Modeling Vaccine Discourse and Polarization on YouTube in Brazil: A Semi-Supervised Stance Detection Approach

Geovana S. de Oliveira[1], **Carlos H. G. Ferreira**[2]

[1]Departamento de Computação e Sistemas – Universidade Federal de Ouro Preto
João Monlevade, MG

[2]Departamento de Computação – Universidade Federal de Ouro Preto
Ouro Preto, MG

geovana.so@aluno.ufop.edu.br, chgferreira@ufop.edu.br

Abstract. *Vaccination is a cornerstone of public health, yet the COVID-19 pandemic exposed how online misinformation and polarization can undermine immunization efforts. Studying these dynamics at scale is challenging due to scarce annotated data, class imbalance, and high labeling costs. This work presents a longitudinal analysis of vaccine discourse on YouTube in Brazil (2018–2024), using a semi-supervised stance detection framework to classify 1.4 million comments. The approach enables effective learning from limited supervision and produces a publicly available Portuguese stance classification model alongside a large-scale analysis of online debate. Results show strong performance despite class imbalance and enable the analysis of polarization over time. We find that polarization intensified during the pandemic and became more fragmented afterward, with science communication and digital-native channels concentrating both supportive and opposing engagement.*

1. Introduction

Vaccination remains one of the most effective public health interventions, significantly reducing the burden of infectious diseases worldwide [Knisely and Erbelding 2025]. Historically, fluctuations in vaccine uptake have often followed large-scale epidemiological events, which periodically realign public attention and reshape risk perception [Pan et al. 2017]. The COVID-19 pandemic intensified these dynamics on an unprecedented scale, exposing societies to simultaneous challenges of scientific uncertainty and massive surges of digital communication [Zhao et al. 2023]. Online platforms became central arenas for public deliberation, contestation, and the circulation of health information [Sufi et al. 2022]. Simultaneously, vaccine hesitancy, politicization, and coordinated misinformation emerged as critical global challenges for health systems [Larson 2022]. Several studies have documented how fear-based narratives and distrust in health institutions proliferate online, influencing risk perception and undermining public health interventions [Aygün et al. 2021].

Despite a growing computational literature examining these phenomena on platforms like YouTube [Locatelli et al. 2022, Abao et al. 2021], two important limitations remain. Most studies focus on English-language content or a narrow subset of vaccines, limiting understanding of how public attention evolves across a full immunization schedule. Additionally, few works examine the structural organization of stance in YouTube's health information ecosystem over time. This gap is particularly pronounced for Brazil, which has one of the world's most comprehensive immunization systems [Pércio et al. 2023]

and experienced highly polarized health debates during and after the pandemic [Locatelli et al. 2022]. Large-scale analysis also faces obstacles, including scarce annotated datasets in non-English contexts, severe class imbalance, and high labeling costs. To address these challenges, recent work has explored semi-supervised learning strategies that expand limited labeled datasets through self-labeling mechanisms [Amini et al. 2025].

This work investigates how public discourse and polarization around vaccination evolved in Brazil's YouTube ecosystem from 2018 to 2024, surrounding the COVID-19 pandemic. We develop and apply stance detection models to characterize shifts in user attitudes, engagement dynamics, and content ecosystems across the Brazilian immunization schedule, providing insights for public health communication. Our contributions are: (i) a large-scale longitudinal analysis of vaccine discourse in Brazil; (ii) a semi-supervised stance detection framework robust to class imbalance; (iii) a publicly available Portuguese stance classification model for vaccination discourse; and (iv) insights into polarization across pre-, pandemic, and post-pandemic phases. Conducted as an undergraduate research project, the student led the design and implementation of the semi-supervised pipeline, including data processing, annotation, model fine-tuning, and longitudinal analysis, under advisor supervision. The study also resulted in an accepted full paper at the 18th ACM Web Science Conference (WebSci 2026), supported by an ACM SIGMETRICS student grant.

2. Related Work

Prior work was selected based on relevance to vaccination discourse, stance detection, and social media analysis in online health communication. Research on vaccination discourse across digital platforms spans sentiment and hesitancy, polarization and information flows, and stance modeling. Early studies tracked how public attitudes responded to external events, with negative sentiment clustering around uncertainty and politicized narratives [Qorib et al. 2023, Aygün et al. 2021]. On YouTube, negative or conspiratorial content often achieves higher centrality and algorithmic reinforcement [Song and Gruzd 2017], while toxic comments trigger contagious emotional escalation [Miyazaki et al. 2024]. For automated stance classification, deep learning models, including multilingual Bidirectional Encoder Representations from Transformers (BERT) and hybrid architectures, have substantially improved accuracy over traditional classifiers [Straton 2023, Blanco et al. 2025]. However, these approaches remain constrained by annotation costs and scalability. Prior work typically focuses on narrow time windows and rarely integrates multiple vaccines and platform-specific media structures. This work addresses this gap by combining a scalable semi-supervised stance pipeline, with a seven-year longitudinal analysis covering diverse vaccines in the Brazilian context.

3. Methodology

The methodology follows a multi-stage pipeline comprising corpus construction, manual annotation, supervised stance modeling, confidence-based semi-supervised self-labeling, and descriptive longitudinal analysis of discourse and interaction patterns. Videos, channels, metadata, and user comments related to vaccination in Brazil were collected between January 2018 and July 2024 using search terms derived from the Brazilian National Immunization Program and official vaccine terminology through the YouTube API V3[1]. As the study relies on publicly available data, no personally identifiable information was

[1] https://developers.google.com/youtube/v3

analyzed and all data were processed in aggregate, in compliance with YouTube terms and ethical guidelines. Videos from educational, exam-preparation, or other unrelated contexts were excluded via title-based filtering. Comments were restricted to Portuguese using a transformer-based detector and a statistical language identification tool, retaining those identified as Portuguese by at least one method to cover short and informal texts. After preprocessing, the dataset comprised 3,897 channels, 14,318 videos, and 1,422,406 comments from 591,760 unique users.

To build the labeled dataset used for supervised training, comments were sampled across the full temporal span of the study to preserve temporal diversity and avoid concentrating annotations around salient events. Following established annotation guidelines, three annotators independently labeled each comment as Favorable, Against, or Inconclusive based exclusively on textual content. Inter-annotator agreement was assessed using Fleiss' Kappa ($\kappa = 0.69$) indicating substantial agreement for multi-annotator categorical labeling considering the noisy social media environments [Fleiss et al. 1981]. Final labels were assigned by majority vote, discarding 23 comments with complete disagreement. The resulting labeled dataset contained 3,476 comments: 295 Favorable, 446 Against, and 2,735 Inconclusive, reflecting the empirical structure of large-scale online discussions in which explicitly polarized stances are rarer than ambiguous or weakly opinionated discourse.

The supervised classifier was built by fine-tuning a transformer-based language model through parameter-efficient adaptation. We fine-tuned Llama 3.1 8B using Quantized Low-Rank Adaptation (QLoRA) on the manually annotated dataset. The model used 4-bit NF4 quantization, a context length of 192 tokens (covering 95% of comments), and batch size 128 for up to 20 epochs with weighted cross-entropy to address class imbalance. We used stratified 5-fold cross-validation with early stopping (patience of three epochs) and evaluated performance using Accuracy, Precision, Recall, F1-score, and Macro F1 [Dubey et al. 2024].

Because the labeled dataset is highly imbalanced, we incorporated a semi-supervised self-labeling stage guided by predictive uncertainty following the literature [Amini et al. 2025]. In this setting, the model first trained on manually annotated data generates pseudo-labels for unlabeled comments, expanding supervision beyond the original dataset. However, naive pseudo-labeling can reinforce the majority class and propagate early model errors, particularly when ambiguous comments dominate. To reduce this risk, pseudo-label selection was guided by the Shannon entropy of the predicted class probabilities, computed as $H(\mathbf{p}) = -\sum_i p_i \log p_i$. Selecting only low-entropy predictions prioritizes examples where linguistic evidence is clearer and labels more reliable.

This strategy is particularly important for social media stance detection, where the majority Inconclusive class contains heterogeneous fragments and weakly opinionated statements. Adding pseudo-labels indiscriminately would therefore amplify majority-class noise. Instead, predictions were filtered through an entropy threshold and sampled with inverse-frequency emphasis, prioritizing high-confidence examples from minority classes (Favorable and Against). This process generated 2,004 additional pseudo-labeled comments, increasing the training dataset by 57.65% relative to the manually annotated set while limiting semantic drift during self-training [Amini et al. 2025].

The enriched dataset, combining annotated and selected pseudo-labeled comments, was used to train a second-stage classifier, which was subsequently applied to the entire

corpus. For longitudinal analysis, the timeline was divided into three phases: pre-pandemic (January 2018–March 2020), pandemic (March 2020–May 2023), and post-pandemic (May 2023–July 2024). Vaccine-specific discourse was identified using a lexicon mapping vaccine names to abbreviations, disease names, and colloquial expressions. Interaction patterns were analyzed through reply trees by comparing the predicted stance of parent comments and direct replies to estimate conditional reply patterns between supportive and opposing positions. Finally, channels were characterized by stance distribution, engagement, and polarization, enabling comparisons between institutional and digital-native environments.

4. Results

The results indicate strong performance, particularly for minority classes, which are typically challenging in stance detection tasks. Due to space constraints, the results are summarized in-text. Compared to the supervised baseline trained on 3,476 comments, the low-entropy model increased F1-scores from 0.67 ± 0.02 to 0.88 ± 0.02 (Against) and from 0.57 ± 0.06 to 0.91 ± 0.01 (Favorable). Overall accuracy reached 0.90, with the Inconclusive class achieving precision of 0.95 ± 0.01, recall of 0.94 ± 0.02, and F1-score of 0.94 ± 0.01. The supervised baseline achieved 0.83 ± 0.02 accuracy, with F1-scores of 0.67 ± 0.02 (Against), 0.57 ± 0.06 (Favorable), and 0.92 ± 0.01 (Inconclusive), reported as 95% confidence intervals. The enriched training set combined annotated data with 2,004 high-confidence pseudo-labeled comments (1,133 Favorable, 749 Against, 122 Inconclusive), increasing the dataset by 57.65%. This strategy counterbalanced the dominance of the Inconclusive class through inverse-frequency sampling, enabling the model to learn stronger discriminative features for minority stances.

Applying the final model to the full corpus of 1.4 million comments showed that the Inconclusive class dominates the dataset (1,039,538 comments, 74.4%). Among polarized stances, Against comments (204,179) outnumbered Favorable ones (152,940). Against comments received slightly more likes on average (4.1 vs. 3.9), whereas Favorable comments generated more replies (0.44 vs. 0.36).

Temporal analysis segmented the dataset into pre-pandemic (January 2018–March 2020), pandemic (March 2020–May 2023), and post-pandemic (May 2023–July 2024) phases. During the pre-pandemic period, Favorable replies to Against parents reached probability 0.41 and Against replies to Favorable parents reached 0.31. During the pandemic, in-group alignment decreased to 0.57 for both stances and cross-stance replies reached 0.43. The polarized share of comments increased from 13.62% (pre-pandemic) to 26.02% (pandemic). In the post-pandemic phase, Against $\rightarrow$ Against interactions reached 0.67, while Favorable $\rightarrow$ Favorable interactions reached 0.50. The polarized proportion reached 28.84%.

Channel-level analysis classified channels as certified (affiliated with organizations recognized by the Associação Nacional de Jornais[2]) or non-certified. Non-certified channels hosted 82.4% of anti-vaccine comments. Among the 15 channels with the highest volume of polarized comments, only two were ANJ-certified legacy media (BBC News Brasil and CNN Brasil), while thirteen were non-certified. Additionally, 14 appeared among the top channels for both Favorable and Against comments. Science communicators accumulated large volumes of both supportive and opposing comments. Legacy media showed lower cross-stance reply probabilities than science communicators.

[2] https://www.anj.org.br/

5. Discussion

The results indicate that semi-supervised enrichment effectively improves minority class performance, suggesting that low-entropy self-training mitigates class imbalance in stance detection.

The predominance of Inconclusive comments indicates that most interactions are not explicitly polarized. Among polarized stances, the higher number of Against comments and slightly higher likes suggest stronger passive validation, whereas more replies to Favorable indicate greater conversational engagement. These patterns suggest that pro-vaccine discourse is more dialogical, while anti-vaccine messages tend to be short and assertive.

Temporal dynamics reveal substantial shifts in discourse structure. In the pre-pandemic, discourse shows pro-vaccine cohesion and asymmetric engagement. The pandemic marks a turning point, with increased polarization and more balanced cross-stance interactions. In post-pandemic, persistent high polarization, stronger skeptical in-group reinforcement, and weaker pro-vaccine cohesion indicate a reorganization toward asymmetry. Overall, these findings suggest that vaccine polarization is dynamic, intensifying during health crises and reorganizing afterward.

At the channel level, the concentration of polarized discourse in non-certified channels indicates their central role in shaping debate. The presence of the same channels among the top venues for both stances suggests that highly visible spaces concentrate cross-stance interactions. The centrality of science communicators, combined with lower engagement asymmetry in legacy media, suggests that polarization on YouTube is driven mainly by interactional dynamics in digital-native and science communication channels rather than traditional media alone.

6. Conclusion

This undergraduate study examined how vaccine-related discourse evolved on YouTube in Brazil across pre-pandemic, pandemic, and post-pandemic periods. Methodologically, we show that semi-supervised stance modeling with low-entropy self-training enables scalable, robust analysis of highly imbalanced user-generated content. Substantively, vaccine polarization is dynamic, intensifying during crises and restructuring into asymmetric patterns post-crisis. Furthermore, polarization is anchored in a hybrid media ecosystem dominated by science communicators and digital-native channels, highlighting structural vulnerabilities in health communication that must be addressed to improve immunization campaigns.

This study has limitations. Reliance on pseudo-labeling may propagate biases without external validation, and the dominance of the Inconclusive class limits fine-grained interpretation. The analysis is restricted to YouTube in Brazil. Future work should explore cross-platform validation, alternative semi-supervised strategies, and error analysis, particularly for the Inconclusive class, where ambiguity is more frequent.

References

[Abao et al. 2021] Abao, R. P., Estuar, M. R. J. E., Cataluña, A. A. M., Aureus, J. P., and Mapua, D. C. (2021). Emotion analysis of comments from vaccine-related youtube videos: Understanding the public's response to covid-19 vaccination. In *IEEE SNAMS*, pages 1–7.

[Amini et al. 2025] Amini, M.-R., Feofanov, V., Pauletto, L., Hadjadj, L., Devijver, E., and Maximov, Y. (2025). Self-training: A survey. *Neurocomputing*, 616:128904.

[Aygün et al. 2021] Aygün, I., Kaya, B., and Kaya, M. (2021). Aspect based twitter sentiment analysis on vaccination and vaccine types in covid-19 pandemic with deep learning. *IEEE J-BHI*, 26(5):2360–2369.

[Blanco et al. 2025] Blanco, G., Yáñez Martínez, R., and Lourenço, A. (2025). Leveraging deep learning to detect stance in spanish tweets on covid-19 vaccination. *JAMIA open*, 8(1):ooaf007.

[Dubey et al. 2024] Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al. (2024). The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

[Fleiss et al. 1981] Fleiss, J. L., Levin, B., and Paik, M. C. (1981). The comparison of proportions from several independent samples. *Statistical methods for rates and proportions*.

[Knisely and Erbelding 2025] Knisely, J. M. and Erbelding, E. (2025). Vaccines for global health: Progress and challenges. *The Journal of Infectious Diseases*.

[Larson 2022] Larson, H. J. (2022). Defining and measuring vaccine hesitancy. *Nat. Hum. Behav.*, 6(12):1609–1610.

[Locatelli et al. 2022] Locatelli, M. S., Caetano, J., Meira Jr, W., and Almeida, V. (2022). Characterizing vaccination movements on youtube in the united states and brazil. In *ACM HT*.

[Miyazaki et al. 2024] Miyazaki, K., Uchiba, T., Kwak, H., An, J., and Sasahara, K. (2024). The impact of toxic trolling comments on anti-vaccine youtube videos. *Sci. Rep.*, 14(1):5088.

[Pan et al. 2017] Pan, Y., Wang, Q., Yang, P., Zhang, L., Wu, S., Zhang, Y., Sun, Y., Duan, W., Ma, C., Zhang, M., et al. (2017). Influenza vaccination in preventing outbreaks in schools: A long-term ecological overview. *Vaccine*.

[Pércio et al. 2023] Pércio, J., Fernandes, E. G., Maciel, E. L., and Lima, N. V. T. d. (2023). 50 years of the brazilian national immunization program and the immunization agenda 2030. *Epidemiologia e Serviços de Saúde*, 32:e20231009.

[Qorib et al. 2023] Qorib, M., Oladunni, T., Denis, M., Ososanya, E., and Cotae, P. (2023). Covid-19 vaccine hesitancy: Text mining, sentiment analysis and machine learning on covid-19 vaccination twitter dataset. *Expert Syst. Appl.*

[Song and Gruzd 2017] Song, M. Y.-J. and Gruzd, A. (2017). Examining sentiments and popularity of pro-and anti-vaccination videos on youtube. In *Social Media + Society*, pages 1–8.

[Straton 2023] Straton, N. (2023). Covid vaccine stigma: detecting stigma across social media platforms with computational model based on deep learning. *Appl. Intell.*, 53(13):16398–16423.

[Sufi et al. 2022] Sufi, F. K., Razzak, I., and Khalil, I. (2022). Tracking anti-vax social movement using ai-based social media monitoring. *IEEE-TTS*.

[Zhao et al. 2023] Zhao, S., Hu, S., Zhou, X., Song, S., Wang, Q., Zheng, H., Zhang, Y., Hou, Z., et al. (2023). The prevalence, features, influencing factors, and solutions for covid-19 vaccine misinformation: systematic review. *JPHS*, 9(1):e40201.