FakenewsBR: A Unified Portuguese Misinformation Corpus for Reproducible Data Integration Research

Kauan Divino Pouso Mariano[1,2]*, Fabrycio Leite Nakano Almada [1,2]*, Maykon Adriell Dutra [1,2]*, Victor Emanuel da Silva Monteiro [2]*, Juliana Resplande Sant'Anna Gomes [2], Arlindo Rodrigues Galvão Filho [1,2], Anderson da Silva Soares [1,2]

[1]Institute of Informatics, Federal University of Goiás, Goiânia-GO - Brazil

[2]Advanced Knowledge Center in Immersive Technology (AKCIT), Federal University of Goiás, Goiânia-GO - Brazil

{kauan,fabrycio,maykonadriell}@discente.ufg.br

{victor_emanuel,juliana.resplande}@egresso.ufg.br

{arlindogalvao,andersonsoares}@ufg.br

Abstract. *The growing impact of misinformation in digital environments has intensified the demand for structured and reusable datasets to support empirical research. This paper presents FakenewsBR, a consolidated Portuguese-language fake news resource comprising 51,205 instances integrated from six previously curated corpora rather than newly collected raw data. Through a reproducible pipeline, heterogeneous sources were harmonized under a unified schema, subjected to controlled cleaning procedures, and processed using near-duplicate analysis to flag highly similar records for manual review. The final dataset preserves provenance information, spans a temporal window from 2005 to 2022, and includes fact-check metadata for 18.18% of instances. Rather than proposing new detection models, the contribution focuses on systematic integration, structural normalization, and transparent documentation to facilitate reuse in machine learning, database research, and misinformation analysis. The dataset and its construction pipeline are publicly available at:* `https://github.com/AKCIT-FN/fakenews-data`.

1. Introduction

Misinformation on digital platforms challenges information ecosystems and has driven research on automated fake news detection. In computational contexts, fake news is often defined as intentionally misleading news-like content shared online [Lazer et al. 2018, Gomes et al. 2023]. Progress in detection depends on high-quality, representative datasets [Shu et al. 2017], yet many available resources remain structurally heterogeneous, inconsistently labeled, and poorly documented.

In Portuguese-language contexts, dataset availability remains limited. Although important initiatives have introduced curated corpora for regional misinformation dynamics [Monteiro et al. 2018], they often differ in annotation criteria, schema design,

*These authors contributed equally to this work.

and metadata granularity, hindering interoperability and comparative experimentation [Shu et al. 2017]. Documentation of construction procedures and links to external verification evidence is also often sparse, limiting reproducibility and broader methodological analysis.

To address these limitations, we present *FakenewsBR*, a Portuguese-language misinformation *corpus* with 51,205 instances integrated from six curated datasets. It is built through a reproducible pipeline combining schema harmonization, controlled cleaning, near-duplicate analysis, and verification-assisted review, drawing on procedures described in [Gomes 2025]. By unifying heterogeneous sources under a canonical schema while preserving provenance, the dataset supports cross-source evaluation, structured querying, and reproducible experimentation.

We also provide an analysis of structural integrity, class distribution, temporal coverage from 2005 to 2022, and enrichment availability. By documenting integration decisions and limitations, the work offers a reusable resource aligned with transparency and reproducibility in misinformation research [Hovy and Spruit 2016].

2. Related Work

Research on misinformation has yielded several datasets to support automated detection. However, these resources are often constrained by domain or platform specificities. Disparities in collection methodologies and schema designs further hinder interoperability and direct reuse across studies [Wang 2017, Shu et al. 2020].

In the Portuguese-language context, several datasets have addressed the sources integrated into our study. *Fake.br* provided the first reference corpus for Brazilian Portuguese, featuring aligned true and false news articles [Monteiro et al. 2018, Silva et al. 2020]. Subsequent efforts shifted focus toward social media and specific domains: *FakeWhatsApp-BR* introduced manually labeled messages from public groups [Cabral et al. 2021], while the Portuguese subset of *MuMiN* contributed fact-checked Twitter/X threads within a multilingual graph [Nielsen and McConville 2022, Gomes et al. 2023]. More recently, *LLM4BR-300* leveraged political news to evaluate Large Language Models (LLMs) [Gôlo et al. 2024], and *Portuguese Fake News* offered a broader longitudinal Web news collection [Selau 2021].

Complementary efforts include repositories of fact-checking stories [Couto et al. 2021], niche-specific data such as COVID-19 misinformation [Martins et al. 2021], and even multimodal synthetic content [Santos et al. 2023]. While these resources demonstrate the growing maturity of the field, they remain fragmented—released as standalone corpora with heterogeneous metadata. This fragmentation limits cross-dataset evaluation and the development of robust, source-agnostic models.

Data redundancy and near-duplicate narratives often bias evaluations, yet systematic deduplication and standardized documentation [Bender and Friedman 2018] remain scarce in Portuguese resources. We address these gaps by consolidating corpora under a canonical schema, applying semantic similarity controls, and ensuring reproducibility through transparent documentation

3. Dataset Construction Method

The dataset integrates publicly available Portuguese-language misinformation corpora from Web news, Twitter/X, and public WhatsApp messages. The following subsections summarize the source datasets, selection criteria, processing pipeline, and harmonization procedures used to build the final corpus.

3.1. Data Sources

Table 1 summarizes the integrated datasets by domain, collection strategy, temporal coverage, and reference. Together, they span political news, electoral content, COVID-19 misinformation, and other public-interest narratives.

Table 1. Integrated source datasets included in FakenewsBR.

Dataset	Source Type	Period
MuMiN (PT) [Nielsen and McConville 2022, Gomes et al. 2023]	X/Twitter	2020–2022
COVID-19.BR [Martins et al. 2021]	WhatsApp	2020
LLM4BR-300 [Gôlo et al. 2024]	Political news	2018–2022
FakeWhatsApp-BR [Cabral et al. 2021]	WhatsApp	2018
Fake.br [Monteiro et al. 2018]	Web news	2016–2018
Portuguese Fake News [Selau 2021]	Web news	2005–2022

Datasets were selected for public research availability, explicit false/true labels, and sufficient textual completeness for normalization. No additional scraping was performed; the goal was to consolidate curated corpora while preserving their original annotations.

Combining multiple corpora reduces single-platform and single-domain bias, increasing platform heterogeneity, topical diversity, and temporal breadth [Ribeiro et al. 2018]. Because the sources differ in schema structure, metadata granularity, and attribute naming, explicit harmonization is required. Each record therefore retains provenance metadata to support reproducibility, cross-source analysis, and controlled experimentation [Buneman et al. 2001].

3.2. Data Collection Pipeline

The pipeline incorporates elements of the evidence-enrichment workflow described in [Gomes 2025], while adapting them to a broader corpus-integration setting. Filtering criteria, similarity thresholds, and enrichment settings are externally configurable and version-controlled, so identical configurations yield identical outputs. The workflow runs end-to-end via a command-line interface and records logs, metadata manifests, and cryptographic hashes to support integrity checks and artifact traceability.

As shown in Figure 1, the workflow is deterministic and modular: raw ingestion, schema normalization, cleaning and filtering, near-duplicate analysis, and verification-assisted review. Each stage produces explicit intermediate artifacts, so the process can be inspected, rerun, or partially reconfigured without altering the source data.

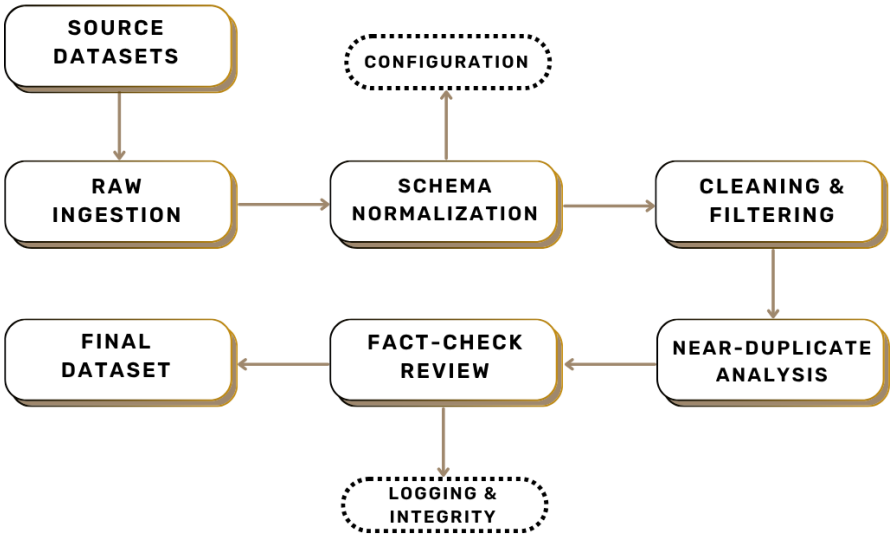

Figure 1. Overview of the data construction pipeline from source ingestion to final integrated dataset.

Schema Normalization. Dataset-specific mappings standardize labels, reconcile data types, merge title and body fields when necessary, and preserve provenance metadata. Nullable attributes accommodate source-specific variation without forcing artificial completeness. Table 6 (Appendix A) summarizes the resulting schema.

Data Cleaning and Filtering. Records are structurally normalized, labels are harmonized to a binary *fake/true* scheme whenever possible, and entries with missing or invalid labels are removed. The cleaning stage extracts URLs, builds normalized text fields, flags null records, and marks texts shorter than five whitespace-delimited tokens in the released configuration (*min_tokens*=5, but externally configurable) as *too_short*. Exact duplicates of the normalized text are also flagged within each source dataset. Rather than applying irreversible changes without traceability, the pipeline stores processing metadata through flags such as *is_null*, *too_short*, and *is_duplicated*, enabling reproducible alternative dataset variants.

Near-duplicate Analysis. Because misinformation narratives often reappear across platforms with minor lexical changes, highly similar records are detected with MinHash approximations of Jaccard similarity over character n-grams combined with locality-sensitive hashing [Broder 2000]. In the released implementation, the detector uses character 5-grams, a Jaccard similarity threshold of 0.70, MinHash seed 3, and 128 permutations; the configuration also retains an LSH band parameter of 50 for compatibility. The goal of this stage is not automatic deduplication, but rather to flag candidate near-duplicates for manual inspection and correction of inconsistent cases, such as items with the same meaning but opposite labels.

Fact-check Review. Candidate verification reports are retrieved through the Google Fact Check Tools API primarily to identify records that may require correction. Because the API returns usable matches for only a minority of records, this stage yields fact-check metadata for 18.18% of the final corpus. Matching records receive optional fields such as *factcheck_rating*, *factcheck_claimant*, and *factcheck_url*. Cases in which

the verdict returned by the API differs from the original dataset label are programmatically identified through stable `rid` values for manual review; when the evidence confirms an inconsistency, the released label may be corrected, while records with no reliable match remain unchanged. The current release does not include a standalone recoverable `corrections.log` or `corrections.csv`; therefore, these interventions are treated as review flags rather than auditable correction counts in the analyses below. Thus, external verification supports targeted review and complementary metadata attachment rather than automatic relabeling.

Table 2. Traceable treated cases by source dataset.

Dataset	Raw	Too short	Exact dup.	Removed	Final	FC match
Portuguese Fake News	23,198	175	101	276	23,081	7,617
Fake.br	7,200	7,200	7,199	7,200	14,359	4
FakeWhatsApp-BR	282,601	215	14,692	276,219	6,383	29
COVID-19.BR	2,899	2,899	2,898	2,899	4,277	27
MuMiN-PT	1,404	1,404	1,403	1,404	2,806	1,537
LLM4BR-300	300	0	1	1	299	93

Table 2 summarizes the number of cases treated per dataset using only the auditable pipeline artifacts. We report records removed for being too short, records removed as exact duplicates after cleaning, the total removed in the traceable pipeline, and the final number of records in the released corpus. Because the currently available artifacts do not include a recoverable log of manual corrections, the table intentionally excludes non-auditable counts such as post-review label corrections; instead, potential conflicts remain traceable at the record level through `rid` identifiers.

The values in Table 2 should be interpreted conservatively. “Raw”, “Too short”, “Exact dup.”, and “Removed” come from the auditable intermediate artifacts, whereas “Final” and “FC match” correspond to the released corpus used in the paper. This distinction is important because some datasets show discrepancies between the traceable pipeline outputs and the final release; therefore, the table is intended to document observable treatment operations rather than claim a fully reconstructable end-to-end history for every record.

Table 3 presents representative treated cases in two separate subtables. Subtable 3a groups traceable pipeline treatments, whereas Subtable 3b isolates label/fact-check mismatches flagged during verification-assisted review. Because the repository does not provide an explicit revision log, the latter are reported as `rid`-level conflicts, not as directly auditable corrections or official correction counts.

Regarding fact-check mismatches, a notable case involved a news item initially labeled as *true* in the source dataset, but later identified as *fake* by a high-confidence match in the Google Fact Check API. Since our pipeline maintains a unique `rid` (record identifier) through an SQLite-backed join, these divergences can be traced and inspected at the record level, but they are not counted as confirmed correction events unless supported by recoverable revision artifacts.

Reproducibility and Data Integrity. To facilitate reuse, the dataset is distribu-

Table 3. Illustrative examples of treated or flagged cases.

(a) Traceable pipeline treatments.

Dataset	Excerpt	Action
Portuguese Fake News	NEVOU EM CURITIBA	removed (*too_short*)
FakeWhatsApp-BR	No dia 07 de Outubro, quando for votar, não esqueça do que os esquerdopatas pensam da classe média.	removed (*is_duplicated*)

(b) Label/fact-check mismatches flagged during review.

Dataset	Excerpt	Status
COVID-19.BR	Vacina contra coronavirus é aprovada em primeiro estudo com humanos.	kept; mismatch flagged
MuMiN-PT	Vacina provoca surto de poliomielite no Sudão.	kept; mismatch flagged

ted with a dedicated Python toolkit. This CLI allows users to reproduce the entire pipeline—from raw ingestion to fact-check enrichment—using simple commands such as `fakenews-br-data pipeline`. Integrity is guaranteed through SHA-256 cryptographic hashes for every file, recorded in an automated `manifest.json`. This approach ensures that any third-party modification to the data can be immediately detected, a critical feature for scientific reproducibility.

4. Dataset Description and Statistics

The final corpus contains 51,205 Portuguese-language instances represented under the canonical schema. Labels remain naturally imbalanced, with 35,150 *fake* instances (68.65%) and 16,055 *true* instances (31.35%). Table 4 reports the contribution of each source corpus.

Table 4. Final dataset composition by source corpus and label distribution.

Dataset	Fake	True	Total	%
Portuguese Fake News	20,364	2,717	23,081	45.08
Fake.br	7,180	7,179	14,359	28.04
FakeWhatsApp-BR	3,055	3,328	6,383	12.47
COVID-19.BR	1,723	2,554	4,277	8.35
MuMiN-PT	2,680	126	2,806	5.48
LLM4BR-300	148	151	299	0.58
Total	**35150**	**16055**	**51205**	**100.00**

The integrated corpus is dominated by news-based content, but it also preserves Twitter/X and WhatsApp records, maintaining cross-platform heterogeneity. Valid publication dates are available for 26,415 instances (51.59%), spanning from September 2005 to December 2022. Text length remains highly variable, as expected for an integrated multi-source corpus. Measured over *text_no_url*, the average length is 89.70 tokens, the median is 25, the interquartile range is 13–125 tokens, and the maximum is 5,938 tokens.

Figure 2 summarizes the most salient terms in the integrated corpus. The prominence of words such as *lula*, *bolsonaro*, *presidente*, *governo*, and *federal* indicates that

Figure 2. Word cloud generated from the integrated corpus.

Table 5. Availability of selected metadata fields across the integrated dataset.

Field Name	Available (n)	%
`date_iso`	26,415	51.59
`url_review`	23,202	45.31
`factcheck_url`	9,307	18.18
`factcheck_rating`	9,307	18.18
`factcheck_claimant`	9,134	17.84
`extracted_urls`[a]	1,594	3.11

[a] Only considering non-empty lists.

the aggregated resource is strongly shaped by political and institutional discourse. In this respect, our result differs from [Couto et al. 2021], whose word clouds are dominated by fact-checking markers and verdict expressions in verification-story titles. At the same time, it is consistent with the lexical perspective adopted in [Gomes 2025], where frequent terms help characterize the dominant thematic profile of each dataset. Here, the integrated view reveals this concentration at the corpus level, showing that multi-source aggregation yields a politically salient vocabulary rather than the editorial language typical of fact-check repositories. This thematic concentration is useful for studying misinformation in electoral and institutional contexts, but it also represents a potential source of topic bias: models trained without topic-aware controls may learn names, actors, and agenda-specific lexical cues instead of more general misinformation patterns. Consequently, applications targeting health, science, climate-related misinformation, or financial scams should validate generalization with external test sets, source-aware splits, or explicit topic controls.

Table 5 summarizes the availability of key metadata fields. Structured fact-check metadata is available for 9,307 instances (18.18%), whereas *url_review* appears in 23,202 records (45.31%) because it is inherited from specific source datasets. All released instances correspond to the post-cleaning corpus after near-duplicate review; processing conditions remain documented through boolean flags for transparency and reproducibility.

Within this metadata group, the Google Fact Check label is preserved in *factcheck_rating* as the original textual verdict returned by the external review, rather than

being collapsed into the canonical target label. This design clarifies how our resource relates to two adjacent lines of work. [Couto et al. 2021] organizes fact-checking stories themselves, so verdict expressions and verification-oriented titles naturally become central descriptive signals. In contrast, [Gomes 2025] enriches Portuguese misinformation corpora with externally retrieved evidence to support semi-automated fact-checking. Our contribution lies between these perspectives: we normalize heterogeneous source datasets into a unified schema and a binary *fake*/*true* label for cross-corpus reuse, while preserving the Google-derived verdict separately as metadata. In this way, normalization supports interoperability without discarding the richer fact-check taxonomy attached to matched records.

Potential Uses. The unified schema and preserved provenance fields make *FakenewsBR* suitable for supervised fake news classification and for cross-domain evaluation across Web news, Twitter/X, and WhatsApp sources. The *dataset_name* and *source_type* attributes support robustness analyses across sources, while fact-check metadata and review URLs enable claim matching, evidence retrieval, and verification-oriented ranking tasks. Beyond predictive modeling, the corpus also provides a reproducible case study for data integration, schema harmonization, provenance tracking, and dataset versioning workflows.

Usage Limitations. These uses should account for the corpus characteristics documented in this paper. The natural class imbalance may bias supervised models if evaluation ignores class-sensitive metrics or stratified protocols. The politically salient vocabulary observed in Figure 2 can favor models that overfit to electoral and institutional discourse, reducing generalization to other misinformation topics. In addition, incomplete temporal metadata restricts unrestricted longitudinal analyses, and the 18.18% coverage of Google Fact Check metadata means that evidence-based tasks should either focus on the enriched subset or explicitly handle missing verification links.

5. Data Quality Analysis

The released corpus corresponds to the post-cleaning version of the integration pipeline. Structurally invalid and excessively short records were excluded, while highly similar records were reviewed to identify potential inconsistencies; the corresponding boolean flags were retained to preserve traceability and reproducibility.

At the same time, the dataset intentionally preserves the aggregate characteristics of its source corpora. Class imbalance remains unchanged, and the coexistence of news articles, tweets, and WhatsApp messages introduces substantial cross-domain heterogeneity in length, style, and discourse structure. This diversity is useful for robustness studies, but it also requires controlled evaluation to reduce stylistic overfitting; for this reason, the *dataset_name* and *source_type* fields are preserved.

Metadata completeness is partial rather than uniform. Valid timestamps are available for only slightly more than half of the corpus, which limits unrestricted longitudinal analysis, and fact-check metadata is concentrated in sources closer to verification workflows. For time-sensitive or enrichment-based studies, researchers should therefore filter explicitly by *date_iso* and fact-check fields.

Temporal heterogeneity is another relevant characteristic of the corpus. Because *FakenewsBR* spans 2005–2022 and aggregates datasets collected under distinct socio-

political conditions, earlier and later portions of the corpus reflect different actors, public agendas, and lexical patterns. As a result, part of the variation observed in the integrated dataset is associated with historically situated events, such as electoral cycles, political crises, and the COVID-19 pandemic, rather than only with more stable linguistic or discursive properties of misinformation. This should be taken into account in downstream analyses, since temporal effects are partially intertwined with source composition and topic prevalence across periods.

Finally, although near-duplicate analysis helps identify inconsistencies among highly similar records, related narratives still recur across platforms in ways that reflect realistic dissemination patterns. Overall, the corpus combines high structural integrity with transparent documentation of its remaining limitations, supporting informed reuse in database, machine learning, and misinformation research.

6. Conclusion

This work presented *FakenewsBR*, a standardized Portuguese-language fake news integration resource with 51,205 instances consolidated from six previously curated datasets. Through a reproducible pipeline, heterogeneous sources were harmonized under a canonical schema, cleaned, reviewed for highly similar cases, and partially enriched with fact-check metadata, while preserving provenance and transformation traceability. In this sense, the paper broadens the evidence-enrichment perspective discussed in [Gomes 2025] to a reproducible multi-source corpus-integration setting, emphasizing integration infrastructure and reusable documentation rather than the creation of newly collected raw data.

The resulting resource supports classification, cross-domain robustness evaluation, claim matching, retrieval, and data integration studies, including downstream claim normalization scenarios [Almada et al. 2025]. At the same time, it intentionally preserves class imbalance, cross-source heterogeneity, and partial metadata coverage, making its limitations explicit rather than hidden. By releasing both the dataset and its construction pipeline, this work contributes a reusable Portuguese-language benchmark aligned with reproducibility, transparency, and structured data integration.

Use of Artificial Intelligence

In this work, generative Artificial Intelligence technologies were used to support the writing, technical documentation, and presentation of the manuscript. Specifically, Prism, ChatGPT, Claude, and the Gemini website, in Reasoning mode, were employed for paraphrasing text excerpts, generating LaTeX table structures, and improving the clarity of the pipeline diagrams. Additionally, these tools assisted in the development and optimization of the Python scripts used for the MinHash LSH implementation and the SQLite data integration layer.

These tools were used only as writing, coding assistance, and visual-organization support and were not responsible for the conception, analysis, or interpretation of the results presented. All final content was reviewed, validated, and approved by the authors, who assume full responsibility for the work.

Acknowledgments

This work has been fully funded by the project Computational Techniques for Multimodal Data Security and Privacy supported by Advanced Knowledge Center in Immersive

Technologies (AKCIT), with financial resources from the PPI IoT/Manufatura 4.0 / PPI HardwareBR of the MCTI grant number 057/2023, signed with EMBRAPII.

References

Almada, F. L. N., Mariano, K. D. P., Dutra, M. A., da Silva Monteiro, V. E., Gomes, J. R. S., Filho, A. R. G., and da Silva Soares, A. (2025). Akcit-fn at checkthat! 2025: Switching fine-tuned slms and llm prompting for multilingual claim normalization.

Bender, E. M. and Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Broder, A. Z. (2000). Identifying and filtering near-duplicate documents. In Giancarlo, R. and Sankoff, D., editors, *Combinatorial Pattern Matching*, pages 1–10. Springer Berlin Heidelberg, Berlin, Heidelberg.

Buneman, P., Khanna, S., and Wang-Chiew, T. (2001). Why and where: A characterization of data provenance. In Van den Bussche, J. and Vianu, V., editors, *Database Theory — ICDT 2001*, pages 316–330, Berlin, Heidelberg. Springer Berlin Heidelberg.

Cabral, L., Monteiro, J. M., Franco da Silva, J. W., Mattos, C. L., and Mourão, P. J. C. (2021). Fakewhastapp.br: Nlp and machine learning techniques for misinformation detection in brazilian portuguese whatsapp messages. In *Proceedings of the 23rd International Conference on Enterprise Information Systems - Volume 1: ICEIS*, pages 63–74. INSTICC, SciTePress.

Couto, J., Pimenta, B., de Araújo, I. M., Assis, S., Reis, J. C. S., da Silva, A. P., Almeida, J., and Benevenuto, F. (2021). Central de fatos: Um repositório de checagens de fatos. In *Anais do III Dataset Showcase Workshop*, pages 128–137, Porto Alegre, RS, Brasil. SBC.

Gomes, J., Neto, V., Barbosa, J., and de Lima, E. (2023). A rapid tertiary review at the fake news domain. In *Anais da XI Escola Regional de Informática de Goiás*, Porto Alegre, RS, Brasil. SBC.

Gomes, J. R. S. (2025). Verificação semi-automática de fatos em português: enriquecimento de corpus via busca e extração de alegação. Dissertação (mestrado em ciência da computação), Universidade Federal de Goiás, Goiânia, Brasil. Instituto de Informática.

Gôlo, M., Mori, A., Oliveira, W., Barbosa, J., Graciano-Neto, V., Lima, E., and Marcacini, R. (2024). On the use of large language models to detect brazilian politics fake news. In *Anais do XXI Encontro Nacional de Inteligência Artificial e Computacional*, pages 1–12, Porto Alegre, RS, Brasil. SBC.

Hovy, D. and Spruit, S. L. (2016). The social impact of natural language processing. In Erk, K. and Smith, N. A., editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics.

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., and Zittrain, J. L. (2018). The science of fake news. *Science*, 359(6380):1094–1096.

Martins, A. D., Cabral, L., Mourão, P., de Sá, I., Monteiro, J., and Machado, J. (2021). Covid19.br: A dataset of misinformation about covid-19 in brazilian portuguese whatsapp messages. In *Anais do III Dataset Showcase Workshop*, pages 138–147, Porto Alegre, RS, Brasil. SBC.

Monteiro, R. A., Santos, R. L. S., Pardo, T. A. S., de Almeida, T. A., Ruiz, E. E. S., and Vale, O. A. (2018). Contributions to the study of fake news in portuguese: New corpus and automatic detection results. In *Computational Processing of the Portuguese Language: 13th International Conference, PROPOR 2018, Canela, Brazil, September 24–26, 2018, Proceedings*, page 324–334, Berlin, Heidelberg. Springer-Verlag.

Nielsen, D. S. and McConville, R. (2022). Mumin: A large-scale multilingual multimodal fact-checked misinformation social network dataset. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 3141–3153, New York, NY, USA. Association for Computing Machinery.

Ribeiro, M., Calais, P., Santos, Y., Almeida, V., and Meira Jr., W. (2018). Characterizing and detecting hateful users on twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1).

Santos, Y., Silva, M., and Reis, J. C. S. (2023). Imagefactck.br: Repositório de imagens para a detecção de desinformação disseminada em plataformas digitais. In *Anais do V Dataset Showcase Workshop*, pages 87–98, Porto Alegre, RS, Brasil. SBC.

Selau, F. (2021). Fakes news in portuguese. `https://www.kaggle.com/datasets/fabioselau/fakes-news-portuguese`. Accessed: April 14, 2026.

Shu, K., Mahudeswaran, D., Wang, S., Lee, D., and Liu, H. (2020). Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big Data*, 8(3):171–188. PMID: 32491943.

Shu, K., Sliva, A., Wang, S., Tang, J., and Liu, H. (2017). Fake news detection on social media: A data mining perspective. *SIGKDD Explor. Newsl.*, 19(1):22–36.

Silva, R. M., Santos, R. L., Almeida, T. A., and Pardo, T. A. (2020). Towards automatically filtering fake news in portuguese. *Expert Syst. Appl.*, 146(C).

Wang, W. Y. (2017). "liar, liar pants on fire": A new benchmark dataset for fake news detection. In Barzilay, R. and Kan, M.-Y., editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

A. Appendix

This appendix presents the canonical schema referenced in the schema-normalization stage. Table 6 details the attributes used to harmonize heterogeneous source datasets under a single representation. For clarity, the fields are grouped into identification and metadata, textual content, fact-checking details, and quality and processing flags, making explicit the information preserved for provenance tracking, reproducibility, and downstream analysis.

Table 6. Canonical schema of the integrated FakenewsBR dataset.

Attribute	**Type**	**Description**
Identification & Metadata		
rid	str	Unique internal record identifier.
dataset_name	str	Canonical name of the original source dataset.
source_type	str	Origin of content (e.g., Web, WhatsApp, X/Twitter).
source_description	str	Brief context about the source corpus.
label	str	Standardized veracity label (e.g., true, fake).
date_iso	str	Publication date in ISO 8601 format (YYYY-MM-DD).
Textual Content		
text	str	Raw unified textual content.
text_clean	str	Preprocessed text (lowercased, special chars removed).
text_no_url	str	Text version with all URLs stripped.
extracted_urls	list	List of URLs found within the content.
Fact-checking Details		
url_review	str	Reference URL for the fact-check article.
factcheck_rating	str	Original verdict assigned by the fact-checker.
factcheck_claimant	str	Entity or person who made the claim.
factcheck_url	str	Direct link to the external verification.
Quality & Processing Flags		
is_duplicated	bool	True if the entry is a semantic duplicate.
is_null	bool	True if the entry was flagged as empty or invalid.
too_short	bool	True if the text fell below the minimum length threshold.