EXECUTIVE SUMMARY:
Reducing the Impact of Anti-Rights Groups in Chile

This ESRC-Impact Acceleration project [2402-SDG-1829], titled Reducing Extremists' Threats to Democracy in Latin America: Chile Pilot Project, was to explore strategies to reduce the impact of anti-rights groups through hate speech monitoring. The project developed a pilot online hate speech monitoring system based on the Chilean case. The system used sophisticated AI tools based on large language models for the automated identification of hate speech. Based on its preliminary results, the project also devised strategies to mitigate the impact of anti-rights actors through hate speech and other potential threats (strategic litigation/negotiation), which are used to a far lesser extent around the world.

Detection of Hate Speech

Hate speech in Latin America has proliferated in a context marked by long-standing social inequalities, political polarization, and digital transformation. Rooted in histories of marginalization, violence, and exclusion, contemporary expressions of hate speech—especially on digital platforms—have become a growing concern for democratic institutions and social cohesion. While the concept of hate speech remains contested, it is broadly understood as discourse that targets individuals or groups on the basis of protected characteristics and/or social categories such as race, ethnicity, gender, or political affiliation.

The pilot model in Chile employed a hybrid methodology combining expert manual coding with large language model (LLM)-based classification to monitor hate speech in over half a million YouTube comments between 2019 and 2024. While the majority of analyzed content did not contain hate speech, a concerning 14.82% did. These comments targeted migrants, LGBT+ individuals, people from lower socio-economic classes, women, and Indigenous peoples, among others.

When disaggregated by year, a general downward trend is observed in the average percentage of comments containing hate speech per channel between 2019 and 2022, decreasing from 11.69% to 6.95% (Table 8). This suggests a decline in the average prevalence of such comments across the analyzed channels. However, in 2023 and 2024, the average rises again, reaching 9.58% in 2024, indicating a resurgence in certain spaces. Despite this general trend, maximum values reveal that some channels consistently register exceptionally high levels of hate speech—up to 50% in 2021, 2023, and 2024—demonstrating the persistence of communities where this type of interaction constitutes a substantial share of total comments, far exceeding the sample's average.

The results of this project demonstrate that the combination of generative language models and advanced machine learning techniques is an effective strategy for hate speech detection, despite the significant operational costs involved. The superior performance of gemini-flash-002 in terms of precision and sensitivity underscores the importance of investing in technologies capable of

capturing linguistic complexity in critical contexts. At the same time, exploring lower-cost approaches and implementing time series methods open new possibilities for improving the predictive and operational capacities of monitoring systems.

Other strategies: Strategic Litigation and Negotiation

This project has approached hate speech in Chile by both pinpointing hate-motivated offences and proposing ways to curb and prevent them. Strategic litigation allows stakeholders to join legal proceedings and secure rulings whose impact surpasses the immediate parties to the crime (Amnesty International, 2025). Chile's Antidiscrimination Statute, known as the Zamudio Law, emerged after the hate-driven killing of Daniel Zamudio and the subsequent public, media and judicial outcry (Fundación Iguales, 2012). Covering an extensive list of protected attributes—such as sex, gender, sexual orientation, race, religion and nationality—the law shows how one person's tragedy can inspire legislation that safeguards many others.

Yet the Zamudio Law concentrates on hate crimes rather than hate speech. Strategic lawsuits could be used to amend the statute or spotlight a fresh test case that targets hateful expression specifically. Although the law is a major step forward, lawyers and advocates may rely on strategic litigation to widen its reach, tackling hate speech alongside violent acts. Crafting legal tools to deter hateful expression is expensive and contentious—often colliding with free-speech principles—but clear definitions and systematic monitoring remain essential for tracing patterns and societal effects. Stronger links must also be drawn between surges in hate speech and subsequent discrimination or violence.

The project also analysed negotiation as a preventive tool, recognising the difficulty of talking with online offenders who hide behind anonymity. When direct engagement is impossible, alternatives include partnering with the private sector to boost counter-speech pages and encouraging companies to share search-engine expertise with anti-extremist groups (Bipartisan Policy Center 2012, 29). Highlighting and supporting individuals and organisations devoted to countering hate—such as Argentine-Brazilian academic and blogger Lola Aronovich—is another vital avenue.

Conclusion

This report on hate speech lays the groundwork for future research and for the development of public policies that strengthen democratic resilience and protect vulnerable communities in an increasingly polarized and digitalized global environment. It also represents a unique methodological approach, combining advanced machine learning techniques with manual coding by experts to better train the coding model and ensure its effectiveness. Such methods can be easily adapted to other contexts to ensure that AI and machine learning tools produce the most accurate results.

REPORT:
Reducing the Impact of Anti-Rights Groups in Chile

Fear of democratic backsliding has prompted efforts to reduce threats from extremist groups around the world. This project addresses the threat to peace, justice, and democratic institutions (SDG 16) from violent backlash mobilisations by extremist groups in Latin America. These groups attack historically, culturally, socially, economically, and politically excluded groups: women, BIPOC groups, LGBT+ communities, immigrants, the economically-disadvantaged, environmentally vulnerable populations, and victim-survivors of human rights violations.

Latin America is a critical region for reducing the threats of extremist anti-rights groups. Those groups have engaged in violent attacks against the LGBT+ community in Brazil, Chile, and Colombia; environmentalists in Brazil and Colombia; migrants in Argentina and Mexico; Indigenous communities in Honduras and Argentina; street children in Brazil and Colombia; women in El Salvador and Argentina; and against Afro-Colombians. Considering the alarming prevalence of such hate crimes in the region and systemic impunity, these forms of violence threaten fragile peace, justice, and democratic institutions. Designing threat-reduction systems to protect against democratic backsliding and safeguard marginalized communities is an urgent necessity.

Despite Latin America having the highest level of violence against human rights defenders in the world (Human Rights Watch), no systematic efforts for addressing threats from anti-rights extremist groups exist. The three models for threat reduction this project seeks to design and implement have proved effective in other world regions. The US and UK governments use the work of monitoring organisations – such as HateLab, Moonshot, the Centre for the Analysis of the Radical Right-CARR, and the Southern Poverty Law Center – to track and intervene, to reduce threats from, extremist groups. The logic behind such monitoring systems is ‘that if extremist right-wing groups are tracked, and their activities rendered public and visible, the violence they perpetrate can be identified, exposed, and guarded against before murder or other criminal acts are committed’ (Zulver and Payne 2023, 245).

Monitoring experts have also advanced models to redirect online users to sites with alternative and factual information to counter the appeal of extremist groups. The data these monitoring organisations have collected have been used to prosecute extremists’ criminal activity, such as the 6 January 2021 insurrectionists in the US. A third strategy is the use of negotiation and mediation models for reducing extremist and criminal organisations’ threats to democracy. Despite coordinated efforts of governments and nongovernmental organisations to develop and implement tracking systems and other threat-reduction strategies in the global north, to date, no ‘solidified or systematised organisation…undertakes this exercise in Latin America’ (Zulver and Payne 2023, 245).

In response to these grave concerns, this ESRC-Impact funded project built a bespoke threat-reduction system for fragile democracies, with a particular focus on online hate-crime monitoring through the use of AI tools. To co-design and co-deliver the threat-reduction system, the academic team adopted a ‘responsible and inclusive’ approach by collaborating with government agencies, monitoring organisations, tech and machine learning experts, and a negotiation leader. The

project united academic experts on extremist groups, monitoring and negotiation practitioners, and government officials to create a bespoke monitoring and threat-reduction system that was piloted in Chile. The project design involved high-level stakeholder meetings, practitioner workshops, and policy and evidence seminars to develop and deploy tools and resources to advance the objective of reducing threats to democracy and human rights. The final aim was a monitoring and threat-reduction system that can be adopted by other governments in the region to reduce the harm to democracy and human rights by extremist groups.

This report begins with a brief background on hate speech in the Latin American region and elsewhere. It then presents the hate speech tracking model and its design, as well as its preliminary findings, such as the different groups targeted by hate speech. It delves into the unique novelty of the use of AI for the project, particularly the vast potential of AI to analyze large datasets when combined with human oversight. The report then explores potential challenges with the model and how to troubleshoot them, and concludes with a discussion of negotiation strategies and relevant strategic litigation tactics that can be used to counter hate speech in Chile and beyond.

BACKGROUND ON HATE SPEECH

In Latin America, hate speech has evolved in response to pronounced inequality, social exclusion, and historical conflicts. For decades, Latin American societies have grappled with tensions arising from deep-rooted socio-economic disparities and the marginalisation of specific groups. These historical antecedents have contributed to the normalisation of discriminatory attitudes and the proliferation of hostile rhetoric directed at vulnerable communities, including women, LGBT+ individuals, migrants, and Indigenous peoples. The advent of digital technologies and the rise of social media have amplified the dissemination of extremist messages, generating new dynamics of radicalisation (Human Rights Watch, 2019).

At the global level, extremism and hate speech have undergone a transformation driven by globalisation and digital interconnectedness. In developed countries, sophisticated systems of monitoring and analysis have been implemented to identify and counter extremist content. Organisations such as the Southern Poverty Law Center (Southern Poverty Law Center, n.d.), HateLab (HateLab, n.d.), and Moonshot (Moonshot, n.d.) have developed robust methodologies that serve as reference points for adapting strategies in other contexts, including Latin America. Furthermore, political, economic, and social crises have fuelled polarisation and the spread of misinformation. In turn, this contributes to the growing radicalisation of certain groups and the social legitimisation of extremist discourse.

The impact of hate speech and extremism manifests at multiple levels. Socially, such discourse fosters division and stigma, increasing the vulnerability of marginalised communities and weakening social cohesion. Politically, the proliferation of extremist narratives undermines trust in democratic institutions, complicating dialogue and negotiation among various actors. Reports by international organisations highlight how polarisation and media manipulation can erode the rule of law and impair the effective functioning of democracy (Freedom House, 2022; Amnesty International, 2023).

Hate speech is a contested concept, and there is no universally agreed-upon definition in international law, but it is broadly understood as any form of expression that denigrates or incites hostility against individuals or groups on the basis of their membership in a particular social category (Article 19, 2020; United Nations, 2019). In general usage, the term refers to communication in speech, writing, or behavior that attacks or uses derogatory or discriminatory language toward someone because of who they are – for example, due to their race, ethnicity, religion, nationality, gender, sexual orientation, or other inherent characteristics (United Nations, 2019). Hate speech typically targets an "out-group" identified by such attributes – often a historically marginalized or minority community – and portrays that group as inferior, evil, or threatening, thereby justifying hostility against them (Matsuda, 1993; Parekh, 2012). Because of these features, hate speech is widely regarded as a uniquely pernicious form of discourse that undermines the dignity and equal standing of its targets in society (Matsuda, 1993; Waldron, 2012).

From a legal standpoint, definitions of hate speech tend to be narrower and tied to specific harms. International human rights law does not supply a single, comprehensive definition of hate speech, but it does prohibit the most extreme cases of hateful expression. For example, the International Covenant on Civil and Political Rights obligates states to ban "any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence" (United Nations, 1966). Similarly, the Council of Europe has defined hate speech as encompassing "all forms of expression which spread, incite, promote or justify racial hatred, xenophobia, anti-Semitism or other forms of hatred based on intolerance" (Council of Europe, 1997).

Legal definitions focus on speech acts that provoke harm or pose a threat to the rights and security of others – in particular, incitement to discrimination, hostility, or violence against a protected group – thereby distinguishing punishable hate speech from speech that is merely offensive or hateful in a colloquial sense (Council of Europe, 1997; UNESCO, 2024). In many jurisdictions, this means that only expression reaching a certain threshold of severity (such as explicit calls for harm, or direct vilification of a group) is legally classified as hate speech, whereas less extreme forms of prejudiced or insulting speech may remain lawful but problematic (UNESCO, 2024). Legal frameworks thus draw a boundary between hate speech and free expression, aiming to proscribe the most dangerous forms of hateful advocacy while respecting the value of open debate (Article 19, 2020).

Beyond the realm of law, sociological and communication theory perspectives adopt a broader view of what constitutes hate speech. From these perspectives, hate speech includes not only overt incitement but any language or symbolic act that stigmatizes, demeans, or vilifies a person because of their group identity, even if it falls short of urging immediate violence (Delgado & Stefancic, 2017; Gelber, 2019). Scholars note that hate speech serves as a mechanism of "othering": it reinforces an in-group vs. out-group division by ascribing negative traits or stereotypes to the target group and thereby legitimizing their marginalization (Douglas et al., 2017; Gelber, 2019).

Despite variations in definition, most characterizations of hate speech identify several common features of its content and form. First, hate speech is group-directed: it targets people because of their real or perceived membership in a social group or category, rather than personal actions or individual traits (United Nations, 2019; Ruscher, 2025). The targeted characteristics are typically those protected in human rights frameworks – for instance, race, ethnicity, religion, gender, sexual orientation, nationality, or disability – which are fundamental to one's identity (United Nations, 2019). Second, hate speech is derogatory, demeaning, or threatening in tone. It often employs epithets, slurs, insults, or pejorative stereotypes to mark the targeted group as inferior or dangerous (Parekh, 2012; UNESCO, 2024).

Hate speech can be expressed through a wide range of communicative forms. While it often occurs in spoken or written language (such as chants, online posts, articles, or speeches), it may also appear in visual or symbolic representations – for example, memes, graffiti, music lyrics, or images that convey hatred toward a group (United Nations, 2019). Even seemingly subtle or coded expressions (sometimes called "dog whistles") can function as hate speech if they are understood to convey a hateful meaning to an in-group audience (Brown, 2018). In essence, what defines hate speech is not the medium or explicitness of the message, but its substance: the deliberate targeting of a group with content that is inherently degrading, hostile, or aimed at instigating harm against that group (Waldron, 2012; Matsuda, 1993).

The impact of hate speech is a critical part of its definition and social significance. Hate speech is not only offensive – it causes tangible harm to individuals and undermines societal values. On an individual level, victims of hate speech often experience psychological and emotional harm, including fear, stress, reduced self-worth, and a sense of being unwelcome or unsafe in their community (Gelber, 2017; Waldron, 2012). Repeated exposure to hate speech can create a hostile environment for members of the targeted group, potentially impairing their opportunities and willingness to participate in civic life, such as engaging in political discussion or using online forums (Gelber, 2017). In a broader societal sense, pervasive hate speech erodes the norms of inclusivity and mutual respect that underpin a pluralistic society (Waldron, 2012; UNESCO, 2024).

Hate speech has a corrosive impact on social cohesion, fostering intergroup distrust and hostility, and legitimizing discrimination by making prejudiced ideas appear socially acceptable or even mainstream (Douglas et al., 2017; UNESCO, 2024). Substantial historical and empirical evidence shows that, if left unchecked, hate speech can escalate into or fuel acts of physical violence against targeted groups (Rosenberg, 2012). Dehumanizing propaganda has repeatedly preceded mass atrocities—for instance, anti-Tutsi radio broadcasts played a key role in inciting violence during the Rwandan genocide, and Nazi propaganda portraying Jews as subhuman helped lay the ideological foundation for the Holocaust (Rosenberg, 2012). Recognizing this potential for real-world harm, international criminal law classifies the intentional incitement of genocide as a punishable offense, emphasizing that extreme hate speech is not merely a consequence of conflict, but often a precursor to it (United Nations, 1948; Rosenberg, 2012).

Hate speech is especially harmful because it both inflicts individual harm and sustains broader patterns of discrimination, disproportionately affecting historically marginalized populations. It is

therefore essential to examine not just the overall prevalence of hate speech, but also **which groups are targeted**. In this study, although a general definition of hate speech is applied, particular attention is paid to discourse directed at historically vulnerable groups—such as those defined by race, ethnicity, religion, gender, socioeconomic status, or sexual orientation—because these groups have faced systemic exclusion and are often afforded specific legal protections under national and international hate speech frameworks (Harel, 2021; Mossie & Wang, 2020).

CHILE PILOT MODEL DESIGN

The methodology implemented in this project is structured across several key pillars, ranging from the collection and selection of data to the advanced analysis of hate speech using machine learning and natural language processing techniques. Each of these pillars is addressed through a series of processes that, taken together, enable the development of a robust system for the monitoring and mitigation of extremist threats within the Chilean context.

As illustrated in Figure 1, the initial stage involved collecting a substantial volume of user interaction texts from digital platforms. Given the large scale of the data collected, and the need for metrics to ensure the validity and reliability of automatic classification, a representative sample was manually annotated by expert human coders. The remaining texts were then labelled using a large language model (gemini-flash-002). These annotated data were subsequently evaluated through the calculation of the performance metrics outlined in Section 2.2.2.4. Based on these results, the system was iteratively refined, with discrepancies reviewed and corrected in accordance with expert-defined criteria. Following the implementation of these corrections, the entire dataset was subjected to automatic classification.

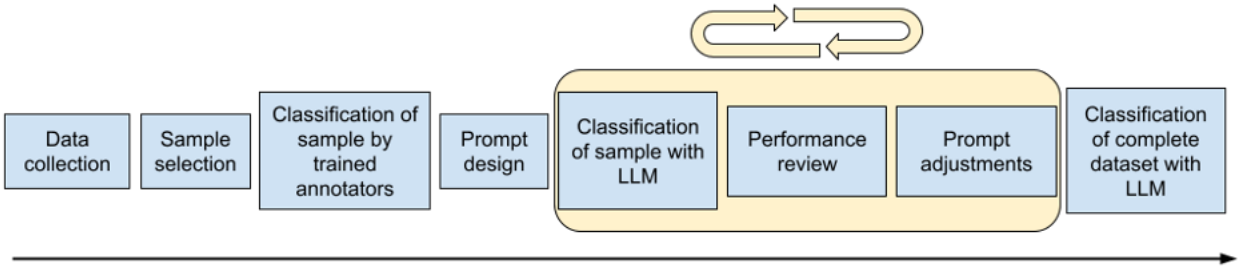
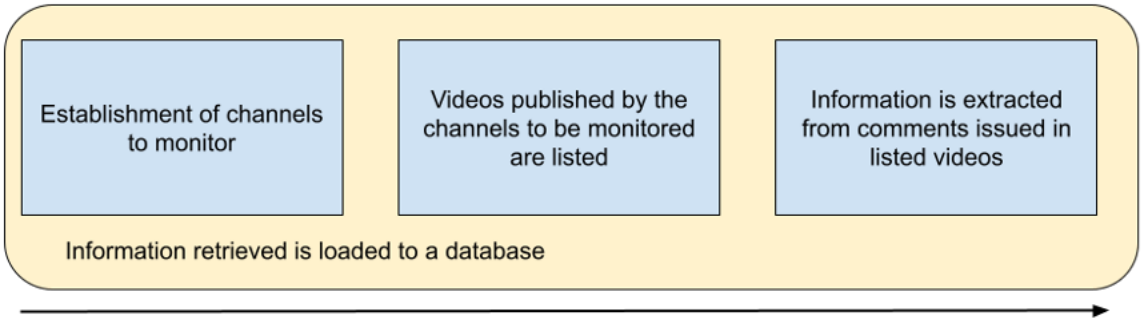


Figure 1. Stages of implementation of automatic classifier based on LLM.

The focus of this research is the hate speech disseminated by users on digital platforms. Accordingly, the user-generated comments constitute the unit of analysis. For the purpose of data collection, an infrastructure was developed that automatically extracts the comments from videos uploaded on a set of YouTube channels and systematically stores them in a database (see Figure 2).

Figure 2. Phases of monitoring system data extraction.



Selection and Collection

The project began with the identification and collection of data from digital platforms. However, due to the restrictions on accessing user activity information imposed by the majority of the country's most widely used social networks – such as Instagram, WhatsApp, Facebook and X/Twitter (Reuters Institute, 2021, 2022, 2023, 2024) – the analysis was focused on YouTube. This platform, in addition to having a broad reach in terms of users domestically, offers an official API that permits access to comments posted on videos, thereby facilitating the structured and automated collection of data (YouTube Data API, n.d.).

To select the channels to be monitored, a preliminary qualitative analysis was conducted to identify those where content that infringes on fundamental rights is disseminated, both within the videos and in the user comments. Although this categorisation does not strictly equate to hate speech, it is employed as a proxy to detect discriminatory patterns. The selection was based on criteria derived from previous studies and the expertise of human rights and communications specialists (Human Rights Watch, 2019; United Nations, 2020), thus ensuring a representative sample of the phenomenon. In total, 82 channels were listed, as detailed in Appendix 8.1.

The collection process was carried out using scraping techniques via the YouTube API provided by Google to automatically download the comments posted on videos published by the monitored channels for the period 2019–2024. This strategy ensured the compilation of an updated and coherent database. Moreover, the establishment of the selection criteria allowed for the delimitation of the data universe and focused the analysis on sources with a high incidence of messages containing hate speech.

As a result of the extraction process, 564,411 comments were collected between 3 January 2019 and 31 December 2024. A total of 82 channels and 9,062 videos with at least one associated comment were recorded. Additionally, 118,815 users were identified as the authors of the collected comments.

Table 1 presents the number of comments collected per year, with 2019 being the complete year with the fewest observations and 2024 registering the most observations among the selected channels. Furthermore, an analysis of the evolution of the users who posted these comments—as well as the number of videos and channels in which they were published—reveals sustained

growth throughout the period. This trend is only interrupted in 2021 in the case of the number of videos, a year that recorded a slight decrease relative to the previous year.

	2019	**2020**	**2021**	**2022**	**2023**	**2024**
Comments	22,047 (3.91%)	73,253 (12.98%)	61,598 (10.91%)	86,907 (15.40%)	95,860 (16.98%)	224,746 (39.82%)
Users	7,110	24,413	24,529	25,023	27,259	55,589
Videos	483	926	863	1,602	2,022	3,225
Channels	21	44	54	58	65	69

Table 1. Comments, author users, source channels and source videos, by year.

Once the data were collected, the next step was their preparation and labelling. A representative sample of comments was manually classified by experts according to the presence or absence of hate speech, following pre-established criteria. Subsequently, a generative language model, gemini-flash-002, was employed to automatically label the remainder of the comments. This model was fine-tuned through the use of a specific prompt designed to capture both explicit and implicit expressions of hate.

After the labelling stage, an evaluation and tuning process of the model was conducted. The results generated by the automatic model were compared with the manual labels to verify their accuracy. Based on this comparison, iterative adjustments were made to the prompt and system parameters in order to enhance performance and minimise errors. Finally, the system was integrated into a cloud-based infrastructure. The technological infrastructure utilised, which was based on cloud services such as Cloud Run and Cloud SQL, ensures that the system operates continuously, securely, and at scale, without the need for constant manual intervention.

FINDINGS FROM THE PRELIMINARY MODEL

An analysis of the 564,411 comments labeled using the gemini-flash-002 model and the prompt detailed in Annex 2 reveals a clear predominance of comments without hate speech. Specifically, 85.18% of the comments were classified as not containing hate speech, while only 14.82% were identified as containing it.

Hate speech throughout the period

The overall percentage of comments labeled as hate speech across all channels (Table 1) shows a sustained decline from 21.22% in 2019 to 11.13% in 2023, before rising again to 15.93% in

2024. This aggregate indicator highlights the general evolution of the phenomenon in the full dataset, capturing the relative weight of hate speech in the total conversation each year. Yet, while informative, this measure alone does not reveal how hate speech is distributed across different channels. To examine this variation, Table 3 presents yearly descriptive statistics at the channel level.

2019	2020	2021	2022	2023	2024
21.22%	17.52%	13.05%	13.38%	11.13%	15.93%

Table 2. Percentage of comments labeled as hate speech, by year

When looking at the average percentage of comments containing hate speech per channel, disaggregated by year, a clear downward trend emerges between 2019 and 2022, from 11.69% to 6.95% (Table 3). This indicator is useful because it captures the general prevalence of hate speech across all analyzed channels, smoothing out the influence of isolated extreme cases and allowing us to observe broader temporal shifts. In 2023 and 2024, however, the average rises again, reaching 9.58% in 2024, pointing to a resurgence in certain spaces. Despite this overall trend, the maximum values show that some channels consistently register exceptionally high levels of hate speech—up to 50% in 2021, 2023, and 2024—demonstrating the persistence of communities where this type of interaction constitutes a substantial share of total comments, far above the sample's average.

Year	Minimum	Mean	Standard Deviation	First Quartile	Median	Third Quartile	Maximum
2019	0	11,69	12,41	0	9,88	23,56	35,40
2020	0	8,73	10,44	0	5,03	15,37	34,29
2021	0	8,50	11,39	0	5,60	12,03	50,00
2022	0	6,95	7,93	0	4,96	12,05	33,33
2023	0	7,98	9,98	0	6,69	11,32	50,00
2024	0	9,58	10,25	0	7,72	14,29	50,00

Table 3. Descriptive statistics of the percentage of comments labeled as hate speech by channel and year.

These findings are reflected in the temporal distribution of channels with the highest percentages of hate speech (Figure 3). Some, such as Padres Objetores Chile and Pastor Ev. Javier Soto, appear recurrently across multiple years, although with varying levels. For instance, Padres Objetores Chile registers 32.2% in 2019, peaks at 50% in 2021, and continues to appear in subsequent years. In contrast, other channels show high levels only in specific periods—for example, Oikonomos Centro de Liderazgo Político Cristiano does not appear in earlier years but reaches 50% in both 2023 and 2024, while Pinochet Inmortal emerges in 2022 and maintains high levels thereafter. This suggests that while some channels consistently contribute to high proportions of hate speech, others emerge and fade depending on contextual and discursive dynamics each year.

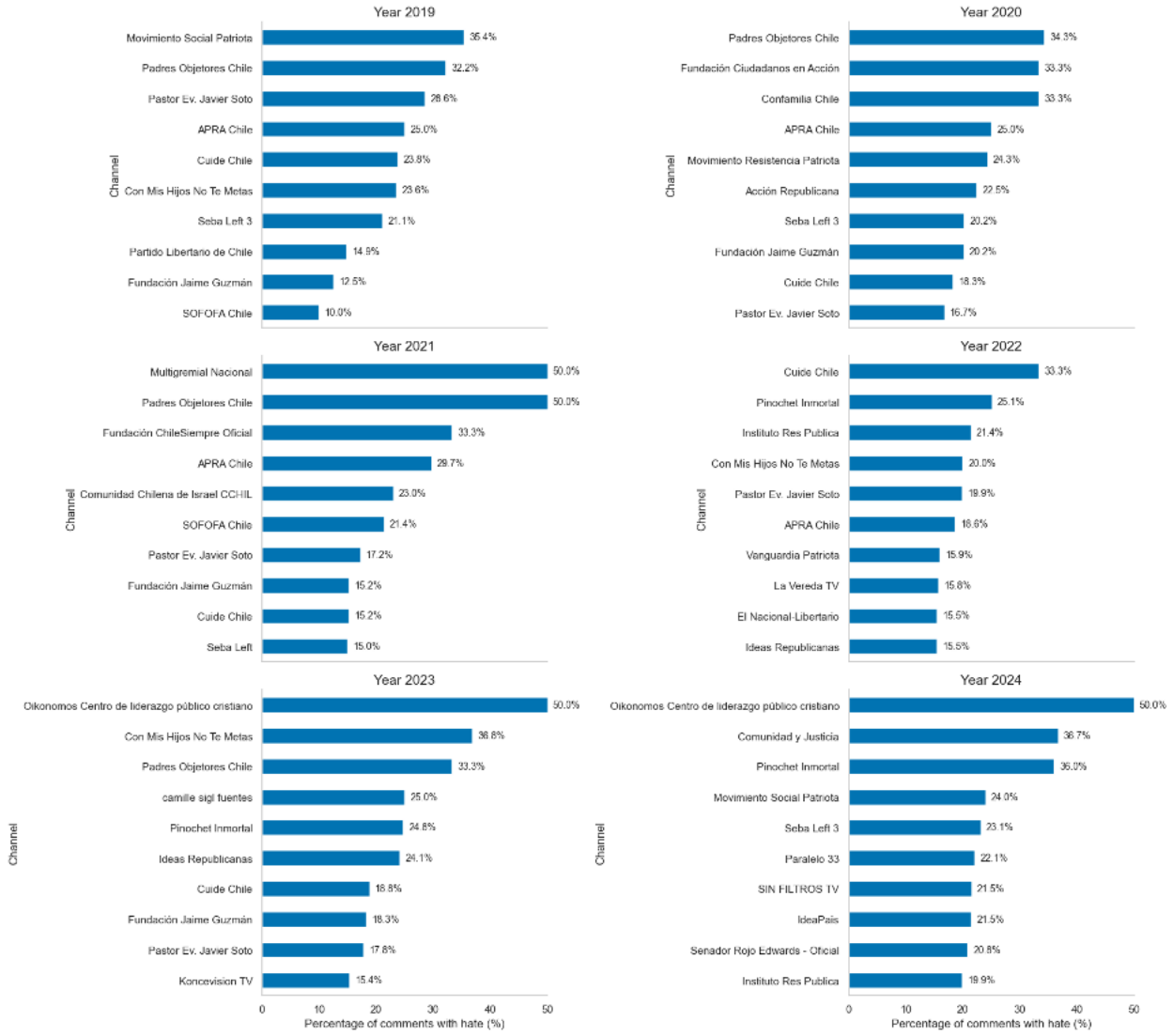


Figure 3. Ten channels with the highest percentage of comments with hate, by year.

The proportion of comments containing hate speech varies over time, displaying monthly fluctuations within each year (Figure 4). Overall, 2019 and 2020 register the highest levels, with several months surpassing 20% of comments labeled as hate speech. From 2022 onward, the values tend to decline, with most months remaining below 15%, and some even falling below 10%. However, in 2024, although the year begins with relatively low percentages, there is a noticeable increase toward the end of the year, peaking in December at 19.2%, suggesting a resurgence of the phenomenon.

The highest peaks in the entire period occur in November 2019 (23.07%) and February 2020 (22.24%), marking moments of heightened intensity in the presence of hate speech. Conversely, the lowest levels are observed in December 2021 (5.16%) and July 2023 (8.47%), indicating a

diminished presence during those periods. These dynamics are also reflected in the annual distribution (Figure 5), where the median percentage of hate speech comments shows a downward trend from 2019 through 2023, followed by an increase in 2024. Notably, variability is greater in the earlier years, progressively narrowing through 2023, but rises again in 2024—suggesting renewed fluctuations in the prevalence of hate speech over time.

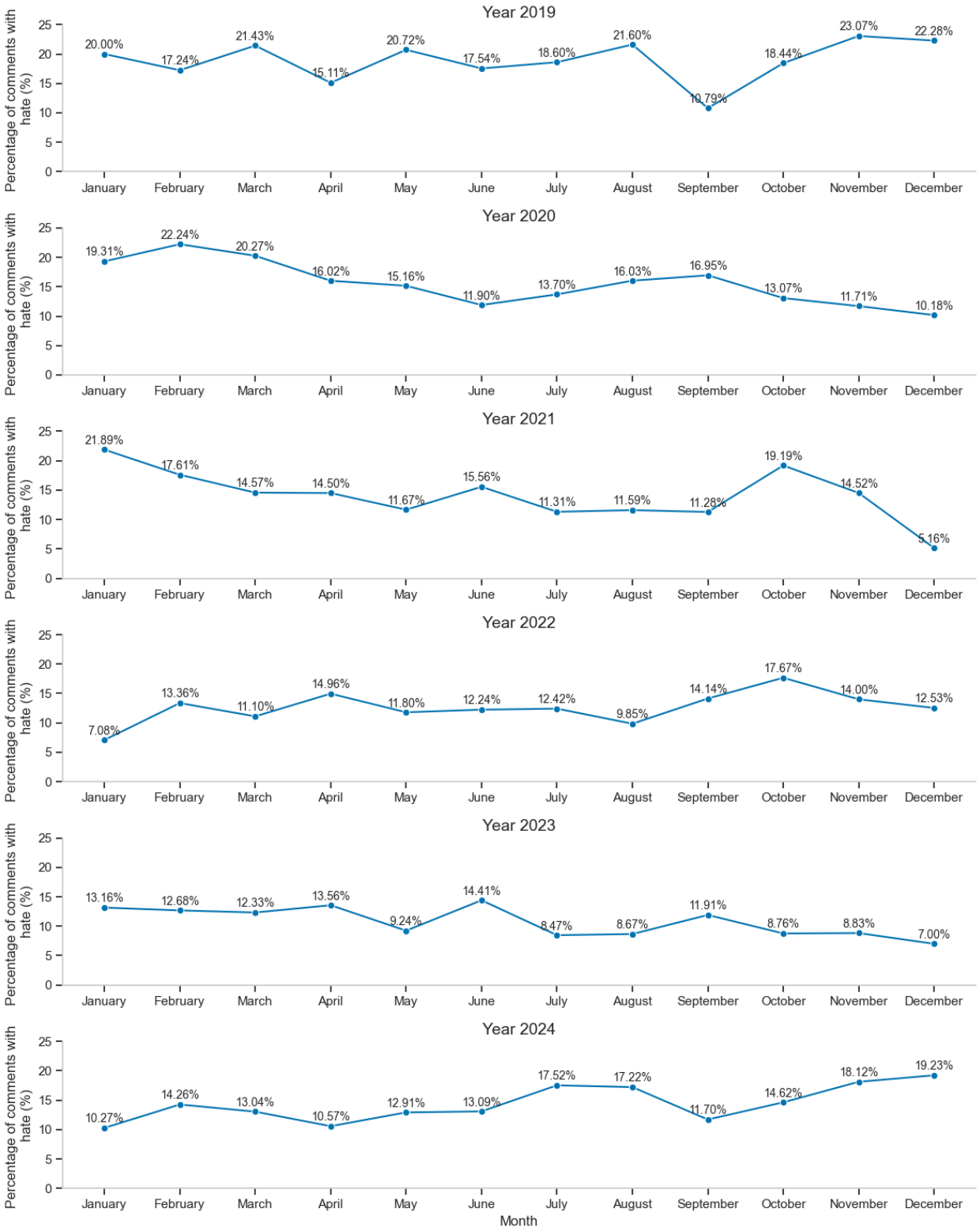


Figure 4. Percentage of monthly comments with hate, by year.

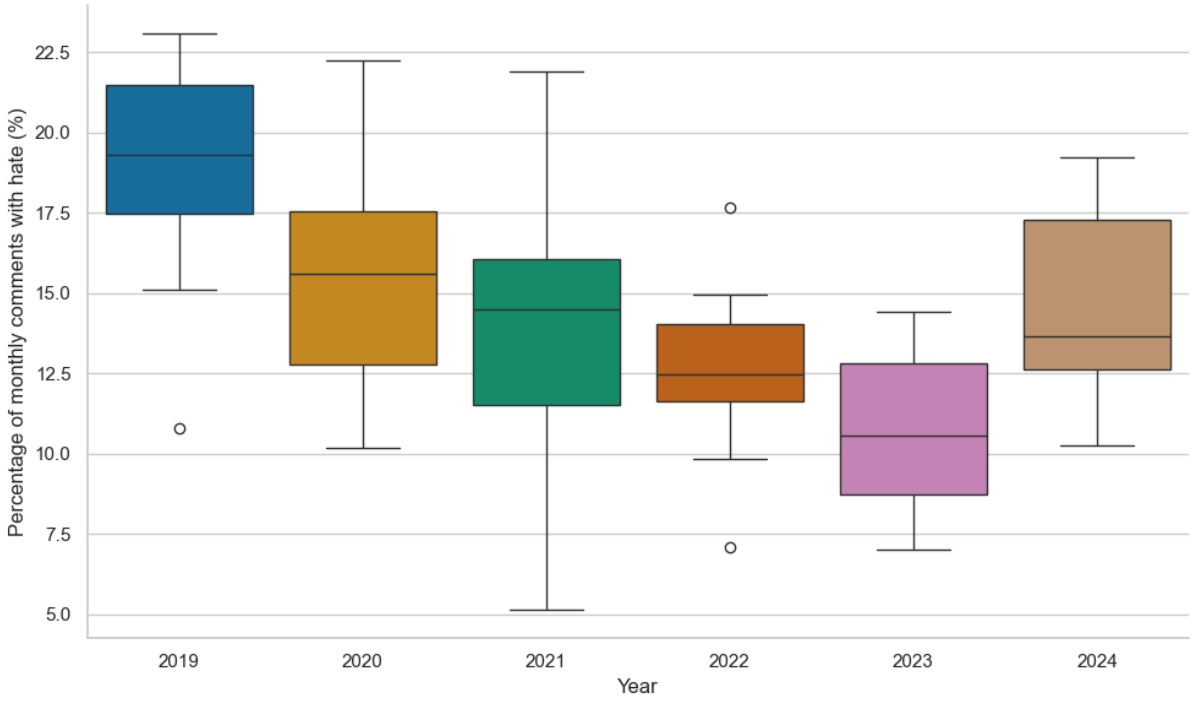


Figure 5. Distribution of monthly hate speech comments, by year.

Targets of hate speech

Following the narrowed definition of hate speech that guides this study—focused on discourse directed at historically marginalized groups—the classification prompt (Annex 8.1) also requested the identification of hate speech targets in comments where such discourse was detected. This allows for a subsequent filtering of comments containing hate speech, considering only those that are directed at the relevant target groups of interest.

Of the 83,635 comments labeled as containing hate speech, at least one target was identified in 99.93% of cases (83,577 comments). Based on a ***general definition of hate speech***, Figure 6 shows that the "left" is consistently the most frequently targeted group in hate-labeled comments. This trend aligns with the fact that the analyzed sample predominantly consists of channels associated with the right wing of the political spectrum. The proportion of comments targeting the "left" peaks in 2020 (50.72%) and remains the most prominent across all years, albeit with some fluctuations. This indicates a persistent pattern in which a substantial share of hate speech is directed at individuals or collectives affiliated with the political left.

Another consistently targeted group is "politicians," who show a marked increase in 2022 (36.94%) and 2023 (39.82%), suggesting a rise in polarized discourse directed at political figures during those years. The group "migrants" also stands out, with a notable upward trend from 3.61% in 2020 to 13.50% in 2024, reflecting growing hostility toward this population. Other groups—including the "religious community," the "LGBT community," and "women"—are also consistently targeted, although at lower levels, with annual proportions ranging between 2.53% and 9.91%. The "others" category—which includes a range of unspecified targets—remains among the most

frequently mentioned, reaching 14.19% in 2019 and 7.90% in 2024. In recent years, new categories such as “children and adolescents” and “state professionals” have emerged, suggesting a diversification of hate speech targets in response to evolving public and political discourses.

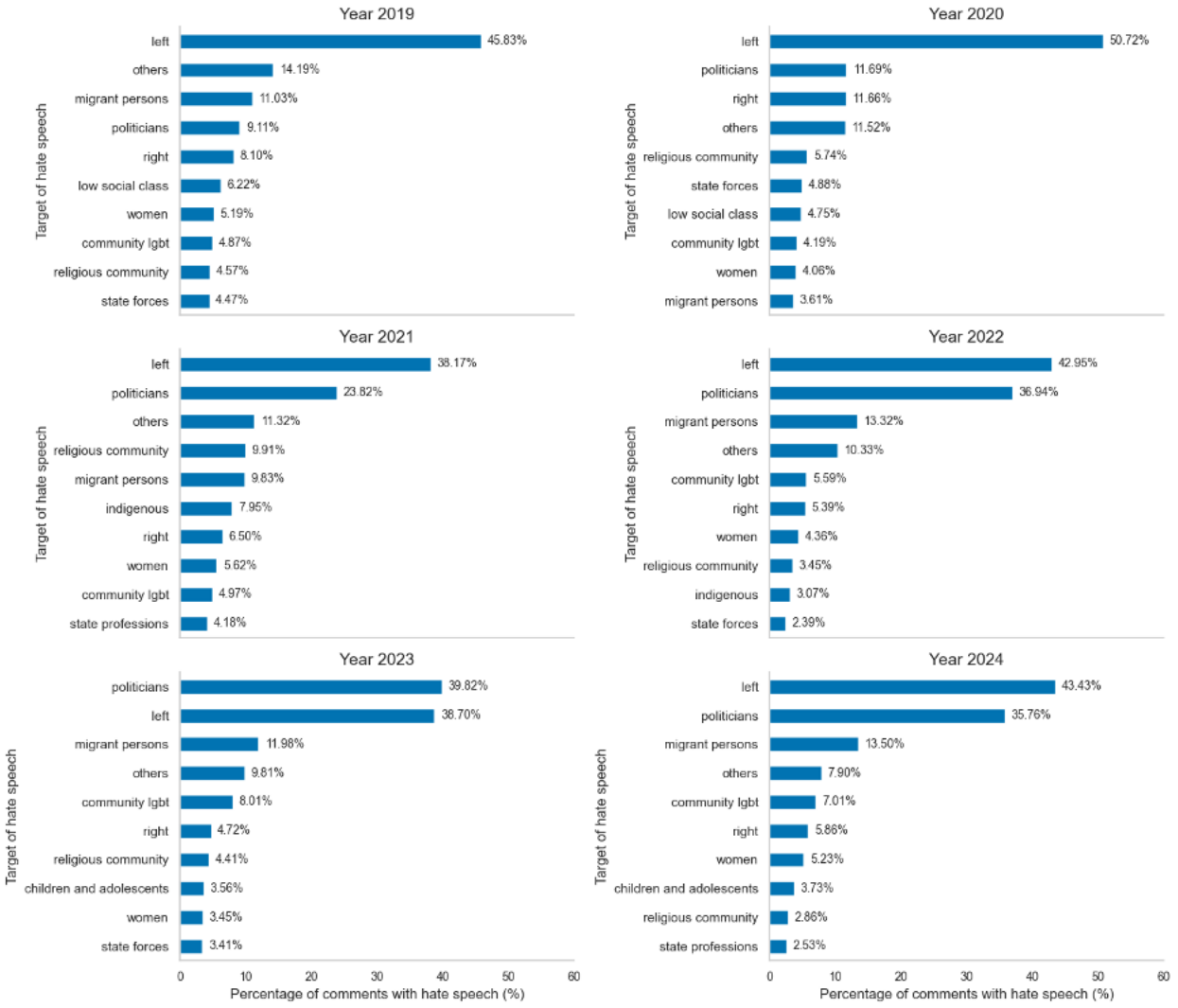


Figure 6. Ten most frequent hate targets in hate comments, by year.

Additionally, in 28.91% of comments containing hate speech (24,176 cases), more than one target was identified. To examine the co-occurrence of targets within this subset, the Jaccard index was calculated for all target pairings. As shown in Table 9, most pairs exhibit low or no similarity, with an average score of 0.01 and a median of 0.00, indicating that targets rarely appear together in the same comment. However, some combinations display significantly higher values, revealing specific patterns of shared hate speech toward certain groups. These findings suggest that, while hate speech typically focuses on individual targets, there are identifiable clusters of groups that are frequently attacked together.

Minimum	Mean	Standard Deviation	Quantile 1	Quantile 2	Quantile 3	Maximum
0.00	0.01	0.06	0.00	0.00	0.00	1.00

Table 4. Descriptive statistics of Jaccard Index.

When examining the targets that most frequently appear together (Figure 7), the strongest associations are observed between agnostics and atheists, Mapuche and police (Carabineros), and cultures and nations, all of which reach a Jaccard index of 1.0. This indicates that, within the subset of comments containing more than one target, whenever one of these groups is mentioned, the other is always mentioned as well. Other notable associations include animal rights activists and communists (0.50), communists and socialists (0.50), the left and politicians (0.46), Cubans and terrorists (0.33), human rights defenders and terrorists (0.33), the LGBT community and children and adolescents (0.17), and family and children and adolescents (0.17).

These combinations suggest specific patterns of hate speech, in which certain ideological or social categories tend to be targeted together in comments containing more than one hate speech target. Although exploratory, these findings reveal the presence of clusters of targets that are frequently attacked in tandem, highlighting the need to consider intersectionality when analyzing whom hate is directed toward.

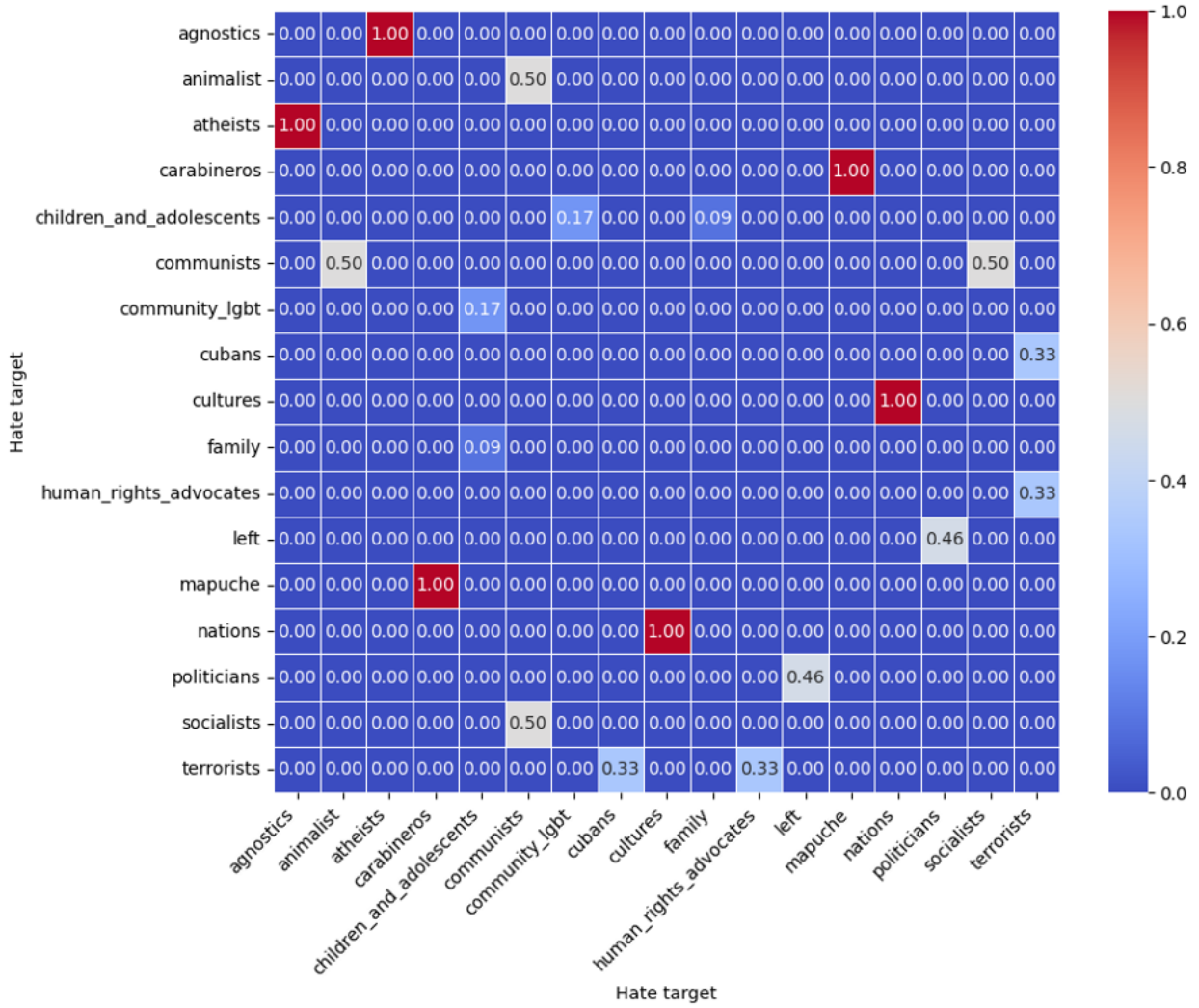


Figure 7. Ten combinations of targets with the highest Jaccard Index.

Based on the ***narrowed definition of hate speech*** adopted in this study—focused on historically marginalized groups—4.07% of all comments in the sample contain hate speech directed at one or more of the following groups: migrants, the LGBT+ community, women, people from low social classes, incarcerated individuals, Indigenous peoples, people with diseases, individuals with addictions, people with physical differences, people in precarious jobs, and homeless individuals.

Furthermore, 26.70% (22,329) of the comments labeled as containing hate speech target one of these groups. Among them, migrants consistently receive the highest proportion of hate comments across all six years analyzed, peaking at 13.50% in 2024. The LGBT+ community also ranks among the most frequently targeted groups, especially in recent years, reaching 8.01% in 2023. Women remain a prominent target throughout the period, with percentages ranging from 3.45% to 6.52%.

While most target groups appear with relatively low and stable proportions, Indigenous people stand out in 2021, when they received 7.96% of all hate-labeled comments—their highest level in

the series. This spike suggests a context-specific intensification of hateful discourse toward this group during that year. Other groups, such as people with diseases, incarcerated individuals, and homeless people, appear consistently but remain below 1% in most years.

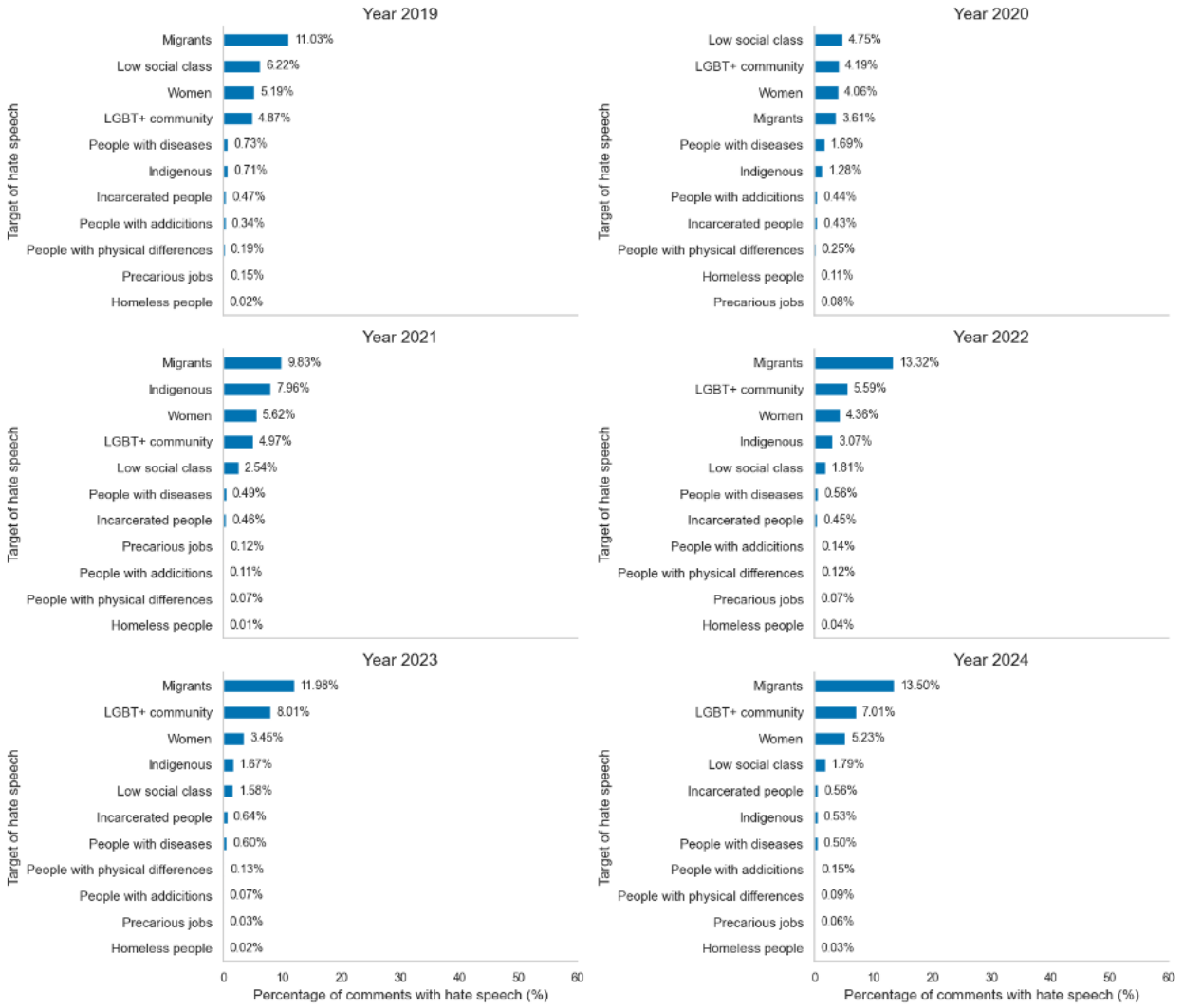


Figure 8. Percentage of comments with hate targeted towards historically marginalized groups, by year.

To shed light on how hate speech can simultaneously target multiple marginalized groups, Figure 9 illustrates the strongest pairwise associations based on the Jaccard Index. The heatmap is derived from a subset of 1,220 comments in which more than one historically marginalized group was mentioned alongside hate speech.

The highest co-occurrence is observed between women and the LGBT+ community, with a Jaccard Index of 0.47, indicating that nearly half of the instances where one group is targeted also involve the other. Other notable associations include migrants and low social class (0.26),

indigenous peoples and migrants (0.14), and people with diseases and people with addictions (0.14), suggesting specific clusters where hate speech narratives converge on multiple axes of marginalization.

Although most combinations show relatively low overlap—reflected in the overall low values of the matrix—these higher-scoring pairs reveal patterns of joint stigmatization. The association of migrants with multiple groups, including low social class, indigenous peoples, and women, suggests that this category often acts as a cross-cutting axis of exclusion.

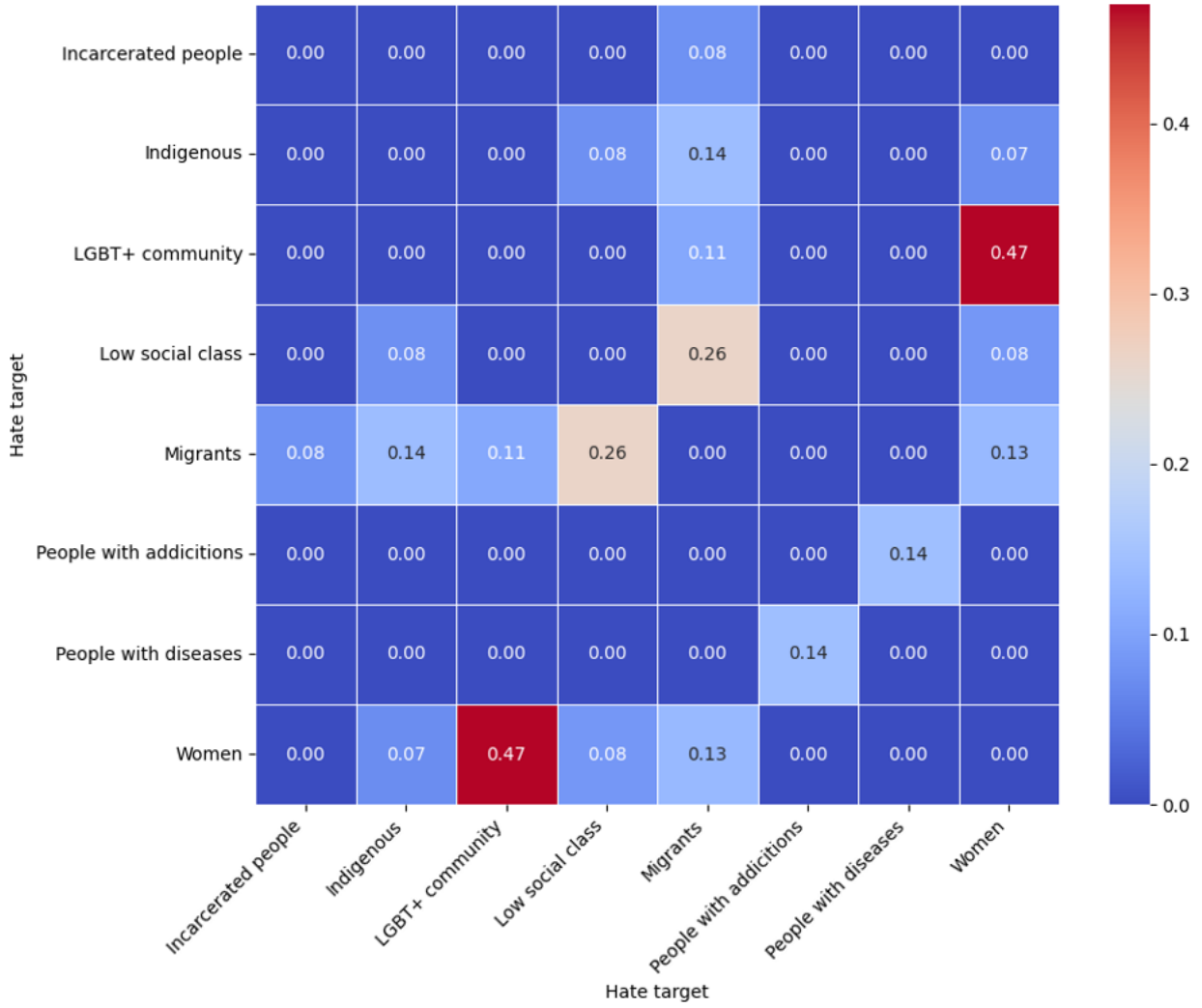


Figure 7. Ten combinations of targets with the highest Jaccard Index,considering only historically marginalized groups.

Strengths and Potential Challenges of the Model

One of the primary strengths of this approach is its ease of replication and scalability in other contexts. To do so, certain components must be adapted. Firstly, the data collection process should be adjusted according to the characteristics of each social network. For example, platforms such as Twitter, Facebook, or TikTok have different access mechanisms and various interaction formats, so it is necessary to employ the specific APIs or extraction tools for each.

Furthermore, it is essential that the model takes into account the linguistic and cultural particularities of the new environment. This involves modifying the classification prompt and, if necessary, conducting additional manual annotations with local examples. Incorporating expressions and idioms specific to the country or community helps to enhance the system's sensitivity, thereby avoiding misclassifications due to language or contextual differences.

Another key element for scaling up the replication is the technological infrastructure. The use of cloud platforms allows for the flexible handling of large data volumes. Services such as Cloud Run offer automatic scalability, which means the system can easily adjust to an increased workload without compromising efficiency. Additionally, tools like Terraform facilitate the configuration and maintenance of technological resources across different environments.

Finally, to ensure the system's sustainability, it is recommended to establish a process of continuous updating and feedback. This includes periodically reviewing the model's results, incorporating new data, and adjusting its parameters according to changes in social and discursive dynamics. It may also be beneficial to create collaborative networks with local teams, human rights organisations, and communication specialists who can help validate the findings and provide appropriate contextualisation for the system's decisions.

This study also highlights several important limitations, which can be grouped into three main areas: (1) the challenge of obtaining data that is representative of the context of interest; (2) the difficulties associated with automatic classification and prediction; and (3) the translation of substantive analytical insights into effective public policies that address the phenomenon.

About the first set of challenges, currently, some of the most widely used platforms in Chile—such as X (formerly Twitter), Instagram, TikTok, and Facebook—either impose high costs or require cumbersome authorization processes to access their official data (Reuters, 2021, 2022, 2023, 2024). Since each platform facilitates different modes of interaction and reaches distinct user demographics, these restrictions significantly constrain the generalizability of findings.

Moreover, opaque content moderation practices further compromise the reliability of available data. Hate speech content may be removed before data extraction, and the lack of transparent metrics on moderation activity limits the interpretability of the remaining data. Even on more accessible platforms such as YouTube, identifying which channels should be monitored to obtain content representative of the general population remains a methodological challenge. Channel relevance is tightly linked to the thematic nature of the content, and in this study, the focus was placed on conservative and far-right channels, which exhibited consistent hate speech targeting specific marginalized groups. Extending these findings to reflect broader hate speech dynamics across Chile would require a separate sampling strategy—one that incorporates a more diverse

range of channels and expands iteratively based on patterns observed in an initial exploratory sample.

The second group of challenges pertains to technical and computational limitations. The use of generative language models is resource-intensive, both in terms of computational power and energy consumption, raising concerns about scalability and sustainability. In addition, the imbalance in the dataset—with significantly more comments lacking hate speech than those containing it—affects model performance, requiring the application of inverse weighting techniques and the exploration of class-balancing strategies. In the case of temporal prediction, the low intrinsic variability of hate speech frequency and the absence of relevant external variables limit the models' forecasting capacity. These limitations suggest the need for future studies to integrate complementary data sources capable of explaining abrupt changes in hate speech activity.

The lessons learned through this project open several avenues for future research. First, it is essential to explore solutions that combine the high interpretive capacity of generative models with strategies for improving energy efficiency and reducing operational costs. This may involve the development of optimized versions of large language models (LLMs) or the use of model compression and quantization techniques that retain performance while reducing computational footprint. Second, incorporating external covariates into time series models emerges as a promising direction to enhance predictive capacity. Data on political events, shifts in media agendas, and socio-economic indicators could offer a more comprehensive view of hate speech dynamics and support more precise and timely interventions.

Finally, fostering dialogue among researchers, policymakers, and human rights practitioners is essential to translate these technological advances into effective public policy. Interdisciplinary collaboration can not only help optimize detection and prediction methods, but also ensure that interventions are ethically grounded and oriented toward the protection of fundamental rights. In this regard, the pilot project developed in Chile may serve as a replicable model for other Latin American contexts, contributing meaningfully to the reduction of extremist threats and the strengthening of democratic institutions.

MOVING FORWARD: NEGOTIATION AND LITIGATION STRATEGIES

Recognition and tracking of hate speech threats and patterns of extremist groups is a key first step to preventing future actions against vulnerable groups across Chilean society. In addition to this preliminary tracking and recognition phase, there are also more direct prevention and mitigation solutions to counter hate speech, and, more broadly, hate crimes. Below, we explore strategic litigation and negotiation strategies.

Strategic Litigation

Strategic litigation is a specific form of legal action through which a party deliberately initiates or intervenes in a case to achieve outcomes that extend beyond the interests of the immediate parties, influencing broader social and institutional structures (Amnesty International, 2025). Put differently, the aim is not only to obtain a judgment that punishes the individuals who have

committed violent or discriminatory acts or to sanction the lack of protection by the State, but to use individual cases for the courts to call for extensive institutional and social changes through concrete measures and policies. In other words, this tool not only helps establish important judicial precedents—often referred to as "leading cases"—but also contributes to structural reforms to prevent, investigate, and prosecute serious crimes.

Chile has existing legislative measures that can be drawn on to curb hate crimes and possibly even hate speech. In 2012, Chile passed Law 20609, also known as the Zamudio Law or Antidiscrimination Law. This resulted from a process that can be viewed through the lens of strategic litigation and mobilization. The Zamudio Law is an example of how one individual's case can be strategically adapted into a legislative measure that protects a much broader population. The law was passed as a result of the hate homicide against Daniel Zamudio (killed due to his sexual orientation) and the subsequent social, media and judicial actions that were unleashed as a consequence of Zamudio's death (Fundación Iguales, 2012). Three of the four young men who murdered Daniel Zamudio had a history of hate crimes against immigrants or other people of the LGTBI community, and had manifested some kind of closeness to extreme ideologies such as Nazism (Clarín, 2012). In addition, the crime revealed the cruelty towards the victim because of his sexual orientation, manifesting itself in acts of extreme violence such as torture, beatings, and even causing a scar in the shape of a swastika (MOVILH, 2021). For more than a decade the Zamudio Law has been deployed strategically—mostly by LGBT+ litigants and allies, but occasionally by opponents—in three broad ways: (1) iconic test cases that opened new doctrinal ground; (2) actions that forced public agencies and private insurers to deliver concrete services; and (3) a handful of "reverse-discrimination" suits that tried, with little success, to turn the statute against sexual-diversity rights. Far-right politician José Antono Kast, for instance, used the law to claim ideological discrimination after a union left him off a debate agenda. The civil court ultimately dismissed the action, but the case shows how conservative actors attempt "reverse-Zamudio" tactics.

The Zamudio Law is a comprehensive law against any act of discrimination and hatred against people on a wide range of grounds, including sex, gender, sexual orientation, race, religion, nationality, among others. This law creates a specific judicial process for acts of discrimination and modifies the Chilean Penal Code to regulate and typify different actions considered hate crimes. At the same time, however, the focus of the law is on hate when it takes the shape of acts of discrimination, which is not necessarily the case of hate speech. Hate speech, understood as stigmatizing and discriminatory expressions against certain groups of people due to their sex, gender, race, national origin or religion, among other factors, can be part of a broader context of social hostility that enables the commission and increase of hate crimes (IACHR Court, 2020). It is precisely this context of structural and persistent violence—manifested through speech targeting historically marginalized and vulnerable groups—that has sparked debates about the criminalization of certain forms of hate speech and their regulation in virtual spaces and public media. This is a groundbreaking frontier for strategic litigation, although a challenging one due to "free speech" limitations that exist in many countries. Balancing free speech with rights protections for minority groups from hate speech and extremism will require careful legal framing.

Another important distinction is that strategic litigation, at least in some parts of the world, can be brought locally and also internationally before human rights protection bodies, for example, before the Inter-American Human Rights System. The latter option has its advantages when it comes to obtaining sentences with concrete transformative orders and reparations. This is because through this type of litigation there is greater breadth and evidentiary flexibility to judge the commission of hate crimes, compared to a common criminal process.

The judicial precedents and reparations that can result from hate crime litigation are also important, and some of these have a symbolic-social application that goes beyond typical economic reparations (ReLeG, 2021). Examples of these measures are the changes to how court files are named and how the crimes are referred to within the sentences. These changes aim to make hate crimes visible by giving concrete names to these aberrant actions, such as "femicides", "transvesticides", "transfemicides", or "lesbicides", among others (UFEM MPF, 2023). Such symbolic reparations and visibilization measures occurred in judicial litigation in Latin American countries such as Colombia (Gordon, 2018), Mexico (OpenDemocracy, 2023) and Uruguay (La Cuarta, 2022). In the specific case of hate speech, these concrete names could be strategically adapted to the hate speech context.

Strategic Litigation in Chile: Recommended Takeaways
Through strategic litigation, parties can intervene in a legal case to achieve outcomes that extend beyond the interests of the individuals or groups directly involved in a crime (Amnesty International, 2025). Chile's Zamudio Law (or Antidiscrimination Law) was passed in response to the hate homicide of Daniel Zamudio and the social, media and judicial responses that occurred after his death (Fundación Iguales, 2012). The law is comprehensive and covers discrimination against people on a wide range of grounds, including sex, gender, sexual orientation, race, religion, and nationality, among others. It is an example of how one individual's case can be strategically adapted into a more extensive legislative measure that protects a much broader population.

At the same time, however, the Zamudio Law does not fully cover hate speech and instead was developed to focus more broadly on hate crimes. Strategic litigation practices can be implemented to adapt the Zamudio Law or draw attention to a new individual case that focuses specifically on hate speech. While the Zamudio law represents a valuable advance, legal actors and advocates can use strategic litigation as a tactic to further expand the parameters and reach of prevention measures by covering hate speech (in addition to broader hate crimes). Legal strategies to disincentivize hate speech are costly and can be controversial given their potential clash with people's freedom of speech. However, legislation and policies that clearly typify and systematically monitor hate speech are needed to keep track of its underlying strategies and public impact. Furthermore, strategic litigation and legislation could make more efforts to connect hate speech peaks with acts of discrimination and violence.

Negotiation

In addition to strategic litigation, we also explored negotiation strategies and contemplated how these might play out with perpetrators of hate crimes and hate speech. In conversation with Mark Freeman, Founder and Executive Director of the Institute for Integrated Transitions (IFIT), we

learned about negotiation possibilities. Mark discussed various aspects of negotiation strategies, emphasizing that negotiations often begin secretly, can be direct or indirect, and typically start on an exploratory basis. Mark also explained that negotiation is a means to reach agreements and should be combined with other strategies like litigation when necessary. He advised testing the willingness of the other side to negotiate through engagement, as assumptions about their interests may be unfounded. Lastly, he stressed the importance of negotiating with leadership-level representatives who can implement agreements, and cautions against negotiating with decentralized organizations where command and control structures are unclear.

With few exceptions, most of the cases of negotiation processes come from the US and Europe. While none of those cases occurred in Chile, the available evidence provides seven useful strategic dimensions:

1. **Humanisation and personal dialogue** rest on the insight that hateful ideologies often crumble when their holders meet the people they vilify. A vivid illustration is the work of African-American blues musician Daryl Davis, who in the 1990s began attending rallies of the white-supremacist Ku Klux Klan (KKK), asking members why they hated him and sustaining respectful, face-to-face conversations—sometimes even in his own home. Over time more than two hundred Klansmen surrendered their robes and renounced the organisation, demonstrating that patient, non-judgemental engagement can erode prejudice at its roots. Scandinavia's "Exit" deradicalisation programmes, launched in Norway (1997), Sweden (1998) and later Germany (2000), institutionalise the same principle: counsellors—many themselves reformed neo-Nazis—meet young skinheads in confidential settings, listen to grievances and help them build new identities through jobs, housing and peer support. In both settings the negotiation is intensely personal, converting an abstract enemy into a relatable human being and producing measurable local drops in hate activity.
2. **Mediators and credible messengers** form the relational bridge that makes such dialogue possible. Extremists usually distrust state officials, yet they will listen to voices they perceive as authentic. During Northern Ireland's peace process, Protestant clergy and community leaders with historical links to Loyalist paramilitaries carried government assurances that the region's British status would not change without majority consent; their credibility persuaded groups such as the Ulster Volunteer Force to enter talks and eventually disarm. In London, the counter-extremism think tank Quilliam enlisted its co-founder Maajid Nawaz, a former Islamist radical, to approach Tommy Robinson, leader of the far-right English Defence League (EDL). Because Nawaz "spoke the language of radicalism," Robinson accepted a face-saving public exit that briefly crippled the EDL's street movement. Exit Sweden relies on the same logic: reformed neo-Nazis whose "street pedigree" grants them authority with would-be defectors. Insiders-turned-moderates or respected community figures can therefore open doors that uniformed officials cannot.
3. **Balanced incentive structures** provide the practical leverage behind persuasion. Successful engagements pair carrots—legal leniency, vocational training, public recognition—with sticks such as looming prosecutions, military pressure or social isolation. In Colombia, commanders of the right-wing Autodefensas Unidas de Colombia (AUC, United Self-Defense Forces of Colombia) accepted reduced sentences and

reintegration stipends, but they also knew that refusal meant continued war and likely extradition to U.S. courts. Participants in Scandinavian Exit programmes escape prison prospects or stigma only if they show genuine disengagement. Robinson quit the EDL under the twin pressures of criminal charges and an increasingly violent base, yet gained media legitimacy and policy coaching. Crafting an honourable off-ramp allows extremists to preserve pride while abandoning violence, turning disengagement into a principled step rather than a humiliating defeat.

4. **Formal and informal dialogue** operate along a mutually reinforcing continuum. Informal contacts—quiet church meetings with Loyalist gunmen, Daryl Davis's private coffees with Klansmen, Quilliam's back-channel sessions—soften attitudes away from public scrutiny. Once trust exists, formal negotiations can codify commitments. The 1998 Good Friday Agreement locked in ceasefires and weapons decommissioning for Northern Ireland; Colombia's 2005 Justice and Peace Law set legal terms for AUC demobilisation. When sequencing works, private rapport paves the way for public accords, and the official framework supplies resources and verification for promises first floated in secret. If formal talks proceed without preparatory trust, or if informal deals never obtain legal backing, relapse risks rise sharply.
5. **Short-term and long-term outcomes** diverge depending on the strength of follow-up. Klan defections, the plunge in EDL rallies after 2013, the immediate drop in Loyalist bombings post-1998 and the sharp fall in Colombian massacres after 31,000 paramilitaries disarmed all testify to rapid gains. Yet durability is uneven. Northern Ireland's peace has largely held for over two decades, buttressed by power-sharing institutions and community programmes. Scandinavian Exit graduates seldom relapse thanks to continuing mentorship and social support. Conversely, Robinson soon resurfaced in new anti-Muslim projects, and many demobilised AUC mid-commanders re-armed as criminal bands. Lasting success requires sustained monitoring, economic reintegration and ideological work long after the headlines fade.
6. **Latin American specifics** underline how context shapes feasibility. Beyond Colombia's negotiations—first with the AUC and later, under President Juan Manuel Santos (2012-2016), with the Fuerzas Armadas Revolucionarias de Colombia (FARC-EP) guerrillas—the region offers few examples because hate violence there is frequently intertwined with decentralised gangs and organised crime devoid of cohesive leadership. Still, Colombia shows that when extremist violence has identifiable commanders and a political logic, structured negotiations backed by credible justice mechanisms can dramatically curb atrocities. For other Latin American countries, the lesson is to adapt these core principles—credible messengers, balanced incentives, dual-track dialogue and long-term reintegration—to local criminal dynamics, institutional capacity and social cleavages rather than importing foreign templates wholesale.

In the case of perpetrators of hate speech and hate crimes that are clearly identifiable, these strategies may work and may be useful. If and when, however, these groups become invisible and hide behind the anonymity of the internet, negotiation can become much more of a challenge. In these contexts, other strategies can be employed. A report by the Bipartisan Policy Center (2012) underscores that "[r]ather than hiding violent extremist content, a more productive

approach would be to promote websites and messages that counter it" (Bipartisan Policy Center 2012, 28). This can mean engagement with the private sector, such as encouraging internet companies "to donate sponsored links and share their knowledge about search-engine optimization with groups that oppose extremism" (Bipartisan Policy Center 2012, 29). This is also an option where the government has limited ability to monitor or take down content online due to freedom of speech acts. Another option is to promote websites that counter the content of the extremist pages and to widely share this information. In this regard, the Bipartisan Policy Center (2012, 29) report recommends that "Internet companies should be encouraged to donate sponsored links and share their knowledge about search-engine optimization with groups that oppose extremism." This is a solution that can be employed when blocking content or direct negotiations prove challenging.

Lola Aronovich, an Argentine academic and blogger working in Brazil, is one example of an activist who countered hate content through the promotion of the opposite content. Aronovich is known as "one of the pioneers of cyberfeminism on the Brazilian internet" (Tewari and Dasarathy 2023, 38). She runs an online blog called Escreva Lola Escreva ("Write, Lola, Write"), where she wrote about feminism and gained a large following, including "heated debate in the comments' section" (Nunes 2020, 8). Aronovich started to receive threats online and carefully documented them (Declercq 2018), but faced challenges when the local police refused to investigate the crimes (Lu et al 2022, 29). Despite these threats, Aronovich's blog page shows that she continued to post content from 1998 to 2025 (Aronovich, n.d.). This form of consistent engagement represents resistance even in the face of violence. Eventually, a new law called "Lei Lola" (Lola's Law) was signed in 2018, "allow[ing] police to investigate cases of online misogyny and render[ing] hate speech towards women illegal" (Declercq 2018). Luizianne Line, a federal representative and member of the Worker's Party, developed the law. This is another example of how the case of a specific individual can be leveraged to create legal protection mechanisms that extend far beyond the individual threats and violence that person faced. Lola's Law will help other women facing misogyny for years to come.

For Chile, where no recorded negotiations have yet been attempted with organised hate actors, the international evidence suggests a clear, staged pathway. First, the state and civil society would need to identify interlocutors whom potential extremists already respect—whether former militants, religious leaders, community elders, or respected figures from indigenous or nationalist sectors—because such messengers carry an authority that administrative officials rarely replicate. Once credible bridges exist, authorities can prepare "exit-with-honour" packages that blend legal leniency for non-violent infractions with concrete incentives such as vocational training, counselling, or educational grants, thereby offering an appealing alternative to continued extremism. Quiet pilot dialogues—modelled on Scandinavian hotlines or Daryl Davis-style face-to-face conversations—could be launched before violence hardens and public opinion polarises; these informal tracks would aim to chip away at prejudice and build trust that formal negotiations can later codify. Any Chilean approach must also embed strong accountability: truth-telling mechanisms, transparent monitoring, and meaningful reparations would help avert the Colombian experience of relapse and re-armament, while signalling to victims that justice is not being sacrificed. Finally, success depends on synchronising informal trust-building with the legal and political frameworks of formal engagement, ensuring that promises made in living-room

dialogues are ultimately underwritten by enforceable agreements, sustained funding, and visible governmental commitment. In short, adapting proven strategies to Chile's social fabric could transform nascent hate dynamics into structured opportunities for disengagement before they become entrenched.

Negotiation in Chile: Recommended Takeaways
Negotiation can complement litigation against hate actors when it follows clear rules: begin discreetly, test willingness, and involve leaders who control followers, as Mark Freeman advises. Evidence—mainly from the US and Europe—shows seven success factors. Humanising dialogue, exemplified by Daryl Davis's talks with Ku Klux Klan members and Scandinavia's "Exit" counselling, erodes prejudice. Credible messengers such as Protestant clergy in Northern Ireland or ex-radical Maajid Nawaz with the English Defence League open doors officials cannot. Balanced incentives—leniency plus pressure—underpinned Colombia's AUC demobilisation and Tommy Robinson's exit. Informal talks must feed enforceable accords; without long-term monitoring, gains fade, as Colombia's criminal "bandas" show. Latin America offers few cases, yet Colombia proves negotiation works when leadership is clear, while decentralised online hate demands counter-messaging tactics like those urged by the Bipartisan Policy Center and illustrated by Brazil's "Lei Lola." For Chile, credible interlocutors, honourable off-ramps, pilot dialogues and robust accountability could turn incipient hate dynamics into sustainable disengagement.

CONCLUSIONS

This project has tackled hate speech in Chile by not only identifying hate crimes, but by proposing mitigation and prevention solutions. The three methodological angles covered are: 1) a hate speech tracking system; 2) strategic litigation; 3) negotiation strategies.

First, the project demonstrates that it is possible to construct an effective tracking system for detecting hate speech by combining artificial intelligence technologies with local knowledge. Moreover, the model is not only applicable in Chile but can be relatively easily adapted to other digital platforms and countries, provided that the contextual particularities are respected and a continuous evaluation of its performance is maintained.

The results of this project demonstrate that the combination of generative language models and advanced machine learning techniques is an effective strategy for hate speech detection, despite the significant operational costs involved. The superior performance of gemini-flash-002 in terms of precision and sensitivity underscores the importance of investing in technologies capable of capturing linguistic complexity in critical contexts. At the same time, exploring lower-cost approaches and implementing time series methods open new possibilities for improving the predictive and operational capacities of monitoring systems.

Second, the project examined strategic litigation as a possible mitigation and prevention strategy in the Chilean case. Through strategic litigation, parties can intervene in a legal case to achieve outcomes that extend beyond the interests of the individuals or groups directly involved in a crime (Amnesty International, 2025). Chile's Zamudio Law (or Antidiscrimination Law) was passed in response to the hate homicide of Daniel Zamudio and the social, media and judicial responses

that occurred after his death (Fundación Iguales, 2012). The law is comprehensive and covers discrimination against people on a wide range of grounds, including sex, gender, sexual orientation, race, religion, and nationality, among others. It is an example of how one individual's case can be strategically adapted into a legislative measure that protects a much broader population.

At the same time, however, the Zamudio Law does not fully cover hate speech and instead was developed to focus more broadly on hate crimes. Strategic litigation practices can be implemented to adapt the Zamudio Law or draw attention to a new individual case that focuses specifically on hate speech. While the Zamudio law represents a valuable advance, legal actors and advocates can use strategic litigation as a tactic to further expand the parameters and reach of prevention measures by covering hate speech (in addition to broader hate crimes). Legal strategies to disincentivize hate speech are costly and can be controversial given their potential clash with people's freedom of speech. However, legislation and policies that clearly typify and systematically monitor hate speech are needed to keep track of its underlying strategies and public impact. Furthermore, strategic litigation and legislation could make more efforts to connect hate speech peaks with acts of discrimination and violence.

Third, as another mitigation and prevention measure, the project examined negotiation strategies and how these can be carefully constructed to engage with perpetrators of hate speech and hate crimes. Negotiation can complement litigation against hate actors when it follows clear rules: begin discreetly, test willingness, and involve leaders who control followers. Negotiation with perpetrators of online hate speech in Chile is a challenging feat, particularly because many of these groups or individuals are hiding behind the anonymity of the internet. When groups are difficult to identify, other strategies can be employed, such as engaging with the private sector to promote pages that counter hate speech and encouraging dialogue with the private industry to "share their knowledge about search-engine optimization with groups that oppose extremism." (Bipartisan Policy Center 2012, 29). This includes promoting the activities of particular individuals and organizations who do this important counterextremist and counter hate speech work. The example of Lola Aronovich, an Argentine academic and blogger working in Brazil, is offered as one such example. For Chile, credible interlocutors, honourable off-ramps, pilot dialogues and robust accountability could turn incipient hate dynamics into sustainable disengagement.

These conclusions, along with the limitations and recommendations identified, lay the groundwork for future research and for the development of public policies that strengthen democratic resilience and protect vulnerable communities in an increasingly polarized and digitalized global environment.

REFERENCES

Amnesty International. (2023). *Amnesty International Report 2023*. Amnesty International.

Amnesty International (2025). *Strategic litigation*. Amnesty International. https://www.amnesty.org/es/strategic-litigation/

Aronovich, L. (n.d.). Escreva, Lola, Escreva. Blogspot. Retrieved July 21, 2025, from https://escrevalolaescreva.blogspot.com/

Article 19. (2020). *Free expression in an age of hate*. Article 19.

Bipartisan Policy Center. (2012). *Countering online radicalization in America*. Bipartisan Policy Center.

Brown, M. (2018). *Understanding political hate: Language and symbolism in the digital age*. Journal of Communication, 68(1), 35–56.

Clarín (2012, March 30). *Neo-Nazi killers of young Chilean gay man, complicated*. Clarín. https://www.clarin.com/mundo/asesinos-neonazis-joven-chileno-complicados_0_rJWDxpS2D7g.html

Council of Europe. (1997). *Countering hate speech in Europe: A unified approach*. Council of Europe Publishing.

Declercq, M. (2018, May 10). *The most notorious misogynist in Brazil is behind bars, again*. VICE. https://www.vice.com/en/article/the-most-notorious-misogynist-in-brazil-is-behind-bars-again

Delgado, R., & Stefancic, J. (2017). *Must we defend hate speech?* New York University Press.

Douglas, K. M., Sutton, R. M., & Cichocka, A. (2017). *The psychology of conspiracy theories*. Current Directions in Psychological Science, 26(6), 538–542.

Freedom House. (2022). *Freedom in the World 2022*. Freedom House.

Fundación Iguales (2012). *Ley Antidiscriminación - Iguales.* https://iguales.cl/incidencia-politica/ley-antidiscriminacion/

Gelber, D. (2017). *The dynamics of hate speech: Rhetoric and polarization in society*. Journal of Communication, 67(3), 459–478.

Gelber, D. (2019). *Hate speech in the digital age: Policy and research perspectives*. MIT Press.

Gordon, D. (2018, December 19). "This was a person who deserved to die. She didn't belong in this life" Analysis of the first femicide conviction of a trans woman in Colombia-Race and Equality. *Race and Equality - The Institute on Race, Equality and Human Rights Works with Counterparts in the Afro-Descendant and Indigenous Communities and the LGBTI Movement in Brazil, Colombia, Cuba, and Peru, as Well as with Regional Afro-Descendant and LGBTI Networks in the Americas.* https://raceandequality.org/es/resources/esta-era-una-persona-que-merecia-morir-no-debia-estar-en-esta-vida-analisis-de-la-primera-condena-por-feminicidio-de-una-mujer-trans-en-colombia/

Harel, A. (2021). Hate Speech. En A. Stone & F. Schauer (Eds.), The Oxford Handbook of Freedom of Speech. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198827580.013.25

HateLab. (s.f.). *Homepage*. Retrieved from https://hatelab.net/

Human Rights Watch. (2019). *World Report 2019*. Human Rights Watch.

I/A Court H.R. (2020). *CASE OF AZUL ROJAS MARÍN Y OTRA VS. PERU.* https://www.corteidh.or.cr/docs/casos/articulos/seriec_402_esp.pdf

La Cuarta (2022, March 9). *Uruguayan justice condemns for the first time the crime of transfemicide | Crónica*. La Cuarta. https://www.lacuarta.com/cronica/noticia/justicia-uruguaya-condena-por-primera-vez-el-delito-de-transfemicidio/K7ATSGG2XFF6LJAZ6UVWNKBV5E/

Lu, A., Posetti, J., & Shabbir, N. (2022, November). *Legal and normative frameworks for combatting online violence against women journalists*. UNESCO.

Matsuda, M. J. (1993). Public response to racist hate speech: Considering the victim's perspective. *University of Chicago Legal Forum, 1993*(1), 139–168.

Moonshot (s.f.). Moonshot Team. Retrieved from https://moonshotteam.com/

Mossie, Z. & Wang, J. (2020). Vulnerable community identification using hate speech detection on social media. Information Processing & Management, 57(3), 102087. https://doi.org/10.1016/j.ipm.2019.102087

MOVILH. (2021). *Daniel Zamudio is remembered 9 years after the homophobic crime that took his life - Movilh Chile.* https://www.movilh.cl/recuerdan-a-daniel-zamudio-a-9-anos-del-crimen-homofobico-que-le-quito-la-vida/

OpenDemocracy (2023). *Transfeminicide in Mexico: Justice for Natalia Lane*. openDemocracy. https://www.opendemocracy.net/es/transfeminicidio-mexico-viaje-justicia-natalia-lane/

Nunes, R. (2020). Outlining the history of cyberactivism in Brazil. *Internet Histories*, *4*(1), 1–17.

Parekh, B. (2012). *Rethinking multiculturalism: Cultural diversity and political theory* (Updated ed.). Harvard University Press.

ReLeG. (2021). *Strategic Litigation in Gender Violence: Experiences from Latin America*. ReLeG. https://www.releg.red/informes

Rosenberg, B. (2012). *The dangerous echo: How hate speech can lead to violence*. Palgrave Macmillan.

Ruscher, J. B. (2025). *Hate speech*. Cambridge University Press. (Elements in Applied Social Psychology).

Southern Poverty Law Center. (s.f.). *Homepage*. Retrieved from https://www.splcenter.org/

Tewari, A., & Desarathy, A. (2023, February). *Feminist perspectives on social media governance*. IT for Change & Internet Lab.

UFEM MPF. (2023). *Transfemicides, transvesticides and bias crimes in Argentina (2016-2021) - Analysis of 12 sentences 10 years after the Gender Identity Law (2022).* https://www.fiscales.gob.ar/acciones-genero/transfemicidios-travesticidios-y-crimenes-por-prejuicio-en-argentina-2016-2021-analisis-de-12-sentencias-a-10-anos-de-la-ley-de-identidad-de-genero-2022/

UNESCO. (2024). Hate speech and misinformation: Toward a framework for policy and education. UNESCO Publishing.

United Nations. (1948). Convention on the Prevention and Punishment of the Crime of Genocide. https://www.un.org/en/genocideprevention/genocide-convention

United Nations. (1966). International Covenant on Civil and Political Rights.

United Nations. (2019). Guidelines on hate speech and hate crime. United Nations Human Rights Office of the High Commissioner.

United Nations. (2020). *United Nations strategy and plan of action on hate speech: Detailed guidance on implementation for United Nations field presences*. https://digitallibrary.un.org/record/3889286/files/UN_Strategy_and_PoA_on_Hate_Speech_Guidance_on_Addressing_in_field.pdf

Waldron, J. (2012). *The harm in hate speech*. Harvard University Press.

Reuters Institute. (2021). *Chile: Digital News Report 2021*. Reuters Institute for the Study of Journalism, University of Oxford. https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2021/chile

Reuters Institute. (2022). *Chile: Digital News Report 2022*. Reuters Institute for the Study of Journalism, University of Oxford. https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2022/chile

Reuters Institute. (2023). *Chile: Digital News Report 2023*. Reuters Institute for the Study of Journalism, University of Oxford. https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2023/chile

Reuters Institute. (2024). *Chile: Digital News Report 2024*. Reuters Institute for the Study of Journalism, University of Oxford. https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2024/chile

YouTube Data API. (s.f.). *YouTube Developers*. Retrieved from https://developers.google.com/youtube/v3

Zulver, J., & Payne, L. (2023). Righting rights, righting wrongs: Final reflections. In L. A. Payne, J. Zulver, & S. Escoffier (Eds.), *The right against rights in Latin America* (pp. 229–245). Oxford University Press.

**AI was used in the development of the automated tracking tool and in the elaboration of the final report.