TaCertoIssoAI: A Data Lake for Misinformation Analytics from WhatsApp in Brazil

Cauê Paiva Lira[1], **Luiz Felipe Diniz Costa**[1], **Pedro Henrique Ferreira Silva**[1]

[1]Institute of Mathematics and Computer Sciences – University of São Paulo (USP)
São Carlos – SP – Brazil
{cauepaivalira, lfediniz, pedrohfsilva}@usp.br

Abstract. *This paper presents the analytics data lake of TaCertoIssoAI*[1]*, a WhatsApp-based fact-checking platform operating in Brazil. Verified messages are anonymized and stored in a hybrid data architecture on Google Cloud, combining BigQuery for analytical workloads (OLAP) and Firestore for operational access (OLTP). The stored fact-checks follow a claim-oriented data model that supports multimodal content representations, together with vector embeddings for semantic search. A public analytics front-end allows users to filter, explore, and export fact-checking data. With over 7,200 users and 14,000 verified messages stored, the system provides a live dataset for analyzing misinformation patterns on messaging platforms and social media. The project has also won national technology competitions, been presented at a conference held at MIT, and received coverage from national news outlets such as G1.*

1. Introduction

WhatsApp is installed on over **99% of Brazilian smartphones** and serves as a primary channel for news consumption [Hale et al. 2024]. This widespread adoption has also made WhatsApp a fertile ground for misinformation. Its end-to-end encryption prevents representative sampling of circulating content, making the platform's information flows largely unobservable [Hale et al. 2024]. Hale et al. [Hale et al. 2024] analyzed claims submitted to fact-checking organizations during the 2022 Brazilian election, showing that misinformation dynamics on WhatsApp differ substantially from those on open platforms. With the 2026 Brazilian presidential elections approaching, these concerns become even more pressing.

Automated fact-checking (AFC) has emerged as a promising approach to cope with this volume, organizing the problem around claim detection, evidence retrieval, and veracity classification [Guo et al. 2022], with extensions to multimodal settings [Akhtar et al. 2023]. The retrieval-augmented generation (RAG) paradigm [Lewis et al. 2020] has become central to systems that must ground their outputs in traceable external evidence, though resources for Portuguese remain scarce [Gomes et al. 2025]. However, existing efforts focus almost exclusively on the verification pipeline itself. While verification answers whether a single claim is true or false, a data infrastructure layer is necessary to aggregate verified claims over time, identify recurring narratives, reveal thematic trends, and enable researchers and institutions to study misinformation as a systemic phenomenon rather than a collection of

[1]Project repository: `https://github.com/TaCertoIssoAI`. Demo video: `https://youtu.be/dLqpMPrUp8c?si=_hCJ5lD-uFC9jznt`.

isolated cases. Data lake architectures are well established for social media analytics [Sawadogo and Darmont 2021], but their application to structured misinformation intelligence from private messaging platforms remains underexplored.

TaCertoIssoAI [TaCertoIssoAI 2025a] was built with this missing capability in mind. The project provides an AI-powered chatbot on WhatsApp that allows users to forward suspicious messages for multimodal fact-checking (text, audio, images, videos, and links) through an LLM pipeline with curated data sources. The contribution of this paper, however, is not the fact-checking pipeline itself, but the analytics data lake that it feeds. Every verified message is anonymized, with explicit user consent via a first-interaction disclaimer, and stored in a data lake on Google Cloud Platform (GCP). This infrastructure transforms individual fact-checks into a queryable knowledge base about misinformation patterns in Brazil, featuring claim-oriented modeling, multimodal content representations, topic classification, and vector embeddings for semantic search.

2. Architecture and Data Flow

Figure 1 summarizes the TaCertoIssoAI architecture, focusing on the data flow from message ingestion to analytical storage.

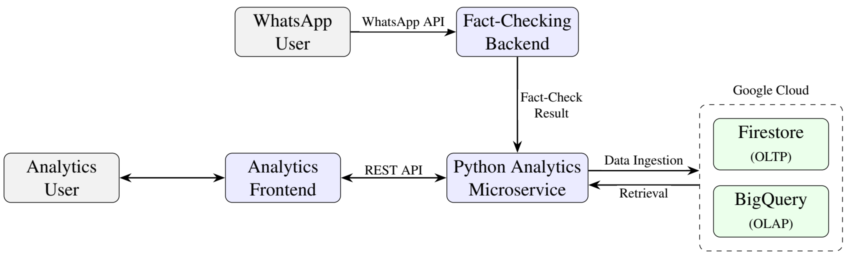

Figure 1. System architecture: data flows from WhatsApp through the AI fact-checking pipeline to the Python analytics microservice.

When a user forwards a message to the TaCertoIssoAI WhatsApp contact, the fact-checking system runs the content through its AI pipeline and produces a structured analysis with identified claims, verdicts, and supporting sources. This analysis is sent to the Python analytics microservice [TaCertoIssoAI 2025b], which normalizes verdicts, computes aggregate metrics, assigns topics based on the IPTC (International Press Telecommunications Council) taxonomy, and generates vector embeddings. The microservice then stores the resulting record in the hybrid data lake, which the analytics frontend queries via REST APIs for search, filtering, dashboards, and export.

3. Data Lake Architecture

3.1. Hybrid Storage Design

The data lake follows a hybrid HTAP (Hybrid Transactional and Analytical Processing) design [Song et al. 2024, Zhang et al. 2024], separating OLAP and OLTP workloads into purpose-built stores. Google BigQuery serves as the OLAP layer, hosting verified analyses in a columnar schema optimized for aggregation queries, time-series analysis, dashboards, exports, and vector similarity search. Firestore serves as the OLTP layer, providing low-latency document reads for the frontend and access patterns such as retrieving individual analyses. This separation keeps large analytical queries from degrading interactive read performance.

3.2. Data Model and Schema

The data model is designed around a central principle: claim-level granularity. Rather than reducing an entire message to a single verdict, each verification record contains multiple claims, each with its own classification. This design reflects the reality of misinformation, where a single message often mixes true statements, distortions, out-of-context claims, and outright falsehoods [Guo et al. 2022]. Figure 2 shows the BigQuery schema.

Field name	Type	Mode
claims	RECORD	REPEATED
final_comment	STRING	NULLABLE
analysis_metrics	RECORD	NULLABLE
overall_verdict	STRING	NULLABLE
scraped_links	RECORD	REPEATED
processed_at	TIMESTAMP	NULLABLE
source_type	STRING	NULLABLE
user_message_text	STRING	NULLABLE
media_info	RECORD	NULLABLE
analysis_title	STRING	NULLABLE
full_combined_text	STRING	NULLABLE
document_id	STRING	NULLABLE
embedding	FLOAT	REPEATED

Figure 2. BigQuery table schema, full documentation can be found in[2].

3.3. IPTC Thematic Classification

Each verification ingested is automatically classified using the IPTC Media Topics taxonomy [IPTC 2024], a standardized hierarchical vocabulary widely adopted in the news industry. The classification relies on dense vector embeddings: every concept in the IPTC hierarchy is represented as a numerical vector computed offline using a language model, capturing its semantic meaning. At ingestion time, the claim text is embedded with the same model and compared against all concept vectors via cosine similarity to identify the most related topics.

The top candidates are then used to build full classification paths through the IPTC hierarchy, from root categories down to specific subcategories, with each path scored by a weighted combination of leaf and path-level similarities. A language model then acts as a judge, ranking the candidate paths and selecting the best match. While Kuzman and Ljubešić [Kuzman and Ljubešić 2025] demonstrated that LLMs alone can perform IPTC classification without manual annotation, our hybrid approach uses embeddings for efficient candidate retrieval and reserves the LLM for final ranking. The resulting labels support filtering, trend analysis, and hierarchical exploration of the IPTC taxonomy.

3.4. Vector Embeddings and Semantic Search

The `embedding` field stores a dense 3072-dimensional vector computed at ingestion time from each verification's `full_combined_text` with a sentence-embedding model. At query time the search text is embedded with the same model, and BigQuery's `VECTOR_SEARCH` function returns the nearest neighbors by cosine distance, retrieving recurring narratives even when they share no exact keywords with the query.

[2] https://github.com/TaCertoIssoAI/analytics/blob/main/docs/schema-bigquery.md

3.5. Anonymization and Privacy

The system follows privacy by design principles aligned with Brazil's LGPD [Governo Digital, Brasil 2024]. A disclaimer displayed upon first interaction with the WhatsApp contact explains data collection and obtains consent. Personal identifiers are removed before data enters the analytical layer, and the data lake never stores phone numbers, names, or conversation metadata.

4. Analytics Platform

The analytics platform [TaCertoIssoAI 2026] is a React-based web application that consumes the data lake through REST APIs exposed by the Python analytics microservice.

The platform provides a feed of recent verifications with filters by verdict, media type, IPTC category, and date range. It also supports semantic search via vector embeddings and detailed views of individual analyses, including claims, sources, and media. Additional features include summary dashboards and CSV export for query results.

The platform also involves qualified human reviewers, such as volunteer journalists and researchers. Through the platform, they can audit AI-generated fact-checks, flag questionable verdicts, and recommend sources the system missed.

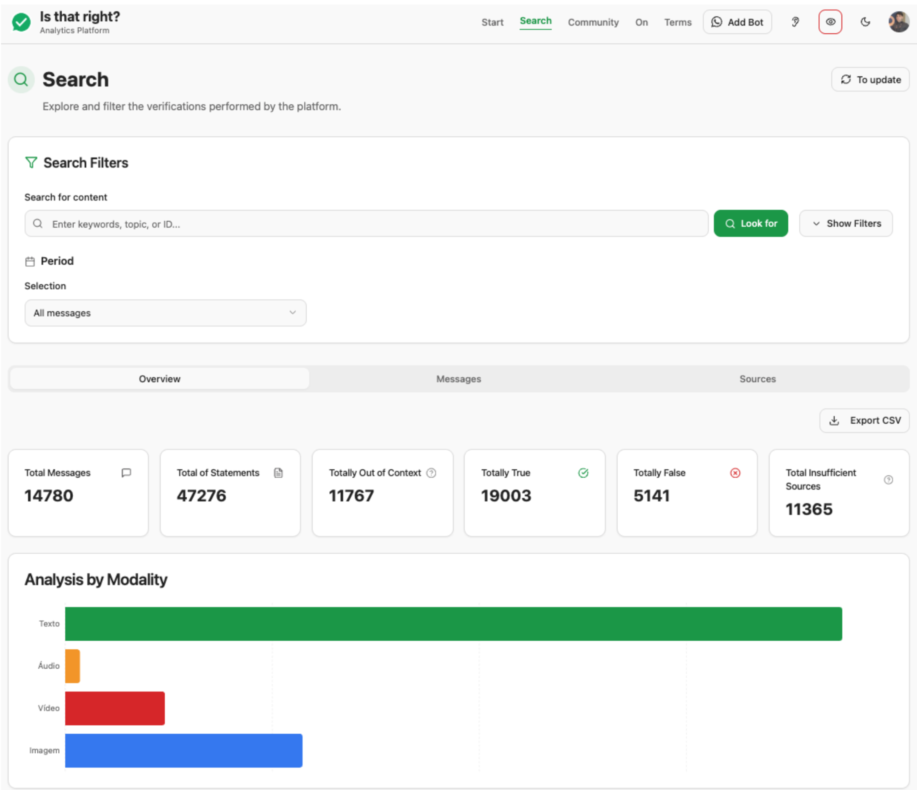

Figure 3. Analytics dashboard showing verification statistics, media modality distribution, and search filters.

5. Results and Impact

Since its launch in September 2025, TaCertoIssoAI has reached over 7,200 distinct WhatsApp users and processed more than 14,000 verified messages stored in the data lake [TaCertoIssoAI 2026]. The project participated in the AI4Good program organized by the Brazil Conference and was presented at MIT and Harvard[3]. It was also covered by Brazilian news outlets such as G1[4], Veja[5], and Jornal da USP[6].

In a snapshot of the data lake collected on May 6, 2026 [TaCertoIssoAI 2026], non-text content (audio, images, and video) accounted for approximately 23% of all verified messages, confirming the multimodal nature of misinformation on WhatsApp. Across the more than 46,000 individual claims extracted by the AI system, *out-of-context* verdicts substantially outnumber outright *false* ones (approximately 11,000 vs. 5,000). As these labels are model-generated and not yet fully validated by human reviewers, they should be read as an observed trend rather than a definitive finding. Nonetheless, the pattern suggests that misinformation circulating on WhatsApp in Brazil may rely more on reframing and decontextualization than on outright fabrication.

6. Limitations and future work

The 2026 Brazilian presidential elections are expected to sharply increase platform traffic, posing scalability and cost challenges across the microservices, web application, and data architecture. To keep costs sustainable, we are evaluating query caching, batch ingestion, and tiered storage for older records.

While TaCertoIssoAI detects AI-generated media during fact-checking [TaCertoIssoAI 2025c], the data lake does not yet store the associated signals, such as confidence scores and model uncertainty. In future work, we plan to persist these fields to support evaluation of deepfake detection models on real-world WhatsApp data.

7. Conclusion

This paper presented the data lake and analytics platform for the TaCertoIssoAI project, including claim-level modeling, HTAP architecture, multimodal representations, topic classification, vector embeddings, and privacy-preserving anonymization. As of May 2026, the platform has served over **7,200 users** and stored over **14,000 verified messages**. By turning individual fact-checks into a queryable dataset available on the web, the data lake supports analysis of misinformation trends on WhatsApp ahead of the Brazilian 2026 presidential elections.

References

Akhtar, M., Schlichtkrull, M., Guo, Z., Cocarascu, O., Simperl, E., and Vlachos, A. (2023). Multimodal automated fact-checking: A survey. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.

Gomes, J., Garcia, E., and Galvão Filho, A. R. (2025). Portuguese automated fact-checking: Information retrieval with claim extraction. In *Proceedings of the FEVER Workshop*.

[3]https://www.instagram.com/p/DVbrBhtlY6Y/

[4]https://g1.globo.com/sp/sao-carlos-regiao/videos-jornal-da-eptv-1-edicao/video/estudantes-da-usp-de-sao-carlos-desenvolvem-chat-para-verificar-fakenews-no-wha ghtml

[5]https://veja.abril.com.br/coluna/planeta-ia/ia-contra-fake-news-criada-por-alunos-

[6]https://jornal.usp.br/universidade/chatbot-contra-desinformacao-criado-por-alunos-

Governo Digital, Brasil (2024). Guia sobre privacidade desde a concepção e por padrão. https://www.gov.br/governodigital/pt-br/privacidade-e-seguranca/guias/guia_privacidade.pdf. Accessed: 5 May 2026.

Guo, Z., Schlichtkrull, M., and Vlachos, A. (2022). A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206. DOI: 10.1162/tacl_a_00454.

Hale, S. A., Humprecht, E., Esau, K., and Ferrara, E. (2024). Analyzing misinformation claims during the 2022 Brazilian general election on WhatsApp, Twitter, and Kwai. *International Journal of Public Opinion Research*.

IPTC (2024). IPTC media topics. https://iptc.org/standards/media-topics/. Accessed: 5 May 2026.

Kuzman, T. and Ljubešić, N. (2025). LLM teacher-student framework for text classification with no manually annotated data: A case study in IPTC news topic classification. *IEEE Access*, 13:38772–38786. DOI: 10.1109/ACCESS.2025.3544814.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., and Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 9459–9474.

Sawadogo, P. and Darmont, J. (2021). On data lake architectures and metadata management. *Journal of Intelligent Information Systems*, 56(1):97–120. DOI: 10.1007/s10844-020-00608-7.

Song, H., Zhou, W., Cui, H., Peng, X., and Li, F. (2024). A survey on hybrid transactional and analytical processing. *The VLDB Journal*, 33:1485–1515. DOI: 10.1007/s00778-024-00858-9.

TaCertoIssoAI (2025a). TaCertoIssoAI – AI-powered fact-checking platform for WhatsApp. https://github.com/TaCertoIssoAI. Accessed: 5 May 2026.

TaCertoIssoAI (2025b). TaCertoIssoAI analytics – microservice and web application. https://github.com/TaCertoIssoAI/analytics. Accessed: 5 May 2026.

TaCertoIssoAI (2025c). TaCertoIssoAI deep fake detector. https://github.com/TaCertoIssoAI/deep-fake-detector. Accessed: 5 May 2026.

TaCertoIssoAI (2026). TaCertoIssoAI – analytics platform. https://tacertoissoai.com.br/. Accessed: 6 May 2026.

Zhang, C., Li, G., Zhang, J., Zhang, X., and Feng, J. (2024). HTAP databases: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 36:6410–6429. DOI: 10.1109/TKDE.2024.3389693.