Latin American Data in Science, v.5, n.1
datainscience.com.br
ISSN: 2763-9290

FactPolCheckBr: a dataset of fake news fact-checked during the 2022 Brazilian presidential elections

Recebido :30/05/25|Aceito: 23/10/25 | Publicado: 09/12/25
https://doi.org/10.53805/lads.v5i1.76

Sylvia Iasulaitis[1], Eloize R. M. Seno[2], Mariana C. de Souza[3], Alan D. B. Valejo[4], Isabella Vicari[5], Ian Victor R. Ruiz[5], Yanni Marcela Gameiro[1], Eanes T. Pereira[6], Guilherme Henrique Messias[4], Bruno C. Greco[4], Rafaela de A. B. Silva[6]

ABSTRACT

The use of social media and instant messaging applications in contemporary society has amplified the dissemination of misleading content, reaching audiences on an unprecedented scale. Many news outlets have made considerable efforts to check the accuracy of online content, especially during election periods, which are critical moments for spreading fake news. However, the fake news verification process is labor-intensive for humans, given the volume and speed at which misinformation circulates. Numerous studies in natural language processing and machine learning have emerged in recent years seeking to investigate and develop computational models capable of detecting fake news. Algorithm training is primarily based on supervised machine learning, which relies on labeled datasets to learn the characteristic patterns of misinformation. Labeled fake news datasets in Brazilian Portuguese are scarce. This research addresses this gap developing the first fact-checked fake news dataset related to the 2022 presidential elections in Brazil, which was widely regarded as the most polarized in the country's political history and marked by a large-scale disinformation campaign. The dataset, called FactPolCheckBr, includes 1,873 news items categorized as fake news, which were manually collected from online fact-checking platforms. The full texts of the fake news items were subsequently retrieved from the web using a scraping algorithm. Next, a clustering algorithm was applied to group similar news items, which enabled the identification of the main topics targeted by fake news during the elections. Each news item in the dataset also includes information on the candidate favored by the misinformation in that electoral context. The information was provided by political scientists who employed content analysis to examine the news texts carefully. This article presents an exploratory study of the FactPolCheckBr dataset, highlighting its key features and potential applications across various domains.

Keywords: Curated database; Fake news; Fact-checking; Elections; Machine learning..

PRIOR PUBLICATIONS

SILVA, R. de A. B.; PEREIRA, E. T.; IASULAITIS, S. Automatic detection of fake news in Tweets about the Elections 2022 Brazilians. AtoZ, 2025.

[1] Federal University of São Carlos, Department of Social Sciences, São Paulo, Brazil. si@ufscar.br
[2] Federal Institute of São Paulo, Computing Area, São Paulo, Brazil.
[3] Federal University of Mato Grosso do Sul, School of Computing, Mato Grosso do Sul, Brazil.
[4] Federal University of São Carlos, Department of Computer Science, São Paulo, Brazil.
[5] Federal University of São Carlos, Postgraduate Program in Science, Technology and Society, São Paulo, Brazil.
[6] Federal University of Campina Grande, Academic Unit of Systems and Computing, Paraíba, Brazil.

MATSUDA, W. T. M.; CASELI, H. de M.; VALEJO, A. D. B.; IASULAITIS, S. Câmaras de eco políticas durante os atos antidemocráticos: topologia de interação no Twitter/X. P2P & INOVAÇÃO, Rio de Janeiro, v. 11, n. 2, p. 1-29, e-7398, jan./jun. 2025. DOI: https://doi.org/10.21728/p2p.2025v11n2e-7398

VICARI, I. A urna eletrônica brasileira: entre controvérsias e desinformação. 2024. Dissertação (Mestrado em Ciência, Tecnologia e Sociedade) – Universidade Federal de São Carlos, São Carlos, 2024. Disponível em: https://repositorio.ufscar.br/handle/20.500.14289/19404.

CAPELLARO, L. ToMAS: Sumarização abstrativa multinível baseada em tópicos usando LLMs. Dissertação de Mestrado - Programa de Pós-graduação em Ciência da Computação da UFSCar. São Carlos, SP. Disponível em: https://repositorio.ufscar.br/handle/20.500.14289/21352

DATA IMPORTANCE

- In the context of machine learning, deep learning and natural language processing, the dataset can serve the following purposes (VASWANI, A. et al., 2017; KUNTUR, S. et al., 2024):
    - (i) Development of predictive models: The dataset can be used to train computational models that accurately identify fake news or to enhance the performance of existing models by allowing them to generalize the patterns found in misinformation.
    - (ii) Fine-tuning of deep learning models: It supports the adjustment of internal parameters in transformer-based models (e.g., BERT) specifically for fake news detection.
    - (iii) Development of transfer learning models: A model trained on this political dataset can be adapted to identify and classify fake news in other domains, which is particularly valuable in contexts where curated domain-specific fake news datasets are scarce.
    - (iv) Development of few-shot approaches: In prompt-based generative models (LLMs), which take a task description as input and generate outputs accordingly, the inclusion of fake news examples can serve as a conditioning mechanism to guide model behavior.
    - (v) Development of retrieval-augmented generation (RAG) approaches: In RAG, document retrieval is combined with text generation. The inclusion of fake news into the pipeline can help train more effective discriminators, thereby improving the system's ability to distinguish between reliable and misleading information.
- In terms of social and political relevance, this dataset stands out for:
    - (vi) Providing contextualized sociopolitical artifacts for researchers across various disciplines, as well as for legal experts and policymakers. It enables analyses of disinformation dissemination patterns, their influence on social behavior, disinformation strategies, and other related phenomena.

MATERIAL AND METHODS

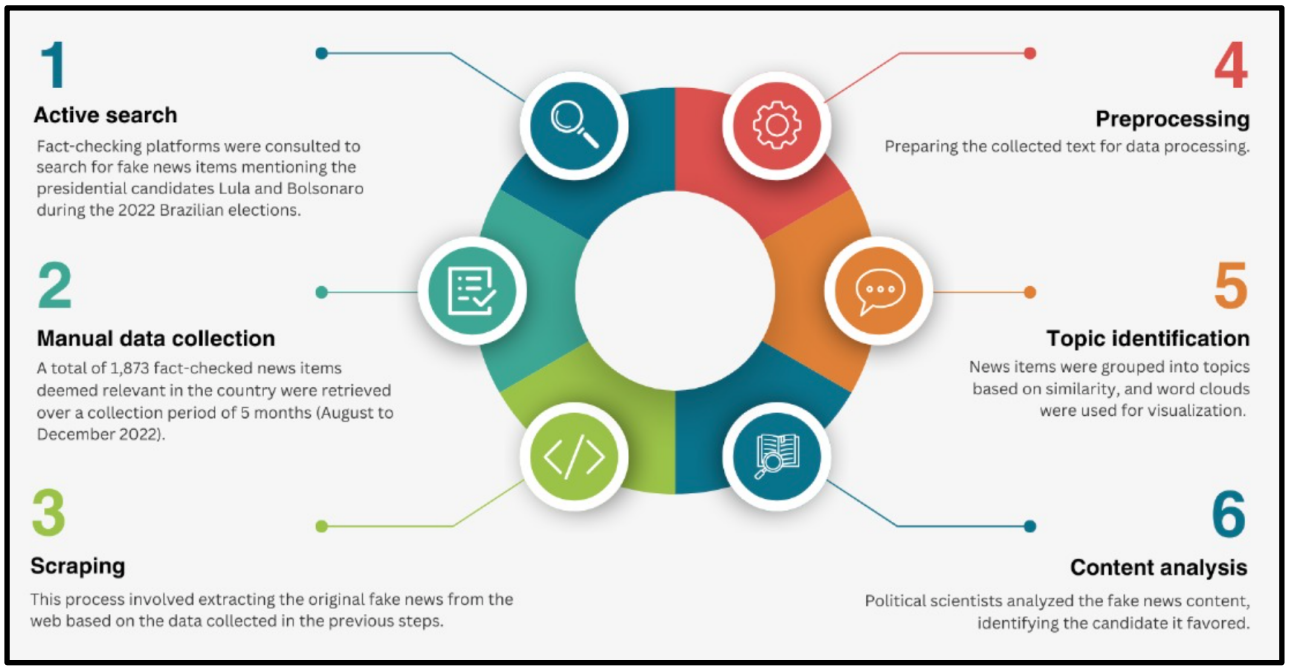
The fake news fact-checked dataset FactPolCheckBr was created by gathering verified news from online fact-checking platforms. These platforms follow a rigorous and methodical process to determine the veracity or falsehood of news stories, public statements, and other content shared online and on social media. The process typically involves selecting information circulating on social media, identifying the source and context in which the content was published, consulting official data, public documents, and subject-matter experts, and critically analyzing the content based on available evidence to assess its accuracy. A verification report is produced after

the analysis, explaining the findings and citing the criteria and sources used. The analysis is then published, and the verification may be updated if new evidence emerges.

As we can see in Figure 1, the first step of the dataset construction involved manually searching for fact-checked news items related to the then-presidential candidates Luiz Inácio Lula da Silva (left-wing) and Jair Messias Bolsonaro (far-right) during the 2022 Brazilian presidential election. These candidates were selected because they advanced to the second round of voting and were the primary subjects of circulating fake news.

The search (Step 2) was conducted manually across ten highly relevant Brazilian online fact-checking platforms, as listed in Table 1. All news items verified by the platforms and published in their politics sections between August 1, 2022, and December 1, 2022, involving the two candidates and classified as fake news were collected. The time frame was selected based on the official calendar of the Electoral Court, which designates August as the official start of the election campaign period, including authorization for online campaigning. Additionally, the data collection period was extended through November due to the continued circulation of misinformation following the announcement of the second-round results on October 30. The post-election period was marked by intense online activity, with numerous misleading publications reverberating across social media throughout November. In total, 1,873 verified fake news items were collected. Table 1 details the number of news items retrieved from each fact-checking platform.

Figure 1.- Main steps in the construction of the FactPolCheckBr dataset.



The texts published by fact-checking platforms usually do not preserve the original structure of the fake news, as they incorporate information from the verification process used to determine content accuracy. Moreover, these texts are significantly longer than the original publications.

Therefore, Step 3 (Fig. 1) involved extracting the original fake news content from the web through a web scraping process, using the titles of the fact-check reports published on the platforms as search queries. The step was performed

automatically using the Python libraries Requests[1] and BeautifulSoup[2].

To identify the main topics associated with the fake news included in the dataset, each news item underwent preprocessing (Step 4), which involved removing HTML tags, punctuation marks, and stopwords, followed by tokenization. These operations were performed using the Python NLTK library[3]. Subsequently, vector embeddings[4] were generated using the Word2Vec model available in the Python Gensimx library[5]. With these vector embeddings, the Affinity Propagation (AP) clustering algorithm[6], was applied using the following parameters: affinity Euclidean, convergence_iter 10, damping 0.7.

Table 1.- Fact-checking platforms used to gather the dataset and the number of fact-checked news items retrieved.

Fact-checking platform	**Description of the fact-checking agency**	**Number of fact-checked fake news items**
Projeto Comprova https://projetocomprova.com.br/	*Projeto Comprova* brings together several media outlets to verify information that has been widely circulated online since June 2018.	175
AFP Checamos https://checamos.afp.com/	*AFP Checamos* was founded in 2017 and is a department of the French news agency Agence France-Presse.	196
E-farsas https://www.e-farsas.com/	The website *E-farsas* has been investigating rumors on the internet since 2002.	50
CNJ https://www.cnj.jus.br/	Website of the Brazilian National Council of Justice.	01
Fato ou Fake https://g1.globo.com/fato-ou-fake/	*Fato ou Fake* is a fact-checking service from Globo Group launched in 2018.	171
Lupa https://lupa.uol.com.br/	*Lupa* is a fact-checking and media education agency founded in 2015.	245
Boatos.org https://www.boatos.org/	*Boatos.org* is an initiative created in June 2013 that brings together several journalists committed to fact-checking.	313
Aos Fatos https://www.aosfatos.org/	*Aos Fatos* is an independent fact-checking agency created in 2015. It was inspired by the initiatives Chequeado, from Argentina, and PolitiFact, from the United States.	310
UOL Confere https://noticias.uol.com.br/confere/	*UOL Confere* is a UOL Group fact-checking service.	228
Fato ou Boato https://www.justicaeleitoral.jus.br/fato-ou-boato/	*Fato ou Boato* is a platform conceived in 2020 as part of the Brazilian Superior Electoral Court's (TSE) Disinformation Combat Program.	184
TOTAL		1,873

[1] https://pypi.org/project/requests/
[2] https://pypi.org/project/beautifulsoup4/
[3] https://www.nltk.org/
[4] Vector embeddings are numerical representations of words that capture their meanings and relationships derived from texts.
[5] https://pypi.org/project/gensim/
[6] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AffinityPropagation.html

In Step 5, the AP algorithm grouped news items based on similarity using the generated embeddings. Topic identification was performed by calculating the centroid of each cluster, using the mean of the vector embeddings. The topics were then visualized through word clouds, where the largest words represent terms that are closest to the group's centroid. This procedure was developed and implemented in the study by Silva, Pereira, and Iasulaitis (2025) to support the creation of fake news labeling functions.

Finally, in Step 6, political scientists analyzed the fake news content and categorized it using content analysis, identifying which candidate each item favored within the electoral context. As proposed by Bardin (2011), the method involves a systematic and objective set of procedures for describing and categorizing textual content, allowing for the inference of relevant information from the analyzed data. In this case, political scientists manually examined the fact-checked version of each news story and classified it based on whether it favored Luiz Inácio Lula da Silva or Jair Messias Bolsonaro. This classification took into account both the content's implicit intent and its potential impact on the candidate's public image.

DATA DESCRIPTION

Dataset

The database consists of 11 .csv files in total. The first 10 files contain the following attributes: the title of the fact-check, the publication date, the classification of the news item, the candidate favored by the news, the platform that published the fact-check, and the URL of the publication containing the full analysis with evidence-based explanations, as illustrated in the example shown in Table 2. Each of these 10 files corresponds to a different fact-checking agency (with exception to the file attributed to the "CNJ" agency, which shared links to fact-checks, performed by other agencies, that were collected during the study). The remaining .csv file joins the entries in the previously described files, and has one additional attribute, containing the text of the aforementioned analysis of each respective entry.

Figure 2 shows the distribution of fact-checks, per classification ("False", "Partly true" or 'True") performed by each agency, as present in our dataset; while Figure 3 displays the overall distribution of fact-check instances in our dataset, inside a period of time, delimited by the dates of the oldest and newest fact checks collected during the study. The colors denote the distribution of fact-checks, belonging to each bucket inside said period, per agency that performed the fact-checks.

The database includes fact-checks of fake news disseminated in textual, image, multimedia, multimodal, and deepfake formats. The textual format refers to written content, while the image format involves manipulated or out-of-context images. The multimedia format combines elements such as video, audio, and images to reinforce the misleading message. The multimodal format integrates various modes of communication — text, images, emojis, and others — that work in a complementary manner. The deepfake format employs artificial intelligence to generate realistic videos or audio recordings that simulate individuals saying or doing things that never occurred.

Table 2.- Example of fake news from the database.

Fact-check title: *É #FAKE print em que Lula fala em atacar igrejas evangélicas, aprovar leis pró-aborto e ideologia de gênero.* [The screenshot where Lula allegedly discusses attacking evangelical churches, endorsing pro-abortion laws, and promoting gender ideology is #FAKE.]
Original news item: Lula said: "We have to confront the growth of evangelical churches that insist on getting involved in politics. Everyone knows that churches should not get involved in politics. We will increase taxes on churches, making their existence unviable. We will also approve laws in favor of abortion and gender ideology in schools. Children should be able to be whoever they want to be. If a boy wants to wear a dress or a girl wants to wear boys' clothes, it is not the parents who should get in the way of these children's desires."
Date of the verification: August 03, 2022
Classification of the news item: Fake news
Fact-checking agency: Fato ou Fake | Portal G1
Link: https://g1.globo.com/fato-ou-fake/noticia/2022/08/03/e-fake-print-em-que-lula-fala-em-atacar-igrejas-evangelicas-aprovar-leis-pro-aborto-e-ideologia-de-genero.ghtml
Result of the fact-checking process: An image (Fig. 4) disseminated on social media shows a screenshot of an alleged post by former president Luiz Inácio Lula da Silva, in which he purportedly expresses an intention to attack evangelical churches, increase taxes on them until they become unviable, and approve laws supporting abortion and the teaching of "gender ideology" in schools. This claim is #FAKE. According to Lula's advisors, such a post never existed, and the image is fabricated. They issued the following statement: "The former president [Lula was president for two mandates, from 2003 to 2011] never said, posted, or even thought this. On the contrary, he sanctioned the Religious Freedom Law when he was president. We regret that Bolsonarism [refers to a political movement tied to Jair Bolsonaro] resorts to daily lies against Lula to cover up the disastrous Bolsonaro administration and to deceive the public for political gain." A search of the former president's official social media profiles reveals no record of the statement in the screenshot. Such a sensational declaration would not go unnoticed by the professional media. Yet, searches on major news platforms and search engines yield no results confirming the existence of such content. Lula has publicly stated that he is personally against abortion and believes the issue should be addressed as a matter of public health. *Fato ou Fake* has previously debunked other false claims aimed at creating conflict between Lula and evangelical communities. The term "gender ideology" is not recognized in academic discourse. It is commonly used by conservative groups, including evangelical churches, to oppose gender studies, which started in the United States and Europe during the 1960s and 1970s, that explore the distinction between biological sex and gender. These studies propose that being a man or a woman is not solely determined by genitalia or chromosomes, but also by social and cultural factors acquired throughout life. Conservative groups argue that the conclusions of gender studies lack validation from the exact and biological sciences. However, Brazil's Federal Supreme Court (STF), as guardian of the Constitution, has consistently ruled unconstitutional laws that attempt to ban what is referred to as "gender ideology".

Figure 2.- Category of News by Agency.

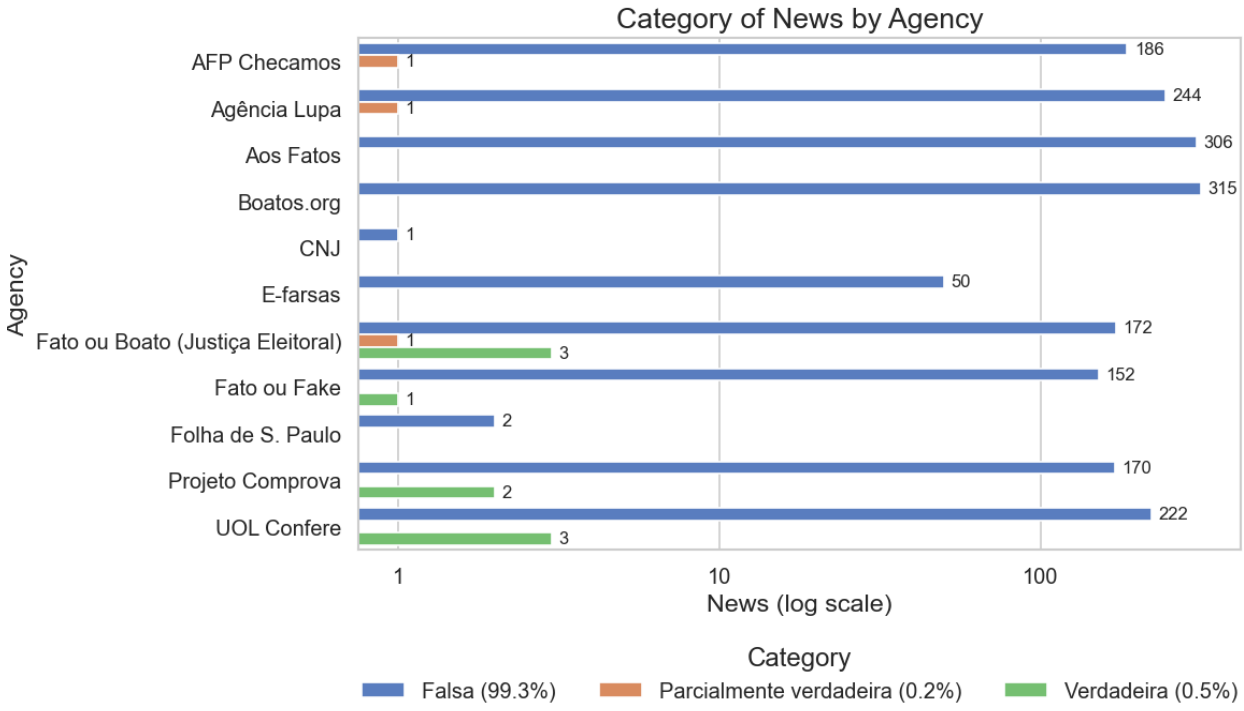


Figure 3. Distribution of Fact-checks by Date and Agency.

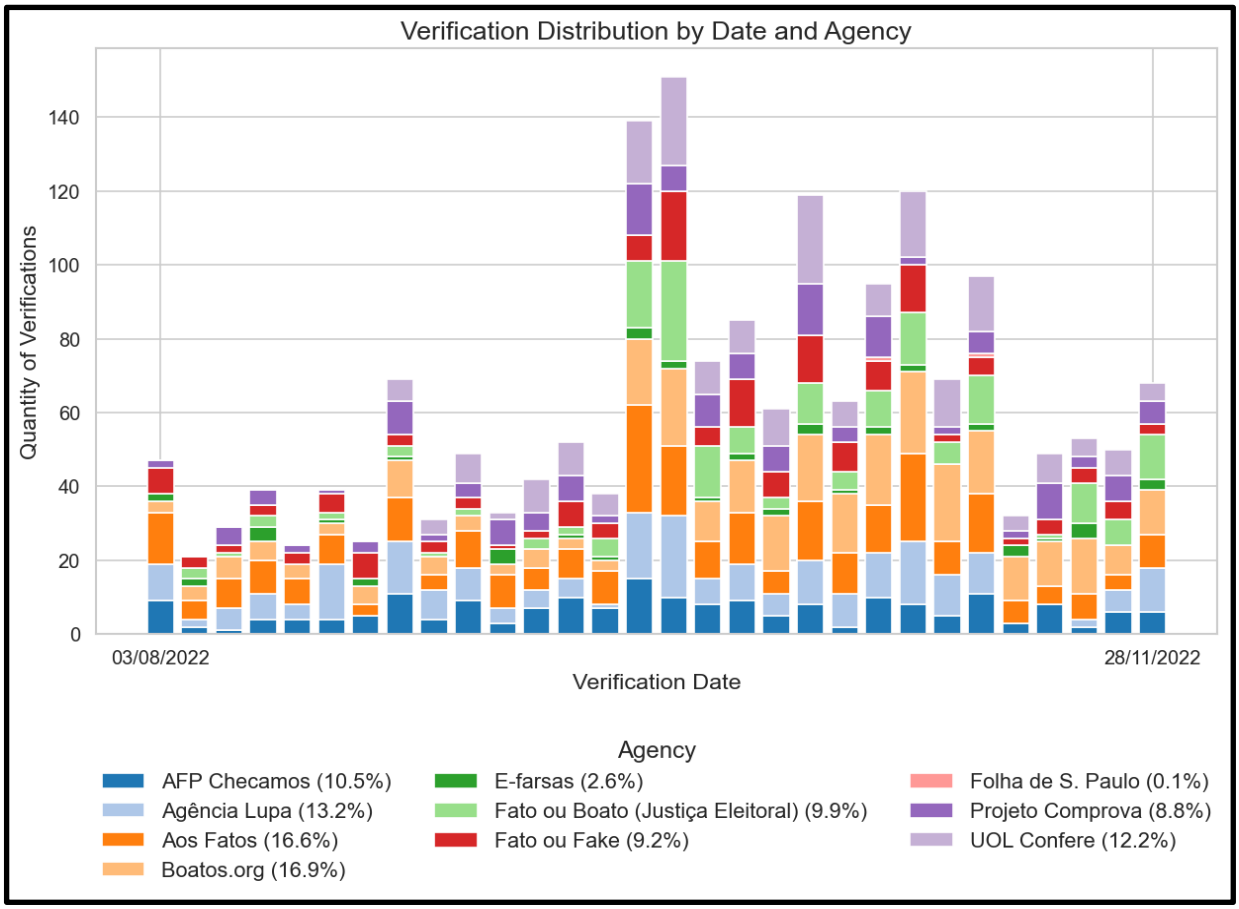


Figure 4.- Image reporting fake news

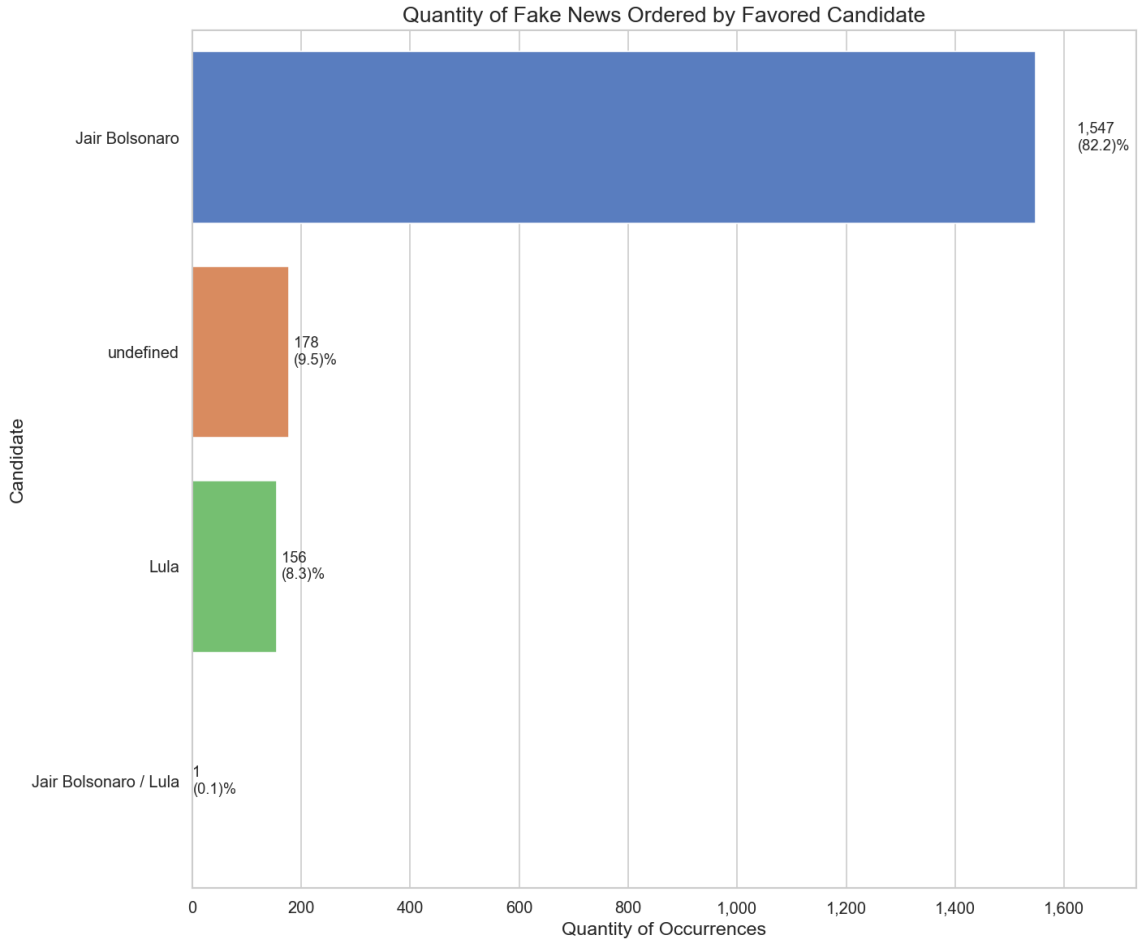
Source: Fato ou Fake | G1

The main themes characterizing the fake news agenda during the campaigns included (Fig. 5): moral and religious values (e.g., abortion, persecution of churches, gender ideology, Satanism, Freemasonry, cannibalism), economic issues (e.g., the extinction of PIX [real-time instant payment system], privatization of state-owned enterprises, reduction of the minimum wage, and dismantling of redistributive policies), attacks on institutional legitimacy (including allegations of fraud in electronic voting machines), false endorsements by public figures, associations between political opponents and organized crime, and scientific denialism, particularly regarding vaccines.

Figure 5.- Word cloud of fact-checking titles

The content analysis demonstrated that the majority of fake news (82,2%) benefited the presidential candidate Jair Bolsonaro, as observed in Figure 6.

Figure 6.- Quantity of fake news ordered by favored candidate.



SUPPLEMENTARY MATERIALS

Repository name: GitHub
DOI of the dataset (when available): 10.5281/zenodo.15284980
Link to access the data: https://github.com/Interfaces-UFSCAR/Dataset-FactPolCheckBr

ACKNOWLEDGEMENTS

This work was funded by FAPESP under grant number 2022/03090-0, coordinated by Prof. Dr. Sylvia Iasulaitis.

This research was supported by the Coordination for the Improvement of Higher Education Personnel - Brazil (CAPES) - Finance Code 001, through the award of a master's scholarship and by the National Council for Scientific and Technological Development (CNPq) under grant number 420025/2023-5. The funders had no involvement in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

REFERENCES

BARDIN, L. Análise de conteúdo. 1. ed. São Paulo: Edições 70, 2011.

KUNTUR, S. et al. Under the Influence: A Survey of Large Language Models in Fake News Detection. IEEE Transactions on Artificial Intelligence, 2024.

VASWANI, A. et al. Attention is all you need. Advances in neural information processing systems, v. 30, 2017.