

Portuguese Automated Fact-checking: Information Retrieval with Claim Extraction

Juliana Gomes, Eduardo Garcia, Arlindo Rodrigues Galvão Filho

Advanced Knowledge Center for Immersive Technologies - AKCIT
Institute of Informatics, Federal University of Goiás

Correspondence: juliana.resplande@discente.ufg.br

Abstract

Current Portuguese Automated Fact-Checking (AFC) research often relies on datasets lacking integrated external evidence crucial for comprehensive verification. This study addresses this gap by systematically enriching Portuguese misinformation datasets. We retrieve web evidence by simulating user information-seeking behavior, guided by core claims extracted using Large Language Models (LLMs). Additionally, we apply a semi-automated validation framework to enhance dataset reliability.

Our analysis reveals that inherent dataset characteristics impact data properties, evidence retrieval, and AFC model performance. While enrichment generally improves detection, its efficacy varies, influenced by challenges such as self-reinforcing online misinformation and API limitations. This work contributes enriched datasets, associating original texts with retrieved evidence and LLM-extracted claims, to foster future evidence-based fact-checking research.

The code and enriched data for this study is available at https://github.com/ju-resplande/pt_afc.

1 Introduction

News fact-checking agencies, such as Agência Lupa and Boatos.org in Brazil, manually investigate the veracity of claims (Faustini and Covões, 2019; Couto et al., 2021). However, the speed at which viral messages spread surpasses the capacity of investigative journalists to verify the accuracy of the information.

This inherent limitation has spurred the development of Automated Fact-Checking (AFC) systems, which aim to verify claims by leveraging external knowledge sources through techniques from Information Retrieval (IR) and Natural Language Processing (NLP) (Guo et al., 2022).

Despite advances in AFC, a significant gap persists within the Portuguese language context. Our

investigation reveals that existing approaches and datasets for misinformation detection in Portuguese predominantly focus on intrinsic content analysis, such as writing style or bias (Monteiro et al., 2018), rather than incorporating the crucial step of external evidence verification.

This work aims to directly address this lacuna. Our objective is to develop, apply, and analyze a methodology for enriching existing Portuguese-language misinformation datasets with relevant external evidence retrieved through web search mechanisms. This process mimics how users might seek corroborating information online. To guide the search effectively, especially when initial queries yield suboptimal results, we employ Large Language Models (LLMs) to extract the main claim from each news item, which then serves as an optimized query.

The key contributions of this work are: (i) A comparative analysis of Portuguese-language misinformation-related datasets, detailing their characteristics, including domains, sources, and data collection methodologies (top-down vs. bottom-up (Hangloo and Arora, 2022)); (ii) A semi-automatic data validation process, addressing near-duplicates, instances referencing the same URL, and cross-verification using the Google FactCheck Claim Search API; (iii) A methodology for enriching Portuguese misinformation datasets with external evidence retrieved via web search, guided by LLM-based claim extraction, alongside an experimental evaluation of this enrichment’s impact on the performance of misinformation detection models.

2 Related work

Large Language Models (LLMs), such as Gemini and ChatGPT, are increasingly central to fact-checking. They offer a cost-effective alternative to manual methods like crowdsourcing or expert

review (Ali et al., 2022; Aimeur et al., 2023), with performance often comparable to human crowd-sourcers (Maia and da Silva, 2024; Ni et al., 2024). A key advantage of LLMs is their proficiency in zero-shot or few-shot learning and their ability to generate explanations, frequently without the need for supervised fine-tuning (Gangi Reddy et al., 2022; Chen and Shu, 2024). This marks a notable distinction from smaller language model architectures (SLMs), which typically require substantial supervised training for specific tasks (Qiu and Jin, 2024).

In automated fact-checking pipelines (Guo et al., 2022), LLMs are employed at various stages. These include claim identification (extracting main claims (Kotitsas et al., 2024; Ni et al., 2024), multiple claims (Tang et al., 2024), or generating search queries (Cho et al., 2022)), claim verification (the detection task itself) (Tan et al., 2023; Choi and Ferrara, 2024), and the generation of justifications for verification outcomes (Kim et al., 2024; Zeng and Gao, 2024).

While prior work has utilized LLMs for claim extraction (Kotitsas et al., 2024; Ni et al., 2024) or for zero-shot classification of claims (Tan et al., 2023; Choi and Ferrara, 2024), our research distinctly focuses on enriching existing Portuguese datasets. We achieve this by retrieving web-based evidence, where the search process is triggered by main claims extracted from the news items using LLMs.

Numerous Portuguese-language datasets for misinformation detection exist (Monteiro et al., 2018; Moreno and Bressan, 2019; Vargas et al., 2023; Nielsen and McConville, 2022). However, a significant challenge is that few of these datasets provide directly accessible external evidence. Many only offer tweet identifiers or indirect references, which hinders the direct use of evidence for verification (Cordeiro and Pinheiro, 2019; da Silva et al., 2020; Shahi and Nandini, 2020). Our work aims to address this critical gap by systematically augmenting these resources with verifiable, web-retrieved evidence.

3 Methods

The core of our methodology involves enriching misinformation datasets with external evidence. This process, depicted in Figure 1, begins with an input text from a selected misinformation dataset. If an initial web search for this text (or a query

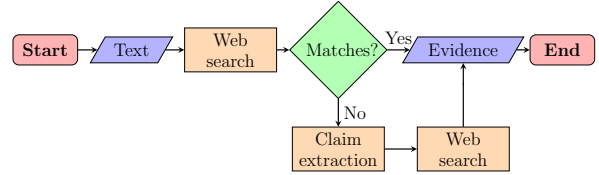


Figure 1: Flowchart for evidence retrieval from an input text. The web search was conducted using Google Custom Search Engine (CSE), and claim extraction was performed by Gemini 1.5 Flash.

derived from it) yields results with strong lexical correspondence to the query, claim extraction is bypassed. Otherwise, an LLM extracts the main claim from the text, and this extracted claim is used for a subsequent, more targeted web search. Each component is detailed below.

3.1 Dataset Selection

We selected datasets for public availability, a minimum of 1000 examples, and a peer-reviewed publication. We excluded those solely comprising fact-checking reports (not original misleading content). Table 1 details the three selected datasets:

Fake.br (Monteiro et al., 2018). Relevant for its canonical status in Portuguese misinformation research, this dataset consists of news articles from web pages, covering general domains, collected between 2016 and 2018. It uses a bottom-up collection approach, directly gathering content from these web pages (Hangloo and Arora, 2022).

COVID19.BR (Martins et al., 2021; de Sá et al., 2021). Contains messages from the WhatsApp platform from 236 public groups, focusing on the health topic (COVID-19), collected between April and June 2020. It also follows a bottom-up collection approach.

MuMiN-PT (Nielsen and McConville, 2022) is a Portuguese corpus of X (formerly Twitter) posts from general domains (2020-2022). It provide a crucial methodological contrast to the others, having been collected using a top-down approach (Hangloo and Arora, 2022) by finding posts corresponding to claims verified by fact-checkers. It is a subset of the multilingual MuMiN corpus, extracted with Lingua (Stahl, 2023).

3.2 Dataset Validation and Cleaning

To ensure the quality and integrity of our datasets, we implemented a rigorous semi-automatic validation and cleaning pipeline. This multi-stage process combined automated scripts with targeted manual

Dataset	Domain	Source	Collection Approach	Size per label	Temporal Coverage	Fact-Checking Sources
Fake.BR (Monteiro et al., 2018)	General	News Websites	Bottom-up	3,600 true 3,600 fake	2016–2018	estadao.com.br folha.uol.com.br g1.globo.com
COVID19.BR (Martins et al., 2021)	Health	WhatsApp	Bottom-up	1,987 true 912 fake	2020	boatos.org lupa.uol.com.br
MuMiN-PT (Nielsen and McConville, 2022)	General	Twitter (now X)	Top-down	1,339 fake 65 true	2020–2022	afp.com/pt aosfatos.org projetoacomprova.com.br observador.pt oglobo.globo.com piaui.folha.uol.com.br uol.com.br

Table 1: Characteristics of datasets enriched with external verification evidence in this study. Collection approaches follow the classification by [Hangloo and Arora \(2022\)](#), where **top-down** involves collecting posts for known, long-standing rumors (often starting from fact-checking websites), and **bottom-up** involves gathering all relevant posts from a given time frame to identify emerging misinformation. While original publications list verification sources, they typically do not provide the specific evidence snippets used for fact-checking each item.

review by one of the authors to identify and rectify issues ranging from formatting errors to label inconsistencies. The primary steps in our pipeline were as follows:

1. **Automated Initial Filtering:** Removal of exact duplicates and entries with fewer than 15 tokens after removing emojis, URLs, stopwords, and punctuation¹.
2. **Language Filtering:** Automatic Portuguese detection via Lingua ([Stahl, 2023](#)), followed by manual verification and exclusion of non-Portuguese examples.
3. **Contradictory Duplicate Removal:** Manual review of near-duplicate pairs that were automatically identified via MinHash LSH (Section 4.1.1) with inconsistent labels. The conflicting instances were subsequently removed to maintain dataset integrity.
4. **External Label Verification:** Manual review of instances where a dataset’s label potentially conflicted with results from the Google FactCheck Claim Search API. The original labels were corrected when the external evidence provided a clear refutation or confirmation.
5. **URL Reference Removal:** Manual review of instances that referenced the exact same URL within their text but possessed differing veracity labels. The conflicting instances were subsequently removed to maintain dataset integrity.
6. **Random Inspection:** Manual check of a random subset from each dataset for overall qual-

ity.

7. **Specific Treatment for Fake.br:** (a) Removed near-duplicates from the same source URL, as the source in this corpus is a URL; (b) Examples lacking full text were removed from the original authors’ normalized set²; (c) Maintained pair integrity by removing items whose pair was previously excluded.

The number of instances corrected or removed at each stage is quantified in Table 2. For a detailed breakdown of case counts and illustrative examples, please see Appendix A.

Validation Stage	COVID19.BR	Fake.br	MuMiN-PT
Initial Automated Filtering	804	1	0
Language Filtering	8	0	11801
Contradiction Resolution	20	0	0
External Label Verification	23	0	4
Subset Inspection	88	0	0
Fake.br Specific Treatment	0	61	0

Table 2: Number of examples corrected or removed in the corpora during the semi-automatic validation workflow.

As a final preprocessing step, all explicitly mentioned URLs were removed from the original texts to mitigate potential biases from URL domains, as shown in Appendix B.

3.3 LLM-based Claim Extraction

For claim extraction, we utilized the Gemini 1.5 Flash model. To simulate a user’s initial scan of a text for its core assertion, we provided the LLM with up to the first 75 words of the news item or message.

¹using cl100k_base tokenizer from <https://github.com/openai/tiktoken>

²<https://github.com/roneysco/Fake.br-Corpus/issues/7>

We experimented with variations on a small data subset. We opted for single main claim extraction, excluding multi-claim splitting to avoid generating multiple queries per input item. The final prompt used is shown in Figure 2. This zero-shot prompt proved effective for our goal of generating concise search queries, without requiring specific role instructions or few-shot examples.

What is the main fact presented in the text?

1. Extract a passage of up to 20 words from the following text.
2. Return only the claim, without any title or preamble.

Text: [INPUT TEXT]
Claim:

Figure 2: Prompt template used with Gemini 1.5 Flash for main claim extraction.

3.4 Evidence Retrieval via Web Search

Our evidence retrieval process involves up to two stages of web searching.

Initial Web Search Strategy. The initial input text (from the dataset) is queried using the Google Custom Search Engine (CSE) API³. Query parameters for CSE were `gl=pt-BR` (geographical restriction to Brazil), and `lr=lang_pt` (language restriction to Portuguese).

To simulate how a user might create a concise search query, we developed a set of empirically-determined heuristics. The logic was designed to quickly extract the core assertion from a given text. For short texts of 20 words or less, the entire text was used as the query. For longer texts, we first extracted the initial sentence. If this sentence was long enough to likely contain the main claim (7 words or more), it became the query. However, if the first sentence was too brief (fewer than 7 words), the query was then formed by taking the longer of either the full first paragraph or the first 20 words of the text. Emojis and specific Unicode quote characters were removed from queries.

Correspondence Check. The lexical containment of the query (excluding stopwords) within CSE snippets is assessed. The proportion of the query’s non-stopword terms found in the snippet is calculated. Empirically, if $\geq 80\%$ of the query’s non-stopwords are present in the snippet, claim extraction is bypassed.

³<https://developers.google.com/custom-search/v1/reference/rest/v1/cse/list>

FactCheck API Search. The input text is also searched using the Google FactCheck Claim Search API⁴ with parameters `languageCode=pt-BR`, and `pageSize=5`. The same query preprocessing from CSE search was applied.

Claim-based Fallback Search. If the initial CSE search is unsuccessful, a claim is extracted, and the claim initiates a second search using both CSE and FactCheck Claim Search APIs with original parameters.

If a CSE result links to a page indexed by Google FactCheck, a `ClaimReview` schema might be present in the CSE result’s metadata. While this schema typically does not include the veracity label itself directly in the CSE metadata, we store the first such linked result if found via CSE.

4 Cleaned and Enriched Data Analysis

4.1 Cleaned Dataset

After applying the validation and cleaning steps from Section 3.2, we analyzed the remaining data. Table 3 shows statistics for the cleaned datasets.

Stats.	MuMiN-PT		Fake.br		COVID19.BR	
	fake	true	fake	true	fake	true
count	1339	65	3580	3580	848	1139
% URL	0.3%	0.0%	1.0%	0.7%	28.9%	56.5%
# words	18.9	16.3	181.4	183.1	167.7	111.1
word len (chars)	5.0	4.9	4.8	5.0	4.9	6.6
# sents	1.4	1.4	10.4	9.0	10.9	5.8
# words/sent	14.5	12.3	18.6	22.1	19.2	22.9

Table 3: Cleaned dataset statistics: overall size and word/sentence counts per label.

Table 3 summarizes the textual statistics. The average text length varies significantly, reflecting the source platform: posts on X (MuMiN-PT) are the shortest, followed by WhatsApp messages (COVID19.BR), and web news articles (Fake.br). The average word length, however, was remarkably stable across datasets and labels, hovering around 4.8-5.0 characters.

In terms of labels, MuMiN-PT does not show variation in word and sentence statistics between true and false labels, possibly due to the low character limit imposed on X. Fake.br’s true news is

⁴<https://developers.google.com/fact-check/tools/api#the-google-factcheck-claim-search-api>

typically longer, so its original paper (Monteiro et al., 2018) normalized text length by truncation for fair classification. This work uses this normalized version (see Section 3.1). Conversely, in COVID19.BR, fake texts are longer and more variable, which its authors attribute to different writing styles (Martins et al., 2021).

COVID19.BR leads in link count (889 examples, 44.7%), likely because WhatsApp environments foster more link sharing than websites or X. True news in both COVID19.BR and Fake.br feature more links. MuMiN-PT, as expected for X posts, has virtually no links.

4.1.1 Near-Duplicates Analysis

Near-duplicate detection used the Akin library⁵ (v0.1.0) with MinHash LSH (char 5-grams, 128 hash bits, 50 bands, Jaccard threshold 0.7, seed=3).

COVID19.BR had the most (271 examples, 13.6% of cleaned data), likely due to easy sharing of short, similar messages on WhatsApp. Fake.br had the fewest (6 examples, 0.08%), while MuMiN-PT had 28 (2.0%). In Fake.br, near-duplicates were all true news articles, often minor updates republished by the same outlet (see Appendix C). Conversely, near-duplicates in MuMiN-PT (all examples) and COVID19.BR (186 examples) were predominantly false. This pattern, especially the volume in COVID19.BR, likely reflects viral misinformation spread via short, easily shared messages on WhatsApp

4.1.2 Observed Data Limitations

Beyond near-duplicates, we noted other challenges:

Topic Temporality: Misinformation topics and associated vocabulary are volatile and time-dependent. Evidence found today might not reflect the context when the claim first appeared.

Label Temporality: The truthfulness of a claim can change over time (e.g., “Mask use is mandatory in Goiás” depends on the specific date).

Degrees of Veracity: Fact-checking employs a spectrum of veracity labels beyond a simple true/false dichotomy, such as “misleading,” “partly false,” and “unproven” (Wang, 2017; Hangloo and Arora, 2022; Couto et al., 2021).

Claim Verifiability: Some texts express personal experiences or uncheckable anecdotes (e.g., “I want to stay quietly at home. Today a gentleman with coronavirus passed away here.” from COVID19.BR).

URL Dependence: Analyzing text without linked content misses crucial verification data. Additionally, URL domains can act as strong veracity heuristics due to varying frequencies in true/false claims (Appendix B).

4.2 Claim Extraction Outcomes

Per Section 3, Gemini 1.5 Flash extracted claims when the initial CSE search failed to find strong correspondence, which was for 94.0% (1,868/1,987) of COVID19.BR, 80.1% (5,738/7,160) of Fake.BR, and 70.7% (992/1,404) of MuMiN-PT examples.

The lower need for extraction in MuMiN-PT might relate to its bottom-up construction (starting from verified claims, potentially leading to posts that closely match searchable claims) and the concise nature of posts on X. Conversely, Fake.BR often required extraction, possibly because full news articles are less likely to directly match concise search results without summarizing the core claim. COVID19.BR (WhatsApp messages) frequently required extraction, perhaps due to conversational context or less structured claims.

Analysis of the successfully extracted claims (using the prompt in Figure 2) showed an average length of 11-12 words per claim, with an average word length of 4.9 characters, consistent across the corpora. Appendix E shows examples of the full enrichment process.

4.3 Search Engine Results Analysis

The top three domains found in the Google CSE results (combining results from both initial searches and searches using extracted claims) were consistent across searches for all three datasets: Brazilian government sites (‘gov.br’), Globo (major media network), and the BBC. Brazilian government sites accounted for 34.0% of results for COVID19.BR, 25.0% for Fake.BR, and 23.1% for MuMiN-PT. Globo represented 3.9%, 11.0%, and 7.0% respectively, while the BBC appeared in 2.8%, 3.7%, and 3.7% of results for the same datasets.

Table 4 presents the domains corresponding to fact-checking search results. These results refer to instances where a CSE search result links to a page containing a ClaimReview schema, as indicated in the CSE result’s metadata. However, this metadata from CSE typically does not include the actual veracity label (e.g., true, false) of the claim.

In particular, the domains e-farsas.com, sbt.com.br, and sapo.pt noted in our search results (Table 4) do not appear as source agencies

⁵<https://github.com/justinbt1/Akin/tree/v0.1.0>

Domain	COVID19.BR	Fake.br	MuMiN-PT
afp.com	8	10	154
uol.com.br	5	11	95
observador.pt	4	4	80
estadao.com.br	4	17	11
e-farsas.com	0	1	8
sbt.com.br	3	5	0
globo.com	0	0	7
projetoconprova.com.br	0	0	7
sapo.pt	0	1	0

Table 4: Count of fact-checking agency domains found within CSE search results containing ClaimReview metadata. Parent domains aggregate related subdomains (e.g., uol.com.br includes lupa.uol.com.br).

in Table 1. Alternatively, boatos.org is indicated as a source agency there but is absent from our CSE-retrieved ClaimReview findings.

Some CSE results from agency domains included publication dates. Figure 3 plots these dates. Results associated with COVID19.BR peaked in 2020, and MuMiN-PT results clustered in 2020-2022, aligning well with their collection periods (Table 1). Fake.br results showed a broader date distribution, often more recent than the original 2016-2018 collection period.

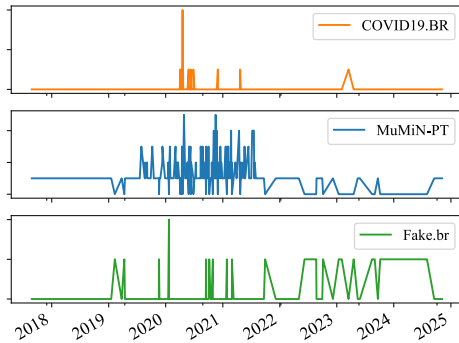


Figure 3: Publication dates of agency results from the CSE search that included ClaimReview metadata.

The Google FactCheck Claim Search API yielded results in 58.8% of MuMiN-PT, 5.0% of COVID19.BR, and 0.9% of Fake.br cases. The high rate for MuMiN-PT is expected, as Google FactCheck Claim Search was involved in its creation process. The low rate for Fake.br might be due to its age and website source (less likely indexed by the ClaimReview schema compared to social media or recent news). COVID19.BR’s rate is intermediate, reflecting its WhatsApp origin and timing during the pandemic.

The FactCheck API directly provides the veracity label assigned by the reviewing organization. Table 5 summarizes the labels found across 986

claims retrieved successfully via this API for all datasets combined. The labels predominantly indicate falsity or related categories. Only 11 claims were explicitly labeled “true”. This supports the observation that fact-checking efforts predominantly target content perceived as problematic or false.

Compared to the CSE results with ClaimReview metadata (Table 4), the direct FactCheck API query retrieved results from boatos.org but not from e-farsas.com or sbt.com.br.

The label distribution generally aligns with findings by Couto et al. (2021), although direct comparison is limited by differences in data collection scope and time period.

4.4 Evidence Patterns from Search Results

Analysis of CSE and FactCheck API search results revealed distinct evidence patterns for true/false claims:

For TRUE claims:

T1: Corroboration: Links to original/reliable sources confirming the information (Appendix E, T1 Ex.).

T2: Absence of Explicit Confirmation: Fact-checkers rarely explicitly confirmed true claims via CSE/API (as they primarily debunk, Table 5).

For FALSE claims:

F1: Direct Debunk: Reputable articles/fact-checks directly refute the claim (Appendix E, F1 Ex.).

F2: Misinformation Reinforcement: Results often surface the same misinformation from unreliable sources, amplifying it (Appendix E, F2 Ex.).

F3: Academic Recognition as Misinformation: Scholarly articles identify the claim as known misinformation (Appendix E, F3 Ex.).

While T1/F1 provide clear veracity signals, F2 can propagate misinformation, and F3 offers meta-evidence of falsity. (Examples: Appendix E). While T1 and F1 offer clear veracity signals, F2 can propagate misinformation. F3 provides unique meta-evidence of a claim’s dubious nature. Illustrative examples are in Appendix E.

5 Experimental Strategy and Setup

To assess the impact of our data enrichment, we established the following experimental configurations, which were applied to the COVID19.BR and Fake.BR datasets. While MuMiN-PT was inval-

Fact-checking Source	Total Claims Retrieved	Assigned Veracity Labels (Count)
aosfatos.org	283	false (246) , distorted (29), not quite so (2), unsustainable (2), exaggerated (2), contradictory (2)
uol.com.br	150	false (130) , misleading (10) , unsustainable (7), distorted (2), satire (1)
observador.pt	147	wrong (124) , misleading (23)
lupa.uol.br	114	false (100) , true (5), true but (5), exaggerated (2), under review (1), too early to tell (1)
boatos.org	99	false (99)
afp.com/pt	71	false (48) , misleading (17) , unverified (2), true (1), no evidence (1), out of context (1), lacks context (1)
estadao.com.br/estadao-verifica	47	misleading (25) , false (20) , out of context (2)
projetoacomprova.com.br	45	false (24) , misleading (21)
g1.globo.com/fato-ou-fake	11	fake (11)
bol.uol.com.br	2	false (2)
folha.uol.com.br	2	misleading (1) , false (1)
piaui.folha.uol.com.br	2	false (2)

Table 5: Veracity labels assigned by fact-checking agencies as retrieved via the Google FactCheck Claim Search API across all datasets. The most frequent labels, primarily indicating falsity or deception, are highlighted in bold. These results were used to guide our manual “External Label Verification” step (Section 3.2), where conflicts with an original dataset’s binary label prompted a review. Appendix A.4 details which labels were considered evidence for FAKE NEWS or TRUE.

able for the preceding comparative analyses of data characteristics and evidence retrieval patterns (Sections 4.1 to 4.4), it was excluded from classification experiments due to severe class imbalance (Table 3).

1. **Validated-only Data (Baseline):** Datasets after our full semi-automatic validation workflow (Section 3.2). This configuration serves as the baseline to measure the impact of enrichment.
2. **Validated and Enriched Data:** Validated datasets supplemented with external web search evidence, evaluated under two conditions:
 - (a) **Complete Enrichment:** Using the first Google API CSE search result as context.
 - (b) **Filtered Enrichment (No Social Media):** Excluding social media results (e.g., Twitter/X, Facebook) from CSE searches.

For comparability, both COVID19.BR and Fake.br used a standardized 80/10/10 train/validation/test split. For Fake.br, news pairs from the original dataset were kept together within each partition to ensure methodological consistency.

Model performance per data configuration was measured via: 1) supervised fine-tuning of the Portuguese SLM Bertimbau base (Souza et al., 2020), and 2) few-shot learning with the Gemini 1.5 Flash

LLM (Reid et al., 2024).

5.1 Evaluation Environment

For few-shot learning, the prompt for each test instance included the same set of 15 randomly selected examples from the training set. All experiments were run 3 times to be averaged, but the result between each experiment was the same. Figure 4 presents the base prompt used for inference by the model.

The following are texts from messages and news in Portuguese. Your task is to classify each text as containing FAKE NEWS or as being TRUE.
To assist in the classification, extra context will also be provided, corresponding to a Google search for the terms of the text to be classified.
Respond **only** with one of the following tags: “FAKE NEWS” or “TRUE”.

Figure 4: Base prompt for misinformation detection. The underlined section is included when external context (retrieved evidence) is incorporated.

Fine-tuning experiments used PyTorch with SimpleTransformers (Rajapakse et al., 2024) on NVIDIA V100 (32GB VRAM) or A100 (80GB VRAM) GPUs. Runs typically required 8-12 GB VRAM, varying with data configuration and batch size. Hyperparameter grid search details are in Appendix F. Best hyperparameters were chosen via validation F1-score.

6 Experimental Results and Discussion

This section presents misinformation detection results for fine-tuned Bertimbau base and few-shot Gemini 1.5 Flash, comparing performance on validated-only data against data enriched with external evidence (Table 6).

The introduction of external evidence through enrichment generally led to performance gains over the validated-only baseline, although the impact varied by dataset and model. For COVID19.BR, complete enrichment improved the F1-macro score for both Bertimbau (+1.0) and Gemini (+2.4).

The introduction of external evidence through enrichment generally led to performance gains over the validated-only baseline, though the impact varied. On COVID19.BR, it improved the F1-macro score for both Bertimbau (+1.0) and Gemini (+2.4). In contrast, on Fake.BR, enrichment provided a modest improvement for Bertimbau (+0.3) but degraded performance for Gemini, a result potentially explained by temporally mismatched evidence for this older dataset (Figure 3).

A notable result is Bertimbau’s near-perfect performance on the canonical Fake.br dataset, a score suggesting the dataset may be ‘saturated’ for modern models. The original corpus authors (Monteiro et al., 2018; Silva et al., 2020) explain this by noting significant, systematic differences between the classes: true news was sourced from professional journalistic outlets while fake news came from amateur sites, resulting in disparate writing quality and style. Although we use a length-normalized version of the dataset, these other strong intrinsic signals likely contribute to the exceptionally high baseline performance, making it difficult to measure the marginal impact of our evidence enrichment. This underlines the importance of evaluating on more varied datasets like COVID19.BR, which proved to be a more challenging benchmark for assessing the value of external evidence.

Filtering social media from enriched data (Filtered vs. Complete) generally reduced performance for Bertimbau on both datasets and for Gemini on COVID19.BR. This suggests that the excluded social media content contained relevant signals for the models.

Overall, the fine-tuned Bertimbau base model consistently outperformed the few-shot Gemini 1.5 Flash across all configurations and datasets, which is expected given its supervised training on a larger volume of task-specific data.

7 Conclusion

Our analysis included three distinct datasets: the canonical Fake.br, the topic-specific COVID19.BR, and the bottom-up collected MuMiN-PT. While Fake.br and MuMiN-PT presented limitations for our final classification task, they were instrumental, alongside COVID19.BR, for our broader investigation. Fake.br, while foundational, exhibited signs of saturation, with models reaching near-perfect accuracy, making it a less sensitive benchmark for measuring evidence enrichment. MuMiN-PT was excluded from classification entirely due to severe class imbalance. This left COVID19.BR as our primary dataset for evaluating the impact of enrichment, as it provided a more challenging and less saturated baseline. Nevertheless, the inclusion of all three was invaluable for a comprehensive comparative analysis: they provided contrasting examples of collection methodologies (top-down vs. bottom-up), domains (general vs. health), and highlighted challenges such as temporal evidence misalignment (most prominent in the older Fake.br) and the varying effectiveness of evidence retrieval APIs across different data sources.

While a validation pipeline improved dataset reliability by removing various issues, models trained on the cleaned data sometimes performed slightly worse. This suggests the cleaning process, despite enhancing quality, might remove useful heuristic signals, thereby increasing the task’s intrinsic difficulty before external contextual evidence is applied.

A qualitative analysis of the retrieved evidence revealed significant variability in the usefulness of recovered content for fact-checking purposes. While true claims often found corroboration through reliable news sources and official statements, false claims presented more complex evidence patterns. The ideal evidence for false claims consisted of direct refutations from established fact-checking organizations, but such explicit debunking was not consistently available through our search methodology. Crucially, the retrieved evidence cannot be uniformly considered golden evidence for automated verification. Our analysis identified several quality issues: (1) the self-reinforcing nature of online misinformation; (2) varying coverage by the Google FactCheck Claim Search API; and (3) temporal misalignment between claims and available evidence, particularly affecting older datasets like Fake.BR. These findings highlight

Model	Processing	COVID19.BR		Fake.br	
		Accuracy	F1-macro	Accuracy	F1-macro
Bertimbau base (Souza et al., 2020)	Validated-only	81.1	81.4	98.9	98.9
	+Enriched (Complete)	82.1	82.4	99.2	99.2
	+Enriched (Filtered)	77.9	78.3	98.7	98.8
Gemini 1.5 Flash (Reid et al., 2024)	Validated-only	76.9	76.9	81.0	80.4
	+Enriched (Complete)	79.3	79.3	77.6	76.7
	+Enriched (Filtered)	79.0	78.9	78.1	77.2

Table 6: Accuracy and F1-Macro results for Bertimbau base (fine-tuned) and Gemini 1.5 Flash (few-shot) on the test set. Performance is compared between validated-only data (baseline) and validated data enriched with external evidence. The best scores for each model and dataset are in bold.

that evidence retrieval quality significantly impacts the potential effectiveness of evidence-based fact-checking systems.

Fine-tuned Bertimbau (SLM) consistently outperformed few-shot Gemini 1.5 Flash (LLM). While external web content enrichment generally improved performance over validated-only data (notably for COVID19.BR), its benefits varied. Efficacy depended on search quality, claim extraction accuracy, API coverage, and temporal alignment of claims and evidence.

The enriched datasets generated in this study, which associate original claims with retrieved web evidence snippets and LLM-extracted main claims, offer valuable resources for future research. Potential applications include developing more sophisticated evidence-based fact-checking models, exploring stance detection between claims and evidence, and investigating the temporal dynamics of misinformation and its verification.

Our findings underscore that robust misinformation detection likely requires a hybrid approach, combining sophisticated textual analysis with a critical and nuanced evaluation of external evidence.

Limitations

The inherited binary veracity labels (true/false) do not capture the full veracity spectrum (e.g., “misleading”) used by fact-checkers (Hangloo and Arora, 2022; Wang, 2017), although our FactCheck API retrieval offered some nuance (Table 5). Some claims, such as personal anecdotes or opinions, are inherently unverifiable via external web evidence. Furthermore, the temporality of claims and evidence remains a challenge, as current evidence may not reflect past contexts.

A core limitation of our work is the absence of semantic verification between claims and retrieved evidence; no explicit stance detection or Natural

Language Inference (NLI) was performed to determine if evidence supports or refutes a claim. Additionally, the quality of the LLM-based claim extraction itself was not directly evaluated against a human-annotated ground truth, which is a key methodological limitation. Our claim extraction focused on a single main claim per text, while texts with multiple claims would require more advanced splitting techniques (Tang et al., 2024; Vargas et al., 2023). Results from claim extraction (Gemini 1.5 Flash) and evidence retrieval (Google Search API) are dependent on these specific LLMs and APIs and might vary with alternatives.

Furthermore, our study relied on Google’s Search APIs for evidence retrieval. A valuable direction for future work would be to compare this approach against a broader range of state-of-the-art information retrieval techniques. This could include employing dense retrieval models or more advanced query expansion strategies to assess how different retrieval methods impact the quality of the evidence found and, consequently, the performance of the fact-checking model.

The lack of a fixed random seed for the Gemini API version used limits exact reproducibility of claim extraction. Finally, evidence retrieval analyzed search snippets and metadata, not in-depth content of full URLs, potentially missing richer contextual information.

Ethics Statement

The authors have reviewed and commit to upholding the ACL Code of Ethics. We have considered the ethical implications of our research, data usage, and potential applications throughout this work. Our research focuses on the Fact Extraction and VERification (FEVER) task, specifically by enriching Portuguese-language misinformation datasets with external evidence.

Data Handling and Potential Biases:

The datasets used in this study (Fake.BR, COVID19.BR, MuMiN-PT) are publicly available. Our semi-automated validation process (Section 3.2) aimed to improve data quality and consistency. However, we acknowledge that these datasets may contain inherent biases stemming from their original collection methodologies (e.g., source, topic selection as described in Table 1), temporal context, and the nature of misinformation itself, which often includes sensitive or controversial topics. While our cleaning process addresses some structural issues (e.g., near-duplicates, label inconsistencies), it does not eliminate all potential underlying biases in the data. The texts themselves, being misinformation, can contain harmful or offensive content; our work analyzes this existing data and does not generate new harmful content. To prevent models from relying on superficial source-based heuristics, URLs were removed from texts before classification experiments, as discussed in Appendix B. We acknowledge this methodological choice may be misplaced for real-world applications. In practice, links often provide essential context for fact-checking, and developing models that learn to evaluate source trustworthiness could prove more beneficial and realistic than simply removing this information. This represents a trade-off between controlled experimental conditions and ecological validity that should be reconsidered in future work.

Models and Algorithmic Bias: We utilize Large Language Models (Gemini 1.5 Flash) for claim extraction (Section 3.3) and web search APIs (Google Custom Search Engine, Google FactCheck Claim Search API) for evidence retrieval (Section 3). We recognize that both LLMs and search engine results can exhibit biases (e.g., reflecting dominant viewpoints, perpetuating stereotypes, or being influenced by algorithmic filtering) and are not infallible. Our analysis (e.g., Pattern F2, Section 4.4, Appendix E) shows that search can sometimes reinforce misinformation. The performance and fairness of Automated Fact-Checking (AFC) systems heavily depend on the quality and representativeness of the data and the underlying models. The lack of a fixed random seed for the Gemini API version used, as mentioned in Section 7, presents a limitation for exact reproducibility of claim extraction.

Potential for Misuse and Mitigation Strategies:

Automated Fact-Checking systems, if inaccurate or misused, could inadvertently lead to the mislabeling of information, potentially suppressing legitimate speech or failing to identify harmful misinformation. Our work aims to improve the robustness of AFC by focusing on evidence-based verification. By making our enriched datasets and Python code publicly available upon publication (as stated in the Abstract and Conclusion), we aim to foster transparency, reproducibility, and further research into more reliable and fair fact-checking systems for the Portuguese language.

Scope and Limitations: This work operates primarily with binary veracity labels (true/false) inherited from the source datasets, which is a simplification of the nuanced reality of information veracity, as discussed in Sections 4.1.2 and 7. Some claims are inherently difficult to verify through automated web searches (e.g., personal opinions, unverifiable anecdotes). The temporality of claims and evidence also poses a challenge (Section 4.1.2), as the truthfulness or relevance of information can change over time. Our evidence retrieval is based on search snippets and metadata, not full-page analysis for all results, which is a limitation in depth.

Broader Impact: The intended broader impact of this research is to contribute to the development of more effective tools for combating misinformation in Portuguese, a significant societal challenge. By providing enriched, evidence-linked datasets (associating original texts with retrieved evidence and LLM-extracted claims, as mentioned in the Abstract and Conclusion), we hope to facilitate advancements in evidence-aware AFC systems. We believe that responsible development and deployment of such technologies are crucial for a well-informed public discourse.

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 EM-BRAPIL.

References

- Zahra Abbasiantaeb and Mohammad Aliannejadi. 2024. [Generate then retrieve: Conversational response retrieval using llms as answer and query generators](#). *Preprint*, arXiv:2403.19302.
- Esma Aimeur, Sabine Amri, and Gilles Brassard. 2023. Fake news, disinformation and misinformation in social media: a review. *Social Network Analysis and Mining*, 13(1):30.
- Ihsan Ali, Mohamad Nizam Bin Ayub, Palaiahnakote Shivakumara, and Nurul Fazmidar Binti Mohd Noor. 2022. Fake news detection techniques on social media: A survey. *Wireless Communications and Mobile Computing*, 2022.
- Gilmara Joanol Arndt, Milena Tarcisa Trindade, Juliana de Oliveira Alves, and Raquel de Barros Pinto Miguel. 2021. [Quem é de direita toma cloroquina, quem é de esquerda toma... vacina](#). *Revista Psicologia Política*, 21(51):608–626.
- Andréa Barbieri. 2021. [Tem dúvida? não compartilhe! o uso de fake news por professores de língua portuguesa do ensino fundamental ii com o propósito de desenvolver habilidades em educação midiática com seus alunos](#). Dissertação (mestrado), Universidade Tuiuti do Paraná, Curitiba, 10. Orientadora: Mônica Cristine Fort.
- Suelen Mazza Batista. 2020. [Onde os fatos não têm vez: uma análise foucaultiana das fake news relativas à cultura](#). Dissertação (mestrado em administração), Universidade Federal de Pernambuco, Recife, 2. Orientador: Sérgio Carvalho Benício de Mello.
- Marcia Borin da Cunha and Beatriz Tilschneider Garcia Rosa. 2022. [Fake science: proposta de análise](#). *Góndola, Enseñanza y Aprendizaje de las Ciencias*, 17(3):520–538.
- Francisco Frank Dourado Capistrano. 2022. [Fake news sobre a covid-19 nas aulas de química: uma abordagem didática na monitoria das práticas de ensino](#). Trabalho de conclusão de curso (licenciatura em química), Universidade Federal do Pará, Ananindeua, 8. Orientadora: Janes Kened Rodrigues dos Santos.
- Canyu Chen and Kai Shu. 2024. [Combating misinformation in the age of llms: Opportunities and challenges](#). *AI Magazine*, 45(3):354–368.
- Sukmin Cho, Soyeong Jeong, Wonsuk Yang, and Jong Park. 2022. [Query generation with external knowledge for dense retrieval](#). In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 22–32, Dublin, Ireland and Online. Association for Computational Linguistics.
- Eun Cheol Choi and Emilio Ferrara. 2024. [Automated claim matching with large language models: Empowering fact-checkers in the fight against misinformation](#). In *Companion Proceedings of the ACM on Web Conference 2024, WWW '24*, page 1441–1449, New York, NY, USA. Association for Computing Machinery.
- Paulo Roberto Cordeiro and Vladia Pinheiro. 2019. Um corpus de notícias falsas do twitter e verificação automática de rumores em língua portuguesa. In *Proceedings of the Symposium in Information and Human Language Technology*, pages 219–228.
- Joao MM Couto, Breno Pimenta, Igor M de Araújo, Samuel Assis, Julio CS Reis, Ana Paula C da Silva, Jussara M Almeida, and Fabrício Benevenuto. 2021. Central de fatos: Um repositório de checagens de fatos. In *Anais do III Dataset Showcase Workshop*, pages 128–137. SBC.
- Flávio Roberto Matias da Silva, Paulo Márcio Souza Freire, Marcelo Pereira de Souza, Gustavo de A. B. Plenamente, and Ronaldo Ribeiro Goldschmidt. 2020. [Fakenewssetgen: A process to build datasets that support comparison among fake news detection methods](#). In *Proceedings of the Brazilian Symposium on Multimedia and the Web, WebMedia '20*, page 241–248, New York, NY, USA. Association for Computing Machinery.
- Ulisses Matheus Braga de Freitas Melo. 2022. [Feita sob medida: a estrutura de uma notícia falsa e seu papel no convencimento do eleitor](#). Dissertação (mestrado em ciência política), Universidade Federal de Pernambuco, Recife, 2. Orientador: Sérgio Carvalho Benício de Mello.
- Mônica Chaves de Melo. 2019. [A pauta da desinformação: “fake news” e categorizações de pertencimento nas eleições presidenciais brasileiras de 2018](#). Dissertação (mestrado), Pontifícia Universidade Católica do Rio de Janeiro, 4. Orientadora: Adriana Andrade Braga.
- Mônica Chaves de Melo. 2021. [A pauta da desinformação: as ideias por trás das “fake news” nas eleições de 2018](#), 1 edition. Fafich/Selo PPG-COM/UFGM, Belo Horizonte.
- Ivandro Claudino de Sá. 2021. [Digital lighthouse: a platform for monitoring misinformation in whatsapp public groups](#). Dissertação (mestrado), Universidade Federal do Ceará. Orientação: Prof. Dr. José Maria da Silva Monteiro Filho.
- Ivandro Claudino de Sá, José Monteiro, José Franco da Silva, Leonardo Medeiros, Pedro Mourão, and Lucas Carneiro da Cunha. 2021. [Digital lighthouse: A platform for monitoring public groups in whatsapp](#). In *Proceedings of the 23rd International Conference on Enterprise Information Systems - Volume 1: ICEIS*,., pages 297–304. INSTICC, SciTePress.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for*

- Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Jia Ding, Yongjun Hu, and Huiyou Chang. 2020. [Bert-based mental model, a better fake news detector](#). In *Proceedings of the 2020 6th International Conference on Computing and Artificial Intelligence, ICCAI '20*, page 396–400, New York, NY, USA. Association for Computing Machinery.
- Felipe dos Santos Gusmão. 2023. [Estudo comparativo de modelos de classificação textual aplicados na classificação de fake news](#). Trabalho de conclusão de curso (bacharelado em engenharia da computação), Universidade Federal do Amazonas, Manaus, 6. Orientador: José Luiz de Souza Pio.
- Salma El Anigri, Mohammed Majid Himmi, and Abdelhak Mahmoudi. 2021. How bert’s dropout fine-tuning affects text classification? In *Business Intelligence*, pages 130–139, Cham. Springer International Publishing.
- Pedro Faustini and Thiago Covões. 2019. [Fake news detection using one-class classification](#). In *2019 8th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 592–597.
- Revanth Gangi Reddy, Sai Chetan Chinthakindi, Yi R. Fung, Kevin Small, and Heng Ji. 2022. [A zero-shot claim detection framework using question answering](#). In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6927–6933, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Daniel Griebhaber, Johannes Maucher, and Ngoc Thang Vu. 2020. [Fine-tuning BERT for low-resource natural language understanding via active learning](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1158–1171, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. [A survey on automated fact-checking](#). *Transactions of the Association for Computational Linguistics*, 10:178–206.
- Sakshini Hangloo and Bhavna Arora. 2022. [Combating multimodal fake news on social media: methods, datasets, and future perspective](#). *Multimedia Systems*, 28(6):2391–2422.
- Arthur Ituassu, Sergio Lifschitz, Letícia Capone, and Vivian Mannheimer. 2019. [De donald trump a jair bolsonaro: democracia e comunicação política digital nas eleições de 2016, nos estados unidos, e 2018, no brasil](#). In *Anais do 8º Congresso da Associação Brasileira de Pesquisadores em Comunicação e Política*, Brasília. Universidade de Brasília, Associação Brasileira de Pesquisadores em Comunicação e Política - Compolítica.
- Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. 2023. [WiCE: Real-world entailment for claims in Wikipedia](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7561–7583, Singapore. Association for Computational Linguistics.
- Kyungha Kim, Sangyun Lee, Kung-Hsiang Huang, Hou Pong Chan, Manling Li, and Heng Ji. 2024. [Can llms produce faithful explanations for fact-checking? towards faithful explainable fact-checking via multi-agent debate](#). *Preprint*, arXiv:2402.07401.
- Sotiris Kotitsas, Panagiotis Kounoudis, Eleni Koutli, and Haris Papageorgiou. 2024. [Leveraging fine-tuned large language models with LoRA for effective claim, claimer, and claim object detection](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2540–2554, St. Julian’s, Malta. Association for Computational Linguistics.
- Eulália Vera Lúcia Fraga Leurquin and Chloé Leurquin. 2021. [Fake news, desinformação e necessidade de formar leitores críticos](#). *Scripta*, 25(54):265–295.
- Arthur Guimarães Lima. 2019. [A propagação de fake news e seus impactos: um estudo sobre a onda conservadora na política ocidental contemporânea](#). Trabalho de conclusão de curso (bacharelado em comunicação social com habilitação em relações públicas), Universidade de São Paulo, São Paulo. Orientador: Luiz Alberto de Farias.
- Guilherme Lima, Marcos Calazans, and Luciana MASSI. 2021. [Mensagens falsas sobre o novo coronavírus: legitimidade e manipulação na luta de classes](#). *Chasqui. Revista Latinoamericana de Comunicación*, 1(147):259–280.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [Roberta: A robustly optimized bert pretraining approach](#). *Preprint*, arXiv:1907.11692.
- Dyonatan Maia and Nádia Félix Felipe da Silva. 2024. [Enhancing stance detection in low-resource Brazilian Portuguese using corpus expansion generated by GPT-3.5](#). In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 503–508, Santiago de Compostela, Galicia/Spain. Association for Computational Linguistics.
- Antônio Diogo Forte Martins, Lucas Cabral, Pedro Jorge Chaves Mourao, Ivandro Claudino de Sá, José Maria Monteiro, and Javam Machado. 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. SBC.
- Rafael A. Monteiro, Roney L. S. Santos, Thiago A. S. Pardo, Tiago A. de Almeida, Evandro E. S. Ruiz,

- and Oto A. Vale. 2018. Contributions to the study of fake news in portuguese: New corpus and automatic detection results. In *Computational Processing of the Portuguese Language*, pages 324–334. Springer International Publishing.
- João Moreno and Graça Bressan. 2019. [Factck.br: A new dataset to study fake news](#). In *Proceedings of the 25th Brazilian Symposium on Multimedia and the Web*, WebMedia '19, page 525–527, New York, NY, USA. Association for Computing Machinery.
- Jaqueline Gonçalves do Nascimento. 2021. [Disseminação de desinformação sobre a covid-19 em um núcleo familiar: um estudo de caso](#). Trabalho de conclusão de curso (bacharelado em biblioteconomia), Universidade Federal do Ceará, Fortaleza. Orientador: Antonio Wagner Chacon Silva.
- Jingwei Ni, Minjing Shi, Dominik Stambach, Mrinmaya Sachan, Elliott Ash, and Markus Leippold. 2024. [Afacta: Assisting the annotation of factual claim detection with reliable llm annotators](#). *Preprint*, arXiv:2402.11073.
- Dan S Nielsen and Ryan McConville. 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*, pages 3141–3153.
- Gabriel Nogueira. 2021. Br fake news detection. https://github.com/Talendar/br_fake_news_detection. Accessed on 2/2/2025.
- Luiza Prevedel Pereira and Juliano Desiderato Antonio. 2023. [É verdade ou fake news? estratégias linguísticas de manipulação em textos que promovem a desinformação](#). *Revista USP*, (138):27–38.
- Yunjian Qiu and Yan Jin. 2024. [Chatgpt and finetuned bert: A comparative study for developing intelligent design support systems](#). *Intelligent Systems with Applications*, 21:200308.
- Miguel Quessada. 2022. [Desinformação e esquerda brasileira: o discurso por trás das fake news](#). Dissertação (mestrado), Universidade Federal de São Carlos, São Carlos, 2. Orientador: Thales Haddad Novaes de Andrade.
- Vanessa Daiane Contente Quintanilha. 2021. [Combatendo as fake news sobre o sars-cov-2: o revisor como fact-checker](#). Dissertação (mestrado), Universidade Nova, 9. Orientadora: Matilde Gonçalves.
- Thilina C. Rajapakse, Andrew Yates, and Maarten de Rijke. 2024. [Simple transformers: Open-source for all](#). In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, SIGIR-AP 2024, pages 209–215.
- Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.
- Márcio Moretto Ribeiro and Pablo Ortellado. 2018. [O que são e como lidar com as notícias falsas](#). *Sur - Revista Internacional de Direitos Humanos*, 15(27):71–83.
- Guilherme Yukio Sakurai. 2019. [Processamento de linguagem natural - detecção de fake news](#). Trabalho de conclusão de curso (bacharelado em ciência da computação), Universidade Estadual de Londrina, Londrina. Orientador: Sérgio Montazzolli Silva.
- Mirella Gadelha Santos. 2020. [Detecção de fake news: Um comparativo entre os modelos de aprendizado supervisionado passivo agressivo e multinomial naive bayes](#). Trabalho de conclusão de curso (bacharelado em sistemas de informação, Centro Universitário Christus - Unichristus, Fortaleza, 8. Orientador: Daniel Nascimento Teixeira.
- Alessandro Scirè, Karim Ghonim, and Roberto Navigli. 2024. [FENICE: Factuality evaluation of summarization based on natural language inference and claim extraction](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14148–14161, Bangkok, Thailand. Association for Computational Linguistics.
- Gautam Kishore Shahi and Durgesh Nandini. 2020. [FakeCovid- A Multilingual Cross-domain Fact Check News Dataset for COVID-19](#). ICWSM.
- Renato M. Silva, Roney L.S. Santos, Tiago A. Almeida, and Thiago A.S. Pardo. 2020. [Towards automatically filtering fake news in portuguese](#). *Expert Systems with Applications*, 146:113199.
- Fábio José dos Santos Sousa. 2022. [Transferência de conhecimento para detecção automática de fake news com aprendizagem profunda](#). Trabalho de conclusão de curso (bacharelado em ciência da computação), Universidade Federal do Ceará, Cratêus. Orientador: Livio Antonio de Melo Freire.
- Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. Bertimbau: Pretrained bert models for brazilian portuguese. In *Intelligent Systems*, pages 403–417, Cham. Springer International Publishing.
- Peter M Stahl. 2023. [Lingua-py](#). <https://github.com/pemistahl/lingua-py>. Accessed on 4/3/2024.
- Xin Tan, Bowei Zou, and Ai Ti Aw. 2023. [Evidence-based interpretable open-domain fact-checking with large language models](#). *Preprint*, arXiv:2312.05834.
- Liyan Tang, Philippe Laban, and Greg Durrett. 2024. [MiniCheck: Efficient fact-checking of LLMs on grounding documents](#). In *Proceedings of the 2024*

Conference on Empirical Methods in Natural Language Processing, pages 8818–8847, Miami, Florida, USA. Association for Computational Linguistics.

Francielle Vargas, Kokil Jaidka, Thiago Pardo, and Fabrício Benevenuto. 2023. [Predicting sentence-level factuality of news and bias of media outlets](#). In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 1197–1206, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

William Yang Wang. 2017. [“liar, liar pants on fire”: A new benchmark dataset for fake news detection](#). In *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.

Yuxia Wang, Revanth Gangi Reddy, Zain Muhammad Mujahid, Arnav Arora, Aleksandr Rubashevskii, Jiahui Geng, Osama Mohammed Afzal, Liangming Pan, Nadav Borenstein, Aditya Pillai, Isabelle Augenstein, Iryna Gurevych, and Preslav Nakov. 2024. [Factcheck-bench: Fine-grained evaluation benchmark for automatic fact-checkers](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14199–14230, Miami, Florida, USA. Association for Computational Linguistics.

Fengzhu Zeng and Wei Gao. 2024. [JustiLM: Few-shot justification generation for explainable fact-checking of real-world claims](#). *Transactions of the Association for Computational Linguistics*, 12:334–354.

A Illustrative Examples of the Semi-automatic Validation Workflow

This appendix provides concrete examples to illustrate the different stages of the semi-automatic validation workflow described in Section 3.2.

It is important to note that the removal of URLs from texts, mentioned as a general step to avoid domain bias, was performed after the validation steps exemplified here that might depend on the presence of these URLs (such as URL-based contradiction resolution or Fake.br-specific filtering by source URL). The examples below show the texts before this final URL removal, but with other specific validation processes being illustrated.

A.1 Initial Automated Filtering

This stage removed examples that did not meet basic content or format criteria. Figure 5 illustrates three types of removal in this phase: texts composed solely of a URL, texts considered excessively short after tokenization and removal of *stopwords*, and exact duplicates within the same corpus.

Text composed only of a URL (COVID19.BR) - Removed

<https://fanoticias.com.br/covid-19-jovem-sobre-os-efeitos-parecia-furar-meu-peito/>,

Text with few relevant tokens (COVID19.BR) - Removed

hahaha looks like corona.hahaha

Exact Duplicate (Fake.br) - Example true_0069 removed

Suplicy will participate in Doria’s web program this Thursday

Councilman Eduardo Suplicy (PT-SP) will participate this Thursday (10), at 8:30 PM, in the program “Olho no Olho” (Eye to Eye), broadcast on Mayor João Doria’s (PSDB-SP) social media.

Doria has already hosted allies and personalities on the show such as singer Lobão, presenter José Luiz Datena, former basketball player Oscar, singer Roger from Ultraje a Rigor, and journalist Joice Hasselmann. At the beginning of his term, the tucano (PSDB member) spared criticism of his predecessor Fernando Haddad (PT-SP), but in recent months, he has accused the former mayor of leaving a R\$ 7 billion deficit in the city hall. Haddad denies this and says he left the city’s accounts in order.

Suplicy, in turn, is one of the critical voices against Doria in the City Council.

Figure 5: English-translated examples of removals performed in the initial automated filtering stage.

A.2 Language Filtering

Examples identified as not being primarily in Portuguese were removed. The Lingua tool was used for automatic detection, followed by manual verification in ambiguous cases. Figure 6 shows examples removed for being in English and Spanish.

A.3 Contradiction Resolution

This stage addressed label inconsistencies between semantically very similar examples or those that referenced the same source. Figure 7 illustrates a case of near-identical texts in the COVID19.BR corpus that had conflicting veracity labels (true vs. false). After manual analysis, based on context and external sources when available, both conflicting examples were removed.

A.4 External Label Verification

The Google FactCheck API was used to identify potential labeling errors in the original data. When an API result was found, we performed a manual review by one of the authors if the retrieved verac-

Example in English (COVID19.BR) - Removed
German state minister kills himself as coronavirus hits economy - https://www.aljazeera.com/news/2020/03/german-state-minister-kills-coronavirus-hits-economy-200329165242615.html
Example in Spanish (MuMiN) - Removed
Vídeo de fraude en las urnas de Flint (Michigan) durante las elecciones de Estados Unidos

Figure 6: Examples filtered for not being in Portuguese.



Near-Duplicate Text A (COVID19.BR) - Original Label: False
Amidst the severe CORONAVIRUS crisis, the National Congress refused to give up the money allocated to the electoral fund to aid in combating the pandemic. As always, these bloodsuckers show they only care about their own interests. Let's sign the petition for the closure of the National Congress!  https://peticaopublica.org/fechamento-congresso-nacional/
Excerpt considered true (COVID19.BR)
Amidst the severe CORONAVIRUS crisis, the National Congress refused to give up the money allocated to the electoral fund to aid in combating the pandemic. As always, these bloodsuckers show they only care about their own interests. Let's sign the petition for the closure of the National Congress!  https://peticaopublica.org/fechamento-congresso-nacional/

Figure 7: Example of a pair of near-duplicate texts in the COVID19.BR corpus that had contradictory veracity labels. Manual resolution was necessary to determine the correct label or remove the pair.

ity label conflicted with the original binary label. During this review, retrieved labels that clearly indicated falsity—specifically “false,” “fake,” “misleading,” and “wrong” (highlighted in Table 5)—were considered strong evidence to check for a potential FAKE NEWS mislabel. Conversely, a retrieved label of “true” was used to check for potential TRUE mislabels. Nuanced or uncertain labels like “not quite so” or “too early to tell” did not automatically trigger a label change but were considered by the human annotator to inform the final decision for that specific case. Figure 8 demonstrates a case in the MuMiN-PT corpus where the original label (true) was corrected to fake after external verification and manual analysis confirmed the initial

incorrectness.

Example of Label Correction (MuMiN-PT)
Corpus: MuMiN-PT
Original text: <u>The American Medical Association lifted restrictions and began recommending hydroxychloroquine against covid-19</u>
Label: true fake
Query (Extracted claim): The American Medical Association recommends hydroxychloroquine against covid-19.
CSE Result:
<p>Title: Hydroxychloroquine is not recommended as early treatment ...</p> <p>Snippet: 1 day ago ... From time to time, the drug hydroxychloroquine is again pointed out as an effective early treatment against Covid-19.</p>
FactCheck Result:
<p>Agency: Aos Fatos Label: False Reviewed claim: American Medical Association does not recommend hydroxychloroquine ...'</p>

Figure 8: Illustration of the external label verification process. An example from MuMiN-PT had its original label (true) corrected to fake based on evidence from the Google FactCheck API and subsequent manual confirmation.

A.5 Specific Treatment for Fake.br

The Fake.br corpus has particular characteristics that required additional treatment steps. Figure 9 illustrates the removal of near-duplicate texts that shared the same source URL, a situation specific to this corpus. Additionally, two other specific rules were applied (not illustrated with detailed visual examples for brevity, but described below):

- Removal of incomplete texts:** Examples whose normalized texts (as per 4.1) were identified as truncated or incomplete compared to the original news story (addressed in Issue #7 of the Fake.br repository) were removed.
- Maintenance of pairs:** Given that Fake.br is structured in pairs of news items (true and false about the same event), if one element of the pair was removed by any of the previous

validation criteria, the corresponding element was also removed to preserve the integrity of the dataset’s paired structure.

Near-Duplicates with Same Source URL (Fake.br) - Example true_3023 removed

The offenders’ paradise To get a third of the deputies’ votes, Temer decrees pardon of fines Privatization projects to improve public accounts performance are being undermined in the process of winning a third of the Chamber’s votes to spare President Temer from investigation for crimes of criminal organization and obstruction of justice. More so now, after the convicted Waldemar da Costa Neto, aka Boy, was heeded, who demanded the removal of Congonhas airport from the announced package, which would bring in 6 billion in revenue, in exchange for keeping the country’s second-largest airport under Infraero’s control. It’s obvious that the corruption scheme of previous governments is maintained. ...

Figure 9: Example of Fake.br-specific removal: near-duplicate texts (true_0251 and true_3023) originating from the same source URL. One of them (true_3023) was removed to reduce redundancy originating from the collection.

B URL Dependency

In the context of *links* in the COVID19.BR dataset, Table 7 shows the most mentioned URL domains in the examples. The domains Gazeta Brasil, bit.ly, Globo, WhatsApp, and JapinaWeb almost always represent true examples. However, the presence of a governmental domain is not, in itself, a guarantee of veracity. Brazilian government domains, such as gov.br, can be instrumentalized in misleading contexts, as illustrated by an example from the MuMiN-PT corpus in Figure 10.

Example of Misleading Use of Official URL (MuMiN-PT)

Everyone, you need to register on conectesus to get vaccinated. I suggest doing it now. The site probably won’t handle the traffic when the time comes. <https://conectesus-paciente.saude.gov.br/> It’s a SUS registration. Those who took the Yellow Fever vaccine in 2018 already have it. Or those who have used SUS in recent years. The application works more or less like those driver’s license or voter ID apps.

Figure 10: Example from MuMiN-PT where an official URL is used in a misinformation context. ⁶

Mentioned URL Domains	fake	% fake	true	% true	Total
gazetabrasil	0	0.0	259	100.0	259
bit.ly	6	6.8	82	93.2	88
youtube	47	67.1	23	32.9	70
globo	6	11.3	47	88.7	53
facebook	19	38.0	31	62.0	50
dunapress	0	0.0	46	100.0	46
twitter	25	56.8	19	43.2	44
whatsapp	1	3.1	31	96.9	32
uol	11	37.9	18	62.1	29
conexaopolitica	16	59.3	11	40.7	27
gov.br	6	26.1	17	73.9	23
instagram	5	21.7	18	78.3	23
jornaldacidadeonline	14	63.6	8	36.4	22
japinaweb	0	0.0	21	100.0	21
atrombetanews	7	35.0	13	65.0	20

Table 7: Count of the 15 most referenced URL domains in the COVID19.BR corpus, broken down by fake and true labels (absolute counts and percentages). Domains and their ‘% true’ values are highlighted in bold if over 80% of the mentions for that domain are in examples labeled as true. Brazilian governmental domains were aggregated under “gov.br”.

C Near-Duplicate Examples

This appendix provides examples of near-duplicate texts identified during the validation process (Section 3.2).

Figure 11 shows two near-duplicate examples from the Fake.br dataset, both originally labeled ‘true’. They report the same event (death of journalist Marcelo Rezende) but were published a few hours apart on the same news site (G1) with minor updates regarding the wake.

Figure 12 shows examples of near-duplicate messages from the COVID19.BR dataset, originally labeled ‘false’. These messages promote a non-existent offer of free internet data, varying slightly in the amount offered (10GB vs. 500GB) and the link provided.

D Prompt Patterns Explored

This appendix lists the main types of prompts identified in our literature review (Section 3.3) for claim extraction and related tasks. We experimented with variations based on these patterns before settling on the simpler zero-shot prompt in Figure 2.

- **Simple Claim Detection (Kotitsas et al., 2024):** Basic instruction asking for the main claim.

What is the main claim of the input?

- **Claim Splitting/Decomposition (Scirè et al., 2024; Kamoi et al., 2023; Tang et al., 2024; Wang et al., 2024):** Instruction to break down a

Marcelo Rezende passed away at 65 in São Paulo. He was a victim of multiple organ failure resulting from cancer, as reported by Hospital Moriah.. The journalist Marcelo Rezende died at 5:45 PM on Saturday (16th) in São Paulo, at 65 years old, victim of multiple organ failure resulting from cancer, according to Hospital Moriah...

(a) English-translation of the original news article posted at <https://web.archive.org/web/20220808194736/https://g1.globo.com/sao-paulo/noticia/morre-aos-65-anos-o-jornalista-marcelo-rezende.ghml>

The body of Marcelo Rezende will lie in state at the Legislative Assembly this Sunday. He died from multiple organ failure resulting from cancer, as reported by Hospital Moriah.. The body of journalist Marcelo Rezende will lie in state this Sunday (17th) at the São Paulo Legislative Assembly. The burial arrangements have not yet been announced by the family. ...

(b) English-translation of the updated news article at <https://web.archive.org/web/20190208022401/https://g1.globo.com/sao-paulo/noticia/corpo-de-marcelo-rezende-sera-velado-na-assembleia-legislativa-neste-domingo.ghml>

Figure 11: Two near-duplicate examples (English translations) from the Fake.br dataset ('true' label). The second is an update of the first, published shortly after by the same source (G1). Differences are highlighted.

Free 500 GB 4G Internet Due to the COVID-19 epidemic, we're offering 10 GB of free Internet, valid for 90 days to help you stay home! Get free Internet access and stay home <https://bit.ly/10gbytes>

Free 500 GB 4G Internet Due to the COVID-19 epidemic, we're offering 500 GB of free Internet, valid for 90 days to help you stay home! Get free Internet access and stay home <http://hu5k.com/Covid>

Free 500 GB 4G Internet Due to the COVID-19 epidemic, we're offering 500 GB of free Internet, valid for 90 days to help you stay home! Get free Internet access and stay home <https://bit.ly/500gbytes>

Figure 12: Near-duplicate examples (English translations) of misinformation (clickbait) from the COVID19.BR dataset ('false' label). Minor variations in text and URL are highlighted.

sentence into multiple atomic claims. (Not used in our final approach).

Segment the following sentence into individual facts: [SENTENCE]

- **Role Specification (Kotitsas et al., 2024):** Defining the AI's role to guide its behavior.

You are an AI assistant helping fact-checkers identify check-worthy information. Extract the main claim from: [TEXT]

- **Explicit Fact Categories (Ni et al., 2024):** Providing definitions of what constitutes a factual claim.

Identify claims mentioning specific actions, quantities, correlations, rules, or predictions in the following text: [TEXT]

- **Query Formulation Analogy (Abbasiantaeb and Aliannejadi, 2024):** Framing the task as generating a search query to verify the text.

What would you search on Google to verify this text? Extract the core query: [TEXT]

- **Few-shot Demonstration (Scirè et al., 2024):** Providing input/output examples in the prompt.

Input: [Example Text 1]
Claim: [Example Claim 1]
Input: [Actual Text]
Claim:

Our final prompt (Figure 2) is closest to the Simple Claim Detection pattern, adding constraints on length and output format.

E Evidence Pattern Examples

This appendix provides concrete examples illustrating the evidence patterns T1, F1, F2, and F3, as discussed in Section 4.4. Each example includes metadata from the original dataset (dataset, label, original text with the query portion underlined), the query used for search (either the underlined text or an extracted claim), the primary CSE result snippet obtained, and any relevant FactCheck API result. Translations are provided where necessary.

Pattern T1: Corroboration of True Claims

This pattern involves finding evidence from reliable sources (like news articles) that confirms the factual information in a true claim. The example below (Figure 13) shows a claim extracted from a message about COVID-19 test reliability. The CSE result points to an article from a reputable source (Fiocruz, a Brazilian research institution)

discussing the possibility of false negatives, thus corroborating the claim. The FactCheck API returned no results, consistent with pattern T2.

Example of Pattern T1: Corroboration

Dataset: COVID19.BR

Original Text: We’ve had patients who took COVID-19 tests, including tests at Sabin laboratory that came back negative, but when repeated at Oswaldo Cruz Hospital using rapid tests showed positive results. Remember that nasal swab tests may yield false negatives. Stay alert to avoid false negatives... (truncated)

Original Label: true

Query (Extracted Claim): COVID-19 tests may yield false negatives, even when conducted at reputable laboratories.

CSE Result:

Title: Covid-19: Fiocruz researcher answers questions about testing...

Snippet: Jan 15, 2021 false negatives may occur due to low specificity and analytical sensitivity of the test... no laboratory test is perfect and its...

FactCheck API Result: None

Figure 13: Example illustrating Pattern T1. The search based on the extracted claim found corroborating information from a reliable source. (English translation from COVID19.BR)

Pattern F1: Direct Debunk

This pattern represents the ideal outcome for verifying a false claim, where search results contain a direct refutation from a fact-checking agency or reliable source. Figure 14 shows an example where the CSE search for a claim about coronavirus transmission via parcels returned a result explicitly stating the opposite, effectively debunking the claim. Note that in this specific instance, the FactCheck API did not return a result, but the CSE result itself serves as the debunk.

Pattern F2: Misinformation Reinforcement

This common pattern occurs when searching for a piece of misinformation returns results that simply repeat or amplify the same false claim, often from unreliable sources or social media. Figure

Example of Pattern F1: Direct Debunk

Dataset: MuMiN-PT

Original Text: The new coronavirus can be transmitted through parcels shipped from China.

Original Label: false

Query (Original Text): The new coronavirus can be transmitted through parcels shipped from China.

CSE Result:

Title: The new coronavirus cannot survive in parcels shipped from ...

Snippet: March 4, 2020 coronavirus can be transmitted through parcels shipped by mail from China to other countries. The claim, which served as the basis ...

Source Type: News/Fact-Check Report

FactCheck API Result: None

Figure 14: Example illustrating Pattern F1. The CSE search using the false claim as a query yielded a result directly refuting it. (English translation from MuMiN-PT)

15 demonstrates this: the CSE search for the false claim about the “Popular Pharmacy” program returned a result from an unreliable source echoing the misinformation. Critically, even though the FactCheck API did find a relevant debunk for this claim, the CSE result exemplifies the F2 pattern, highlighting the challenge where search engines might surface reinforcing content alongside (or instead of) corrective information.

Pattern F3: Academic Recognition as Misinformation

This pattern occurs when searching for a known piece of misinformation yields results pointing to academic literature (papers, theses, articles) that cite or discuss the misinformation as an object of study. While not providing a direct debunk snippet in the same way as F1, these results serve as ‘meta-evidence’, confirming that the claim is recognized as problematic or false within the research community. Table 8 provides a list of such academic publications encountered during our searches for misinformation examples present in the datasets used. No figure is provided for F3 as the primary evidence is the existence of these academic refer-

Example of Pattern F2: Misinformation Reinforcement	
Dataset:	Fake.br
Original Text:	Temer government will end the popular pharmacy. Who hasn't heard of or even needed the popular pharmacy, right? The system offers the distribution of free medicines for various types of diseases, these resources are available until the month of August 2017 ... (truncated)
Original Label:	false
Query (Original Text):	Temer government will end the popular pharmacy.
CSE Result:	<div> <p>Title: Temer Government closes Popular Pharmacy and intends to abolish the SUS ...</p> <p>Snippet: Feb 22, 2018 ... Are you going to allow it? The Ministry of Health has just closed the 517 popular pharmacies maintained by the federal government in the country.</p> <p>Source Type: Unreliable Blog/Social Media</p> </div>
FactCheck API Result:	<div> <p>Agency: Lupa - UOL Rating: False </p> <p>Claim Reviewed: Temer does not 'make official the end of the Popular Pharmacy project'</p> </div>

Figure 15: Example illustrating Pattern F2. The CSE search returned a result reinforcing the false claim, despite a debunk existing (found via the FactCheck API). (English translation from Fake.br)

ences rather than a specific search snippet.

F Hyperparameter Details for Bertimbau Fine-tuning

Table 9 presents the hyperparameter grid search space used to fine-tune the base Bertimbau model across each training data configuration. For each configuration and corpus, the best hyperparameters were chosen based on the highest validation F1-score.

Based on prior work with similar models and tasks (Devlin et al., 2019; Liu et al., 2019; Ding et al., 2020; Nogueira, 2021; dos Santos Gusmão, 2023; Souza et al., 2020), we set maximum training epochs to 10. A standard weight decay of 0.01 was

used (Devlin et al., 2019; Souza et al., 2020; Liu et al., 2019).

We employed the AdamW optimizer (standard for BERT) with canonical $\beta_1(0.9)$, $\beta_2(0.999)$ (Devlin et al., 2019; Souza et al., 2020; Rajapakse et al., 2024), and SimpleTransformers' default AdamW epsilon ($1e - 8$) (Rajapakse et al., 2024). The maximum sequence length was set to the standard 512 tokens.

Task layer dropout rates of 0.1, 0.2, and 0.3 were tested, as higher rates can aid regularization with limited data (El Anigri et al., 2021; Griebhaber et al., 2020). Batch sizes of 8 and 16 were explored, consistent with similar applications (Devlin et al., 2019; Liu et al., 2019; Souza et al., 2020; Ding et al., 2020; Nogueira, 2021; dos Santos Gusmão, 2023).

A learning rate scheduler with linear decay and 6% warmup followed (Liu et al., 2019)'s recommendations. Early stopping based on the validation F1-score (halting if no improvement greater than 0.001 for 3 epochs) prevented overfitting and optimized training time.

Reference	Area	Year	Institution	Publication Medium	Theme
(Ribeiro and Ortellado, 2018)	Social Sciences	2018	USP	Journal	Politics
(de Melo, 2019)	Communication	2019	PUC-RIO	Master's Thesis	Politics
(Sakurai, 2019)	Computer Science	2019	UEL	Final Project	General
(Ituassu et al., 2019)	Communication	2019	PUC-RIO	Conference	Politics
(Santos, 2020)	Computer Science	2020	Unichristus	Final Project	General
(de Melo, 2021)	Communication	2021	PUC-RIO	Book	Politics
(Barbieri, 2021)	Communication	2021	UTP	Master's Thesis	COVID-19
(Quintanilha, 2021)	Linguistics	2021	UNL	Master's Thesis	COVID-19
(Lima et al., 2021)	Communication	2021	UFOP	Journal	COVID-19
(Lima, 2019)	Communication	2021	USP	Final Project	Politics
(Arndt et al., 2021)	Politics	2021	UFSC	Journal	COVID-19
(Leurquin and Leurquin, 2021)	Linguistics	2021	UFC	Journal	COVID-19, Politics
(de Sá, 2021)	Computer Science	2021	UFC	Master's Thesis	COVID-19, Politics
(Nascimento, 2021)	Library Science	2021	UFC	Final Project	COVID-19
(Capistrano, 2022)	Chemistry	2021	UFPA	Final Project	COVID-19
(Batista, 2020)	Administration	2021	UFPE	Master's Thesis	General
(de Freitas Melo, 2022)	Politics	2022	UFPE	Master's Thesis	Politics
(Borin da Cunha and Tilschneider Garcia Rosa, 2022)	Education	2022	Unioeste	Journal	COVID-19, Science
(Sousa, 2022)	Computer Science	2022	UFC	Final Project	General
(Quessada, 2022)	Politics	2022	UFSCAR	Master's Thesis	Politics
(Pereira and Antonio, 2023)	Linguistics	2023	UEM	Journal	General
(dos Santos Gusmão, 2023)	Computer Science	2023	UFAM	Final Project	General

Table 8: Selection of Academic Publications Found in Search Results Discussing Misinformation Examples from the Datasets (Illustrating Pattern F3).

Hyperparameters	Search space / Value
Batch size	8, 16
Learning rate	{1e-6, 5e-6, 1e-5, 2.5e-5, 5e-5, 1e-4}
Task layer dropout	0.1, 0.2, 0.3
Seed	2025
Weight decay	0.01
Max. train epochs	10
Max. Seq. Length	512
Learning rate scheduler	Linear w/ 6% <i>warmup</i>
Optimizer	AdamW
AdamW ϵ	1e-8
AdamW β_1	0.9
AdamW β_2	0.999
Early stopping patience	3 epochs
Early stopping threshold (F1-score)	0.001

Table 9: Hyperparameter search space for fine-tuning the base Bertimbau model.