Detecting Misinformation on Telegram Anti-vaccine Communities

Athus Cavalini[1,2], **Thamya Donadia**[2], **Fábio Malini**[3], **Giovanni Comarela**[2]

[1]Núcleo de Informática
Instituto Federal do Espírito Santo (Ifes) – Alegre, ES – Brazil

[2]Programa de Pós-Graduação em Informática
Universidade Federal do Espírito Santo (Ufes) – Vitória, ES – Brazil

[3]Departamento de Comunicação Social
Universidade Federal do Espírito Santo (Ufes) – Vitória, ES – Brazil

athus.cavalini@ifes.edu.br, thamya.donadia@edu.ufes.br,
fabiomalini@gmail.com, gc@inf.ufes.br

Abstract. *Due to the substantial volume of misinformation regarding COVID-19 in Brazil, this paper proposes the application of machine learning methods to identify false or harmful information in anti-vaccine communities on Telegram. To this end, we developed a dataset of 1,500 messages labeled by experts according to three aspects: veracity level, potential for harm, and category. The labeling process achieved an agreement score of 81%. Experiments were conducted using state-of-the-art algorithms such as XGBoost, a BERT-based classifier, and a GPT-based classifier. The models trained on the labeled dataset achieved an F1-Score of 0.83 for detecting falsehood and 0.92 for potential harm, indicating their effectiveness in identifying misinformation in this context.*

1. Introduction

The COVID-19 pandemic has provoked the production of an unprecedented volume of information, leading the World Health Organization to characterize the situation as an "infodemic" [WHO 2022]. This includes false or misleading information about the virus itself, COVID-19 treatments, and public health measures like vaccinations. The spread of such misinformation potentially poses a significant threat to public health efforts, leading to vaccine hesitancy [Lee et al. 2022], decreased adherence to public health guidelines, and ultimately, increased morbidity and mortality [Kılıç et al. 2022]. In this context, Telegram has attracted considerable focus in deliberations concerning information disorder, as it has historically been used by deplatformed and extremist actors [Rogers 2020].

Especially in Brazil, anti-vaccine groups and channels already have a strong presence on Telegram, forming a community that shares an unprecedented volume and variety of misinformation, reaching an audience of millions [Malini et al. 2024]. This underscores the urgent need for effective tools to monitor and combat false information. To address this challenge, we propose a work based on two phases. Firstly, we introduce a novel, curated dataset of short text messages annotated by domain experts to identify false or harmful information related to COVID-19. The dataset comprises 1,500 messages sourced from Telegram channels and groups within the anti-vaccine community. Secondly, we trained and evaluated both classical and state-of-the-art machine learning algorithms for detecting misinformation in these ecosystems, aiming to provide a robust and scalable solution for real-time misinformation detection and intervention.

2. Related Work

Recently, many research efforts involving Telegram have been developed, especially due to its significant growth and interest from extremist communities, such as terrorist organizations [Weimann 2016], white supremacists [Guhl and Davey 2020], and neo-Nazi communities [Callum Jones and Robards 2024].

In this regard, various monitoring tools have been developed to capture and structure data from the platform for subsequent analysis. For instance, [Júnior et al. 2022] developed the "Telegram Monitor," and [Cavalini et al. 2023] presented the "Telegram Observatory," both of which are used for monitoring political communities.

Moreover, many researchers have focused on the topic of misinformation on social networks. Specifically in the context of messaging apps, [Machado et al. 2019] and [Gaglani et al. 2020] have developed frameworks for identifying misinformation in WhatsApp messages.

Despite these advancements, we did not find a comprehensive solution for identifying misinformation within specific communities, which propagates narratives that differ significantly from mainstream misinformation. This gap highlights the need for targeted approaches to effectively detect and mitigate the unique types of false information prevalent in these groups. In that way, this work aims to fill this gap by introducing a methodology that can be easily applied in other contexts to build tools to combat information disorder in specific contexts.

3. Methods

This section outlines the methodology for collecting and labeling the corpus, along with the training and evaluation of the classification models.

3.1. Data Collection

To obtain a representative collection of Telegram channels that compose the anti-vaccine community, we follow the work made by *Observa ICEPi*[1], a project that monitors online actors and discussions related to public health on social platforms. From its public data and following the methodology described in its technical report [ICEPi 2022], we were able to assemble a set of 779 active anti-vaccine Telegram channels and groups.

The collection of the messages from these chats was performed using a script that communicates directly with the official Telegram Database Library API[2], enabling the safe and reliable download of massive data. The main corpus was formed by 9.941.879 messages shared between March 2020 and August 2023.

After the collection, messages were filtered in order to maintain only those semantically relevant, removing those consisting of only links, emojis, or short sentences (in this case, fewer than 60 characters), typical of online conversations.

[1] https://icepi.es.gov.br/projeto-observa-icepi

[2] https://core.telegram.org/tdlib

Additionally, we limited the corpus to topics related to vaccines and/or the COVID-19 pandemic, gathering only the messages containing any of the following keywords (some stemmed): *covid, corona, virus, mrna, spike, vaccin, inject, sting, inoculat, adverse, collateral* and *early* (treatment)[3].

The final set was composed of 276,720 messages. Due to the impracticality of labeling all of them in a timely manner, the messages were sorted in descending order based on the number of forwards they received, prioritizing the labeling of content with the highest circulation (i.e., greater spread within the network).

3.2. Labeling Approach

The goal of this labeling process is to identify content that negatively impacts the informational process, especially through the spread of misinformation. In this work, we chose to consider the framework proposed by [Wardle and Derakhshan 2017], which introduces the concept of Information Disorder. The authors assess three different types of information that comprise Information Disorder (in addition to information itself) from the perspective of two dimensions: veracity and intent to harm.

Therefore, the dataset was labeled using a three-step approach. In the first step, the level of veracity was assigned, quantifying the degree of falsehood of the content in the text on a discrete scale from 1 to 5. This step required annotators to consult high-credibility scientific sources as well as information disseminated by governmental health entities. In the second step, the potential for harm was also defined on a scale from 1 to 5, indicating the likelihood of the content causing harm to the informational process. Finally, the messages were categorized into ten different themes associated with the context of interest: Anti-Vaccine Movement, COVID-19, Vaccine Adverse Reactions, Scientific Denialism, Political/Partisan Discourse, Conspiracy Theories, Xenophobia, Antisemitism, Religious Discourse, and Racism. These categories were defined by the annotators themselves based on an initial perception obtained from the preliminary analysis of the data.

The annotators, in turn, were carefully selected to ensure a rigorous and multidisciplinary evaluation. Thus, we selected specialists from three different areas: three from Applied Data Science, two from Public Health, and two from Social Communication. The process began with the requirement of at least 3 labelers per message, and this number was slightly reduced as the process progressed based on the observation of the high agreement score. The agreement between annotators was calculated using the Cosine Similarity between their evaluations. This metric was applied considering the scales assigned in steps 1 and 2, disregarding the labels assigned in the third stage.

3.3. Classification Models

For the misinformation detection task, we selected three methods: XGBoost, a classical and robust method for classification, along with BERT-based and GPT-based classifiers, both of which are considered state-of-the-art Large Language Models (LLMs).

At this phase, we used the average of labels received by each message to achieve a binary classification. Messages that scored greater than or equal to three (≥ 3) for "falsehood" received the label 1, while the others received the label 0. The same method was applied to "potential harm".

[3] In portuguese: *covid, corona, virus, mrna, spike, vacin, inje, picad, inocul, advers, colatera, precoce*

In all cases, text normalization was performed to mitigate the impact of abbreviations, spelling errors, and internet slang. The Enelvo library [Bertaglia and Nunes 2016] was used for this purpose. Emojis, URLs, and other special characters were also removed.

For XGBoost, we have applied a more comprehensive preprocessing: filtering stopwords and lemmatization, to reduce data dimensionality and retain only relevant words for analysis; transforming texts into feature vectors using TF-IDF [Robertson and Jones 1976]; and balancing classes in training set using Synthetic Minority Over-sampling Technique (SMOTE) which generates synthetic examples of the minority class. For BERT and GPT, no additional textual preprocessing was required, as these models are capable of handling text nuances and linguistic features effectively.

To ensure result reliability, XGBoost and BERT models underwent 30 repetitions. In each repetition, the dataset was randomly split into training (80%) and test (20%) sets, following good Machine Learning practices and ensuring that the test set was never used during training or hyperparameter tuning. For XGBoost, hyperparameter optimization was conducted using grid search and 5-fold cross-validation. For the BERT-based classification, we used the pre-trained model BERTimbau [Souza et al. 2020] with batch sizes of 16, four epochs, and a learning rate of $2e^{-5}$, as recommended by [Devlin et al. 2018].

For the GPT-based classification, we fine-tuned GPT-3.5[4]. In this process, the model is pre-trained[5] on a labeled dataset and subsequently instructed to classify new data. In this case, we split the dataset into training (70%), validation (20%), and test (10%). Considering the high cost of the process as it is a private model running in the OpenAI cloud, no repetitions were performed.

4. Results

4.1. Labeled Dataset

At the end of the process, 1,500 messages were labeled[6]. Figure 1 shows the cumulative distribution of the number of annotators who labeled each message. The process achieved an average agreement score between annotators of 82,30% for falsehood and 79.82% for potential harm. Considering an average score greater than or equal to three (≥ 3) as a "positive" label, 58% of the messages were labeled as false, while 81% were labeled as potentially harmful.

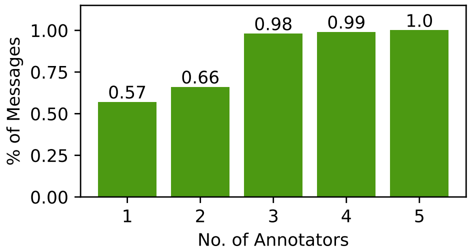


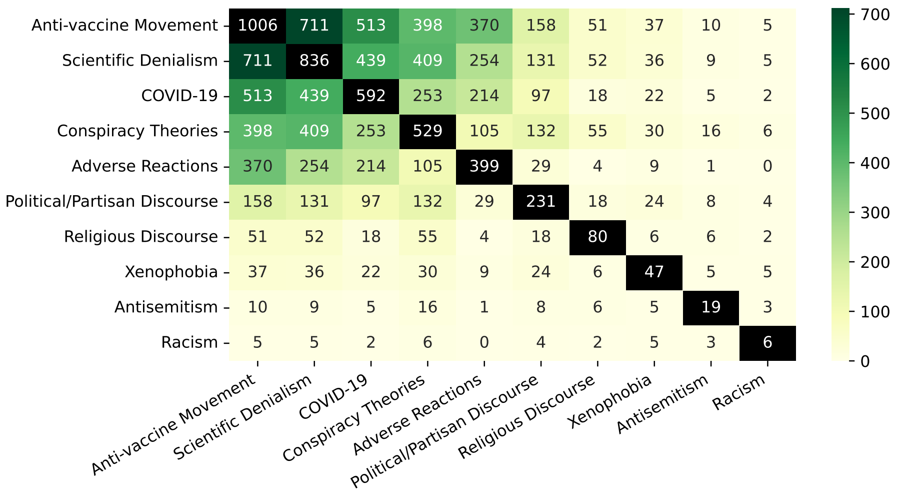
Figure 1. Labelers per message (cumulative distribution).

[4] https://openai.com/

[5] The prompt used to instruct the model was: "You are a model that classifies texts containing misinformation about COVID-19 in a binary manner, where 0 means true content and 1 means false content. Respond with 0 or 1, and nothing else." A similar prompt was used to the harmful classification.

[6] Available at `https://github.com/dsl-ufes/covax-br/`.

For category labeling, all received labels were considered. Figure 2 indicates the percentage of messages that received each label, as well as their co-occurrence, not only detailing the categorization results but also providing significant insights into the anti-vaccine community's topics and how they correlate.

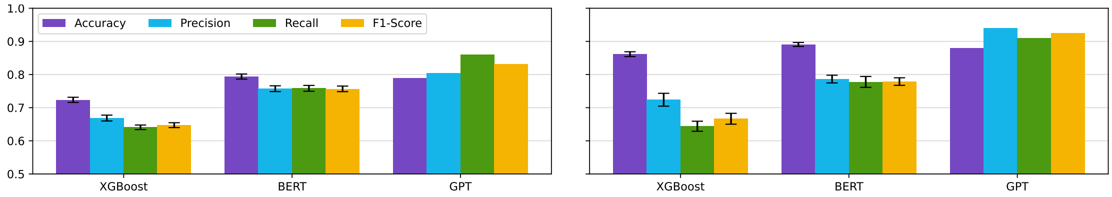


	Anti-vaccine Movement	Scientific Denialism	COVID-19	Conspiracy Theories	Adverse Reactions	Political/Partisan Discourse	Religious Discourse	Xenophobia	Antisemitism	Racism
Anti-vaccine Movement	1006	711	513	398	370	158	51	37	10	5
Scientific Denialism	711	836	439	409	254	131	52	36	9	5
COVID-19	513	439	592	253	214	97	18	22	5	2
Conspiracy Theories	398	409	253	529	105	132	55	30	16	6
Adverse Reactions	370	254	214	105	399	29	4	9	1	0
Political/Partisan Discourse	158	131	97	132	29	231	18	24	8	4
Religious Discourse	51	52	18	55	4	18	80	6	6	2
Xenophobia	37	36	22	30	9	24	6	47	5	5
Antisemitism	10	9	5	16	1	8	6	5	19	3
Racism	5	5	2	6	0	4	2	5	3	6

Figure 2. Categories co-occurrence.

4.2. Classification

All models were trained and evaluated using performance metrics including **Accuracy**, **Precision**, **Recall**, and **F1-Score**. The results showed that the GPT-based model demonstrated significantly better performance for predicting misinformation, surpassing other methods in most of the evaluated metrics. This may be due to its ability to capture complex contexts and semantic relationships in text, along with its pre-existing "knowledge".



(a) Regarding falsehood. **(b) Regarding potential harm.**

Figure 3. Average metrics per model.

For classifying potential harm, the models showed significantly better results, maintaining, in a linear manner, the same standard of performance between them. Figure 3 presents a comparative overview of the results, showing the average and standard error for the XGBoost and BERT methods, highlighting their consistency and reliability. The GPT-based model results are also displayed for comprehensive comparison.

5. Final Remarks

This paper presents a comprehensive approach to addressing COVID-19 misinformation in anti-vaccine communities on Telegram. Our contributions include the creation of a novel, expert-labeled dataset and the application and evaluation of advanced machine learning models for misinformation detection.

We developed a dataset of 1,500 messages from anti-vaccine channels from Telegram, a prominent platform for misinformation in Brazil, labeled by specialists from three distinct areas. The labeling process achieved an agreement score of 81%, demonstrating high-quality results and providing valuable insights about types of misinformation.

In addition to the dataset, we conducted extensive experiments using state-of-the-art machine learning models, including XGBoost, BERT, and GPT-3.5. Our results demonstrate that these models, especially the GPT-3.5-based classifier, are highly effective in detecting misinformation and assessing potential harm within the messages, achieving F1-Scores of 0.83 and 0.92, respectively.

These findings underscore the potential for developing real-time tools to detect and mitigate the spread of misinformation across various social media platforms. The experiments suggest that GPT-based approaches, in particular, can significantly enhance the identification of complex and nuanced misinformation, emphasizing the need to expand the application of GPTs in domain-specific tasks to achieve improved solutions.

Looking forward, it is crucial to address the scalability of our approach by expanding the labeled dataset and testing the models on data from other contexts and platforms. This will help to refine the models and ensure their generalization and adaptability. By doing so, we aim to create a universal methodology for combating misinformation that can be applied across different social media and messaging apps.

Moreover, future experiments and evaluations could leverage the multilabel categorization results to develop more sophisticated models that can handle multiple categories simultaneously, improving the precision of misinformation detection.

We anticipate that this work has the potential to benefit researchers, governments, and journalists in their efforts to combat misinformation, offering valuable insights and effective tools to address this growing challenge. It is important to note that this paper represents a work in progress, opening the door to further questions and investigations. Future research could explore the integration of these models with real-world applications, the ethical implications of automated misinformation detection, and the continuous adaptation of models to emerging misinformation trends.

Acknowledgments

This work was partially supported by: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) – Finance Code 001; FAPES – 368/2022 - P: 2022-NGKM5, 129/2021 - P: 2021-GL60J, 1026/2022 - P: 2022-ZQX6F; Fapesp/MCTI/CGI.br (#2020/ 05182-3); and Instituto Capixaba de Ensino, Pesquisa e Inovação em Saúde (ICEPi).

References

Bertaglia, T. F. C. and Nunes, M. d. G. V. (2016). Exploring word embeddings for unsupervised textual user-generated content normalization. In *2nd WNUT*, pages 112–120.

Callum Jones, S. R. and Robards, B. (2024). White Warriors and Weak Women: Identifying Central Discourses of Masculinity in Neo-Nazi Telegram Channels. *Studies in Conflict & Terrorism*, pages 1–26.

Cavalini, A., Malini, F., Gouveia, F., and Comarela, G. (2023). Politics and Disinformation: Analyzing the Use of Telegram's Information Disorder Network in Brazil for Political Mobilization. *First Monday*, 28(5).

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Gaglani, J., Gandhi, Y., Gogate, S., and Halbe, A. (2020). Unsupervised whatsapp fake news detection using semantic search. In *4th ICICCS*, pages 285–289.

Guhl, J. and Davey, J. (2020). A Safe Space to Hate: White Supremacist Mobilisation on Telegram. Institute for Strategic Dialogue (ISD). `https://www.isdglobal.org/isd-publications/a-safe-space-to-hate-white-supremacist-mobilisation-on-telegram/`.

ICEPi (2022). Metodologia de Coleta e Modelagem das Redes de Desordem Informacional do Telegram. Technical report, Instituto Capixaba de Ensino, Pesquisa e Inovacação em Saúde.

Júnior, M., Melo, P., Kansaon, D., Mafra, V., Sa, K., and Benevenuto, F. (2022). Telegram Monitor: Monitoring Brazilian Political Groups and Channels on Telegram. In *33rd ACM Conference on Hypertext and Social Media*, HT '22.

Kılıç, J., Yıldırım, M. S., Alakuş, O. F., Kiliç, D. K., Y., N. A., and Ebik, B. (2022). Vaccination hesitancy as a cause of covid-related mortality. *International Journal of Research in Medical Sciences*, 10(9):1833–1838.

Lee, S. K., Sun, J., Jang, S., and Connelly, S. (2022). Misinformation of COVID-19 Vaccines and Vaccine Hesitancy. *Scientific Reports*, 12(1):13681.

Machado, C., Kira, B., Narayanan, V., Kollanyi, B., and Howard, P. (2019). A study of misinformation in whatsapp groups with a focus on the brazilian presidential elections. WWW '19, page 1013–1019, New York, NY, USA.

Malini, F., Sodré, F., Cavalini, A., Herkenhoff, G., and Gouveia, F. (2024). Five patterns of vaccine misinformation on telegram. In *16th International Conference on Advances in Social Networks Analysis and Mining, ASONAM*, Italy. (to appear).

Robertson, S. and Jones, S. (1976). Relevance weighting of search terms. *Journal of the American Society for Information science*, 27:129–146.

Rogers, R. (2020). Deplatforming: Following extreme Internet celebrities to Telegram and alternative social media. *Media Studies*, 35(3):213–229.

Souza, F., Nogueira, R., and Lotufo, R. (2020). BERTimbau: Pretrained BERT Models for Brazilian Portuguese. In Cerri, R. and Prati, R. C., editors, *Intelligent Systems*, pages 403–417, Cham. Springer International Publishing.

Wardle, C. and Derakhshan, H. (2017). Information disorder: Toward an interdisciplinary framework for research and policy making. Council of Europe.

Weimann, G. (2016). Terrorist Migration to the Dark Web. *Perspectives on Terrorism*, 10(3):40–44. JSTOR, `http://www.jstor.org/stable/26297596`.

WHO (2022). Health topics: Infodemic. World Health Organization. `https://www.who.int/health-topics/infodemic`. Accessed on: May 9, 2023.