# Collection and analysis of human trafficking news from Brazilian websites

**Ana Rosa A. Nascimento**[1], **Matheus Rigato**[1], **João V.L.P. dos Santos**[1,2], **João V. C. Neres de Sousa**[1], **Caetano Traina Jr.**[1], **Mirela T. Cazzolato**[1]

[1]Institute of Mathematics and Computer Science
University of São Paulo (ICMC-USP) - São Carlos, Brazil

[2]Faculty of Philosophy, Sciences and Letters at Ribeirão Preto
University of São Paulo (FFCLRP-USP) - Ribeirão Preto, Brazil

rosa.nascialmeida@usp.br, matheus_rigato@usp.br, joaovlpss@usp.br, joaovneres@usp.br, caetano@icmc.usp.br, mirela@icmc.usp.br

***Abstract.*** *Human trafficking (HT) is a serious crime, yet structured datasets on the topic remain scarce in Brazil. This paper presents a dataset of 3,556 news articles mentioning HT, collected from three Brazilian online portals. We describe a Python-based pipeline for collecting data from heterogeneous websites and discuss the main challenges encountered during extraction. The corpus was analyzed using AI-assisted labeling, semantic similarity, dimensionality reduction, clustering, and supervised classification. The dataset and collection methodology aim to support future research on HT, natural language processing, open data analysis, and public safety in the Brazilian context.*

## 1. Introduction

Human trafficking (HT) constitutes a severe violation of human rights, affects millions of individuals worldwide, and generates substantial illicit revenues annually [United Nations Office on Drugs and Crime 2024]. Brazil occupies a critical position in this context as both a country of origin and transit for trafficking victims, with reported cases involving labor exploitation, sexual exploitation, and illegal organ trafficking [United Nations Office on Drugs and Crime 2024, Rollo et al. 2022]. Despite the severity of the problem, systematic data collection remains limited, and a considerable portion of the available evidence is dispersed across news media sources [Passos et al. 2022].

News articles therefore represent a valuable complementary source for large-scale analyses of crime-related patterns and social phenomena [Rollo et al. 2022]. However, the Brazilian context poses several challenges, including the absence of structured datasets in Brazilian Portuguese, the heterogeneity of online news portals, and the difficulty of distinguishing directly relevant reports from tangentially related content. Previous studies have predominantly focused on English-language sources or social media data [Saxena et al. 2025], leaving an important gap in the analysis of Brazilian news on HT.

To address this gap, this work presents a dataset comprising 3,556 news articles collected from three Brazilian online portals through a custom web crawling pipeline. In addition to data collection, we conduct an Exploratory Data Analysis (EDA) based on embedding-based similarity analysis, dimensionality reduction, clustering, and supervised classification to characterize the corpus and assess the feasibility of automatically

identifying HT-related content. Supporting materials are publicly available[1] to promote open science.

## 2. Background and Related Work

Structured data on human trafficking remains limited, particularly in Brazil. Passos et al. [Passos et al. 2022] analyzed reported cases in the country between 2009 and 2017 using administrative records, highlighting both the severity of the phenomenon and the limitations of official data sources. These limitations motivate the use of complementary sources, including online news media.

News corpora have been employed for large-scale crime analysis. Rollo et al. [Rollo et al. 2022] proposed an event extraction framework for an Italian crime-news corpus, demonstrating that NLP pipelines can derive structured information from heterogeneous journalistic sources. Although news coverage is subject to editorial and reporting biases, it can provide complementary evidence when official records are incomplete.

Thematic news classification commonly combines embedding-based representations with supervised learning. The survey by da Costa et al. [da Costa et al. 2023] reports competitive performance for Logistic Regression across several text classification settings and discusses the advantages of ensemble methods under class imbalance. Chajia and Nfaoui [Chajia and Nfaoui 2024] further demonstrated the effectiveness of combining LLM-based embeddings with Logistic Regression in imbalanced scenarios. For multilingual settings with limited labeled data, pretrained Sentence-Transformers models, such as all-MiniLM-L6-v2 and paraphrase-multilingual-MiniLM-L12-v2 [Reimers and Gurevych 2021a, Reimers and Gurevych 2021b], provide semantic representations without requiring large task-specific training corpora.

Unlike studies based on Brazilian administrative records [Passos et al. 2022], non-Brazilian crime-news corpora [Rollo et al. 2022], or English-language and social-media sources [Saxena et al. 2025], this work focuses on Brazilian Portuguese news concerning human trafficking. It contributes a corpus of 3,556 articles collected from three online news portals and integrates data collection, AI-assisted labeling, semantic analysis, clustering, and supervised classification.

## 3. Methodology

This section describes the procedures adopted for data collection and exploratory analysis of HT-related news from Brazilian online portals.

### 3.1. Data Collection

Figure 1 summarizes the data collection pipeline, from link discovery to article extraction. For each news portal, the crawler identifies the available search-result pages, retrieves their HTML content using the Requests library[2], and parses the pages with BeautifulSoup 4[3] to identify article URLs. The retrieved articles are then processed through portal-specific extraction templates, and their metadata and full text are stored in a structured CSV dataset.

[1] GitHub repository: *https://github.com/aninha-do-ceu/project-HT*

[2] 'Requests' Python library: *https://pypi.org/project/requests/*

[3] 'Beautiful Soup 4' Python library: *https://beautiful-soup-4.readthedocs.io/en/latest/*

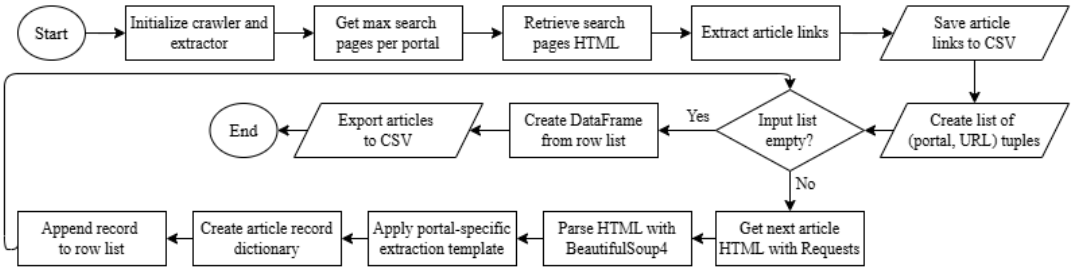


**Figure 1. Data collection pipeline from link discovery to article extraction.**

Because news portals adopt heterogeneous HTML structures, each template defines the tags, attributes, and classes required to extract titles, article bodies, publication dates, and images. When standard selectors are insufficient, the template invokes a custom extraction function, thereby isolating portal-specific processing from the general pipeline and facilitating the inclusion of additional sources. URL retrieval is also separated from the extraction logic, allowing HTML obtained through alternative mechanisms to be processed when anti-bot systems or dynamic JavaScript rendering prevent direct access.

### 3.2. Exploratory Data Analysis

The collected corpus was subjected to an Exploratory Data Analysis (EDA) comprising four stages: LLM-assisted labeling, embedding-based similarity analysis, dimensionality reduction and clustering, and supervised classification.

Articles were initially labeled as either "HT" or "Non-HT" using the `deepseek-chat` model identifier through the DeepSeek API. Titles were used as input because news headlines are intended to provide concise representations of the main content of an article while substantially reducing the amount of text processed [Kanumolu et al. 2024]. Moreover, titles constitute a uniform textual field across the different news portals considered in this study. The model was instructed to assign the HT label to titles explicitly or implicitly referring to human trafficking, trafficking in persons, forced sexual exploitation, forced labor, or closely related concepts. Uncertain cases were assigned to the Non-HT class. The temperature parameter was set to 0.1, and the complete prompt is available in the project repository[1].

To assess label reliability, the authors manually reviewed a subset of the generated labels, prioritizing articles classified as HT because of the strong class imbalance. Approximately 70% of the HT-labeled articles were inspected, together with a smaller, non-systematically sampled subset of Non-HT articles. The review followed the same criteria defined in the labeling prompt. This procedure was exploratory and was not designed as a formal annotation study; therefore, no inter-annotator agreement or complete validation statistics were computed.

Each article was represented by concatenating its title and body text in the format "*title: [], text: []*" and encoding the resulting text with two Sentence-Transformers models: *all-MiniLM-L6-v2* and *paraphrase-multilingual-MiniLM-L12-v2*. Cosine similarity was computed to examine semantic relationships and potential near-duplicates. Principal Component Analysis (PCA) was then applied to obtain two-dimensional projections, followed by HDBSCAN clustering. Finally, Logistic Regression, SVM, and XGBoost classifiers were trained on the embedding representations.

## 4. Results

This section presents the results of the data collection pipeline and the subsequent exploratory analyses. We first describe the corpus and then assess whether HT-related articles can be distinguished through semantic, unsupervised, and supervised methods.

### 4.1. Data Collection

The crawler initially targeted several Brazilian news outlets and collected 3,556 articles from three portals: BNC Amazonas, Carta Capital, and CNN Brasil. Other portals could not be included because of unavailable search interfaces, Cloudflare protection, or JavaScript-dependent content, illustrating the practical challenges of collecting data from heterogeneous news websites.

To support transparency and data provenance without redistributing potentially copyrighted journalistic content, the public repository provides the collection code, article identifiers, source outlets, collection dates, original URLs, assigned labels, and derived experimental results. Full article texts are not publicly available, although requests for additional material may be evaluated subject to legal and institutional constraints.

### 4.2. Analysis of the Collected News Articles

Although the collection process was guided by HT-related search terms, portal search mechanisms also returned articles only tangentially related to the topic. Therefore, the LLM-assisted labeling procedure described in Section 3.2 was applied to distinguish between "HT" and "Non-HT" content. Figure 2 presents the resulting distribution across portals. Of the 3,556 articles, 196 were labeled as HT and 3,360 as Non-HT, corresponding to a positive-class prevalence of approximately 6%.

| Portal | HT | Non-HT | Total |
|---|---|---|---|
| **BNC Amazonas** | 8 | 1 | 9 |
| **Carta Capital** | 6 | 5 | 11 |
| **CNN Brasil** | 182 | 3,354 | 3,536 |
| **Total** | 196 | 3,360 | 3,556 |

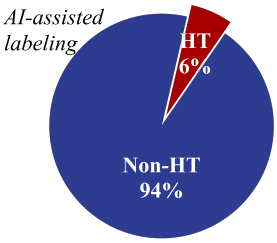


**Figure 2. Distribution of the initial LLM-assisted labels across news portals.**

The strong class imbalance motivated an investigation into whether HT-related articles occupy distinct regions in the semantic representation space. The title and body text of each article were encoded using *all-MiniLM-L6-v2* [Reimers and Gurevych 2021a] and *paraphrase-multilingual-MiniLM-L12-v2* [Reimers and Gurevych 2021b]. Figure 3(a) presents the resulting pairwise cosine-similarity distributions.

For both models, the HT, Non-HT, and cross-class distributions substantially overlap. Similarity values are concentrated around 0.55 for *all-MiniLM* and 0.32 for *paraphrase-multilingual*. Thus, cosine similarity alone does not reveal a clear distinction between the classes.

Because pairwise similarity did not reveal class-specific patterns, PCA was used to examine the global organization of the embedding spaces. As shown in Figure 3(b),

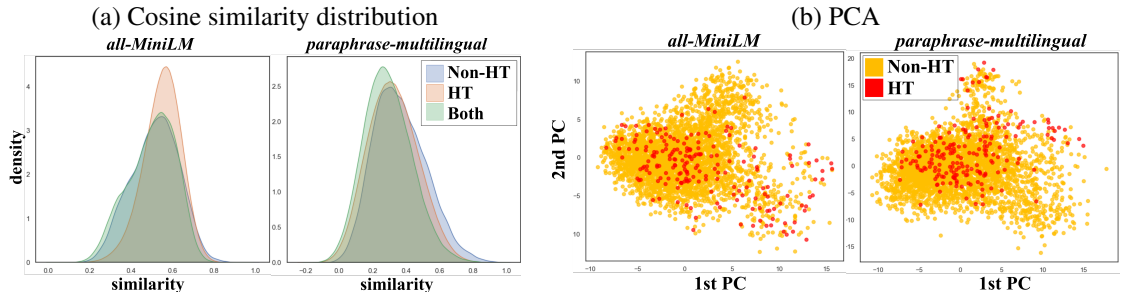


**Figure 3. Cosine-similarity distributions and two-dimensional PCA projections.**

the two-dimensional projections also exhibit substantial overlap, with no clear visual separation between HT and Non-HT articles.

We next examined whether density-based clustering could reveal latent semantic groups. HDBSCAN was applied to the PCA projections, and the resulting clusters are shown in Figure 4. Neither embedding model produced a cluster specifically associated with HT. Instead, the groups primarily reflected recurring high-profile topics, particularly articles concerning Jeffrey Epstein and P. Diddy. This suggests that the dominant semantic organization of the corpus is more strongly associated with specific people and events than with the distinction between HT and Non-HT content.

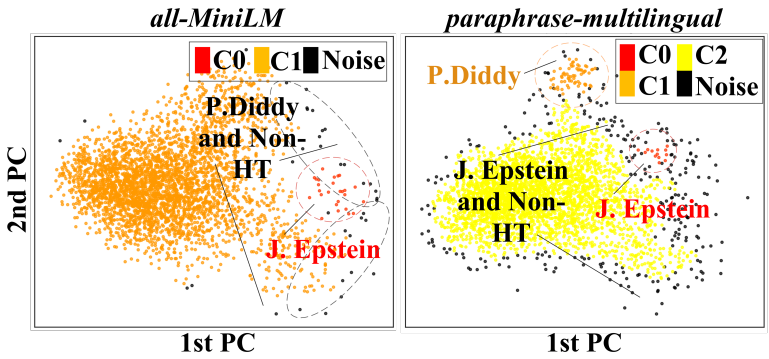


**Figure 4. HDBSCAN clustering over the first two principal components.**

Since the unsupervised analyses did not provide a reliable class separation, Logistic Regression [Chajia and Nfaoui 2024], SVM, and XGBoost [da Costa et al. 2023] were evaluated using a stratified 80/20 train-test split. Table 1 reports accuracy, precision, recall, F1-score, and PR-AUC for the HT class.

**Table 1. Classification results for the HT class.**

| Embedding Model | Classifier | Accuracy | Precision | Recall | F1 | PR-AUC |
|---|---|---|---|---|---|---|
| *all-MiniLM* | **Logistic Regression** | 0.92 | 0.36 | 0.56 | 0.44 | 0.22 |
| | **SVM** | 0.96 | 0.67 | 0.46 | 0.55 | 0.34 |
| | **XGBoost** | 0.95 | 0.73 | 0.21 | 0.32 | 0.19 |
| *paraphrase-multilingual* | **Logistic Regression** | 0.94 | 0.49 | 0.80 | 0.61 | 0.40 |
| | **SVM** | 0.97 | 0.70 | 0.67 | 0.68 | 0.49 |
| | **XGBoost** | 0.96 | 0.79 | 0.38 | 0.52 | 0.34 |

For each classifier, the *paraphrase-multilingual* embeddings yielded better results

than the corresponding *all-MiniLM* representation. Although accuracy ranged from 0.92 to 0.97, it is less informative under the observed class imbalance. PR-AUC provides a more appropriate assessment in this setting, and all configurations exceeded the random-ranking baseline of approximately 0.055.

SVM with *paraphrase-multilingual* achieved the highest F1-score and PR-AUC, reaching 0.68 and 0.49, respectively. XGBoost obtained the highest precision, at 0.79, whereas Logistic Regression achieved the highest recall, at 0.80. Given the objective of minimizing missed HT articles, recall was adopted as the main model-selection criterion.

Figure 5 further examines the false-negative rate as a function of the classification threshold $\tau$, where an article is assigned to the HT class when $P(\mathrm{HT} \mid x) \geq \tau$. Lower thresholds reduce false negatives at the cost of additional false positives. Across the evaluated thresholds, Logistic Regression exhibited the lowest false-negative rate, while *paraphrase-multilingual* generally outperformed *all-MiniLM*. These results support the selection of Logistic Regression with *paraphrase-multilingual* embeddings for the subsequent filtering stage.

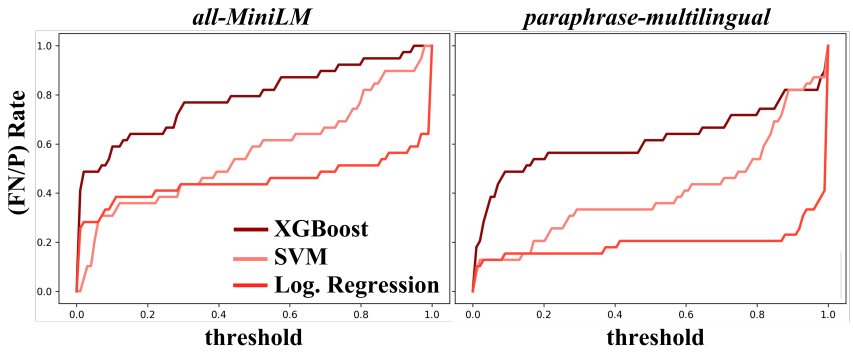


**Figure 5. False-negative rate as a function of the classification threshold.**

## 5. Conclusions

This work presented a pipeline for collecting and analyzing Brazilian news articles related to human trafficking. The resulting corpus comprises 3,556 articles from three online news portals and was examined through LLM-assisted labeling, semantic similarity analysis, dimensionality reduction, clustering, and supervised classification. Among the evaluated configurations, Logistic Regression with *paraphrase-multilingual-MiniLM-L12-v2* embeddings achieved the highest recall, reaching 0.80, and was therefore selected for filtering subsequent batches of articles collected by the pipeline.

The study is limited by the restricted number of supported portals, heterogeneous HTML structures, anti-crawling mechanisms, JavaScript-dependent content, strong class imbalance, and automatically generated labels subjected only to partial human review. Moreover, keyword-based searches retrieved both relevant and tangentially related articles, requiring an additional classification stage. Future work includes expanding the number of supported portals, incorporating dynamic crawling mechanisms, systematically validating the labels, comparing additional LLMs and classification strategies, and investigating the extraction of information such as locations, victims, and *modus operandi*.

**Acknowledgements.** This work was supported by the São Paulo Research Foundation (FAPESP – grants #2026/11130-2, #2024/15430-5, #2024/13328-9), the Coordination for Higher Education Personnel Improvement (CAPES – Finance Code 001), the National Research Council (CNPq – grants #401496/2025-2 and #402987/2025-0), and the University of São Paulo (PRPI 1032 - grant #207, and PUB - grants #4622 and #5192).

**Origin of the work.** This paper reports results obtained from a Scientific Initiation (IC), Extension and Undergraduate Projects, focused on combating HT through data analysis.

**Generative AI usage statement.** Grammarly was used solely to support the linguistic revision and improve the clarity of the manuscript. All scientific content, analyses, interpretations, and conclusions were produced and verified by the authors.

## References

Chajia, M. and Nfaoui, E. H. (2024). Customer Churn Prediction Approach Based on LLM Embeddings and Logistic Regression. *Future Internet*, 16(12). DOI: 10.3390/fi16120453.

da Costa, L. S., Oliveira, I. L., and Fileto, R. (2023). Text classification using embeddings: a survey. *Knowledge and Information Systems*, 65(7):2761–2803. DOI: 10.1007/s10115-023-01856-z.

Kanumolu, G., Madasu, L., Surange, N., and Shrivastava, M. (2024). TeClass: A human-annotated relevance-based headline classification and generation dataset for Telugu. In *Proceedings of the Joint Intl. Conf. on Computational Linguistics, Language Resources and Evaluation*, pages 15711–15720, Torino, Italia. ELRA and ICCL.

Passos, T. S., Santana, M. F. S., Cordero-Ramos, N., and Almeida-Santos, M. A. (2022). Profile of Reported Trafficking in Persons in Brazil Between 2009 and 2017. *Journal of Interpersonal Violence*, 37(11-12):NP8257–NP8273. DOI: 10.1177/0886260520976219.

Reimers, N. and Gurevych, I. (2021a). all-MiniLM-L6-v2: Lightweight Sentence Transformer Model. https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2. Sentence-Transformers model on Hugging Face Model Hub.

Reimers, N. and Gurevych, I. (2021b). paraphrase-multilingual-MiniLM-L12-v2: Multilingual Sentence Embedding Model. https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2. Sentence-Transformers model on Hugging Face Model Hub.

Rollo, F., Po, L., and Bonisoli, G. (2022). Online News Event Extraction for Crime Analysis. In *Proceedings of the 30th Italian Symposium on Advanced Database Systems (SEBD 2022)*, volume 3194 of *CEUR Workshop Proceedings*, pages 257–264. CEUR-WS.org. URL: https://ceur-ws.org/Vol-3194/paper28.pdf.

Saxena, V. K., Ashpole, B., Van Dijck, G., and Spanakis, G. (2025). "MATCHED: Multimodal Authorship-Attribution To Combat Human Trafficking in Escort-Advertisement Data". In *Findings of the Association for Computational Linguistics*, pages 4334–4373, Vienna, Austria. ACL 2025. DOI: 10.18653/v1/2025.findings-acl.225.

United Nations Office on Drugs and Crime (2024). *Global Report on Trafficking in Persons 2024*. United Nations, Vienna. United Nations publication E.24.XI.11.