Proceedings of the Twentieth International AAAI Conference on Web and Social Media (ICWSM 2026)

WhatsApp Vaccine Discourse (WhaVax): An Expert-Annotated Dataset and Benchmark for Health Misinformation Detection

Jonatas Henrique dos Santos[1], Julio C. S. Reis[2], Philipe de Freitas Melo[2], João Francisco Hecksher Olivetti[2], Thales Henrique Silva[1], Matheus Gontijo Guimaraes[1], Glaucio de Souza[1], Marcos Goncalves[1], Fabrício Benevenuto[1], Filipe Belchior Bessa Zanovello[1], Marco Antônio Gonçalves Rodrigues[1], Cristiano Xavier Lima[1]

[1]Universidade Federal de Minas Gerais (UFMG), Brasil
[2]Universidade Federal de Viçosa (UFV), Brasil
jonatashds@ufmg.br, jreis@ufv.br, philipe.freitas@ufv.br, joao.olivetti@ufv.br, thaleshenrique@ufmg.br
guimaraesmatheus22@gmail.com, glaucio.gss@gmail.com, mgoncalv@dcc.ufmg.br, fabricio@dcc.ufmg.br,
filipebessa9@gmail.com, magro.mg@terra.com.br, cxlima@ufmg.br

Abstract

We introduce **WhaVax**, a new expert-annotated dataset of vaccine-related WhatsApp messages collected from large Brazilian public groups spanning multiple pandemic years. The dataset was constructed through a rigorous, carefully designed pipeline that integrates keyword-based data collection, semantic deduplication to remove near-duplicate content, and a multi-stage annotation protocol conducted by medical specialists. This process produced a high-quality gold-standard corpus, characterized by substantial inter-annotator agreement and strong reliability for downstream analysis. Additionally, we provide a detailed characterization of WhatsApp misinformation, revealing distinctive linguistic, structural, lexical, temporal, and group-level patterns, as well as a meaningful layer of ambiguous cases that reflect the complexity of health discourse in private messaging. We also benchmark classical models, fine-tuned Small Language Models, and zero- or few-shot Large Language Models under realistic data-scarcity constraints, demonstrating that strong embeddings and LLM approaches perform competitively, while domain alignment and data availability remain critical factors. This study provides a rare, high-quality resource to support misinformation research and computational modeling in encrypted communication environments.

Introduction

Brazil currently faces a renewed public health alert: the country has returned to the list of the twenty nations with the highest number of unvaccinated children worldwide, according to UNICEF and WHO (World Health Organization and UNICEF 2025). This decline poses a serious risk for a nation historically recognized for successful mass immunization campaigns, and health misinformation has emerged as a key factor undermining confidence in vaccines and public health institutions (Vijaykumar et al. 2021).

Within this ecosystem, WhatsApp plays a central role (Benevenuto and Melo 2024). With more than 120 million users in Brazil and deep penetration across regions and demographics, the platform's encrypted and semi-private group structure enables misinformation (Martins et al. 2021; Reis et al. 2023) to circulate rapidly while limiting monitoring and moderation capabilities (Melo et al. 2019b; Reis et al. 2020a). During the COVID-19 pandemic, WhatsApp became a particularly active arena for rumors, conspiratorial narratives, and alarmist vaccine discourse, amplifying uncertainty and distrust (Sharma et al. 2023).

Despite its undeniable societal impact, misinformation research on WhatsApp faces unique structural and technical barriers. Access restrictions, privacy concerns, informal language, and the highly contextual nature of messaging make systematic analysis exceptionally difficult (Garimella et al. 2025). As a result, the research community still lacks high-quality, expert-validated datasets that capture how health misinformation actually unfolds in WhatsApp conversations. Without such resources, it is difficult to characterize misinformation dynamics, evaluate computational approaches, or inform public health interventions.

This paper addresses this gap by constructing and analyzing a new expert-annotated dataset of vaccine-related WhatsApp messages. Developed through a rigorous multi-stage medical annotation process, the dataset spans multiple years of pandemic discourse and reflects misinformation as it appears in authentic conversational environments, rather than in curated fact-checking portals or public social platforms. Building this dataset raises broader research questions (RQs) that motivate our study:

- **RQ1**: Is it feasible to construct a domain-specific, expert-annotated, high-quality dataset of medical misinformation based on authentic WhatsApp communication?
- **RQ2**: Given the inherent constraints of such environments – limited samples, informal text, and topic specificity – what limitations and opportunities arise when training classifiers on small, specialized datasets?
- **RQ3**: How do different classification strategies (classical machine learning, fine-tuned Transformer-based Small Language Models (SLMs), and zero and few-shot Large Language Models (LLMs)) compare when applied to a medical misinformation classification task in WhatsApp?

To address these questions, we (i) build and publicly release an expert-labeled dataset of WhatsApp vaccine-related messages, called **WhaVax**; (ii) provide a detailed characterization of linguistic, structural, lexical, temporal, and group-level patterns; and (iii) systematically compare classical classification models, Small Language Models (SLMs), and Large Language Models (LLMs) under realistic data-scarcity conditions for misinformation classification.

Our results show that constructing such a dataset is feasible and highly informative. Careful filtering, semantic deduplication, and expert review yield a reliable gold-standard corpus with meaningful agreement. Despite size and access constraints, the dataset enables rich analyses, revealing distinct stylistic and behavioral patterns, as well as meaningful ambiguity zones that reflect the real complexity of health misinformation. Furthermore, our modeling experiments demonstrate that even small, carefully curated datasets can support effective automatic detection. Classical models with strong embeddings remain competitive, fine-tuned Transformers are highly sensitive to domain alignment, and zero-/few-shot LLMs achieve strong performance without supervised tuning, although outcomes vary across architectures.

In sum, this work offers three main contributions: (1) a novel, expert-annotated dataset of vaccine-related WhatsApp messages; (2) an in-depth empirical characterization of misinformation in group messaging environments; and (3) a systematic comparative evaluation of misinformation classifiers under realistic, resource-constrained conditions. Together, these contributions provide a rare, high-quality resource for the community and actionable evidence to support computational research and public health strategies in encrypted messaging platforms.

Related Work and Datasets

The study of online health information has received increasing attention, particularly during the COVID-19 pandemic (Skafle et al. 2022), which amplified previously niche anti-vaccination narratives (Loomba et al. 2021; Zhao et al. 2023). These developments highlighted the societal risks of misinformation, especially in crisis contexts (Caceres et al. 2022), where misleading content directly conflicts with medical guidance and can endanger lives.

Misinformation on Online Social Networks. The ubiquity of online information has reshaped everyday life, influencing behaviors (Storm, Stone, and Benjamin 2017) and democratic processes worldwide (Maweu 2019; Reis et al. 2020b; Mauk and Grömping 2024). However, the volume and speed of information circulation also create conditions that complicate truth-seeking. Research on information overload shows that users exposed to excessive content are more likely to share unverified or incorrect information (Huang, Lei, and Ni 2022; Bermes 2021). Social media platforms, in particular, have been repeatedly identified as major vectors of online misinformation (Suarez-Lledo and Alvarez-Galvez 2021; Chaudhuri et al. 2024).

Datasets on Vaccine Opinion and Misinformation. A systematic review on Instant Messaging Misinformation highlights automatic detection as a key research direction on the field (Olivetti et al. 2025). However, this remains challenging due to the open-ended nature of misinformation, which varies in topic, format, and degree of subtlety. Progress in this area depends critically on access to diverse, high-quality datasets that capture both reliable and misleading information across platforms. To address this need, several studies have released publicly available datasets on vaccine discourse and credibility assessment. Table 1 summarizes these resources, highlighting their scope, content, and annotation strategies.

Research Gap. As summarized in Table 1, most publicly available vaccine misinformation datasets are sourced from Twitter/X. While these resources are valuable, they provide a skewed perspective focused exclusively on public-facing discourse on Twitter/X. As a result, they fail to capture the unique dynamics of group communication on WhatsApp, which have become central to modern misinformation ecosystems. Unlike public social platforms, WhatsApp feature end-to-end encryption and a semi-private group structure that enables the rapid spread of rumors while hindering external monitoring or moderation. Furthermore, by introducing a large-scale corpus in Portuguese, **WhaVax** bridges the gap between public and private platform analysis and enhances linguistic diversity, enabling more robust cross-cultural and cross-lingual investigations into vaccine discourse.

Dataset Construction Methodology

Our dataset is based on the publicly available WhatsApp repository introduced in (Resende et al. 2019; Melo et al. 2019a), which compiles messages from large Brazilian public groups. The content covers 2020–2023 and comes from public groups accessible through invitation links obtained via community-sharing platforms and targeted Web searches, initially guided by Brazilian political keywords.

Data collection was conducted using an automated large-scale pipeline, where multiple registered accounts continuously retrieved group messages. For each message, both text and metadata were collected, including unique identifiers, sender information, group origin, media type, and timestamps. In accordance with standard practices in instant messaging research, all personally identifiable information was anonymized to comply with Brazilian data protection regulations, retaining only area codes to allow coarse-grained geographic analysis without compromising privacy. Processed messages were stored in JSON format to support subsequent analysis.

Health Data Filtering

The original repository contains millions of messages collected over three years, most of which are unrelated to vaccination or health. To isolate relevant content, we applied a keyword-based filtering procedure using Portuguese vaccination terms (e.g., “vacina”, “vacinação”, “vacinado”) along with variant spellings, slang, and obfuscated forms (e.g., “v4c1n4”, “vachina”), which are commonly linked to conspiratorial discourse. Additional terms, including vaccine brand names, were tested but later discarded after yielding a high proportion of off-topic content.

Dataset	Description	Labels	Annotation Method
(Cui and Lee 2020)	COVID-19 healthcare misinformation dataset with news articles, social media posts, and user engagements.	Fake vs. real + metadata labels.	Fact-check verification.
(Hossain et al. 2020)	6.6k tweets annotated for stance regarding curated COVID-19 misconceptions.	Agree/Disagree/No Stance per misconception.	Misconceptions curated by medical experts + human annotation.
(Weinzierl and Harabagiu 2021)	Tweets mapped to specific COVID-19 vaccine misinformation targets; includes stance labels.	Target labels; stance: Agree/Disagree/No Stance.	Curated misinformation targets + human expert annotation.
(Chen, Chu, and Subbalakshmi 2021)	Multimodal repository with news and tweets (text + images + temporal data) focused on COVID-19 vaccines.	News: reliable/unreliable; Tweets: reliable/unreliable/inconclusive.	Media credibility rankings + manual stance/-credibility annotation.
(Hayawi et al. 2022)	Large Twitter dataset (15M tweets) with 15k manually annotated for COVID-19 vaccine misinformation.	Binary: misinformation vs. general vaccine tweet.	Human annotation + expert validation.
(Crupi et al. 2022)	Twitter dataset (16.2M tweets) of the Italian COVID-19 vaccination debate (665k unique users, Sep 2019–Nov 2021).	Stance: Supporter/Hesitant/Other/Pets.	Manual annotation propagated via hierarchical clustering on retweet networks.
(Poddar et al. 2022a)	Large Twitter dataset (15.7M tweets) for long-term (Jan 2018–Mar 2021) analysis of vaccine opinions, focusing on stance change over time.	Tweet Stance: Anti-Vax/Pro-Vax/Neutral (used for user classification).	Human annotation (1.7k tweets) + CT-BERT++ classifier for user stance categorization (97% precision).
(Weinzierl and Harabagiu 2022)	Large Twitter dataset (9.1M original tweets, Dec 2019-Jul 2021) used to identify 113 Vaccine Hesitancy Framings and derive 9 user profiles.	Framing Stance: Accept/Reject/-Doubt. 9 User Hesitancy Profiles.	Tweets are summarized through a QA system; Hesitancy profiles are clustered through sparse k-means.
(Giovanni et al. 2022)	Large Twitter dataset (over 70M tweets) on COVID-19 vaccine conversations in French, German, and Italian (Nov 2020–Nov 2021).	Stance: Pro-vaccines/Anti-Vaccines/Neutral/Out-of-Context.	Annotation by native speakers of 1000 random tweets per language; a third annotator resolved conflicts.
(Poddar et al. 2022b)	Large-scale dataset (10k tweets) for multi-label classification and summarization of COVID-19 anti-vaccine concerns.	12 specific concern classes (multi-label) + explanations for each label.	Manual annotation by human experts (labels and natural language explanations).
(Zarei et al. 2023)	6,373 vaccine-related tweets annotated with stance, misinformation indicator, entities, and message type.	Stance (pro/anti/neutral), misinformation (yes/no), entities, message type.	Manual annotation by trained communicators/journalists.
(Mu et al. 2023)	Dataset of 3,101 tweets in English on attitudes toward COVID-19 vaccination, notable for separating hesitancy from anti-vaccine attitude.	Pro, Anti, Hesitancy, Irrelevant.	Manually annotated into four categories; the process included annotator training and quality testing.

Table 1: Datasets related to vaccine misinformation.

keyword	#Occurrences	Unique users	Unique groups
vacina	65373	10710	876
vacinas	32476	6835	768
vacinação	17773	4450	695
vacinados	7813	2521	533
vacinar	5675	2458	532
vacinado	3504	1758	473
vacinadas	2517	1058	365
vacinada	1928	990	348
vacinal	1686	855	340
vacinou	1637	1012	335
vachina	1487	670	274
vacinei	147	130	94
v4c1n4	14	8	8

Table 2: Key words for the vaccine filter.

After filtering, we obtained 84,640 vaccination-related messages from 15,148 users. Table 2 summarizes keyword coverage, including alternative terms such as "vachina" (1,487 occurrences) and "v4c1n4" (14 occurrences), which indicate pejorative and encoded references commonly associated with ideological and conspiratorial discourse. Although the broader corpus is largely political, those large public WhatsApp groups also served as key communication spaces for many other topics, such as health issues, during the analyzed period. While this dataset represents one of the largest known WhatsApp corpora related to vaccines, the absence of official WhatsApp-access mechanisms prevents precise claims about population representativeness.

Finally, to ensure data quality and reduce redundancy, we applied semantic deduplication. Message embeddings were generated using `SentenceTransformers` (Reimers and Gurevych 2019) and indexed with `Faiss` (Douze et al. 2025) for nearest-neighbor similarity search. Empirically, most near-duplicates appeared below a distance threshold of 0.02, which we adopted to group equivalent messages, retaining only the first instance in each cluster. This step removed 5.18% of messages, resulting in a final dataset of 80,257 unique messages from 14,322 users across 932 public WhatsApp groups.

Expert Medical Annotation

Message annotation was conducted by four medical professionals using a progressive, iterative, protocol-refinement process to ensure consistent criteria and stable decision-making. Because the evaluators had no prior experience with computational labeling tasks, the procedure began with a small pilot phase to align expectations, clarify the study scope, and define what would constitute health misinformation in this context.

In the first stage, each evaluator labeled 50 randomly selected messages. The resulting annotations were analyzed

to assess agreement and identify sources of ambiguity, which led to targeted refinements in the guidelines. A second round of 50 messages was then completed, followed by an intermediate stage with 300 messages to evaluate whether consistency was maintained at a larger scale. At the end of this stage, Fleiss' Kappa reached 0.65, indicating "substantial agreement" according to (Landis and Koch 1977) and validating the protocol for full-scale annotation.

In the final phase, the same four experts annotated 950 randomly selected messages, maintaining thematic diversity and natural temporal distribution. The result is a dataset in CSV format containing approximately 30% misinformation and 70% non-misinformation, with 950 WhatsApp messages consistently labeled by medical specialists, serving as the gold-standard reference for the analyses and classification experiments presented in this work. Each entry includes the message text, all four annotators' labels, anonymized group and sender metadata, temporal information, indicators of whether the message is forwarded or quoted, and the sender's country or area codes.

Limitations

Although **WhaVax** provides a rare, medically validated view of vaccine discourse within encrypted messaging environments, some limitations remain. First, the dataset is derived from Brazilian public WhatsApp groups, which represent only a portion of the broader messaging ecosystem. As a result, it may overrepresent politically engaged, highly vocal communities and underrepresent private, familial, or less politically exposed conversations. Temporal coverage, platform-specific features, and evolving sociopolitical contexts may also limit generalization to other time periods, populations, or countries.

Second, despite expert curation, annotation inevitably involves judgment and may introduce biases, especially in borderline or ambiguous messages, even though it yielded a final Fleiss' Kappa value of 0.621, which is considered good. These cases reflect the inherent uncertainty of real-world health communication and the challenges of interpreting irony, cultural references, or implicitly framed narratives, which can still cause interpretive skew. Even with remaining gray areas, the sustained involvement of qualified medical professionals throughout dataset conception and validation substantially reduces clinical misinterpretation. Beyond strengthening scientific rigor, this participation is increasingly required by emerging regulations for medical datasets, reinforcing ethical robustness, reliability, and legal compliance.

Dataset Characterization

This section characterizes the annotated dataset by describing the label distribution, expert agreement dynamics, and the presence of ambiguous instances that challenge automatic classification. Together, these analyses provide a behavioral perspective on the corpus.

Annotation and Agreement Pattern Distribution

Of the 950 messages evaluated, 286 ($\approx$ 30%) were labeled as misinformation under the majority criterion (at least three

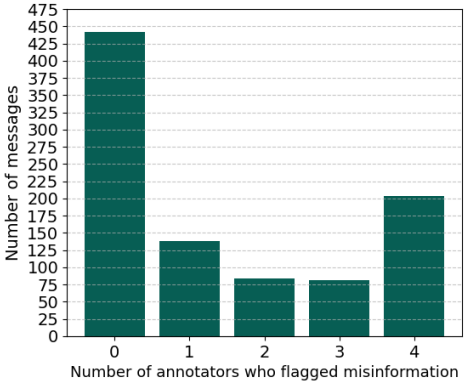

Figure 1: Message distribution of annotators agreements.

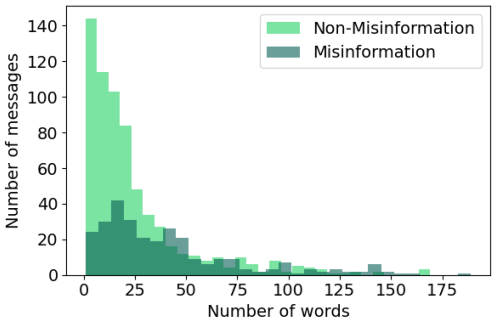

Figure 2: Message size distribution.

out of four votes), while the remaining 664 messages were classified as non-misinformation, resulting in considerable imbalance (skewness). As shown in Figure 1, agreement varies substantially: 442 messages were unanimously classified as non-misinformation and 204 as misinformation, indicating a sizable subset of clearly identifiable cases.

Linguistic and Structural Messages Analysis

Further characterization of the dataset was based on the textual properties of the messages. Clear differences emerge between misinformation and non-misinformation content. As shown in Figure 2, misinformation messages are noticeably longer, containing on average about 290 characters and nearly 40 words, whereas non-misinformation messages contain roughly 211 characters and 22 words.

These findings indicate that misinformation tends to be more elaborate, often relying on extended explanations, causal narratives, or lists of alleged adverse effects. Such verbosity may serve a persuasive role by creating an impression of credibility or overwhelming readers with excessive detail, potentially masking incorrect claims.

In addition to message length, notable stylistic differences appear in the use of expressive punctuation. As shown in Table 3, misinformation messages contain substantially more exclamation points and question marks, suggesting a more emotional or rhetorical discourse style. This pattern

Class	#Char.	#Words	#Excl.	#Quest. Marks	#Cap. letters
Non Misinf.	210.92	22.43	0.22	0.29	12.57
Misinf.	290.39	39.62	0.59	0.46	35.67

Table 3: Textual characteristics.

aligns with engagement strategies commonly observed in misinformation, where emotional appeal is used to enhance persuasion or introduce doubt. A similar trend is seen with emojis: misinformation messages contain approximately four times as many emojis as non-misinformation content, reinforcing affective cues and amplifying the intended emotional impact on readers.

The use of capital letters also emerges as a relevant discriminative feature. Messages labeled as misinformation contain, on average, nearly three times more capitalized text than informational messages. This pattern suggests deliberate emphasis strategies, likely intended to attract attention or convey a sense of urgency, further reinforcing the emotional tone commonly associated with misinformation content.

Lexical Patterns and N-gram Analysis

Analysis of the most frequent n-grams also highlights semantic distinctions between classes. In misinformation messages, terms such as “mRNA”, “efeitos”, “morte”, “não”, and “tomar” are more frequent, usually associated with narratives of risk, fear, or doubt towards treatments and vaccines. This recurrent usage emphasizes potential negative outcomes and distrust of medical interventions. In contrast, non-misinformation messages feature n-grams linked to institutional sources, public policies, and health campaigns—such as “vacinação”, “governo”, and “contra covid”—suggesting a more informative discourse aligned with official communications and public health guidance.

These lexical distinctions reinforce the idea that health misinformation is conveyed not only through factually incorrect claims but also through distinctive linguistic patterns, which classification models can systematically identify.

Group-Level Distribution of Misinformation

Group-level analysis shows that misinformation is unevenly distributed across communities, as illustrated in Figure 3. Some groups have much higher concentrations of misinformation, with proportions exceeding 60% and, in a few cases, approaching 100%. This indicates highly polarized groups where misinformation prevails.

Conversely, a significant portion of messages displayed ambiguity; specifically, 84 messages resulted in tied evaluations among the four experts, representing approximately 8.8% of the dataset. These cases highlight the challenges of assessing health misinformation, particularly in short or opinion-driven messages. Following a conservative strategy, these cases were labeled as non-misinformation. Additionally, no clear linear relationship exists between forwarding and misinformation, indicating that sharing depends more

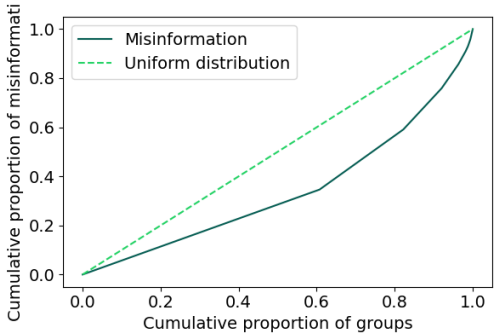

Figure 3: Concentration of misinformation by group.

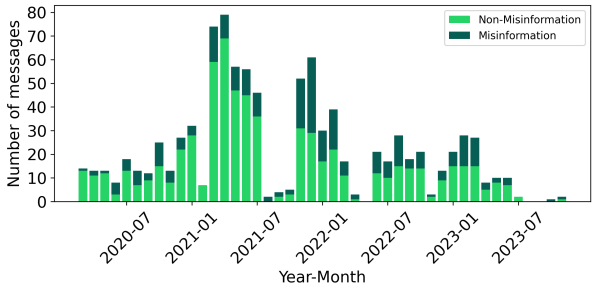

Figure 4: Temporal distribution of messages.

on group-specific social and contextual factors than on content alone.

Temporal Analysis

Figure 4 presents the monthly temporal distribution of messages from 2020 to 2024, distinguishing between misinformation and non-misinformation content. Instead of focusing on specific peaks, this analysis evaluates dataset coverage and the coexistence of informative and misleading content over time. Misinformation is consistently present throughout the entire period, including months with lower activity, indicating its persistent nature. The dataset offers broad temporal coverage, allowing computational analyses to include messages from different phases and discursive contexts of the pandemic, which helps mitigate temporal bias and enhances the robustness of subsequent experiments.

Two major activity peaks emerged in 2021. The first and most pronounced occurred in March 2021, coinciding with the arrival of COVID-19 vaccines in Brazil and the widely publicized first administration (Natalie Cancian, Renato Machado 2021; Vieira 2021). This marked a discursive turning point on WhatsApp, with an intensified presence of supportive, hesitant, and rejection narratives regarding vaccination. A second peak appeared at the end of 2021, aligned with the surge in COVID-19 mortality driven by the Omicron variant (Simões 2021) and ANVISA’s authorization of Pfizer’s vaccine for children aged 5–11 (Ricardo Brito 2021), which reignited debate marked by increased misinformation, moral panic, and conspiratorial narratives (de Albuquerque et al. 2022).

Classification Experiments – Approaches, Experimental Setup and Results

We frame our problem as a binary text classification task, where each of the 950 expert-annotated WhatsApp messages is labeled as either *misinformation* or *non-misinformation*. The distribution is naturally imbalanced (approximately 70% non-misinformation and 30% misinformation), reflecting real-world prevalence and presenting additional challenges due to limited data, informal writing, and highly contextual conversational language.

To thoroughly investigate model behavior under these realistic constraints, we evaluate three families of approaches: (i) classical machine learning models, (ii) fine-tuned Transformer-based Small Language Models (SLMs), and (iii) Large Language Models (LLMs) using in-context learning (ICL). Classical supervised classifiers—Support Vector Machines (SVM), Logistic Regression, Random Forest, Multilayer Perceptron (MLP), and XGBoost—were tested with different embedding strategies to assess the impact of representation quality on performance (Kowsari et al. 2019).

For SLMs, we fine-tuned Portuguese and biomedical-oriented Transformers, including BERTimbau (Souza, Nogueira, and Lotufo 2020), RoBERTa (Conneau et al. 2020), BERTuguês (Mazza Zago and Agnoletti dos Santos Pedotti 2024), and BioBERTpt (Schneider et al. 2020). We optimized hyperparameters (learning rate, weight decay) using nested cross-validation to maximize performance and mitigate overfitting.

We then examine LLMs in zero- and few-shot settings, focusing on in-context classification without parameter training. We evaluate open-source models such as LLaMA 3.1 (8B), LLaMA 3.2 (3B), DeepSeek v3.2 (685B) (Liu et al. 2025), and Qwen 3 (30B) (Team 2025), enabling analysis of model size, architecture, and language alignment effects. For LLaMA 3.1, we also vary the number of in-context examples to assess sensitivity to prompt conditioning. Finally, we include proprietary models accessed via API (GPT-5.1 and GPT-5.2[1]), evaluated in few-shot mode to compare open and proprietary systems and to explore whether more recent knowledge grounding affects performance.

This multifaceted evaluation enables a direct, fair comparison across paradigms, highlighting when simple models suffice, when domain-adapted Transformers help, and when modern LLMs provide advantages.

Experimental Setup

We use 5-fold stratified cross-validation to maintain class balance across folds and improve the robustness of estimates (Raschka 2020). To further reduce sampling variance, we repeat the entire cross-validation process eight times with controlled shuffling, generating distributions of results rather than single-point estimates. For each model, we report the mean and standard deviation across runs.

We compute accuracy, precision and recall for each class, along with macro-F1 as the primary comparison metric, since it assigns equal weight to misinformation and non-misinformation—an essential property in imbalanced scenarios. Statistical significance between competing models is assessed using the paired Wilcoxon test over cross-validation runs, which avoid normality assumptions while ensuring reliable comparative claims. This evaluation framework provides a rigorous and comparable basis for assessing classical models, fine-tuned SLMs, and in-context LLMs under the same conditions. Table 4 summarizes performance using macro-F1 as the main criterion, with model families discussed further in the following sections.

[1]openai.com/api/

Model	Average macro-F1
Bertimbau Embedding	
Support Vector Machine	0.724 ± 0.030
Logistic Regression	0.747 ± 0.036
Random Forest	0.656 ± 0.041
Multi-Layer Perceptron	0.738 ± 0.031
XGBOOST	0.687 ± 0.034
Qwen8b Embedding	
Support Vector Machine	0.784 ± 0.025
Logistic Regression	**0.791 ± 0.026**
Random Forest	0.663 ± 0.035
Multi-Layer Perceptron	**0.789 ± 0.026**
XGBOOST	0.729 ± 0.035
BerTimbau	0.712 ± 0.101
Roberta	0.580 ± 0.116
Bertugues	0.660 ± 0.056
BioBertpt	0.634 ± 0.102
Llama 3.1 8B (ICL Mode - 8 examples)	0.734 ± 0.028
Llama 3.2 3B (ICL Mode - 8 examples)	0.503 ± 0.031
Qwen 30B (ICL Mode - 8 examples)	0.383 ± 0.027
DeepSeek-V3.2 685B (ICL Mode - 8 examples)	0.744 ± 0.024
GPT 5.1 (ICL Mode - 8 examples)	**0.786 ± 0.026**
GPT 5.2 (ICL Mode - 8 examples)	0.780 ± 0.023

Table 4: Macro F1 - Classifications models.

Classical Machine Learning Models with Contextual Embeddings

Messages were represented using embeddings from pre-trained language models to provide classical classifiers with dense semantic encodings. BERTimbau embeddings were initially used, computed as the mean of the last hidden-layer token representations, yielding strong baseline results with Logistic Regression. In a second phase, all models were re-evaluated with Qwen (8B) embeddings under the same pipeline, enabling a controlled assessment of representation effects. Classical models show stable performance, with Logistic Regression and Qwen (8B) achieving the best result (macro-F1 = 0.791). The comparison confirms that richer embeddings consistently outperform BERTimbau (macro-F1 = 0.747), highlighting the central role of representation quality in handling the informal and noisy language of WhatsApp messages.

The gains achieved by Qwen (8B) extend beyond Logistic Regression, delivering consistent macro-F1 improvements across all classical models compared to BERTimbau, including SVM (0.724 $\rightarrow$ 0.784) and MLP (0.738 $\rightarrow$ 0.789). Even models that typically struggle with dense, high-dimensional vectors, such as Random Forest and XGBoost, show modest improvements, reinforcing that embedding quality often has a greater influence than classifier architecture in lexically complex settings. Qwen's superior performance likely

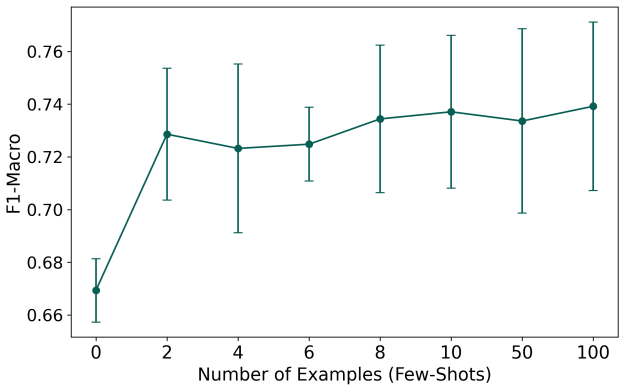

Figure 5: Impact of few-shots on performance.

results from its larger scale and broader pre-training, which better capture semantic nuance, informal language, and implicit references common in WhatsApp messages. These results highlight that robust embeddings can compensate for simpler classifiers, underscoring the central role of textual representation in misinformation detection.

Fine-Tuned Small Language Models (SLMs)

Fine-tuned Small Language Models (SLMs) show more heterogeneous performance. BERTimbau achieves a macro-F1 of 0.712, comparable to classical models but with noticeably higher variance, indicating sensitivity to training splits and hyperparameters under data scarcity. Other SLMs, including BERTuguês, BioBERTpt, and especially RoBERTa, perform below expectations. This likely results from a mismatch between their pre-training domains and the characteristics of our dataset, which features informal writing, abbreviations, typos, and conversational phrasing typical of instant messaging. The limited sample may also be insufficient to adapt these models for further generalization. Fine-tuned SLMs show variable performance and depend heavily on domain alignment and data availability. In our setting, simpler models with strong embeddings were more reliable.

Large Language Models with In-Context Learning

Few-Shot Behavior and Context Sensitivity. We first analyze the few-shot[2] behavior of LLaMA 3.1 (8B), which serves as our reference model. Figure 5 shows its macro-F1 performance as a function of the number of examples provided in the prompt. The x-axis is shown on a logarithmic scale to better capture performance variations in low-context regimes, where changes tend to be most informative. A clear gain is observed when moving from zero-shot to few-shot configurations: even a small number of examples substantially improves performance, indicating that the model quickly internalizes discriminative cues for the task. However, this improvement plateaus around eight examples, after which additional context yields only minor oscillations within the variance margin.

[2]Few-shot learning is a type of in-context learning (ICL) in which typically a few examples (2-50) are used in the prompt.

This pattern suggests that most of the signal for distinguishing misinformation is captured in early examples, with diminishing returns as context increases. Based on this, we use eight shots for all LLM experiments to balance performance, cost, and comparability. The prompt used for Few-Shot consisted of an instruction to the models and eight already labeled examples:

English translation of in-context learning prompt

```
You are a health misinformation
classifier.

Classify messages as:
0 = Non-misinformation or
1 = Is misinformation

Here are some examples:
{few_shot_examples}

Now classify the following message,
respond strictly with 0 or 1,
do not respond in any other way.

Text: [Message]

Answer:
```

Few-Shot Behavior and Context Sensitivity. LLMs evaluated under this standardized few-shot regime present heterogeneous outcomes. LLaMA 3.1 (8B) and DeepSeek-V3.2 (685B) achieve competitive macro-F1 scores near 0.74, comparable to the best classical models. These results indicate that, even without supervised fine-tuning, LLMs can effectively learn the task from limited in-context supervision.

In contrast, smaller or less linguistically aligned models, such as LLaMA 3.2 (3B) and Qwen (30B), perform substantially worse. This indicates that success in in-context learning depends not only on scale, but also on pre-training quality, language coverage, and instruction-following capability. Overall, our findings confirm that in-context learning is powerful but not universally reliable, with performance sensitive to model design and task domain alignment.

API-Based in-context Learning Models. Finally, GPT-based models rank among the strongest performers. GPT-5.1 reaches a macro-F1 of 0.786, and GPT-5.2 achieves 0.780, both with low variance, statistically matching the best classical approach (Logistic Regression with Qwen embeddings). These results position GPT models as highly effective, stable alternatives for misinformation detection in Portuguese WhatsApp messages without fine-tuning.

Their advantage likely stems from large-scale pre-training, strong instruction-following, and broad prior exposure to health-related data. Interestingly, the small gap between GPT-5.1 and GPT-5.2 suggests that temporal knowledge updates play a smaller role, while general reasoning and semantic understanding appear to be more important.

In-depth Analysis of the Best Models

Table 6 presents the effectiveness of the three best-performing models – LR, MLP (both using Qwen embeddings), and GPT-5.1—broken down into Overall Accuracy (MicroF1), and Precision and Recall for non-misinformation (Class 0) and misinformation (Class 1). The table shows that all three models achieve very similar Macro-F1 scores, with GPT-5.1 having a slight disadvantage. However, their

Portuguese	English (Translation)	Correct Label	N Incorrect Models
CENTENAS DE MILHARES ESTÃO MORTOS E ENTERRADOS PORQUE O **** RECUSOU-SE A COMPRAR VACINAS! POR CAUSA DISSO A CONTAMINAÇÃO AUMENTOU E EXPLODIU!!! **** É UM ASSASSINO MILICIANO GENOCIDA!	HUNDREDS OF THOUSANDS ARE DEAD AND BURIED BECAUSE **** REFUSED TO BUY VACCINES! BECAUSE OF THIS, THE CONTAMINATION INCREASED AND EXPLODED!!! **** IS A GENOCIDAL MILITIA MURDERER!	0	18
BÉLGICA EM FÚRIA Bruxelas sob domínio da Guerra Os holandeses e agora os Belgas. A humanidade está saindo do ostracismo das manifestações pacíficas e idiotas para a verdadeira reação contra a tirania sanitária. O alvo são as cabeças dos políticos, dos governantes e também dos jornalistas corruptos que promovem a tirania das vacinas genocidas e patrocinam a ditadura dos passaportes sanitários. Parece que a fúria foi despertada. Agora a Humanidade está acordando para a realidade: de que não há conquistas sem guerras. Bruxelas amanheceu debaixo da ira da população que atacou o sistema com pedras, porretes e muita gasolina. A fúria e a raiva chegarão nos responsáveis, que serão caçados, trucidados e enviados de volta para o inferno. NÃO RECUEM! NÃO SE DOBREM! ARREBEBENTEM COM TUDO!	BELGIUM IN FURY Brussels under the dominion of War The Dutch and now the Belgians. Humanity is emerging from the ostracism of peaceful and idiotic demonstrations into a true reaction against sanitary tyranny. The target is the heads of politicians, rulers, and also corrupt journalists who promote the tyranny of genocidal vaccines and sponsor the dictatorship of health passports. It seems that fury has been awakened. Now Humanity is waking up to the reality: that there are no conquests without wars. Brussels awoke under the wrath of the population that attacked the system with stones, clubs, and plenty of gasoline. The fury and rage will reach those responsible, who will be hunted down, slaughtered, and sent back to hell. DO NOT RETREAT! DO NOT BEND! BREAK EVERYTHING DOWN!	0	17
Pessoal, percam uma horinha de vocês e vejam esta reportagem. É estarrecedora e muito esclarecedora. São dois cientistas, sendo que um deles é umas maiores sumidades sobre vacinas, responsável pelo design de vacinas em grandes órgãos públicos e empresas privadas. Veja o que ele fala sobre a vacinação. É muito esclarecedor.	Folks, take an hour of your time and watch this report. It's shocking and very enlightening. It features two scientists, one of whom is a leading authority on vaccines, responsible for vaccine design in major public agencies and private companies. Watch what he says about vaccination. It's very insightful.	0	17
China admite que eficácia de suas vacinas contra Covid-19 não é alta	China admits that the effectiveness of its Covid-19 vaccines is not high.	1	16
Vai saber se foi alguma vacina que deu errado	Who knows if it was some vaccine that went wrong?	1	16

Table 5: Examples incorrectly classified by most models. The characters **** are used to anonymize personal names.

Model	Accuracy	Precision (0)	Recall (0)	Precision (1)	Recall(1)
LR	0.825 ± 0.02	0.875 ± 0.02	0.876 ± 0.02	0.711 ± 0.04	0.707 ± 0.06
MLP	0.827 ± 0.02	0.865 ± 0.02	0.893 ± 0.02	0.731 ± 0.04	0.674 ± 0.05
GPT 5.1	0.798 ± 0.03	0.964 ± 0.02	0.739 ± 0.04	0.607 ± 0.03	0.935 ± 0.03

Table 6: Metrics for the Top 3 models.

behavior differs in important ways. LR and MLP, both using contextual embeddings, exhibit similar patterns: they achieve higher precision for non-misinformation but are less effective at identifying misinformation, with precision and recall for this class in the 0.67–0.73 range. In contrast, GPT-5.1 demonstrates a different profile. It excels in precision for non-misinformation and, importantly, in recall for misinformation, meaning it identifies most misinformation cases while accepting a higher rate of false positives. In summary, although overall effectiveness is comparable, LR and MLP provide more conservative and balanced decisions, whereas GPT-5.1 prioritizes high recall for misinformation, making it suitable for screening scenarios where missing misinformation is unacceptable.

Model Error Analysis

Table 5 shows common misclassification patterns across models. The first two cases, misclassified by 18 and 17 out of 20 models, contain highly emotional and politicized language. Although medical experts labeled them as non-misinformation, most models flagged them as misinformation, likely because they associate inflammatory rhetoric and pandemic narratives with false content. This suggests confusion between hate speech or extremism and factual misinformation.

A second example appears in messages that appeal to scientific authority or invite reflection. Even without explicit false claims, references to "scientists" and "revelations" triggered misinformation predictions. Models struggle to distinguish skepticism from actual incorrect information. The bottom examples show the opposite difficulty: short, ambiguous statements – sometimes used manipulatively – were often misclassified as non-misinformation. These involve insinuation, framing, or selective facts that subtly foster distrust, which models frequently fail to detect.

Overall, these errors reinforce that misinformation detection goes beyond standard text classification and requires sensitivity to discourse, pragmatics, and context, dimensions that current models still struggle to capture.

Conclusion and Future Work

This paper introduces **WhaVax** – a new expert-annotated dataset of vaccine-related WhatsApp messages, providing one of the first medically validated resources that capture how health misinformation circulates within encrypted messaging environments. Through rigorous filtering and a structured multi-stage labeling process with medical professionals, our dataset was derived from large-scale WhatsApp public group data collection, demonstrating the scientific value

and feasibility of such approaches.

Our analyses revealed distinct linguistic, structural, lexical, temporal, and group-level traits of misinformation, and highlighted a meaningful subset of inherently ambiguous messages that reflect the complexity of real-world health communication. Classical models with strong embeddings remain competitive, fine-tuned SLMs depend on domain alignment and data availability, and zero-/few-shot LLMs perform well without supervision, although with substantial variation across architectures. Collectively, these results provide an empirically grounded resource and key insights for misinformation detection in private messaging platforms. We hope that releasing our dataset and methodology will foster new research directions and help address the scarcity of annotated data in this domain.

Modeling Annotation Disagreement and Ambiguity. Given that the dataset includes multiple annotations per message, there is a valuable opportunity to study disagreement among health misinformation expert annotators. This enables research on ambiguity, borderline cases, and subjectivity in issues that are especially important in health-related contexts, where opinion, uncertainty, and emerging evidence often intersect. Such analyses can inform the design of more robust annotation schemes and evaluation protocols.

Informing Public Health Interventions and Policy. Insights from this dataset may inform public health campaigns by highlighting recurring narratives, misconceptions, and sources of confusion about vaccines. Although not intended for direct operational use, the dataset can support evidence-based discussions on how health authorities and fact-checkers might better address misinformation circulating in instant messaging ecosystems.

Ethics Statement and Dataset Availability

This study was conducted in compliance with established biomedical research ethics standards and was reviewed and approved by an independent Research Ethics Committee (IRB/CEP equivalent). The protocol was approved under the Brazilian national ethics system (Plataforma Brasil).

All data collection, processing, and release protocols were designed to minimize risk and protect participant privacy. Only de-identified data are included in the dataset, and access is restricted to research and educational purposes.

Importantly, the study included active participation of qualified medical professionals throughout dataset conception, curation, and validation. Their involvement ensured clinical relevance, ethical rigor in handling sensitive health-related data, and alignment with evolving regulatory expectations that increasingly require medical oversight in the development and dissemination of healthcare datasets.

Last, the dataset build in this paper is publicly available here: https://zenodo.org/records/18190030.

Acknowledgments

This work was funded by the Center for Innovation and Artificial Intelligence in Health (CI-IA Saúde), with financial support from the FAPESP, grant no. 2020/09866-4; the FAPEMIG, grant no. PPE-00030-21; and UNIMED Belo Horizonte. It was also supported by the National Institute of Science and Technology in Responsible Artificial Intelligence for Computational Linguistics, Information Treatment, and Dissemination (INCT-TILDIAR), funded by the CNPq, grant no. 408490/2024-1. Additional individual support was provided by FAPEMIG, including grants APQ-04803-25 and APQ-03115-24.

References

Benevenuto, F.; and Melo, P. 2024. Misinformation Campaigns through WhatsApp and Telegram in Presidential Elections in Brazil. *Comm. of the ACM*, 67(8): 72–77.

Bermes, A. 2021. Information overload and fake news sharing: A transactional stress perspective exploring the mitigating role of consumers' resilience during COVID-19. *Journal of Retailing and Consumer Services*, 61: 102555.

Caceres, M. M. F.; Sosa, J. P.; Lawrence, J. A.; Sestacovschi, C.; Tidd-Johnson, A.; Rasool, M. H. U.; Gadamidi, V. K.; Ozair, S.; Pandav, K.; Cuevas-Lou, C.; Parrish, M.; Rodriguez, I.; and Fernandez, J. P. 2022. The impact of misinformation on the COVID-19 pandemic. *AIMS Public Health*, 9(2): 262–277.

Chaudhuri, N.; Gupta, G.; Bagherzadeh, M.; Daim, T.; and Yalcin, H. 2024. Misinformation on social platforms: A review and research Agenda. *Technology in Society*, 78: 102654.

Chen, M.; Chu, X.; and Subbalakshmi, K. 2021. Mmcovar: multimodal COVID-19 vaccine focused data repository for fake news detection and a baseline architecture for classification. In *ASONAM*.

Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; and Stoyanov, V. 2020. Unsupervised Cross-lingual Representation Learning at Scale. 8440–8451.

Crupi, G.; Mejova, Y.; Tizzani, M.; Paolotti, D.; and Panisson, A. 2022. Echoes through Time: Evolution of the Italian COVID-19 Vaccination Debate. *ICWSM*, 16(1): 102–113.

Cui, L.; and Lee, D. 2020. CoAID: COVID-19 Healthcare Misinformation Dataset. arXiv:2006.00885.

de Albuquerque, T. R.; Macedo, L. F. R.; de Oliveira, E. G.; Neto, M. L. R.; and de Menezes, I. R. A. 2022. Vaccination for COVID-19 in children: Denialism or misinformation? *J. Pediatr. Nurs.*, 64: 141–142.

Douze, M.; Guzhva, A.; Deng, C.; Johnson, J.; Szilvasy, G.; Mazaré, P.-E.; Lomeli, M.; Hosseini, L.; and Jégou, H. 2025. The faiss library. *IEEE Transactions on Big Data*.

Garimella, K.; Cintaqia, P.; Rojas-Constain, J. J.; Nayak, B. K.; and Vashistha, A. 2025. Global Patterns of Viral Content on WhatsApp. *ICWSM*, 19: 586–601.

Giovanni, M. D.; Pierri, F.; Torres-Lugo, C.; and Brambilla, M. 2022. VaccinEU: COVID-19 Vaccine Conversations on Twitter in French, German and Italian. *ICWSM*, 16(1): 1236–1244.

Hayawi, K.; Shahriar, S.; Serhani, M. A.; Taleb, I.; and Mathew, S. S. 2022. ANTi-Vax: a novel Twitter dataset for COVID-19 vaccine misinformation detection. *Public Health*, 203: 23–30.

Hossain, T.; Logan IV, R. L.; Ugarte, A.; Matsubara, Y.; Young, S.; and Singh, S. 2020. COVIDLies: Detecting COVID-19 Misinformation on Social Media. In *W. on NLP for COVID-19 at EMNLP*.

Huang, Q.; Lei, S.; and Ni, B. 2022. Perceived Information Overload and Unverified Information Sharing on WeChat Amid the COVID-19 Pandemic: A Moderated Mediation Model of Anxiety and Perceived Herd. *Frontiers in Psychology*, Volume 13 - 2022.

Kowsari, K.; Jafari Meimandi, K.; Heidarysafa, M.; Mendu, S.; Barnes, L.; and Brown, D. 2019. Text Classification Algorithms: A Survey. *Information*, 10(4): 150.

Landis, J. R.; and Koch, G. G. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1): 159–174.

Liu, A.; Mei, A.; Lin, B.; Xue, B.; Wang, B.; Xu, B.; Wu, B.; Zhang, B.; Lin, C.; Dong, C.; et al. 2025. Deepseek-v3. 2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*.

Loomba, S.; de Figueiredo, A.; Piatek, S. J.; de Graaf, K.; and Larson, H. J. 2021. Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nature Human Behaviour*, 5(3): 337–348.

Martins, A.; Cabral, L.; Mourão, P. J.; Monteiro, J.; and Machado, J. 2021. Detection of Misinformation about COVID-19 in Brazilian Portuguese WhatsApp Messages Using Deep Learning. In *SBBD*.

Mauk, M.; and Grömping, M. 2024. Online Disinformation Predicts Inaccurate Beliefs About Election Fairness Among Both Winners and Losers. *Comparative Political Studies*, 57(6): 965–998.

Maweu, J. M. 2019. "Fake Elections"? Cyber Propaganda, Disinformation and the 2017 General Elections in Kenya. *African Journalism Studies*, 40(4): 62–76.

Mazza Zago, R.; and Agnoletti dos Santos Pedotti, L. 2024. BERTugues: A Novel BERT Transformer Model Pre-trained for Brazilian Portuguese. *Semina: Ciências Exatas e Tecn.*, 45.

Melo, P.; Messias, J.; Resende, G.; Garimella, K.; Almeida, J.; and Benevenuto, F. 2019a. WhatsApp Monitor: A Fact-Checking System for WhatsApp. *ICWSM*, 13(01): 676–677.

Melo, P.; Vieira, C. C.; Garimella, K.; de Melo, P. O. V.; and Benevenuto, F. 2019b. Can WhatsApp Counter Misinformation by Limiting Message Forwarding? In *Complex Networks*, 372–384.

Mu, Y.; Jin, M.; Grimshaw, C.; Scarton, C.; Bontcheva, K.; and Song, X. 2023. VaxxHesitancy: A Dataset for Studying Hesitancy towards COVID-19 Vaccination on Twitter. *ICWSM*, 17(1): 1052–1062.

Natalie Cancian, Renato Machado. 2021. Anvisa Approves Definitive Registration of Pfizer Vaccine against Covid. Folha de S.Paulo, https://folha.com/e0cxpbns. Accessed on 01/09/2026.

Olivetti, J.; Bomfim, A.; Oliveira, M.; Marques, H.; Avelar, J.; Reis, J.; and Melo, P. 2025. O Fenômeno da Desinformação no WhatsApp, Telegram e Outras Plataformas de Mensagens Instantâneas: Uma Revisão Sistemática da Literatura (in Portuguese). In *WebMedia*.

Poddar, S.; Mondal, M.; Misra, J.; Ganguly, N.; and Ghosh, S. 2022a. Winds of Change: Impact of COVID-19 on Vaccine-Related Opinions of Twitter Users. *ICWSM*, 16(1): 782–793.

Poddar, S.; Samad, A. M.; Mukherjee, R.; Ganguly, N.; and Ghosh, S. 2022b. CAVES: A dataset to facilitate Explainable Classification and Summarization of Concerns towards COVID Vaccines. In *SIGIR*.

Raschka, S. 2020. Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning. arXiv:1811.12808.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *EMNLP*.

Reis, J. C.; Melo, P.; Belém, F.; Murai, F.; Almeida, J. M.; and Benevenuto, F. 2023. Helping fact-checkers identify fake news stories shared through images on whatsapp. In *WebMedia*.

Reis, J. C.; Melo, P.; Garimella, K.; and Benevenuto, F. 2020a. Can WhatsApp benefit from debunked fact-checked stories to reduce misinformation? *Harvard Kennedy School Misinformation Review*.

Reis, J. C. S.; Melo, P.; Garimella, K.; Almeida, J. M.; Eckles, D.; and Benevenuto, F. 2020b. A Dataset of Fact-Checked Images Shared on WhatsApp During the Brazilian and Indian Elections. *ICWSM*, 14(1): 903–908.

Resende, G.; Melo, P.; Sousa, H.; Messias, J.; Vasconcelos, M.; Almeida, J.; and Benevenuto, F. 2019. (Mis)Information Dissemination in WhatsApp: Gathering, Analyzing and Countermeasures. In *WWW*.

Ricardo Brito. 2021. Brazil health regulator approves Pfizer COVID-19 shot for ages 5 to 11. Reuters, https://www.reuters.com/world/americas/brazil-health-regulator-approves-pfizer-covid-19-shot-ages-5-11-2021-12-16/.

Schneider, E. T. R.; de Souza, J. V. A.; Knafou, J.; Oliveira, L. E. S. e.; Copara, J.; Gumiel, Y. B.; Oliveira, L. F. A. d.; Paraiso, E. C.; Teodoro, D.; and Barra, C. M. C. M. 2020. BioBERTpt - A Portuguese Neural Language Model for Clinical Named Entity Recognition. In *Clinical NLP Workshop*.

Sharma, A. E.; Khosla, K.; Potharaju, K.; Mukherjea, A.; and Sarkar, U. 2023. COVID-19-associated misinformation across the South Asian diaspora: Qualitative study of WhatsApp messages. *JMIR Infodemiology*, 3(1): e38607.

Simões, E. 2021. Brazil on alert after third case of Omicron variant. Reuters, https://www.reuters.com/world/americas/brazil-alert-after-third-case-omicron-variant-2021-12-01/.

Skafle, I.; Nordahl-Hansen, A.; Quintana, D. S.; Wynn, R.; and Gabarron, E. 2022. Misinformation About COVID-19 Vaccines on Social Media: Rapid Review. *J Med Internet Res*, 24(8): e37367.

Souza, F.; Nogueira, R.; and Lotufo, R. 2020. BERTimbau: pretrained BERT models for Brazilian Portuguese. In *BRACIS*.

Storm, B. C.; Stone, S. M.; and Benjamin, A. S. 2017. Using the Internet to access information inflates future use of the Internet to access other information. *Memory*, 25(6): 717–723.

Suarez-Lledo, V.; and Alvarez-Galvez, J. 2021. Prevalence of Health Misinformation on Social Media: Systematic Review. *J Med Internet Res*, 23(1): e17187.

Team, Q. 2025. Qwen3 Technical Report. arXiv:2505.09388.

Vieira, B. 2021. 'I Always Want to Do Good for My Race', Says The First Vaccinated against Covid in Brazil. Folha de S.Paulo, https://folha.com/o8t1fdrg. Accessed on 01/09/2026.

Vijaykumar, S.; Rogerson, D. T.; Jin, Y.; and de Oliveira Costa, M. S. 2021. Dynamics of social corrections to peers sharing COVID-19 misinformation on WhatsApp in Brazil. *Journal of the American Medical Informatics Association*, 29(1): 33–42.

Weinzierl, M. A.; and Harabagiu, S. M. 2021. Automatic detection of COVID-19 vaccine misinformation with graph link prediction. *Journal of Biomedical Informatics*, 124: 103955.

Weinzierl, M. A.; and Harabagiu, S. M. 2022. From Hesitancy Framings to Vaccine Hesitancy Profiles: A Journey of Stance, Ontological Commitments and Moral Foundations. *ICWSM*, 16(1): 1087–1097.

World Health Organization; and UNICEF. 2025. Global childhood vaccination coverage holds steady, yet over 14 million infants remain unvaccinated. https://www.who.int/news/item/15-07-2025-global-childhood-vaccination-coverage-holds-steady-yet-over-14-million-infants-remain-unvaccinated-who-unicef.

Zarei, M. R.; Christensen, M.; Everts, S.; and Komeili, M. 2023. Vax-Culture: A Dataset for Studying Vaccine Discourse on Twitter. In *IJCNN*.

Zhao, S.; Hu, S.; Zhou, X.; Song, S.; Wang, Q.; Zheng, H.; Zhang, Y.; and Hou, Z. 2023. The Prevalence, Features, Influencing Factors, and Solutions for COVID-19 Vaccine Misinformation: Systematic Review. *JMIR Public Health Surveill*, 9: e40201.

Ethics Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes!
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes,the claims made in the abstract and introduction accurately reflect the scope, methodology, and contributions of the paper.
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, the methodology section clearly explains why expert medical annotation, semantic deduplication, and comparative modeling under data-scarcity conditions are appropriate to support the paper's claims about misinformation characterization and classification in WhatsApp environments
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, We explicitly acknowledge that the dataset is derived from Brazilian public WhatsApp groups and is not representative of the entire WhatsApp population. Potential biases related to politically engaged groups, temporal context, and platform-specific dynamics are discussed in the Dataset Description and Limitations sections.
   (e) Did you describe the limitations of your work? Yes, a dedicated Limitations section discusses constraints related to data representativeness, annotation subjectivity, platform-specific affordances, and generalizability across countries, time periods, and private conversations.
   (f) Did you discuss any potential negative societal impacts of your work? Yes, we discuss the risks associated with misclassification in health misinformation detection, including false positives and false negatives, and emphasize that the models are intended for research and decision-support purposes rather than autonomous moderation or enforcement
   (g) Did you discuss any potential misuse of your work? Yes, we acknowledge that automated misinformation classifiers could be misused for over-censorship or unjustified content suppression if deployed without human oversight. The paper explicitly frames the dataset and models as tools for research, auditing, and public health support rather than direct punitive action.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, all messages were de-identified prior to analysis, no personally identifiable information is released, and the dataset is intended for controlled research
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes, We have carefully reviewed the ICWSM ethics guidelines and ensured that the study conforms to their requirements.
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? NA
   (b) Have you provided justifications for all theoretical results? NA
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA
   (e) Did you address potential biases or limitations in your theoretical framework? NA
   (f) Have you related your theoretical results to the existing literature in social science? NA
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes, the experimental setup, data splits, evaluation protocol, and prompts are fully described in the paper. The dataset and codes as cited in the text is available in the Zenodo dataset sharing service with a unique digital object identifier (DOI: 10.5281/zenodo.18165547).
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)?
   Yes, we describe the cross-validation strategy, number of repetitions, evaluation metrics, embedding strategies, and hyperparameter optimization procedures.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? Yes, All reported results include mean and standard deviation across repeated cross-validation runs.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? No, detailed hardware specifications are not emphasized, as the focus is on comparative performance rather than efficiency benchmarking
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made?
   Yes, the evaluation design directly reflects the study's goals, comparing classical models, fine-tuned SLMs, and in-context LLMs under identical, realistic data-scarcity conditions

(f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Yes, we analyze precision–recall trade-offs and explicitly discuss the implications of false positives and false negatives in health misinformation contexts

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...

(a) If your work uses existing assets, did you cite the creators? Yes, all datasets, models, and tools used are properly cited.

(b) Did you mention the license of the assets? Yes, licensing is discussed where applicable, and full details will be included in the dataset release documentation.

(c) Did you include any new assets in the supplemental material or as a URL? Yes, the dataset will be released after acceptance, following responsible data-sharing practices.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes, the data originates from publicly accessible WhatsApp groups, consistent with prior work, and was processed under approved ethical protocols without direct interaction with participants.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes, we explicitly state that all personally identifiable information was removed and acknowledge that the dataset may contain offensive or emotionally charged content as part of real-world misinformation discourse

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? Yes, the dataset is designed to be findable, accessible for research purposes, interoperable through standard formats, and reusable with appropriate documentation

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset ? No

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? Yes, annotation guidelines and procedures are described in detail.

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? Yes, the study was reviewed and approved by an independent Research Ethics Committee (IRB/CEP equivalent), as stated in the Ethics Statement

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? No, annotators were medical professionals and also authors participating as experts, not crowdworkers.

(d) Did you discuss how data is stored, shared, and deidentified? Yes, data storage, anonymization, and controlled sharing procedures are described in the Ethics Statement