Radar Social LGBTQIA+: Cooperative AI Against Platform-Mediated Hate

Abstract

The second semester of 2024, the Brazilian LGBTQIA+ podcast *Entre Amigues* became both a national phenomenon and a target of coordinated digital violence. As hate campaigns orchestrated by far-right figures spread across platforms, Big Tech companies' inaction exposed the extractive logics of platform capitalism: violence was not an error but an outcome of engagement-driven infrastructures. In response, the nonprofit Código Não Binário led a collective effort to document, analyse, and seek accountability for the attacks—laying the groundwork for a public civil action against the platforms themselves.

From this lived experience emerges *Radar Social LGBTQIA+*, an open AI initiative designed to detect and categorise hate against LGBTQIA+ communities across platforms. Evolving from a social listening prototype into an ensemble AI model, the project bridges grassroots digital forensics and machine learning. Built on community-annotated datasets and explainable hybrid architecture, it enables advocacy, legal, and protective uses in the Global South. This paper situates *Radar Social LGBTQIA+* within the continuum between human-led mapping and algorithmic detection, proposing a framework for cooperative AI infrastructures that resist the extractive logic of platform capitalism and reclaim technological agency for marginalised communities.

Keywords: cooperative AI, hate speech detection, strategic litigation, platform capitalism, LGBTQIA+, ensemble model, Global South, digital rights.

1. Introduction

The rapid expansion of generative and social media technologies has transformed not only how information circulates but also how violence is organised, amplified, and normalised. In Brazil, the visibility of queer and trans creators has increased dramatically over the past decade, yet their safety within platform infrastructures remains precarious. The case of *Entre Amigues*, a Brazilian LGBTQIA+ podcast that reached millions across platforms in 2024, epitomises this paradox.

Soon after its viral success, *Entre Amigues* faced a coordinated hate campaign driven by far-right influencers and political figures. The campaign mobilised harassment networks, automated accounts, and ideologically motivated amplification loops, turning online abuse into a spectacle. Despite repeated reports, major platforms—Meta, Google, TikTok, and X—failed to act promptly or consistently, allowing the violence to trend and profit.

This failure was not incidental: it revealed structural complicity within the platform economy. Under engagement-based algorithms, hate generates traffic; traffic generates profit. The

systemic absence of safety mechanisms for marginalised groups, particularly trans and queer communities, is therefore not a mere oversight but a form of design violence.

In response, *Código Não Binário*, a nonprofit organisation focused on technology and gender justice, led a multidisciplinary initiative to document, categorise, and litigate against this wave of online hate. Between April 2024 and August 2025, the organisation built a corpus of evidence combining manual annotation, network mapping, and legal analysis. The process culminated in the preparation of a public civil action (*Ação Civil Pública*) to be filed against multiple Big Tech companies in November 2025.

The second phase of this work, beginning in mid-2025, translates these experiences into a cooperative AI system—*Radar Social LGBTQIA+*—capable of scaling community detection and analysis of hate speech in Brazilian Portuguese. This paper traces the trajectory from human-led response to algorithmic tooling, showing how lived struggle informed the design of an ethical, community-rooted AI model.

2. Context and Legal Response (April 2024 – mid-2025)

2.1 Viral Success and Coordinated Hate

In early 2024, *Entre Amigues* emerged as one of Brazil's most prominent LGBTQIA+ podcasts, achieving millions of views across YouTube, Spotify, Instagram, and TikTok. Hosted by trans and non-binary presenters, it offered candid conversations on identity, politics, and everyday life. Its success, however, attracted hostility from extremist sectors that framed queer visibility as ideological threat.

Coordinated attacks began when far-right public figures amplified distorted clips from the podcast, often accompanied by slurs, moral panic, and coded language. These posts triggered waves of harassment across comment sections and social networks. Content moderation systems either failed to detect the hate or downranked the podcast itself for "controversial content," effectively penalising the victims.

2.2 Platform Complicity and Structural Violence

The response from Big Tech companies was marked by silence, delay, and opacity. Reporting mechanisms proved ineffective: flagged posts remained online for weeks, appeals were ignored, and perpetrators faced no consequences. Internal processes prioritised "engagement" metrics, allowing hate to thrive under commercial incentives.

Código Não Binário's assessment concluded that the violence was systemic rather than incidental—an artefact of *platform capitalism*, where harmful content becomes monetised attention. The organisation identified this as both a **social justice failure** and a **legal opportunity** to demand accountability for discriminatory negligence.

2.3 Mapping, Categorisation and Evidence (faithful to the annexed report)

Aligned with the legal strategy, Código Não Binário led a systematic effort to: **identify and categorise** hate content; **trace sources and amplification** across accounts and platforms; and **quantify reach** by follower counts, engagement patterns and share trajectories. The analysis distinguished **modes of expression** (direct slurs, coded dog-whistles, mockery, insinuation), evaluated **context and intent**, and noted where conduct plausibly crossed into **criminal offences**. This evidentiary corpus enabled **targeted platform reports** and **preparation of legal actions** against specific offenders, while also informing a broader **strategic litigation initiative** aimed at platform accountability for enabling and profiting from systemic violence. Importantly, procedures were designed to preserve URLs, timestamps, IDs and screenshots suitable for forensic use in Brazilian courts.

2.4 Legal Framework and Evidentiary Method

The legal strategy integrated criminal, civil, and consumer protection laws, arguing that platforms had violated Brazil's anti-discrimination framework and consumer rights by enabling hate-based harms. The team developed procedures for data preservation, metadata collection, and evidence integrity—hashing URLs, screenshots, and post IDs to maintain forensic value.

Legal advisors translated this evidence into claims that could demonstrate both *individual culpability* (for perpetrators) and *structural responsibility* (for platforms). This approach, rare in Latin America, positioned digital hate as a matter of public interest and collective harm rather than isolated incidents.

2.5 From Evidence to Strategic Litigation

By mid-2025, the evidence supported a broader legal narrative: that digital platforms' business models constitute an environment of predictable, profitable violence. Código Não Binário and partner organisations prepared a public civil action to be filed in November 2025, demanding regulatory and financial accountability.

The mapping exposed a practical bottleneck: **human review at scale** is slow, emotionally costly and financially unsustainable. These constraints, combined with our organisational capacity and our desire for auditable, locally governed tooling, motivated a pivot from the initial social-listening platform concept toward a **lightweight, explainable model**. A system was needed to **automate the first layer of categorisation**, enabling both faster triage and richer analytics. This necessity gave rise to the *Radar Social LGBTQIA+* project.

3. Methodological Bridge: From Human Mapping to AI Design

The transition from manual annotation to automated detection was not an abstraction—it was born from the legal workflow itself. To complete the evidentiary dossier for the public civil action, the team needed to review thousands of social media posts, identify hate speech patterns, and ensure consistent labelling across reviewers.

A key lesson concerns the **emotional and psychological** toll of annotation on LGBTQIA+ contributors. Repeated exposure to hateful content inflicted harm; **some participants left the project** because the burden became intolerable. This reality reinforced the moral imperative to **liberate LGBTQIA+ people from routine re-exposure** wherever possible. By shifting first-pass triage to a model that encodes community knowledge and legal categories, we can reduce the volume and intensity of material requiring human judgement, reserving expert review for borderline or high-impact cases and implementing strong care protocols for those reviewers.

The process revealed recurring linguistic and contextual markers of hate in Brazilian Portuguese that existing global models failed to capture. Words and phrases used as **dog-whistles**—often ironic or coded—were systematically misclassified as neutral by standard hate speech models. Furthermore, global datasets lacked representation of the complex intersectional insults that target gender diversity in Portuguese-speaking contexts.

Faced with these gaps, the team decided to create a *community dataset* that reflected the lived linguistic reality of Brazilian queer communities online. Using approximately 1,500 manually annotated comments from the *Entre Amigues* incident, the organisation designed an initial classification scheme distinguishing explicit, coded, and neutral discourse.

The methodological shift was therefore not merely technical but political. By transforming human annotation protocols into a machine learning model, Código Não Binário transformed **collective trauma into collective intelligence**—a reusable resource for detection, advocacy, and education.

This bridge between legal evidence and AI model design became the foundation for *Radar Social LGBTQIA+*, ensuring that the system's outputs remained grounded in community-defined notions of harm and legality, rather than abstract computational metrics.

4. AI Model and Technical Solution (From mid-2025 Onward)

4.1 System Overview

Radar Social LGBTQIA+ is an open-source ensemble model for detecting hate speech against LGBTQIA+ communities in Brazilian Portuguese. Also available as a Space on Hugging Face

(https://huggingface.co/spaces/Veronyka/radar-social-lgbtqia-v2.1), it integrates three layers of analysis:

1. A **binary classifier** (BERTimbau + ToLD-BR) detecting generic hate signals.
2. A **specialised model**, fine-tuned with the *Entre Amigues* dataset.
3. A **contextual rule engine** containing over 40 manually curated linguistic and semantic rules.

This ensemble allows the system to balance machine learning flexibility with human interpretability. Each detection is accompanied by rationale outputs—tokens, rules, or embeddings that triggered the classification—allowing users to trace reasoning and maintain forensic validity.

4.2 Technical Architecture

The pipeline includes text normalisation, tokenisation, and context-sensitive rule parsing. The ensemble structure enables parallel inference across the binary and specialised models, which are then aggregated through a weighted decision function.

The ruleset layer operates post-classification, refining results based on contextual heuristics. For instance, reclaimed slurs used in affirming contexts are filtered out, while coded expressions like “ideologia de gênero” (gender ideology) or “banheiro unissex” (unisex bathroom) are flagged when co-occurring with derogatory intent markers.

Evaluation uses stratified cross-validation, with metrics for precision, recall, and F1 across explicit and coded hate categories. Initial testing demonstrates improved recall in ambiguous contexts—essential for triage workflows—while maintaining acceptable precision for legal review.

4.3 Dataset and Open Science

The accompanying dataset (*base-dados-ódio-lgbtqia*, https://huggingface.co/Veronyka/base-dados-odio-lgbtqia), with more than 12,000 records, is openly available under an ethical research licence. It documents label schemas, annotation guidelines, and metadata for replication.

Each entry of the main dataset (1,500 records) contains the original text, context indicators, and label categories (explicit hate, coded hate, neutral).

By publishing this corpus, Código Não Binário aims to establish a foundation for future fine-tuning across dialects and identities in the Global South.

4.4 Ethical and Legal Design

The system prioritises **explainability, auditability, and care**. Outputs include rule matches and model rationales; false positives can be reviewed; and datasets are versioned with change logs.

Furthermore, licences restrict misuse—such as profiling or surveillance of LGBTQIA+ communities—and provide contact channels for reporting. This governance model aligns with the principles of *cooperative AI*: collective ownership, transparent iteration, and rights-based accountability.

4.5 Deployment and Collaboration

Unlike corporate "listening" tools optimised for marketing, *Radar Social LGBTQIA+* is **model-centric and decentralised**. Organisations can integrate it into their own data pipelines without sharing sensitive content externally. Deployment options include local Docker containers, Python scripts, or API calls to the Hugging Face Space.

The next phase will focus on **federated collaboration** with advocacy groups, universities, and regulators. Through shared governance, partners will co-develop new rule sets, contribute annotated data, and expand coverage to other languages and identity groups.

5. Discussion

The *Radar Social LGBTQIA+* project challenges the mainstream narrative that AI ethics is the domain of large corporations or academic laboratories. Instead, it demonstrates that **marginalised communities can themselves become producers of AI knowledge and infrastructure**.

First, the model's origin in legal evidence underscores the epistemic authority of lived experience. Rather than treating hate detection as a neutral task, the system embeds social context into its architecture. Its categories—explicit, coded, or neutral—emerge from activist knowledge, not commercial taxonomies.

Second, the ensemble architecture represents a form of *epistemic pluralism*: combining statistical inference with human rules mirrors the coexistence of analytical rigour and community intuition. This hybrid approach embodies the cooperative AI ethos—systems that *learn with* and *answer to* the communities they serve.

Third, by remaining open-source and explainable, *Radar Social LGBTQIA+* resists the extractive tendencies of proprietary AI ecosystems. Transparency is not only an ethical stance but a legal necessity: communities must be able to verify and contest algorithmic outputs when used in litigation or advocacy.

Finally, the project reframes AI from tool of surveillance to instrument of protection. By equipping communities with the means to map and counter hate, *Radar Social LGBTQIA+*

enacts what feminist technoscience calls *infrastructural resistance*: using the same computational logics that enable harm to sustain collective defence.

6. Conclusion

The trajectory from *Entre Amigues* to *Radar Social LGBTQIA+* illustrates a broader principle: when platforms fail to protect, communities innovate their own infrastructures of care, accountability, and evidence.

Since April 2024, Código Não Binário has been confronting the limits of human endurance in the face of algorithmic violence. In mid-2025, that experience had crystallised into a cooperative AI model—an ensemble of technical, legal, and political learning.

As the *Ação Civil Pública* prepares for filing, *Radar Social LGBTQIA+* stands as both a continuation and a transformation of that process. It turns reactive defence into proactive capacity, offering an open tool that others can adapt for their contexts.

The project demonstrates that the future of AI in social justice depends not on scale but on solidarity: on sharing models, datasets, and methods that encode empathy as much as efficiency. Cooperative AI, in this sense, is not merely a research agenda—it is a survival strategy for those most exposed to the violence of platform capitalism.

Acknowledgements and Availability

We acknowledge the *Entre Amigues* community, Código Não Binário workers and partners, that contributed to data collection, legal groundwork, and ethical design.

Model demo: *Radar Social LGBTQIA+* — https://huggingface.co/spaces/Veronyka/radar-social-lgbtqia-v2.1

Dataset: *base-dados-ódio-lgbtqia* — https://huggingface.co/Veronyka/base-dados-odio-lgbtqia

Contact: contato@codigonaobinario.org