more ambiguous cases.

The three experimental setups designed were:

1. **Semantic Model:** In the first setup, only the LOME annotated open-text fields were considered. As LOME identifies frame targets but not lexical units (LUs), an additional procedure to extract the LUs associated with each annotated span was also applied. This step increased the granularity of the representation, allowing the model to capture not only frames and frame elements, but also specific lexical choices. From these annotations, feature vectors were constructed based on the frequency of frames, frame elements, and lexical units, as well as the co-occurrence of frame elements across frames. Co-occurrences were considered only when at least one of the frames belonged to the Healthcare or Violence domains. The Ternary Qualia Relations between lexical units were also incorporated, assigning them a small weight to enrich the semantic connections without overwhelming the representation. To reduce sparsity, the least frequent features were removed (frames with fewer than 50 occurrences and LUs with fewer than 25). The resulting vectors were weighted using TF-IDF and L1-normalized. Given the high dimensionality of the semantic representation (15,456 features), Principal Component Analysis (PCA), was applied to reduce it to 2,000 components while preserving 94.8% of the variance. These components served as input to the classifier.

2. **Mixed Model:** The second setup considered annotated open-text fields and selected annotated parameterized fields. This experiment followed the same pipeline as the first, but additionally incorporated structured parameterized fields into the semantic representation. Categorical variables — such as race, gender identity, sexual orientation, prosthesis need, and age group — were mapped to corresponding lexical units and frames. After TF-IDF weighting, the feature space comprised 15,478 dimensions. PCA again reduced this to 2,000 components, preserving 93.5% of the variance. This combined representation was used as the input to the model.

3. **Demographic Model:** The third setup excluded semantic annotation and relied exclusively on parameterized data. The features included demographic, clinical, and administrative variables such as race, age, ICD codes, marital status, education level, unit location, and referral timing. Categorical variables were transformed using One-Hot Encoding, expanding the original 20 structured variables to 142 binary features. These features were

used directly as input to the classifier.

3.5 Evaluation

After training, the model was evaluated both quantitatively and qualitatively. While the quantitative evaluation provided numerical evidence of model performance, the qualitative analysis aimed to interpret and understand what the models were learning and how FrameNet annotation contributed to GBV identification.

Quantitative (SVM) Evaluation To compare the three experimental setups and assess the relevance of different data sources for GBV identification, the performance of the models was evaluated using five-fold cross-validation. In this procedure, the dataset was divided into five subsets: in each iteration, four subsets were used for training and one for testing, rotating until all subsets had served as the test set. Performance was assessed using precision, recall and F1 score, and the final results correspond to the average across the five folds.

Qualitative Evaluation This process involved two complementary lines of analysis: first, feature importance scores were extracted from the best-performing semantic model and the demographic model; second, the most frequently evoked frames and lexical units in the annotated e-SUS records of confirmed victims were also examined.

1. **Model features:** The 35 most relevant features for both the semantic and the demographic models were analyzed based on their contribution to the classification of the cases[2]. This allowed us to assess the explanatory power of parameterized fields versus semantic features and to identify key frames, frame elements, and lexical units associated with GBV cases.

2. **Frame and lexical pattern analysis:** Next, the frame activation patterns were analyzed in confirmed GBV cases by examining:

    - the 15 most frequently evoked frames in both domains — Healthcare and Violence;
    - the 20 most frequent LUs per domain;
    - the 30 most frequent LUs evoking the `Health_conditions` frame, to explore possible links between health conditions and violence.

    This analysis allowed a better understanding of patterns that could be linked to GBV and pointed towards future investigation.

[2]At this point, it is not possible to identify for which of the classes each feature was most relevant, only that they were relevant for the model's decision.

4 Results and Discussion

In this section, we present the results of the evaluation conducted on the model setups and discuss their implications to the use of frame-based representations to the identification of GBV in e-medical records.

4.1 Quantitative Evaluation: SVM Models

As shown in Table 1, the semantic model that relies solely on open-text fields consistently produced the strongest results. It obtained the highest recall, indicating an effective identification of positive cases. Precision was lower, but this trade-off is acceptable in the context of the study, once the main concern is avoiding false negative cases of violence in a setting where underreporting is an issue. The F1-score reflects this balance. Furthermore, recall values showed little variation across cross-validation splits, while precision and F1 varied more substantially, suggesting that some data splits were more challenging for the model than others.

Adding parameterized data to the semantic model did not lead to a meaningful improvement in the second experiment. Although a small increase in precision was observed, this was achieved without gains in overall performance and is not particularly advantageous for the task at hand, as it increases the likelihood of false negatives. These results reinforce the idea that semantic information extracted from textual descriptions is more important in identifying cases of violence than structured demographic attributes.

Finally, the contrast with the demographic model is clear. Although recall values were relatively high, precision was extremely low, indicating that many cases were incorrectly classified as positive. This imbalance makes the recall results difficult to interpret with confidence. Moreover, recall varied widely across different splits, resulting in consistently low F1-scores. Taken together, these results show that parameterized data alone are not sufficient for reliable case identification, even if they may be useful for descriptive analyzes.

Overall, the results support the initial hypothesis that FrameNet-based semantic analysis

Model	F1	Recall	Precision
Semantic	0.772 (0.113)	0.838 (0.071)	0.756 (0.190)
Mixed	0.771 (0.114)	0.832 (0.078)	0.759 (0.189)
Demographic	0.461 (0.089)	0.701 (0.173)	0.345 (0.057)

Table 1: Model performance comparison

contributes meaningfully to the identification of gender-based violence cases in electronic medical records. Next, these findings are complemented with a qualitative analysis of the patterns identified by this approach.

4.2 Qualitative Evaluation: Models and Domains

Demographic Model The list of the 35 most relevant features for the demographic model consists of the fill-in options associated with the parameterized fields, as shown in Figure 2. These features provide limited insight, as — at this stage — it is not possible to determine whether they contributed to the classification of violence or non-violence cases.

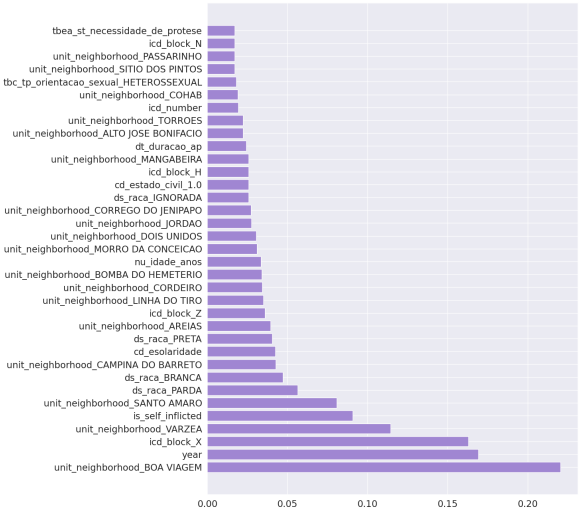

Figure 2: Most relevant features of the Demographic Model

The feature of greatest relevance is the neighborhood in which a health unit is located. In fact, 19 of the 35 selected features are related to the location of healthcare facilities. The location of a health unit does not have a clear explanatory link to the occurrence of violence, particularly since patients may seek care in different facilities, making geographic information a weak and potentially misleading indicator.

Race also appears as a highly relevant feature. Of the six possible values, four — *branca* ('white'), *parda* ('brown'), *preta*('black'), and *ignorada* ('ignored') — were selected, with three among the top ten. This pattern can be influenced by inconsistent field completion, which can amplify the weight of filled values. As a result, the model emphasizes individual characteristics rather than contextual information, reinforcing stereotypes this work aims to avoid.

Despite these limitations, some relevant features are associated with the type of care sought, notably ICD blocks X, Z, H, and N. These correspond to external causes of morbidity and mortality, factors influencing health status and contact with health services, diseases of the eye and adnexa, and diseases of the genitourinary system. Although not predominant, the presence of blocks N and Z is consistent with the patterns identified by the semantic model, as block N may relate to sexual violence and block Z often reflects scenarios requiring follow-up care, such as prenatal monitoring, that are aligned with the findings in the semantic setup.

Thus, this qualitative analysis reinforces the quantitative findings. Parameterized data leads to a model focused primarily on individual attributes, with limited attention to the clinical context. Meaningful interpretation of care-related features was only possible through a comparison with the semantic model, further supporting the use of FrameNet-based semantic analysis for open-text fields. The next section examines the most relevant features of the semantic model that had the best performance in the quantitative analysis.

Semantic Model Figure 3 shows the features that most influenced the semantic model, including frames, frame elements, co-occurrences between frame elements, and lexical units. At this stage, the analysis is exploratory and the discussion focuses on recurring patterns rather than on a detailed interpretation.

Relevant features are not restricted to the Healthcare and Violence domains. Among the most influential features are general vocabulary frames, most notably `Personal_relationships`. This result is unsurprising, given that many cases of gender-based violence involve individuals who are related in some way, which may reflect indirect references to aggressors in the records. Another generic frame

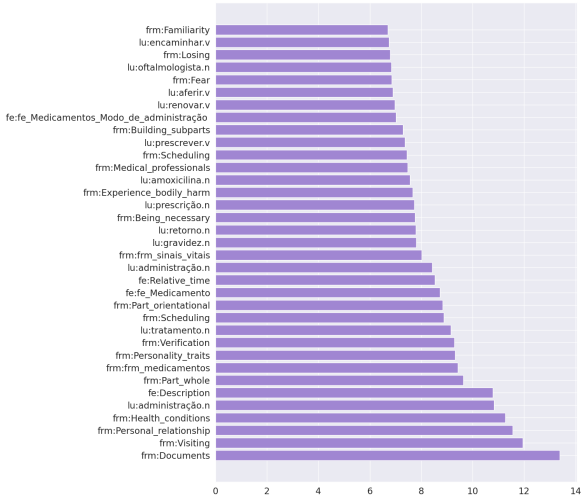

Figure 3: Most Relevant Features of the Semantic Model

that appears prominently is `Fear`. Although it is not part of the Violence domain, its relevance suggests that emotional states expressed in the text may contribute to identifying situations associated with violence.

Healthcare-related frames are more frequent than violence-related ones. Only the `Experience_bodily_harm` frame appears as a relevant frame from the Violence domain. However, the presence of this frame supports the idea that violence-related patterns can be inferred from clinical information. This is particularly relevant for primary care settings, where early identification is critical.

Within the Healthcare domain, `Health_Conditions` ranks among the most relevant frames, which is expected given the nature of the data. Frames related to `Medicines` also appear repeatedly, including more than one of their frame elements. In addition, several lexical units associated with the `Health_Intervention` frame appear as relevant for the model's decisions. Together, these results indicate that routine clinical actions and treatment-related information play an important role in the identification process.

The semantic features identified in the model provide a starting point for understanding how patterns related to gender-based violence can appear in clinical narratives. Rather than examining all relevant patterns, the focus was on the most frequently evoked frames — and lexical units — annotated by LOME in the confirmed cases of GBV that are part of the Healthcare and Violence domains.

Within the Healthcare domain, `Health_conditions` is by far the most frequently evoked frame. This is expected given the nature of clinical records and it was also evident in the model's features' analysis. However, only a small number of its associated lexical units appeared among the most frequent terms, suggesting that not all health conditions carry the same analytical weight. One term that stands out is *gestante.n* ('pregnant'), while *gravidez.n* ('pregnancy') was also a relevant feature for the semantic model. The prominence of pregnancy-related terms raises an interpretive challenge: it is unclear whether this reflects increased vulnerability, patterns specific to the dataset, or a potential bias toward the specific gender healthcare event in focus.

The `Health_service` frame also plays an important role. Terms such as *encaminhar.v* ('referral') and *acompanhamento.n* ('follow-up') occur frequently and often refer to specialized care or ongoing treatment. When considered together with other clinical elements, these references may indirectly signal previous incidents. Similarly, frequent mentions of medical examinations and medications suggest that routine clinical procedures may carry contextual clues. Enriching these elements through semantic relations, such as ternary qualia relations, may allow deeper inferences about the underlying conditions and possible links to violence.

A closer look at the `Health_conditions` frame reveals two notable tendencies. First, pregnancy-related terms appeared with high frequency, as it was already pointed out. Second, mental health conditions — including depression, anxiety, and bipolar disorder — were highly represented. These patterns may reflect the psychological consequences of abuse, although they may also be influenced by broader gendered healthcare-seeking behaviors. In either case, they warrant further investigation.

In contrast, frames from the Violence domain appear less frequently, likely because, in comparison to health issues, explicit references to violence are less frequent in clinical records. Among them, the `Experience_bodily_harm` frame stands out and also contributed to the performance of the model. However, many of its associated terms, such as fall or trauma, are not inherently indicative of violence. More direct signals emerge in references to self-inflicted harm, including self-mutilation and suicide. Although these cases were not analyzed separately, their frequency suggests that self-directed

violence deserves closer attention in future work.

Sexual violence-related patterns are particularly salient. Terms associated with sexual acts, abuse, and related examinations appear consistently, indicating that sexual violence may be a significant factor motivating healthcare visits. References to sexually transmitted infections further reinforce this interpretation. These patterns suggest that even when violence is not explicitly documented, its consequences may be traceable through clinical descriptions.

Therefore, this qualitative analysis also shows that FrameNet-based semantic annotation makes it possible to uncover patterns that would likely remain invisible in structured data alone. Although the findings remain exploratory and require validation in collaboration with healthcare professionals, they support the broader hypothesis that semantic analysis of open-text medical records can contribute to the identification of gender-based violence in primary care settings.

5 Conclusion and Future Work

This study evaluated the use of FrameNet-based semantic annotation to identify Gender-Based Violence (GBV) cases and patterns in open-text fields of e-medical records. Our results show that models using semantic annotation of open-text fields outperform models relying solely on parameterized demographic data, achieving an F1 score 0.31 higher. The addition of parameterized fields to the semantic model provided minimal improvement, highlighting that open-text information carries richer and more relevant insights for detecting GBV. Qualitative analysis confirmed that relying only on parameterized data risks reinforcing stereotypes and provides limited information for pattern discovery, while semantic annotation enables the identification of meaningful patterns that can inform further investigation and policy intervention.

In general, these findings support the hypothesis that FrameNet-based semantic analysis is a valuable tool for identifying potential GBV cases, including those underreported. By revealing patterns in both reported and unreported cases, this approach can assist in early-warning systems and public policies, contributing to improved protection and intervention strategies. Finally, the experimental setups and qualitative assessments presented here provide a baseline for future research on leveraging linguistic analysis for public health surveillance, which is already in development. Progress has already been made on the more systematic identification of GVB patterns (Dutra et al., 2025), and three parallel lines of research are currently being carried out to explore the identification of new patterns. Two of them focus directly on violence-related cases, specifically self-harm and violence against LGBTQ+ individuals. The third broadens the scope of the semantic model beyond violence, with the aim of identifying patterns related to women's health and supporting early detection of potential cancer cases.

Ethics and Limitations

The study presented in this paper was part of a broader project and approved by the Research Ethics Committee of the Federal University of Goiás (CAAE:64733922.3.0000.5083; Approval number: 6.126.995). The research involved highly sensitive information from violence notifications and electronic medical records that could increase the risk for victims of violence. Thus, the research team has extensively studied this issue and consulted data protection specialists before pursuing any implementation of the methodology. To protect the information, all team members signed confidentiality agreements, the data was anonymized - as described-, and access was restricted to anonymized samples only. The methodology was developed to improve the use of health data in Brazil and address the underreporting of health-related events, using frame-based modeling, semantic parsing, and the identification of linguistic pattern in Brazilian Portuguese. Although it can be adjusted and expanded to other languages, it has not been extensively tested yet and may reflect biases specific to Brazilian Portuguese, which is a limitation.

Acknowledgments

This work was supported by the Patrick J. McGovern Foundation's acceleration program, the José Luiz Setúbal Foundation, and the Instituto Galo da Manhã. We also express our gratitude to our partners from the Recife Municipal Health Department — Luciana Caroline, Marcella Abath, Natalia Barros, and Yana Lopes — for their valuable collaboration and continuous support throughout this project. Tiago Torrent is a grantee of the Brazilian National Council for Scientific and Technological Development CNPq – grant 311241/2025-5).

References

Luís Filipe Cunha and José Carlos Ramalho. 2022. Ner in archival finding aids: Extended. *Machine Learning and Knowledge Extraction*, 4(1):42–65.

Lívia Dutra, Arthur Lorenzi, Laís Berno, Franciany Campos, Karoline Biscardi, Kenneth Brown, Marcelo Viridiano, Frederico Belcavello, Ely Matos, Olívia Guaranha, Erik Santos, Sofia Reinach, and Tiago Timponi Torrent. 2025. Frame semantic patterns for identifying underreporting of notifiable events in healthcare: The case of gender-based violence. *Preprint*, arXiv:2510.26969.

Lívia Dutra, Arthur Lorenzi, Lorena Larré, Frederico Belcavello, Ely Matos, Amanda Pestana, Kenneth Brown, Mariana Gonçalves, Victor Herbst, Sofia Reinach, Renato Teixeira, Pedro de Paula, Alessandra Pellini, Cibele Sequeira, Ester Sabino, Fábio Leal, Mônica Conde, Regina Grespan, and Tiago Torrent. 2023. Building a frame-semantic model of the healthcare domain: Towards the identification of gender-based violence in public health data. In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 338–346, Porto Alegre, RS, Brasil. SBC.

Charles J. Fillmore. 1982. Frame Semantics. In Linguistics Society of Korea, editor, *Linguistics in the morning calm*. Hanshin Publishing Co., Seoul, South Korea.

Maucha Andrade Gamonal, Lívia Vicente Dutra, Mariane de Carvalho Pinto, and Tiago Timponi Torrent. Pln e responsabilidade social: Aplicações da framenet-br. In H. M. Caseli and M. G. V. Nunes, editors, *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, 4 edition, volume 3. BPLN. In press.

Cléa Adas Saliba Garbin, Isabella de Andrade Dias, Tânia Adas Saliba Rovida, and Artênio José Ísper Garbin. 2015. Desafios do profissional de saúde na notificação da violência: obrigatoriedade, efetivação e encaminhamento. *Revista Ciência & Saúde Coletiva*, 20(6):1879–1890.

Claudia Garcia-Moreno and Charlotte Watts. 2011. Violence against women: an urgent public health priority. *Bulletin of the World Health Organization*, 89:2–2.

Olívia LC Guaranha, Juliana Rocha Miranda, Fátima Marinho, Renato Teixeira, Erik Santos, Denise Guerra Wingerter, Paola da Costa Silva, Diana Paula de Souza Rego Pinto, Gleidson Paulino Vitório, Sofia Reinach, and 1 others. 2025. Data integration for the prevention of violence against girls and women in northeastern brazilintegración de datos para la prevención de la violencia contra niñas y mujeres en el nordeste de brasil. *Revista panamericana de salud publica= Pan American journal of public health*, 49:e66.

Pierre Guillou. 2021. NER-BERT-Base-Cased-pt-lenerbr: BERT-based NER model for Portuguese (legal domain). https://huggingface.co/pierreguillou/ner-bert-base-cased-pt-lenerbr. Accessed: 2023-08-30.

Luciana Kind, Maria de Lourdes Pereira Orsini, Valdênia Nepomuceno, Letícia Gonçalves, Gislaine Alves de Souza, and Monique Fernanda Félix Ferreira. 2013. Subnotificação e (in) visibilidade da violência contra mulheres na atenção primária à saúde. *Cadernos de Saúde Pública*, 29:1805–1815.

Lorena Larré and Tiago Torrent. 2024. Modelagem baseada em frames para identificação do léxico da violência de gênero. In *Anais do XV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 403–412, Porto Alegre, RS, Brasil. SBC.

Josef Ruppenhofer, Michael Ellsworth, Miriam Petruck, Christopher Johnson, and Jan Scheffczyk. 2016. *FrameNet II: Extended Theory and Practice*.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2019. Portuguese Named Entity Recognition using BERT-CRF. *arXiv preprint arXiv:1909.10649*.

Patricia L Sweet. 2014. Every bone of my body: Domestic violence and the diagnostic body. *Social Science & Medicine*, 122:44–52.

Tiago Timponi Torrent, Ely Edison da Silva Matos, Frederico Belcavello, Marcelo Viridiano, Maucha Andrade Gamonal, Alexandre Diniz da Costa, and Mateus Coutinho Marim. 2022. Representing Context in FrameNet: A Multidimensional, Multimodal Approach. *Frontiers in Psychology*, 13.

Tiago Timponi Torrent, Ely Edison da Silva Matos, Alexandre Diniz da Costa, Maucha Andrade Gamonal, Simone Peron-Corrêa, and Vanessa Maria Ramos Lopes Paiva. 2024. A flexible tool for a qualia-enriched FrameNet: the FrameNet Brasil WebTool. *Language Resources and Evaluation*, pages 1–29.

WHO. 2024. World Health Organization Health Topics: Violence Against Women. Accessed: October 6, 2025.

Patrick Xia, Guanghui Qin, Siddharth Vashishtha, Yunmo Chen, Tongfei Chen, Chandler May, Craig Harman, Kyle Rawlins, Aaron Steven White, and Benjamin Van Durme. 2021. LOME: Large ontology multilingual extraction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 149–159, Online. Association for Computational Linguistics.

Ann Öhman, Marika Burman, Maria Carbin, and Kerstin Edin. 2020. The public health turn on violence against women: analysing swedish healthcare law, public health and gender-equality policies. *BMC Public Health*, 20:1–12.