SocArXiv

Preprint : 2026-04-01 10.31235/osf.io/vpzmq$_v$2

YouPol: A Collaborative Research Infrastructure and Database for Political Content on YouTube and TikTok

Antoine Lemor
Université de Sherbrooke, CIRST, RFICS

Tristan Boursier
Sciences Po Paris & Université du Québec à Montréal

Abstract

We present YouPol (YouTube and TikTok Political Observatory and Longitudinal database), a permanently updated research infrastructure that captures what political content creators actually say on video platforms. As of April 2026 and continuously expanding, the corpus comprises 25,397 videos from 68 channels across France and Quebec, with full speaker-diarized transcripts (645,738 segments, 3.18 million annotated sentences) and 7.7 million archived comments. The infrastructure includes an independent transcription pipeline that produces high-quality transcripts regardless of platform-provided captions, and an LLM-in-the-loop annotation framework built on the open-source LLM Tool platform (Lemor et al., 2025) that can train sentence-level classifiers for any research project, with current projects covering political content detection, far-right ideology, gendered rhetoric, and neo-reactionary discourse. To produce transcription and metadata updates in real time, YouPol also introduces the YouPol Collaborative Computing Network (YCCN), which allows any collaborating researcher to contribute processing capacity from their own machine, freeing the observatory from dependence on institutional computing clusters. YouPol addresses four gaps in the literature: (1) the ideological substance of political video content remains empirically inaccessible through metadata alone; (2) content deletion and deplatforming erase material before researchers can study it; (3) longitudinal engagement dynamics are underexploited; and (4) no existing dataset preserves comments over time or tracks their deletion. The observatory has already preserved 2,305 videos and three entirely deleted channels that are no longer available on the platform. The dataset and API are available at `data.you-pol.com`.

Keywords: computational social science, political discourse, YouTube, speech transcription, NLP, research infrastructure.

Word count: 6091

Introduction

On March 12, 2025, the French far-right YouTube channel FDENEWS, which had accumulated 1,755 videos, 68.3 million views, and 103,000 subscribers over nearly a decade, was permanently deleted by YouTube for repeated violations of its community guidelines. For any researcher who had not anticipated this event, the data was irretrievably lost. Thousands of hours of political commentary, hundreds of thousands of user comments, and a longitudinal record of one of France's most prominent far-right media outlets simply ceased to exist. This type of event underscores a clear need for research infrastructures capable of continuously collecting, processing, and preserving political video content *before* platforms remove it. The observatory presented in this paper, YouPol, addresses this need. Because it had been monitoring, among other channels, FDENEWS since its inclusion in the corpus, the deletion had no effect on the analytical record. Every video had already been downloaded and transcribed, along with every comment and metadata. The channel's disappearance from YouTube was itself logged as a data point.

This example shows, in practice, the difficulties confronting computational social science in the study of political discourse on (video) platforms. The content that researchers most need to study (radical, transgressive, politically consequential) is precisely the content most likely to be removed by platforms. Without infrastructure designed for continuous, pre-emptive collection, researchers are left studying only what platforms allow to remain visible. This introduces a survivorship bias that undermines the validity of any analysis of political extremism online. This is what YouPol addresses by operating as a permanent observatory that continuously collects, transcribes, diarizes, and annotates political video content across platforms, while preserving material that platforms subsequently remove.

Video-sharing platforms have become central venues for political communication, ideological mobilization, and public deliberation (Munger and Phillips 2022). YouTube hosts an ecosystem of political content creators whose collective output rivals traditional media in reach (Rieder, Coromina, and Matamoros-Fernández 2020). TikTok's algorithmic affordances have further transformed how political content circulates, particularly among younger audiences (Gerbaudo 2026; Guinaudeau, Munger, and Votta 2022). Yet despite the political significance of these platforms, the computational social science community lacks transcript-level, longitudinally maintained datasets that capture what is actually *said* in political videos (Lazer, Pentland, Watts, Aral, Athey, Contractor, et al. 2020). Most existing studies rely on titles, metadata, or recommendation patterns and therefore miss the substantive ideological content that only transcription and fine-grained textual analysis can reveal.

This gap is especially acute outside the anglophone context. While English-language political YouTube has received sustained attention, from radicalization pathway audits (Ribeiro, Ottoni, West, Almeida, and Meira 2020) to supply-and-demand analyses of right-wing content (Munger

and Phillips 2022), francophone political ecosystems remain understudied. Prior work on French political content creators has largely relied on video titles and metadata, overlooking the ideological substance (Boursier 2025).

The "post-API age" of social media research (Freelon, Monzer, Jeon, Moy, and Williams 2024) has compounded this challenge. Platform restrictions on data access (Bruns 2019; Tromble 2021), the deprecation of CrowdTangle, and the limitations of TikTok's Research API (Pearson, Silver, Robinson, Azadi, Schillo, and Kreslake 2024) make it increasingly difficult for researchers to build and maintain comprehensive datasets (Chen, Sherren, Lee, McCay-Peet, Xue, and Smit 2024; Ohme, Araujo, Boeschoten, Freelon, Ram, Reeves, and Robinson 2024). When platforms do provide research APIs, systematic audits have revealed significant biases that threaten validity (Rieder, Padilla, and Coromina 2025; Bai and Gu 2026).

In this paper, we present YouPol (YouTube and Tiktok Political Observatory and Longitudinal database), a research infrastructure designed to address these challenges. YouPol is a *permanent observatory* that continuously collects, transcribes, diarizes, and annotates political video content across platforms. Its contributions are threefold. First, YouPol constitutes, to our knowledge, the first transcript-level political video dataset of this magnitude. As of April 2026 and continuously expanding, it comprises 23,712 YouTube and 1,685 TikTok videos with speaker diarization, 7.7 million comments, and 645,738 speaker segments split into 3.18 million annotated sentences from 68 channels across France and Quebec.

Second, the observatory operates as permanent, collaborative infrastructure rather than a one-time collection. Any researcher can join the YouPol Collaborative Computing Network (YCCN) by installing a dedicated application developed by the authors on their machine, which then acts as an autonomous transcription and processing node. The YCCN continuously scans channels, processes new content, and preserves material that platforms subsequently remove, including deleted videos, suppressed channels, and moderated comments. While the initial processing was performed on the high-performance computing infrastructure of the Digital Research Alliance of Canada, production has since moved entirely to the YCCN, making the observatory fully autonomous and independent of institutional computing clusters. Something quite important in the current environment of restricted data access, and sometimes hostility toward researchers.

Third, an LLM-in-the-loop annotation framework built on the open-source LLM Tool platform (Lemor, Dinan, and Gilbert 2025) produces CamemBERT classifiers for sentence-level detection of political ideology, gendered rhetoric, and neo-reactionary discourse. The political detection classifier (`detect_pol`) and the full annotation pipeline are in production. TikTok integration is in progress, with priority given to the TikTok accounts of creators already tracked on YouTube. The approach follows the knowledge-distillation paradigm validated by recent methodological work (Pangakis and Wolken 2024; Gilardi, Alizadeh, and Kubli 2023; Alizadeh, Kubli, Samei, Dehghani, Zahedivafa, Bermeo, Korobeynikova, and Gilardi 2025).

1. Related Work

1.1. Existing Political Discourse Datasets: Coverage and Limitations

Computational social science has produced several important political discourse datasets, but none captures the actual spoken content of political videos on social media. Media Cloud (Roberts, Bhargava, Valiukas, Jen, Malik, Bishop, et al. 2021) collects news articles from 60,000+ sources with continuous updates, yet covers text-based journalism, not video platforms. ParlSpeech V2 (Rauh and Schwalbach 2020) offers 6.3 million parliamentary speeches from nine democracies, and CoCoHD (Hiray, Liu, Song, Shah, and Chava 2024) provides 32,697 U.S. congressional hearing transcripts; both contain full text, but of institutional proceedings rather than informal political communication on social media.

On more informal political communication media such as YouTube, several influential studies have tended to rely exclusively on metadata or comments. Ribeiro et al. (2020) released metadata and comments from over 330,000 videos, Ledwich and Zaitsev (2020) classified roughly 800 political channels, and Rauchfleisch and Kaiser (2020) analyzed far-right audience networks. However, none of these includes video transcripts. Interestingly, Lai, Brown, Bisbee, Tucker, Nagler, and Bonneau (2024) used transcripts from a subset of political YouTube videos to estimate channel ideology, but noted severe data scarcity, since transcripts were available for only a minority of their sample. This data scarcity is particularly acute in non-anglophone contexts: no transcript-level corpus of political video content exists for the French or other language ecosystems, where prior work has relied almost exclusively on video titles and metadata (Gilliotte 2024; Rauchfleisch and Kaiser 2020).

The few studies that do incorporate transcripts rely on platform-provided captions. Sosnovik, Violot, and Humbert (2025) collected over 100,000 French YouTube transcripts from the 2024 elections via platform-provided transcripts and applied LLM-based topic labeling. That corpus, however, covers a single election period, lacks speaker diarization, offers topic-level rather than

Table 1: YouPol compared with existing political discourse datasets.

Dataset	Source	Content type	Scale	Live
YouPol	YT + TikTok	Transcripts + diarization (French)	25.4K videos, 7.7M comments	yes
Sosnovik et al. 2025	YouTube	Auto captions (French)	100K videos	no
Pinto et al. 2025	TikTok	AI transcripts (English)	3.14M IDs	no
Ribeiro et al. 2020	YouTube	Metadata only (English)	330K videos, 72M comments	no
ParlSpeech V2	Parliament	Official transcripts (9 langs.)	6.3M speeches	no
Media Cloud	News	Articles (20+ langs.)	2B stories	yes

sentence-level annotation, and constitutes a one-time snapshot. A fundamental limitation of this approach is that platform-generated subtitles are often of poor quality, are unavailable for many videos (particularly older or less popular content), and do not identify individual speakers. YouPol addresses this through a state-of-the-art independent transcription pipeline that combines audio processing and cleaning via neural source separation (Demucs), state-of-the-art speech recognition (Whisper large-v3), and speaker diarization (pyannote.audio). This produces consistently high-quality, speaker-labeled transcripts regardless of whether the platform provides captions or not.

On TikTok, Pinto, Bickham, Salkar, Menezes, Luceri, and Ferrara (2025) gathered 3.14 million video IDs with AI-generated transcripts on TikTok from the 2024 U.S. election but did so without NLP annotation or continuous updating. Solovev, Drolsbach, Demirel, and Pröllochs (2026) analyzed over 25,000 TikTok videos from the 2025 German federal election and found that negative emotional appeals significantly increase engagement, though their analysis relied on computational content coding rather than independent transcription. Table 1 positions YouPol relative to these resources. No existing dataset combines full video transcripts produced by an independent transcription pipeline, speaker diarization, sentence-level NLP annotation, continuous updating, and content preservation.

1.2. Four Unresolved Gaps: Transcripts, Content Preservation, Engagement Trajectories, and Comment Moderation

Research on political content on video platforms has produced important findings but leaves four fundamental gaps unaddressed. *First, the reliance on metadata means that the ideological substance of political video content remains empirically inaccessible.* Influential studies on YouTube have examined user migration toward extreme content (Ribeiro et al. 2020), the concentration of radical consumption among a small audience (Hosseinmardi, Ghasemian, Clauset, Möbius, Rothschild, and Watts 2021), the supply-and-demand dynamics of right-wing content (Munger and Phillips 2022), algorithmic recommendations (Haroon, Wojcieszak, Chhabra, Liu, Mohapatra, and Shafiq 2023), audience overlap in far-right networks (Rauchfleisch and Kaiser 2020), and migration between anti-feminist and far-right communities through 300 million comments (Mamié, Horta Ribeiro, and West 2021). On TikTok, research has addressed the emergence of algorithmically formed "social interest clusters" (Gerbaudo 2026), the role of virality over follower counts (Guinaudeau et al. 2022), and the effect of negative emotional appeals on engagement (Solovev et al. 2026). While each of these contributions advances our understanding of political video platforms, they all rely on metadata, comments, or behavioral traces. None has access to what creators actually say in their videos, which leaves the ideological content of political speech empirically out of reach.

The inability to access video content is especially consequential for studying metapolitics, a

strategy through which far-right actors seek to normalize their ideology by gradually shifting mainstream cultural and discursive frames to make their ideas more popular (Boursier 2026; Schilk 2025). Rather than pursuing direct electoral outcomes, actors that pursue a metapolitical strategy operate through sustained discursive work that builds alternative vocabulary, reframes political issues, and circulates ideological narratives across platforms such as YouTube or TikTok (Ganesh 2025). These platforms lend themselves to this strategy precisely because they are not primarily political spaces. Creators embed ideological arguments within formats (commentary, reaction videos, interviews) that audiences consume as entertainment, not as exclusively as political per se. This strategy is inherently discursive and "textual", in the sense that its effects are legible only in what creators *actually say*, not in view counts, subscriber metrics, or recommendation graphs. Existing datasets, because they rely on metadata, render metapolitical strategies empirically invisible.

Second, content deletion and deplatforming remain invisible in existing datasets. When YouTube permanently deleted FDENEWS in March 2025, it erased 1,755 videos, 282,000 comments, and 68.3 million cumulative views. Rauchfleisch and Kaiser (2024) analyzed the removal of over 11,000 YouTube channels between 2018 and 2019. They showed that deplatforming effectively reduces the reach of far-right content, but noted that the content itself is lost to researchers. Jhaver, Boylston, Yang, and Bruckman (2021) demonstrated similar dynamics on Twitter: deplatforming reduces both the activity and toxicity of supporters, yet the suppressed content becomes permanently inaccessible for retrospective analysis. More broadly, Lakic, Rossetto, and Bernstein (2023) found that over 20% of URLs in web-sourced multimedia datasets are no longer accessible, which compromises reproducibility. Rieder et al. (2025) showed that YouTube's own search API exhibits severe temporal decay and makes previously indexed videos progressively undiscoverable. YouPol preserves all content in advance. The corpus currently includes 2,305 individually suppressed videos and three entirely deleted channels, which enables the study of what platforms choose to remove and its effects on public discourse.

Third, longitudinal engagement dynamics are underexploited. Most datasets capture metadata at a single point in time and miss the temporal trajectory of political content. Yet the evolution of view counts, likes, and subscriber growth reveals how political influence develops, how audiences respond to external events, and how platform interventions reshape visibility. Research on content virality has shown that political videos rarely achieve influence through a single exposure spike, but rather through cumulative diffusion patterns that unfold across weeks or months (Vosoughi, Roy, and Aral 2018; Gerrand, Ging, Roose, and Flood 2025; Ganesh 2025). Similarly, audience loyalty, understood as the repeated return of viewers to a creator's channel, is a central mechanism through which ideological communities consolidate over time (Munger and Phillips 2022). This temporal dimension is particularly consequential for metapolitical strategies, which, as noted above, operate through the gradual mainstreaming

of ideological frames rather than discrete viral events (Schilk 2025; Boursier 2026; Norocel 2023). Assessing whether metapolitical narratives gain or lose cultural traction requires observing the same content across time, not a single snapshot. YouPol records timestamped metadata snapshots at each observation and produces longitudinal engagement profiles for every video and channel. Combined with transcript-level content analysis, this allows researchers to study the relationship between what creators say and how audiences respond over time.

Fourth, while comment analysis has produced important findings (Wu and Resnick 2021; Mamié et al. 2021), *no existing dataset preserves comments longitudinally or tracks their deletion.* Wu and Resnick (2021) analyzed 134 million YouTube comments and found that conservatives are much more likely to comment on left-leaning videos than liberals on right-leaning ones, and that cross-partisan interactions are more toxic than co-partisan ones. Mamié et al. (2021) traced ideological migration through 300 million comments. Yet comment moderation practices, the evolution of comment sections over time, and the systematic removal of user discourse remain invisible to research. YouPol extracts and preserves 7.7 million comments with full provenance (author, timestamp, likes, reply count), including comments that were subsequently deleted by creators or platforms. This enables the study of comment moderation as a political communication strategy, and allows researchers to cross-reference audience discourse with the content analysis produced by the annotation pipeline.

The "post-API age" (Freelon et al. 2024) has further exacerbated these difficulties. Platform restrictions on data access (Bruns 2019; Tromble 2021), the deprecation of CrowdTangle, and the limitations of TikTok's Research API (Pearson et al. 2024) make it increasingly difficult for researchers to build and maintain comprehensive datasets (Chen et al. 2024; Ohme et al. 2024). When platforms do provide research APIs, systematic audits have revealed significant biases that threaten validity (Rieder et al. 2025; Bai and Gu 2026). YouPol addresses these four gaps simultaneously by combining independent transcript-level collection, pre-emptive content preservation, longitudinal engagement tracking, and comment archiving within a single continuously updated infrastructure. We also introduce the YouPol Collaborative Computing Network (YCCN), a distributed architecture that allows any collaborating researcher to contribute processing capacity from their own machine, freeing the observatory from dependence on institutional computing clusters.

1.3. Reaching Inside Political Videos: Transcription, Speaker Diarization, and LLM-Based Annotation

The gaps identified above can only be addressed if the spoken content of political videos is converted into structured, analyzable text. Recent advances in speech processing have made this feasible. Proksch, Wratil, and Wäckerle (2019) showed that ASR-generated transcripts do not systematically bias downstream measurements such as sentiment analysis or ideological

positioning. Whisper (Radford, Kim, Xu, Brockman, McLeavey, and Sutskever 2023), trained on 680,000 hours of weakly supervised web audio, brought robust multilingual ASR to 99 languages without task-specific fine-tuning, and achieves a word error rate (WER) of approximately 5.6% on the Fleurs French benchmark. WhisperX (Bain, Huh, Han, and Zisserman 2023) extended this with forced alignment and voice activity detection for word-level timestamps, which is critical for speaker diarization, the task of identifying who speaks when. Park, Kanda, Dimitriadis, Han, Watanabe, and Narayanan (2022) provide a comprehensive review of recent advances in neural diarization. YouPol uses pyannote.audio (Bredin, Yin, Coria, Gelly, Korshunov, Lavechin, Fustes, Titeux, Bouaziz, and Gill 2020; Bredin 2023), which achieved state-of-the-art diarization error rates through powerset multi-class training (Plaquet and Bredin 2023), and preprocesses audio through Demucs neural source separation (Défossez, Usunier, Bottou, and Bach 2019; Rouard, Massa, and Défossez 2023) to isolate the vocal track from background music before transcription.

Once transcripts are produced, researchers face a choice of analytical strategy. *Inductive approaches*, such as the topic modeling applied by Sosnovik et al. (2025) or the network-based clustering used by Rauchfleisch and Kaiser (2020), let patterns emerge from the data without prior theoretical commitments. YouPol follows a *deductive approach* because the constructs it targets (far-right ideology, gendered rhetoric, neo-reactionary discourse) are already well defined in the political science literature, and operationalizing them through explicit codebooks produces annotation schemes that other researchers and peers *can evaluate, contest, and reproduce*. However, applying deductive annotation to millions of sentences requires specific methods (Grimmer and Stewart 2013). Recent work has shown that LLMs can match or exceed crowd-worker and expert-coder performance on political text annotation tasks (Gilardi et al. 2023; Törnberg 2025; Ziems, Held, Shaikh, Chen, Zhang, and Yang 2024; Törnberg 2024). The strategy used in YouPol, in which LLM-generated labels validated against human annotations train lightweight BERT classifiers for corpus-wide deployment, has been validated by Pangakis and Wolken (2024) and Heseltine and Clemm von Hohenberg (2024), who showed that classifiers trained on LLM-coded data yield results comparable to those trained on human-coded data. Alizadeh et al. (2025) further demonstrated that fine-tuned open-source LLMs can approach the performance of proprietary models.

YouPol implements this approach through LLM Tool (Lemor et al. 2025), an open-source hybrid pipeline in which a codebook-guided LLM annotates a stratified sample, human annotators independently code an overlap subset to validate inter-annotator agreement, and the validated labels are then used to fine-tune a lightweight BERT classifier for corpus-wide deployment. Empirical validation of LLM Tool on a bilingual corpus of 38,451 Canadian political texts showed that XLM-RoBERTa classifiers trained on LLM-generated labels reach a mean Micro F_1 of 66.7%, with annotation fidelity rather than model size as the primary driver of downstream performance. Within YouPol, the first deployed classifier (`detect_pol`)

achieves a macro F_1 of 92.2% and an accuracy of 93.5% on held-out data, with a Light's κ of 0.787 between human annotators and the LLM on the 1,000-sentence overlap sample. The hybrid architecture yields a 110 to 1,580× inference speedup over direct LLM annotation, which makes corpus-wide deployment feasible for millions of sentences while remaining reliable with low human and monetary costs. YouPol applies this pipeline using CamemBERTav2 (Antoun, Kulumba, Touchent, Villemonte de la Clergerie, Sagot, and Seddah 2024), a French-language DeBERTaV3-based model, to train sentence-level classifiers for political content detection, ideological scoring, and gendered rhetoric analysis.

2. The YouPol Observatory Design and Scope

2.1. Design Principles

Three principles guide YouPol's architecture. First, *permanence*. The system monitors and collects continuously, content is captured before platforms can remove it, and deleted videos and suppressed channels are flagged but preserved in full. Second, *depth*. Full transcriptions with speaker diarization enable sentence-level analysis of what political creators actually say, rather than inference from titles or platform metadata. Third, *collaborative scalability*. The infrastructure grows with its community of users through the *YouPol Collaborative Computing Network* (YCCN), a distributed architecture that eliminates dependence on institutional computing clusters. Any collaborating researcher can contribute processing capacity by installing a dedicated application on their machine, which then operates as an autonomous processing node on the network. The pipeline is language-agnostic (Whisper for speech recognition, any BERT variant for annotation) and can accommodate new languages, platforms, and annotation projects without disrupting existing processing. The current francophone corpus serves as the foundation, and anglophone expansion is under preparation.

2.2. Corpus Scope

The observatory currently monitors 68 channels across two francophone political ecosystems (France and Quebec) and four ideological orientations. This corpus is designed to grow as new channels are identified and new ecosystems are integrated, soon including the anglophone ecosystem. Far right (*extrême droite*) channels promote ethno-nationalist, anti-immigration, or authoritarian discourse and constitute the largest category. Left (*gauche*) channels are associated with progressive perspectives and are included for comparative analysis. Manosphere (*manosphère*) channels promote anti-feminist or "red pill" ideologies at the intersection of gender politics and political radicalization. Conspiracy (*complotisme*) channels are centered on conspiratorial narratives that frequently intersect with far-right discourse.

Channel selection followed an expert-driven iterative approach consistent with established practices in the field (Lewis 2018; Rauchfleisch and Kaiser 2020). The process began with a seed list of channels known for their role in the francophone political content ecosystem, identified through audience metrics (views, subscribers), prior academic work (Boursier 2025; Gilliotte 2024), and media watchdog reports. Additional channels were incorporated through snowball sampling via YouTube's related-channel and recommendation networks, following the iterative-to-saturation protocol described by Reveilhac and Nchakga (2024) for a comparable corpus of 67 French alternative channels. Each channel was classified along three dimensions: ideological orientation (far right, left, manosphere, conspiracy), country of origin (France or Quebec), and creator gender. The resulting corpus spans the full audience spectrum, from major outlets with over one million subscribers (Mediapart, BLAST) to micro-channels with fewer than 1,000, thus capturing both high-reach influencers and the long tail of political content production. Recent independent mapping of 127 French political YouTube channels by Gilliotte (2024) validates the coverage of the YouPol corpus, as the major clusters he identified (far-right native creators, left-aligned journalism, manosphere) correspond directly to our four orientation categories. Figure 1 shows the distribution across orientations and countries, and Figure 2 presents the cumulative growth of the corpus over time.

3. The Seven-Step Processing Pipeline

Figure 3 presents the complete pipeline architecture. The observatory operates as a continuous loop. Newly discovered videos feed back into the collection stage, and each step runs permanently on the YCCN.

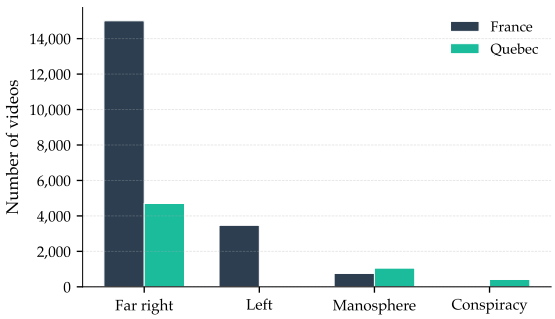

Figure 1: Corpus composition by ideological orientation and country of origin. Far-right content constitutes the majority of the corpus in both France and Quebec.

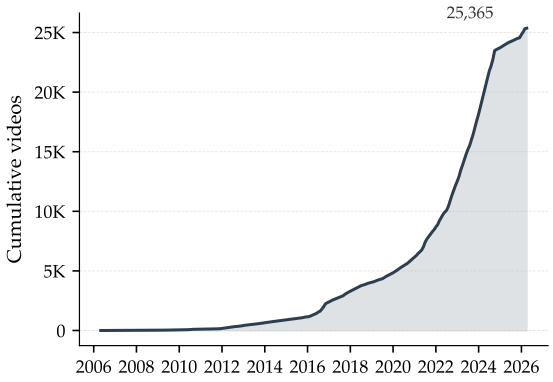

Figure 2: Cumulative growth of the corpus over time. Video production accelerates sharply from 2020.

3.1. Step 1: Channel Selection

Channels are identified through a combination of audience metrics (subscriber and view counts), content analysis, and expert assessment of each channel's role within the francophone political content ecosystem. Each channel is classified along two dimensions: political orientation and country of origin. The corpus currently covers 68 channels across the political spectrum, from investigative left-wing journalism (e.g., Mediapart, 1.1M subscribers) to prominent far-right commentary (e.g., Frontières · Livre Noir, 565K subscribers), encompassing both YouTube and TikTok. Figure 4 presents the 15 channels with the highest cumulative view counts.

3.2. Step 2: Continuous Data Collection

The pipeline continuously scans all tracked channels to detect newly published videos.[1] For each discovered video, the pipeline extracts audio in WAV format, metadata (title, duration, upload date, views, likes, subscriber counts), and comments with full provenance (author, text, timestamp, likes, reply count). The initial batch of over 15 TB of audio data was processed

[1]For YouTube, scanning is performed via `yt-dlp` with automatic cookie rotation. For TikTok, a dedicated headless browser-based scanner handles the platform's restrictive environment through SOCKS5 proxy rotation (68 servers) and anti-detection measures to circumvent rate limiting.

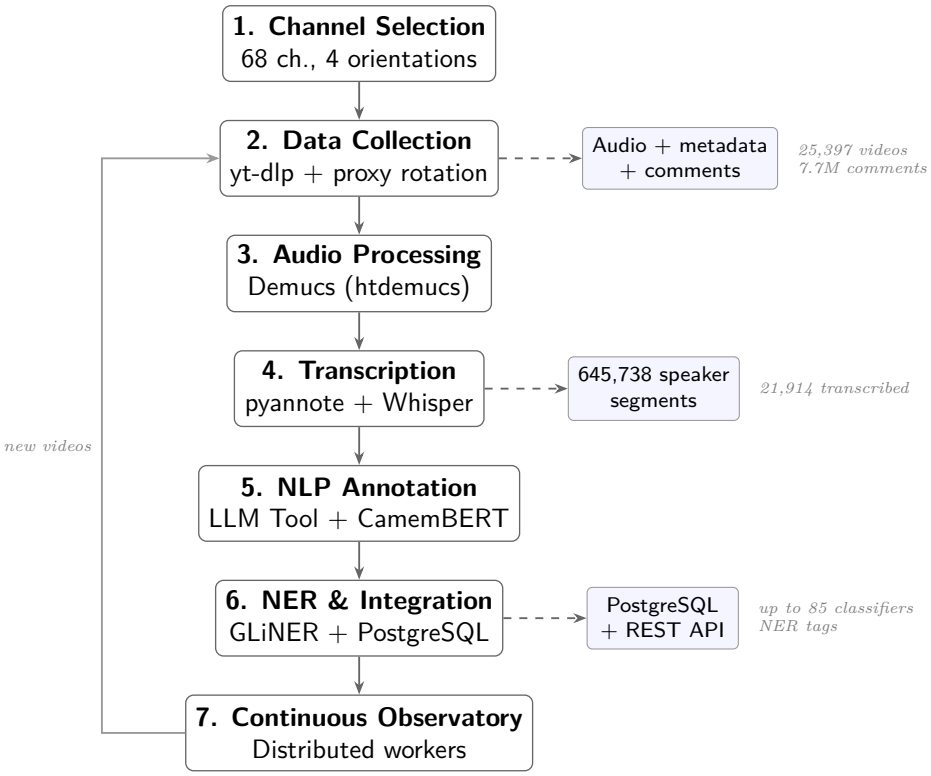

Figure 3: The YouPol seven-step processing pipeline. The observatory operates as a continuous loop: Step 7 feeds newly discovered videos back into Step 2.

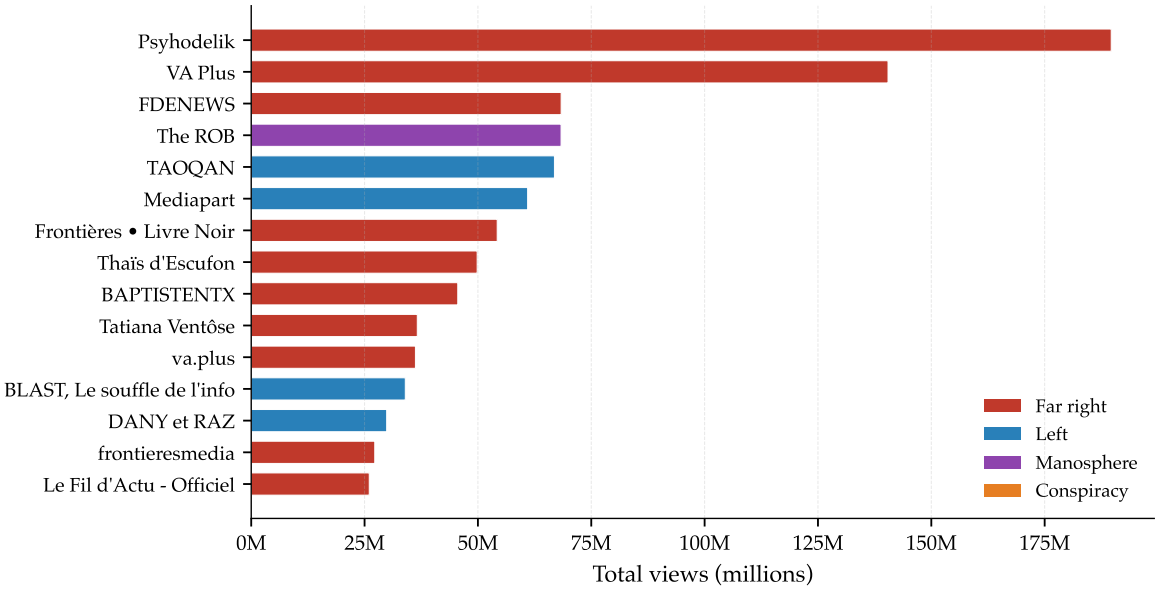

Figure 4: Top 15 channels by cumulative views, colored by ideological orientation. The corpus captures over one billion total views (1,156,113,044).

via the Digital Research Alliance of Canada's high-performance computing infrastructure. Processing has since been transferred to the YCCN described in Section 3.6.

Each scan captures a time-stamped *metadata snapshot* that records view counts, likes, and subscriber metrics at each observation. These snapshots enable longitudinal analysis of engagement dynamics. They capture not only the final state but the full trajectory of each video, and effectively reproduce the dashboard and engagement data that political influencers themselves have access to. The system also tracks *deleted content*, videos removed by creators or platforms are flagged as `suppressed` rather than deleted from the database, and deleted comments are similarly preserved. Section 4.3 details the scale of preserved content, which includes three entirely deleted channels and 2,305 individual video removals.

3.3. Step 3: Audio Processing

Political video content frequently contains background music, jingles, and sound effects. We address this through audio processing using Demucs (Défossez et al. 2019; Rouard et al. 2023), specifically the `htdemucs` model. Demucs decomposes the audio signal into four stems (vocals, bass, drums, other); only the vocal stem is retained. This preprocessing step substantially improves transcription accuracy, particularly for political commentary content where creators systematically use music to frame segments.

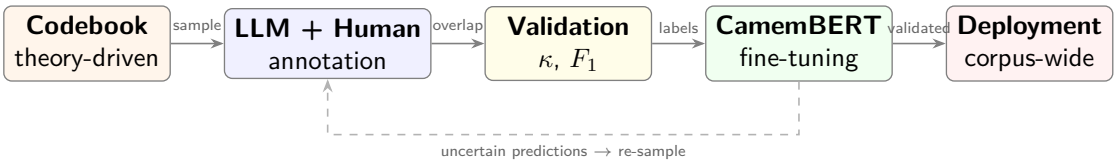

Figure 5: The LLM-in-the-loop annotation framework. A codebook guides LLM annotation of stratified samples. Human annotators validate a representative overlap subset. Validated labels train CamemBERTav2 classifiers, with an iterative refinement loop for uncertain predictions.

3.4. Step 4: Speaker Diarization and Transcription

Transcription proceeds in three stages. First, pyannote.audio (Bredin et al. 2020; Bredin 2023), with a clustering threshold tuned to 0.75 to reduce over-segmentation, segments the vocal track into speaker turns to identify distinct speakers and their temporal boundaries. This is critical for political content, which frequently features interviews, debates, and multi-speaker formats. Second, Whisper large-v3 (Radford et al. 2023) transcribes each segment. On the Fleurs French benchmark, Whisper large-v3 achieves a word error rate (WER) of approximately 5.6%, which places it among the best available open-source ASR models for French.[2] Third, each speaker segment is split into individual sentences using SaT (Segment any Text) (Frohmann, Sterner, Vulić, Minixhofer, and Schedl 2024), a multilingual sentence boundary detection model (`sat-12l-sm`). The resulting transcripts comprise 645,738 speaker-labeled segments split into 3,182,705 individual sentences across 21,914 videos (April 2026). Each sentence is stored with its speaker identifier, segment position, and the original segment context.

3.5. Step 5: NLP Annotation via LLM-in-the-Loop

The annotation of millions of transcript sentences poses a scalability challenge that manual methods cannot address at reasonable cost. YouPol addresses this through an LLM-in-the-loop framework built on the open-source LLM Tool platform (Lemor et al. 2025), following the knowledge-distillation paradigm in which LLM-generated labels, validated against human annotations, are used to train lightweight, deployable classifiers (Pangakis and Wolken 2024; Ziems et al. 2024).

The framework, illustrated in Figure 5, operates in five stages. First, for each annotation dimension, a detailed *codebook* specifies the construct definition, decision rules, and annotated examples, grounded in political science theory (the full codebook prompt for `detect_pol` is provided in Appendix A). Second, a stratified random sample of typically 1,000 sentences is drawn from the corpus and ensures balanced representation across channels, ideological orientations, and content types. A large language model (GPT-5.2) annotates this sample following the codebook instructions, and human annotators independently annotate the

[2]Whisper large-v3 model card, `https://huggingface.co/openai/whisper-large-v3`. The Demucs preprocessing step further reduces WER by isolating the vocal track from background audio.

same sentences. Inter-annotator agreement is then computed between the LLM and human annotations. If agreement falls below acceptable thresholds (Light's $\kappa \geq 0.75$, macro $F_1 \geq 0.90$), the codebook prompt is revised and the process is repeated until satisfactory agreement is reached. Third, once the prompt is validated, a larger stratified sample of over 10,000 sentences is drawn and annotated by the LLM alone. This larger set of LLM-generated labels serves as training data. Fourth, *CamemBERTav2* (Martin, Muller, Ortiz Suárez, Dupont, Romary, de la Clergerie, Seddah, and Sagot 2020) classifiers are fine-tuned on these labels. If classifier performance on held-out data is insufficient, the most uncertain predictions are re-sampled, re-annotated, and used to augment the training set. Fifth, validated classifiers are deployed across the entire corpus. Each annotation project of the YouPol database go through multiple iterations of this loop before deployment.

The annotation pipeline is structured as a cascade. The first and most fundamental classifier, `detect_pol`, operates as a binary gate that identifies transcript sentences containing political content, defined broadly to include current affairs, social issues, political actors, power relations, and social norms. This broad definition is grounded in political science conceptions of "the political" that extend beyond narrow party-political content. Table 2 reports the validation and classifier performance metrics for `detect_pol`. The inter-annotator agreement between two human annotators and the LLM on the 1,000-sentence overlap sample yielded Light's $\kappa = 0.787$, a pairwise F_1 of 89.4%, and a macro F_1 of 92.2%, which confirms the reliability of the LLM-generated labels. The CamemBERTav2 classifier trained on these labels achieved a macro F_1 of 92.2% and an accuracy of 93.5% on the held-out validation set (1,753 sentences). The classifier was deployed across all 645,738 speaker-diarized transcript segments and produces

Table 2: Validation and classifier performance for `detect_pol`.

Stage	**Metric**	**Value**
Human–LLM agreement (1,000-sentence overlap, 2 annotators)		
	Light's κ	0.787
	Krippendorff's α	0.787
	Pairwise F_1	89.4%
	Macro F_1	92.2%
	Weighted F_1	92.0%
	Exact match	88.1%
	Hamming loss	7.7%
CamemBERTav2 classifier (validation set, $n = 1{,}753$)		
	Macro F_1	92.2%
	Accuracy	93.5%
	F_1 (non-political)	94.6%
	F_1 (political)	89.3%
	Training epochs	11 / 25 (early stopping)

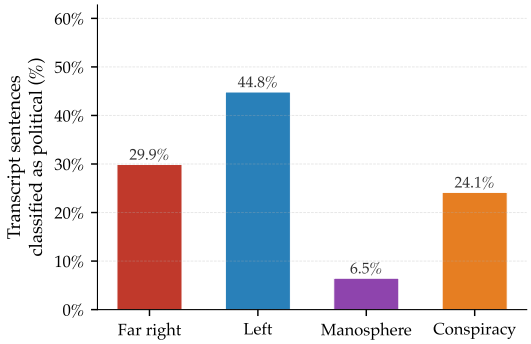

Figure 6: Proportion of transcript sentences classified as political by `detect_pol`, by ideological orientation.

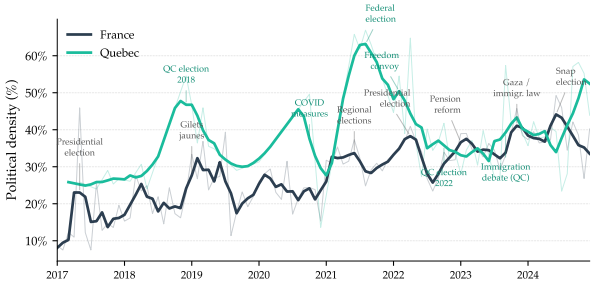

Figure 7: Monthly average political density by country (3-month smoothing for France, 5-month for Quebec due to sparser data). France and Quebec respond to distinct national political events.

a binary label (`1` = political, `0` = non-political) for every sentence. Only sentences classified as political are then routed to project-specific classifiers (depending on the project).

Figure 6 presents preliminary results aggregated by ideological orientation. Political density varies substantially: left-wing channels exhibit the highest proportion of political sentences (44.8%), consistent with their longer, analysis-oriented format. Far-right channels follow at 30.1%, despite constituting the largest share of the corpus. Conspiracy channels show a political density of 24.1%, while manosphere channels exhibit the lowest (6.8%), a result of their hybrid format that mixes political commentary with lifestyle and entertainment content. Figure 7 shows the monthly evolution of average political density separately for France and Quebec from 2017 to 2024. The two curves show that each ecosystem responds to its own national political calendar. For France, clear spikes appear around the 2017 presidential election, the *Gilets jaunes* movement (peaking in January 2019), the 2021 regional elections, the 2022 presidential election, the pension reform protests (early 2023), the Israel–Gaza war and immigration law debate (November 2023, the highest pre-2024 peak at 44%), and the 2024 snap parliamentary election following Macron's dissolution of the National Assembly (reaching 49%, the overall maximum). For Quebec, a first spike follows the 2018 provincial election, as third-party political commentary emerged in the Québécois far right. The COVID-19 public health measures produce a distinct peak in late 2020, driven by anti-restriction content. The sharpest spike corresponds to the 2021 Canadian federal election campaign (June–September 2021, exceeding 60%), followed by the Freedom Convoy movement (February 2022), the 2022 provincial election, and the immigration cap debate of fall 2023. Between these country-specific events, both curves follow a gradual upward trend, which suggests a progressive politicization of both ecosystems over time. This cross-national comparison, made possible by the corpus's transcript-level depth, demonstrates that sentence-level political density is driven by distinct national dynamics rather than a single shared calendar.

Three annotation projects then operate on the subset of sentences classified as political by `detect_pol`. *Project 1: Far-Right Ideological Score (SIED).* Based on Boursier and Lemor (2025), the SIED operationalizes far-right and neo-reactionary ideology through a codebook covering eleven macro-categories: nationalism, immigration, democracy, progress, authority, tradition, equality, technology, libertarianism, ecology, and fictional metaphors (red pill, Lord of the Rings, Star Wars, Cathedral). Each macro-category contains between two and six sub-dimensions, and each sentence is annotated across all of them. This produces a multi-dimensional ideological profile far beyond binary classification. *Project 2: Gender Discourse Analysis (GENRE).* This codebook classifies gendered rhetoric along four dimensions: whether gender discourse is present, its valence (positive, negative, ambivalent, or null), the rationality type invoked (biological/nature, liberal, empirical, or heroic/truth-telling), and the position toward science. *Project 3: Technophile Neo-Reactionary Discourse (NR).* This project targets the intersection of technology enthusiasm and reactionary politics across technology, libertarianism, fictional metaphors, and shared dimensions on equality and hierarchy.

Together with `detect_pol` and shared dimensions, YouPol is designed to deploy up to 85 binary classifiers that produce a multi-dimensional annotation at the sentence level. The cascading architecture, where `detect_pol` gates all downstream annotation, ensures that fine-grained classifiers operate on the relevant subset while providing a standalone research variable that can be analyzed independently.

3.6. Step 6: Named Entity Recognition and Database Integration

All processed sentences undergo Named Entity Recognition using GLiNER (Zaratiana, Tomeh, Holat, and Charnois 2024), a zero-shot token-classification model that identifies entities without task-specific fine-tuning. We use the multilingual variant (`gliner_multi-v2.1`) and extract nine entity types tailored to political discourse analysis: *person*, *political party*, *institution*, *organization*, *media*, *location*, *law*, *event*, and *ideology*. For example, the sentence "Macron announced a 100-billion recovery plan with the support of the RN" ("Macron a annoncé un plan de relance de 100 milliards avec le soutien du RN") yields {person: Macron, political_party: RN, event: plan de relance}. The extracted entities are stored alongside each sentence in the database and are fully integrated into the search and filtering functions of the REST API and the web corpus explorer (see Appendix B), allowing researchers to query the corpus by entity type, name, or co-occurrence. The full dataset is organized into a normalized PostgreSQL schema with over 40 tables (described in Section 4.4), indexed for search, aggregation, and programmatic access via a PostgREST-based REST API at `https://data.you-pol.com/` with JWT authentication and role-based permissions. A Python client library is also available for programmatic corpus queries and data export.[3]

[3] `https://github.com/antoinelemor/youpol-client`

3.7. Step 7: The YouPol Collaborative Computing Network (YCCN)

The seventh step is not a terminal processing stage but the permanent operational mode of the infrastructure. YouPol operates through the YCCN (see Appendix B for a screenshot of the worker application), in which collaborating researchers contribute processing capacity from their own machines rather than relying on centralized institutional computing. This architecture, illustrated in Figure 8, makes the observatory entirely self-sufficient. After the initial processing phase on the Digital Research Alliance of Canada's infrastructure, all production has been transferred to the YCCN. The YCCN eliminates ongoing costs and dependence on institutional or private computing infrastructure.

A central server held by the YouPol team creates processing tasks and places them in a shared queue backed by a PostgreSQL database with exclusive locking, which prevents duplicate processing without centralized coordination. Tasks are organized into four independent types: channel scanning, transcription, comment extraction, and metadata refresh. Each worker machine autonomously claims the next available task, executes the appropriate processing pipeline (download, audio processing, transcription, diarization, annotation), and reports

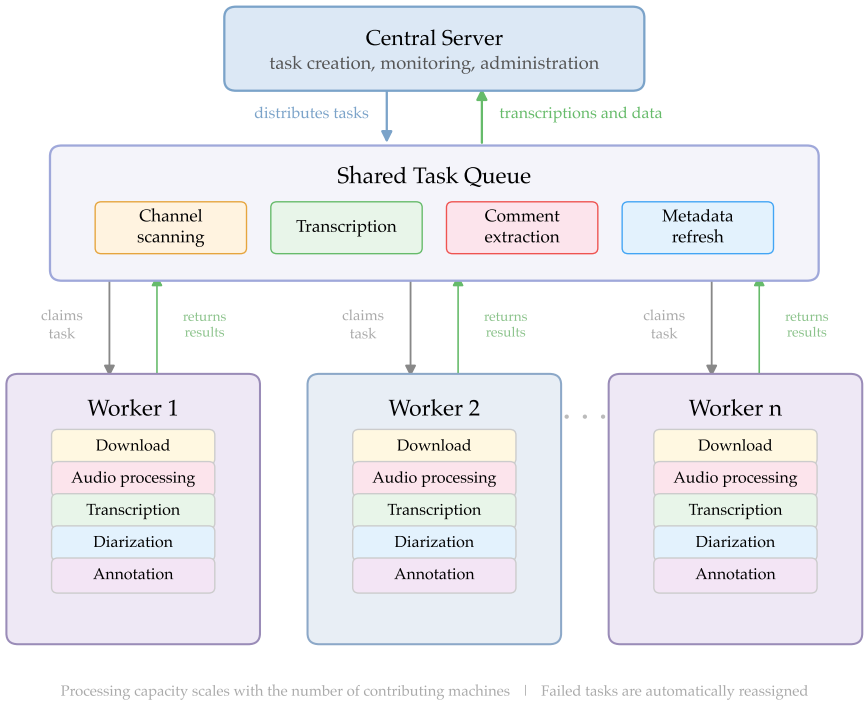

Figure 8: Architecture of the YouPol Collaborative Computing Network (YCCN). A central server distributes tasks to a shared queue organized by type. Each worker independently claims tasks and executes the full processing pipeline. Processing capacity grows as new collaborators contribute machines.

completion. If a worker becomes unavailable, its unfinished tasks are automatically released and reassigned to another worker.

The YCCN rests on a crowdsourcing principle applied to computing resources. Each collaborating researcher who joins the project installs a dedicated application developed by the authors on their machine. Once connected, the machine immediately begins claiming and processing tasks as an autonomous node. The observatory's processing capacity thus grows organically with its community. As the corpus expands to new ecosystems, new collaborators contribute both channels to monitor and computing resources to process them. This model eliminates the traditional bottleneck of institutional or private computing access and its associated costs, and places the infrastructure under the collective control of the research community rather than under the constraints of external resource providers. The system currently processes hundreds of new transcriptions daily alongside continuous comment extraction and metadata updates.

4. The YouPol Dataset

4.1. Summary Statistics

Table 3 presents the key statistics of the dataset as of April 2026. The database contains 23,712 YouTube videos and 1,685 TikTok videos, of which 21,914 have been fully transcribed and annotated. An additional 1,798 YouTube videos await processing in the main database, and the continuous scanner has identified over 16,000 further videos currently being transcribed through the pipeline. The TikTok component of the observatory, currently under active expansion,

Table 3: YouPol dataset: summary statistics (April 2026)

Metric	**Value**
YouTube videos (total)	23,712
Videos fully transcribed	21,914
Videos awaiting processing	1,798
Videos in pipeline (discovered)	>16,000
TikTok videos	1,685
Tracked channels	68
Countries	2 (France, Quebec)
Ideological orientations	4
Temporal span	May 2006 – April 2026
Total video views	1,156,113,044
Total comments extracted	7,703,663
Total likes	14,421,301
Speaker-diarized segments	645,738
Individual sentences	3,182,705
Annotation classifiers (in production)	up to 85
Initial audio data processed	>15 TB

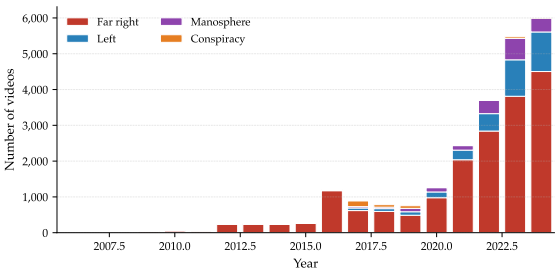

Figure 9: Video counts by year and ideological orientation. Growth accelerates from 2020.

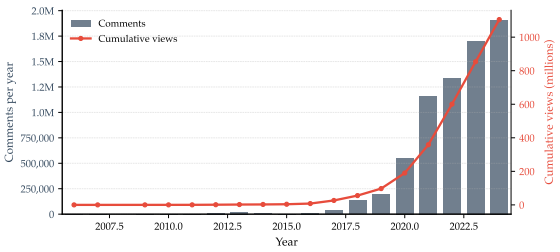

Figure 10: Comment counts per year (bars) and cumulative views (line).

mirrors the full YouTube pipeline (collection, transcription, diarization, annotation) with a dedicated set of tables and scanners. The current priority is to identify and integrate the TikTok accounts of creators already tracked on YouTube, so as to enable cross-platform comparison of the same political actors. As new TikTok accounts are identified, they enter the same seven-step pipeline described in Section 3. An expansion to the anglophone political ecosystem is under preparation. Because the pipeline architecture is language-agnostic (Whisper supports English natively, and CamemBERT can be replaced by any BERT variant via LLM Tool), this extension requires only new channel identification and adapted codebooks, not infrastructural changes.

4.2. Temporal Growth and Engagement Dynamics

Figure 9 shows the temporal distribution of the corpus. Video production has accelerated dramatically since 2020, reflecting the growth of the political YouTube ecosystem. The most active year is 2024 with 6,012 videos, and far-right content dominates across all periods, although left-wing and manosphere content have grown substantially since 2021. Figure 10 presents the parallel growth of user engagement: comment extraction has grown with the corpus to reach 1.88 million comments for 2024 videos alone while cumulative views surpass 1.15 billion.

Figure 11 shows the evolution of total views by ideological orientation since 2012. While far-right channels dominate in absolute view counts, the data shows important structural differences. Left-wing channels achieve substantially longer average durations (48.1 minutes vs. 19.1 minutes for far-right content; Figure 12), reflecting different content formats (investigative journalism vs. commentary). Figure 13 shows the distribution across view count ranges. The bulk of the corpus (10,732 videos) falls in the 10K–100K range, with 68 videos surpassing 1 million views, while the 7,053 videos below 1,000 views represent the "long tail" invisible to studies that rely on algorithmic recommendation.

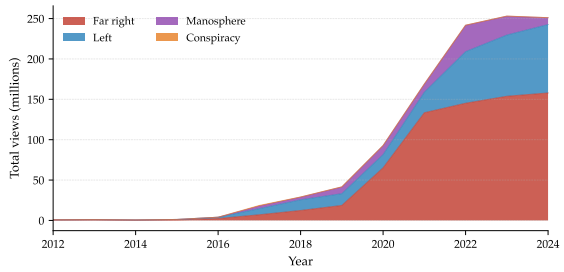

Figure 11: Total views by ideological orientation (2012–2024). Far-right content dominates, with acceleration after 2020.

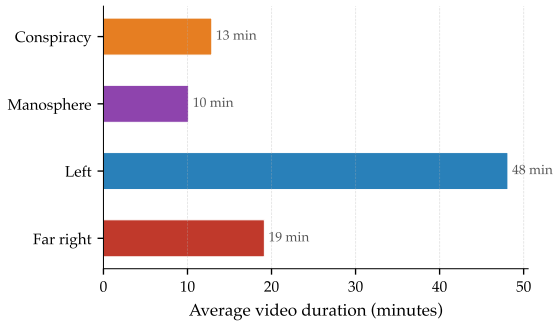

Figure 12: Average video duration by ideological orientation.

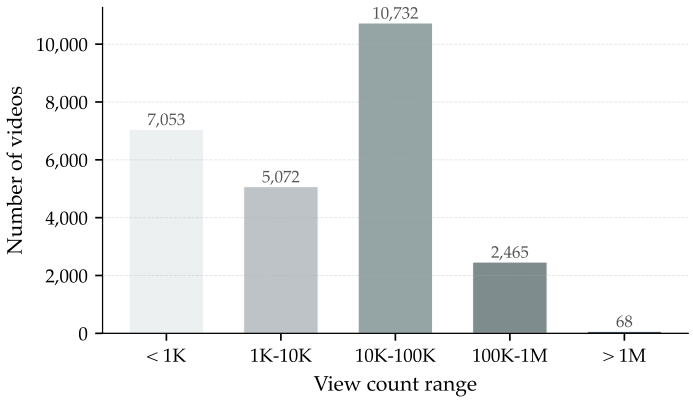

Figure 13: Distribution of videos by view count range, from long-tail content (<1K views) to viral videos (>1M views).

4.3. Preserved Content: Deleted Videos and Channels

A distinctive feature of YouPol is the systematic preservation of content that platforms remove. As of April 2026, the database contains 2,305 suppressed videos (9.1% of the corpus) and 1,578,598 comments that are no longer accessible on the platform, including 312,452 individually suppressed comments and 1,266,146 comments on videos that were subsequently removed. These represent material that YouTube has deleted for community guideline violations, that creators have voluntarily removed, or that has otherwise become unavailable.

Three entire channels have been deleted by YouTube and are preserved in full: FDENEWS (1,755 videos, 68.3M views, far-right, France), Virginie Vota (104 videos, 6.8M views, far-right, France), and Gabriel Duquette (124 videos, 823K views, manosphere, Quebec). For these channels, every video, transcript, speaker segment, comment, and metadata snapshot remains available in the database even though the original content no longer exists on the platform. In addition, 21 active channels have had individual videos removed (Figure 14). Altogether, the suppressed videos represent over 110 million views and 6.4 million likes that would be entirely

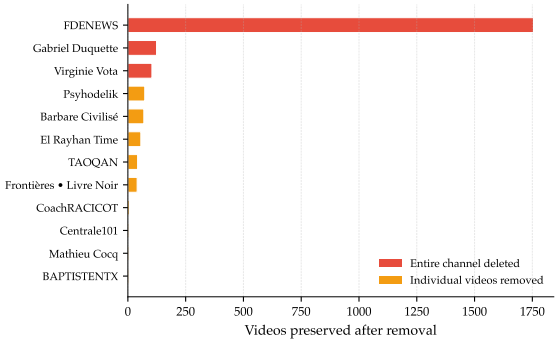

Figure 14: Videos preserved after removal from YouTube, by channel. Red: entire channel deleted. Orange: individual video removals from active channels.

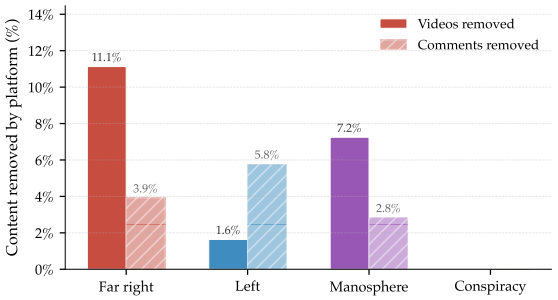

Figure 15: Content removal rates by ideological orientation. Solid: videos removed. Hatched: comments removed. Far-right channels have the highest video removal rate; left-wing channels the highest individual comment removal rate.

lost to research without pre-emptive collection.

Suppression patterns vary across ideological orientations (Figure 15). Far-right channels have the highest video removal rate (11.6%), followed by manosphere channels (8.5%), while left-wing channels experience significantly less video removal (1.6%) and conspiracy channels have no suppressed videos. Individual comment suppression follows a different pattern: left-wing channels show the highest individually suppressed comment rate (5.6%), followed by far-right (3.9%) and manosphere (2.9%). This divergence between video and comment suppression rates across orientations is itself an analytically significant finding that would be invisible without continuous monitoring.

4.4. Database Schema and Data Access

The dataset is organized into a normalized PostgreSQL schema of over 40 tables (see Appendix C for a complete overview). The schema comprises five functional groups: (1) *content tables* storing video metadata and comments for both YouTube and TikTok, with suppression tracking that flags removed content rather than deleting it; (2) *transcript tables* containing raw speaker segments (641,613 for YouTube, 4,125 for TikTok) and processed segments with NLP annotations, political detection labels, and NER tags (3.18 million processed rows); (3) *combined views* aggregating both platforms; (4) *metadata history tables* recording time-stamped engagement snapshots for videos (33,746 video snapshots) and channels (620 snapshots), enabling longitudinal analysis; and (5) *pipeline tables* for YCCN coordination, task queues, and event logs.

The project website at `https://you-pol.com/` presents the observatory and its documentation. Data access is provided through the platform at `https://data.you-pol.com/`, which

offers three modes of interaction (see Appendix B for screenshots). First, an interactive corpus explorer provides full-text search across all 645,738 speaker segments and 7.7 million comments, with filtering by channel, orientation, country, date range, and NER entity type. Visualization dashboards display corpus-level statistics, channel comparisons, and political density trends. Second, a PostgREST-based REST API provides programmatic access with JWT authentication and supports complex SQL-like filtering, pagination, and bulk CSV/JSON exports. Third, a Python client library[4] wraps the API and enables researchers, including those conducting qualitative work, to query the corpus, retrieve full transcripts or comment threads for specific videos, and export results directly into analysis workflows. All interfaces enforce role-based access control with audit logging.

5. Ethical Considerations and Limitations

5.1. Ethics

YouPol collects publicly available content from public figures who have chosen to participate in political discourse on public platforms. All tracked channels are operated by content creators with significant audiences (thousands to hundreds of thousands of subscribers). As established in internet research ethics guidelines, content produced by public figures in public forums carries reduced privacy expectations relative to private communications (Zimmer 2010; Franzke, Bechmann, Ess, and Zimmer 2020). Comments are collected from public threads; no private communications are accessed. The research protocol has been reviewed by the ethics board of the Université de Sherbrooke in accordance with the Canadian Tri-Council Policy Statement on Ethical Conduct for Research Involving Humans (Government of Canada 2023).

This research is conducted across two jurisdictions. In the European Union, where one of the authors is based, the Digital Services Act explicitly recognizes vetted academic research as a legitimate use case for platform data access, including content from very large online platforms (European Parliament and Council of the European Union 2022). YouPol's independent collection infrastructure is consistent with this framework. In Canada and Quebec, ethics governance frameworks for research exist, as reflected in the ethics review conducted by the Université de Sherbrooke, but no equivalent mechanism compels platforms to share their data with researchers (Government of Canada 2023). YouPol's independent collection infrastructure addresses this gap and remains consistent with the institutional frameworks of both jurisdictions.

Commenter identifiers are stored internally for deduplication and network analysis but are not exposed through the public API. Researcher access requires authentication with role-based permissions, usage limits, and audit logging. Content removed by platforms is preserved

[4] https://github.com/antoinelemor/youpol-client

in the database for research purposes but is accessible only to authorized researchers under a data use agreement; it is never publicly displayed. The preservation of removed content raises specific ethical questions; we hold that the research value of maintaining a record of politically consequential speech, particularly given that platform removals introduce systematic survivorship biases that distort the empirical study of political extremism (Ohme et al. 2024; Bruns 2019), outweighs the risk of amplification under these access restrictions.

5.2. Limitations

Several limitations should be acknowledged. Channel selection reflects deliberate research priorities, but it is worth acknowledging the methodological trade-offs this entails. Two broad strategies coexist in the literature for constructing political corpora on video platforms. The first, adopted by YouPol, proceeds *top-down*: channels are identified *a priori* through expert assessment and audience metrics, building on a prior conceptual definition of what constitutes far-right political content (Rauchfleisch and Kaiser 2020; Boursier 2025; Munger and Phillips 2022). The second proceeds *bottom-up*, letting digital content itself (certain words, certain discourses, certain co-engagement patterns) define the actor ecosystem *a posteriori* (Ribeiro et al. 2020; Hosseinmardi et al. 2021; Tainturier 2025).

Each approach involves distinct epistemological risks. Top-down selection ensures analytic clarity and reproducibility but risks confirmation bias. The corpus reflects the researcher's prior categorization of the field, and may systematically exclude actors whose ideological positioning is ambiguous, emergent, or deliberately unmarked. Bottom-up approaches reduce this circularity but introduce their own dependencies (on seed keywords, seed channels, or recommendation graphs) and may miss actors who operate at the margins of the indexed discourse patterns. Both strategies ultimately face what Salganik (2019) describes as the non-representativeness problem inherent to any purposive digital corpus: what is collected reflects collection choices, not the full population of political content.

YouPol's current corpus overrepresents far-right content relative to the broader political YouTube ecosystem. A deliberate choice that reflects both the composition of the francophone political influencer landscape, where far-right channels account for a disproportionate share of total views (Munger and Phillips 2022), and the research questions that motivate the observatory's design. However, the infrastructure YouPol has built creates the conditions for progressively bridging these two paradigms. The sentence-level ideological classifiers, named entity recognition, and discourse annotation tools now make it possible to identify, within existing transcripts, the channels, figures, and outlets that far-right creators actively cite, endorse, or engage with, and consequently actors who may not have been captured by the initial expert-driven selection. A content-based expansion of the corpus, seeded not by audience metrics but by the discursive networks reconstructed from the video transcripts

themselves, would enable a more bottom-up mapping of the francophone far-right ecosystem and provide an empirical check on the boundaries drawn by the initial top-down selection.

Independent of corpus construction decisions, a set of technical limitations constrains what YouPol can currently measure with full reliability. Despite Demucs preprocessing, transcription remains imperfect for overlapping speech, heavy accents, or degraded audio; error rates have not been systematically quantified across the full corpus. The LLM-in-the-loop framework produces silver-standard labels rather than gold-standard human annotations; while the iterative refinement loop mitigates systematic errors, some noise is inevitable. Despite infrastructural independence from platform APIs, YouPol depends on platform availability for initial collection. Content not yet collected when removed cannot be recovered. The current corpus covers francophone content, with anglophone expansion under preparation. Extension to additional linguistic contexts would require new channel identification, language-specific ASR tuning, and culturally adapted codebooks. Finally, TikTok data collection faces greater technical challenges due to more aggressive anti-scraping measures and is currently less comprehensive than YouTube coverage (the collection of TikTok accounts parallel to the already tracked YouTube channels is underway).

Conclusion

YouPol provides the first transcript-level, continuously updated observatory for political discourse on video platforms. The infrastructure addresses the four gaps identified in the literature. First, it makes the ideological substance of political video content empirically accessible through an independent transcription pipeline that combines neural source separation, speaker diarization, and speech recognition to produce full transcripts regardless of platform-provided captions. Second, it preserves content that platforms remove, including 2,305 suppressed videos and three entirely deleted channels, and thus allows retrospective analysis of material that would otherwise be permanently lost to research. Third, it tracks engagement dynamics longitudinally through timestamped metadata snapshots that capture the full trajectory of each video and channel. Fourth, it archives 7.7 million comments with full provenance, including those subsequently deleted by creators or platforms, and enables the study of comment moderation as a political communication strategy.

An LLM-in-the-loop annotation framework, in which large language model labels validated against human annotations are distilled into lightweight CamemBERT classifiers, provides scalable sentence-level annotation across the full corpus. The first deployed classifier, `detect_pol`, assigns a binary political label to each of the 3.18 million sentences in the database. Political density, the proportion of transcript sentences classified as political within a given channel or time period, varies from 44.8% for left-wing channels to 6.8% for manosphere channels. The monthly evolution of this measure in France and Quebec responds to distinct national

political calendars rather than a shared temporal dynamic. These results would be inaccessible through metadata analysis alone. The full annotation pipeline across three projects (far-right ideology, gendered rhetoric, neo-reactionary discourse) will extend this analysis to finer-grained ideological dimensions.

The YouPol Collaborative Computing Network (YCCN) represents a distinct contribution to research infrastructure. By allowing any collaborating researcher to contribute processing capacity from their own machine, the observatory operates independently of institutional computing clusters. The language-agnostic pipeline requires only new channel identification and adapted codebooks to extend to new linguistic and political contexts, and an anglophone expansion is under preparation. TikTok integration is also in progress, with priority given to the accounts of creators already tracked on YouTube in order to enable cross-platform comparison of the same political actors.

A question that the dataset is well positioned to address in future work is whether the political density or ideological intensity of a video predicts its subsequent removal by the platform. More broadly, YouPol offers a foundation for research on political communication that takes the actual spoken content of political videos as its primary object of analysis.

The project is described at `https://you-pol.com/`, with data available via the REST API at `https://data.you-pol.com/`.

Acknowledgements

The initial processing of over 15 TB of audio data was performed on the infrastructure of the Digital Research Alliance of Canada, access to which was provided through the Centre interuniversitaire de recherche sur la science et la technologie (CIRST). We thank François Claveau for his support and guidance throughout this project.

References

Alizadeh, Meysam, Kubli, Maël, Samei, Zeynab, Dehghani, Shirin, Zahedivafa, Mohammadmasiha, Bermeo, Juan D., Korobeynikova, Maria, and Gilardi, Fabrizio. Open-source LLMs for text annotation: A practical guide for model setting and fine-tuning. *Journal of Computational Social Science*, 8(1):17, 2025. doi:10.1007/s42001-024-00345-9.

Antoun, Wissam, Kulumba, Francis, Touchent, Rian, Villemonte de la Clergerie, Éric, Sagot, Benoît, and Seddah, Djamé. CamemBERT 2.0: A smarter French language model aged to perfection. *arXiv preprint*, 2024. doi:10.48550/arXiv.2411.08868.

Bai, Dan and Gu, Yan. Harnessing big data, hindered by bias: Evaluating TikTok research API for fair and optimal social sciences. *Social Science Computer Review*, 2026. doi:10.1177/08944393251413277.

Bain, Max, Huh, Jaesung, Han, Tengda, and Zisserman, Andrew. WhisperX: Time-accurate speech transcription of long-form audio. In *Proceedings of INTERSPEECH 2023*, 2023. doi:10.21437/interspeech.2023-78.

Boursier, Tristan. La banalisation du suprémacisme blanc sur YouTube: Analyse des convergences et des influences idéologiques au sein de l'extrême droite française. *Politique et sociétés*, 44(1):35–62, 2025. doi:10.7202/1114896ar.

Boursier, Tristan. Métapolitique d'extrême droite : Usages et limites d'un concept. *Canadian Journal of Political Science*, pages 1–21, 2026. doi:10.1017/S0008423926101103.

Boursier, Tristan and Lemor, Antoine. Mesurer la pénétration des idées d'extrême droite dans les discours gouvernementaux français. *Revue française de science politique*, 75(2):261–291, 2025. doi:10.3917/rfsp.752.0261.

Bredin, Hervé. Pyannote.audio 2.1 speaker diarization pipeline: Principle, benchmark, and recipe. In *Proceedings of INTERSPEECH 2023*, pages 1983–1987, 2023. doi:10.21437/Interspeech.2023-105.

Bredin, Hervé, Yin, Ruiqing, Coria, Juan Manuel, Gelly, Gregory, Korshunov, Pavel, Lavechin, Marvin, Fustes, Diego, Titeux, Hadrien, Bouaziz, Wassim, and Gill, Marie-Philippe. Pyannote.audio: Neural building blocks for speaker diarization. In *Proceedings of ICASSP 2020*, pages 7124–7128, 2020. doi:10.1109/ICASSP40776.2020.9052974.

Bruns, Axel. After the 'APIcalypse': Social media platforms and their fight against critical scholarly research. *Information, Communication & Society*, 22(11):1544–1566, 2019. doi:10.1080/1369118X.2019.1637447.

Chen, Yan, Sherren, Kate, Lee, Kyung Young, McCay-Peet, Lori, Xue, Shan, and Smit, Michael. From theory to practice: Insights and hurdles in collecting social media data for social science research. *Frontiers in Big Data*, 7:1379921, 2024. doi:10.3389/fdata.2024.1379921.

Défossez, Alexandre, Usunier, Nicolas, Bottou, Léon, and Bach, Francis. Music source separation in the waveform domain, 2019.

European Parliament and Council of the European Union. Regulation (EU) 2022/2065 on a single market for digital services (digital services act), 2022. OJ L 277, 27.10.2022, p. 1–102.

Franzke, Aline Shakti, Bechmann, Anja, Ess, Charles Melvin, and Zimmer, Michael. Internet Research: Ethical Guidelines 3.0. Report, AoIR (The International Association of Internet Researchers), 2020.

Freelon, Deen, Monzer, Cristina, Jeon, Gayoung, Moy, Cameron, and Williams, Natasha. The post-API age of social media data access: Past, present, and future. *The Annals of the American Academy of Political and Social Science*, 715(1):16–37, 2024. doi:10.1177/00027162251372557.

Frohmann, Markus, Sterner, Igor, Vulić, Ivan, Minixhofer, Benjamin, and Schedl, Markus. Segment any text: A universal approach for robust, efficient and adaptable sentence segmentation. In *Proceedings of EMNLP 2024*, pages 11908–11941, 2024. doi:10.18653/v1/2024.emnlp-main.665.

Ganesh, Bharath. The Western Far Right and Digital Technology: Fuzzy Collectivity From Translocal Whiteness to Networked Metapolitics. *Sociology Compass*, 19(2):1–15, 2025. doi:10.1111/soc4.70038.

Gerbaudo, Paolo. TikTok and the algorithmic transformation of social media publics: From social networks to social interest clusters. *New Media & Society*, 28(3):1019–1036, 2026. doi:10.1177/14614448241304106.

Gerrand, Vivian, Ging, Debbie, Roose, Joshua M., and Flood, Michael. Mapping the Neo-Manosphere(s): New Directions for Research. *Men and Masculinities*, page 1097184X251350277, 2025. doi:10.1177/1097184X251350277.

Gilardi, Fabrizio, Alizadeh, Meysam, and Kubli, Maël. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30): e2305016120, 2023. doi:10.1073/pnas.2305016120.

Gilliotte, Quentin. Identifier et cartographier les producteurs d'analyses politiques sur YouTube. *RESET. Recherches en sciences sociales sur Internet*, 13, 2024. doi:10.4000/12cn7.

Government of Canada, Interagency Advisory Panel on Research Ethics. Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans – TCPS 2 (2022), 2023.

Grimmer, Justin and Stewart, Brandon M. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21(3):267–297, 2013. doi:10.1093/pan/mps028.

Guinaudeau, Benjamin, Munger, Kevin, and Votta, Fabio. Fifteen seconds of fame: TikTok and the supply side of social video. *Computational Communication Research*, 4(2):463–485, 2022. doi:10.5117/CCR2022.2.004.GUIN.

Haroon, Muhammad, Wojcieszak, Magdalena, Chhabra, Anshuman, Liu, Xin, Mohapatra, Prasant, and Shafiq, Zubair. Auditing YouTube's recommendation system for ideologically congenial, extreme, and problematic recommendations. *Proceedings of the National Academy of Sciences*, 120(50):e2213020120, 2023. doi:10.1073/pnas.2213020120.

Heseltine, Michael and Clemm von Hohenberg, Bernhard. Large language models as a substitute for human experts in annotating political text. *Research & Politics*, 2024. doi:10.1177/20531680241236239.

Hiray, Arnav, Liu, Yunsong, Song, Mingxiao, Shah, Agam, and Chava, Sudheer. CoCoHD: Congress Committee Hearing Dataset. In *Findings of EMNLP 2024*, pages 15529–15542, 2024. doi:10.18653/v1/2024.findings-emnlp.911.

Hosseinmardi, Homa, Ghasemian, Amir, Clauset, Aaron, Möbius, Markus, Rothschild, David M., and Watts, Duncan J. Examining the consumption of radical content on YouTube. *Proceedings of the National Academy of Sciences*, 118(32):e2101967118, 2021. doi:10.1073/pnas.2101967118.

Jhaver, Shagun, Boylston, Christian, Yang, Diyi, and Bruckman, Amy. Evaluating the effectiveness of deplatforming as a moderation strategy on Twitter. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–30, 2021. doi:10.1145/3479525.

Lai, Angela, Brown, Megan A., Bisbee, James, Tucker, Joshua A., Nagler, Jonathan, and Bonneau, Richard. Estimating the ideology of political YouTube videos. *Political Analysis*, 32(3):345–360, 2024. doi:10.1017/pan.2023.42.

Lakic, Viktor, Rossetto, Luca, and Bernstein, Abraham. Link-rot in web-sourced multimedia datasets. In *MultiMedia Modeling (MMM 2023)*, volume 13833 of *Lecture Notes in Computer Science*, pages 476–488. Springer, 2023. doi:10.1007/978-3-031-27077-2_37.

Lazer, David, Pentland, Alex, Watts, Duncan J., Aral, Sinan, Athey, Susan, Contractor, Noshir, et al. Computational social science: Obstacles and opportunities. *Science*, 369 (6507):1060–1062, 2020. doi:10.1126/science.aaz8170.

Ledwich, Mark and Zaitsev, Anna. Algorithmic extremism: Examining YouTube's rabbit hole of radicalization. *First Monday*, 25(3), 2020. doi:10.5210/fm.v25i3.10419.

Lemor, Antoine, Dinan, Shannon, and Gilbert, Jeremy. LLM Tool: A hybrid pipeline for automated large-scale text annotation using local language models and BERT classifiers, 2025.

Lewis, Rebecca. Alternative influence: Broadcasting the reactionary right on YouTube. Technical report, Data & Society Research Institute, 2018.

Mamié, Robin, Horta Ribeiro, Manoel, and West, Robert. Are anti-feminist communities gateways to the far right? Evidence from Reddit and YouTube. In *Proceedings of the 13th ACM Web Science Conference (WebSci '21)*, pages 139–147, 2021. doi:10.1145/3447535.3462504.

Martin, Louis, Muller, Benjamin, Ortiz Suárez, Pedro Javier, Dupont, Yoann, Romary, Laurent, de la Clergerie, Éric, Seddah, Djamé, and Sagot, Benoît. CamemBERT: A tasty French language model. In *Proceedings of the 58th Annual Meeting of the ACL*, pages 7203–7219, 2020. doi:10.18653/v1/2020.acl-main.645.

Munger, Kevin and Phillips, Joseph. Right-wing YouTube: A supply and demand perspective. *The International Journal of Press/Politics*, 27(1):186–219, 2022. doi:10.1177/1940161220964767.

Norocel, Ov Cristian. Research bricolage on far-right metapolitics: Superordinate intersectionality perspectives on digital identities. *Innovation: The European Journal of Social Science Research*, 24(5):1–17, 2023. doi:10.1080/13511610.2023.2292954.

Ohme, Jakob, Araujo, Theo, Boeschoten, Laura, Freelon, Deen, Ram, Nilam, Reeves, Byron B., and Robinson, Thomas N. Digital trace data collection for social media effects research: APIs, data donation, and (screen) tracking. *Communication Methods and Measures*, 18(2):124–141, 2024. doi:10.1080/19312458.2023.2181319.

Pangakis, Nicholas and Wolken, Sam. Knowledge distillation in automated annotation: Supervised text classification with LLM-generated training labels. In *Proceedings of the 6th Workshop on NLP and Computational Social Science (NLP+CSS @ EMNLP 2024)*, pages 113–131, 2024. doi:10.18653/v1/2024.nlpcss-1.9.

Park, Tae Jin, Kanda, Naoyuki, Dimitriadis, Dimitrios, Han, Kyu J., Watanabe, Shinji, and Narayanan, Shrikanth. A review of speaker diarization: Recent advances with deep learning. *Computer Speech & Language*, 72:101317, 2022. doi:10.1016/j.csl.2021.101317.

Pearson, George D. H., Silver, Nathan A., Robinson, Jessica Y., Azadi, Mona, Schillo, Barbara A., and Kreslake, Jennifer M. Beyond the margin of error: A systematic and

replicable audit of the TikTok research API. *Information, Communication & Society*, 28(3): 452–470, 2024. doi:10.1080/1369118X.2024.2420032.

Pinto, Gabriela, Bickham, Charles, Salkar, Tanishq, Menezes, Joyston, Luceri, Luca, and Ferrara, Emilio. Tracking the 2024 US presidential election chatter on TikTok: A public multimodal dataset. In *Companion Proceedings of the ACM Web Conference 2025*, pages 773–776, 2025. doi:10.1145/3701716.3715291.

Plaquet, Alexis and Bredin, Hervé. Powerset multi-class cross entropy loss for neural speaker diarization. In *Proceedings of INTERSPEECH 2023*, 2023. doi:10.21437/interspeech.2023-205.

Proksch, Sven-Oliver, Wratil, Christopher, and Wäckerle, Jens. Testing the validity of automatic speech recognition for political text analysis. *Political Analysis*, 27(3):339–359, 2019. doi:10.1017/pan.2018.62.

Radford, Alec, Kim, Jong Wook, Xu, Tao, Brockman, Greg, McLeavey, Christine, and Sutskever, Ilya. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning (ICML), PMLR 202*, pages 28492–28518, 2023.

Rauchfleisch, Adrian and Kaiser, Jonas. The German far-right on YouTube: An analysis of user overlap and user comments. *Journal of Broadcasting & Electronic Media*, 64(3): 373–396, 2020. doi:10.1080/08838151.2020.1799690.

Rauchfleisch, Adrian and Kaiser, Jonas. The impact of deplatforming the far right: An analysis of YouTube and BitChute. *Information, Communication & Society*, 27(7):1478–1496, 2024. doi:10.1080/1369118X.2024.2346524.

Rauh, Christian and Schwalbach, Jan. The ParlSpeech V2 data set: Full-text corpora of 6.3 million parliamentary speeches in nine representative democracies, 2020. doi:10.7910/DVN/L4OAKN.

Reveilhac, Maud and Nchakga, Camille. How French alternative media channels on YouTube portray the government and mainstream media on YouTube. *Frontiers in Communication*, 9:1517963, 2024. doi:10.3389/fcomm.2024.1517963.

Ribeiro, Manoel Horta, Ottoni, Raphael, West, Robert, Almeida, Virgílio A. F., and Meira, Wagner, Jr. Auditing radicalization pathways on YouTube. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*, pages 131–141, 2020. doi:10.1145/3351095.3372879.

Rieder, Bernhard, Coromina, Òscar, and Matamoros-Fernández, Ariadna. Mapping YouTube: A quantitative exploration of a platformed media system. *First Monday*, 25(8), 2020. doi:10.5210/fm.v25i8.10667.

Rieder, Bernhard, Padilla, Adrián, and Coromina, Òscar. Forgetful by design? A critical audit of YouTube's search API for academic research. *Information, Communication & Society*, 2025. doi:10.1080/1369118X.2025.2591767.

Roberts, Hal, Bhargava, Rahul, Valiukas, Linas, Jen, Dennis, Malik, Momin M., Bishop, Cindy, et al. Media Cloud: Massive open source collection of global news on the open web. In *Proceedings of the 15th International AAAI Conference on Web and Social Media (ICWSM)*, volume 15, pages 1034–1045, 2021. doi:10.1609/icwsm.v15i1.18127.

Rouard, Simon, Massa, Francisco, and Défossez, Alexandre. Hybrid transformers for music source separation. In *Proceedings of ICASSP 2023*, pages 1–5, 2023. doi:10.1109/ICASSP49357.2023.10096956.

Salganik, Matthew J. *Bit by Bit: Social Research in the Digital Age*. Princeton University Press, Princeton Oxford, first paperback printing edition, 2019. ISBN 978-0-691-15864-8 978-0-691-19610-7.

Schilk, Felix. The Metapolitics of Crises: How the New Right Weaponises Narratives to Mainstream Far-Right Ideology. *International Journal of Politics, Culture, and Society*, 2025. doi:10.1007/s10767-025-09519-3.

Solovev, Kirill, Drolsbach, Chiara, Demirel, Emma, and Pröllochs, Nicolas. Engagement with political videos on TikTok during the 2025 German federal election. *EPJ Data Science*, 15, 2026. doi:10.1140/epjds/s13688-026-00632-7.

Sosnovik, Vera, Violot, Caroline, and Humbert, Mathias. In times of crisis: An exploratory study of media and political discourse on YouTube during the 2024 French elections, 2025.

Tainturier, Benjamin. *"'Dire Sur Le Web Ce Que Les Français Pensent Tout Bas" : Les Expressions Numériques de La Droite Radicale*. Theses, Institut d'études politiques de Paris - Sciences Po, 2025.

Törnberg, Petter. Best practices for text annotation with large language models. *Sociologica*, 18(2):67–85, 2024. doi:10.6092/issn.1971-8853/19461.

Törnberg, Petter. Large language models outperform expert coders and supervised classifiers at annotating political social media messages. *Social Science Computer Review*, 43(6): 1181–1195, 2025. doi:10.1177/08944393241286471.

Tromble, Rebekah. Where have all the data gone? A critical reflection on academic digital research in the post-API age. *Social Media + Society*, 7(1), 2021. doi:10.1177/2056305121988929.

Vosoughi, Soroush, Roy, Deb, and Aral, Sinan. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018. doi:10.1126/science.aap9559.

Wu, Siqi and Resnick, Paul. Cross-partisan discussions on YouTube: Conservatives talk to liberals but liberals don't talk to conservatives. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, volume 15, pages 808–819, 2021. doi:10.1609/icwsm.v15i1.18105.

Zaratiana, Urchade, Tomeh, Nadi, Holat, Pierre, and Charnois, Thierry. GLiNER: Generalist model for named entity recognition using bidirectional transformer. In *Proceedings of NAACL 2024*, pages 5364–5376, 2024. doi:10.18653/v1/2024.naacl-long.300.

Ziems, Caleb, Held, William, Shaikh, Omar, Chen, Jiaao, Zhang, Zhehao, and Yang, Diyi. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291, 2024. doi:10.1162/coli_a_00502.

Zimmer, Michael. "But the data is already public": On the ethics of research in Facebook. *Ethics and Information Technology*, 12(4):313–325, 2010. doi:10.1007/s10676-010-9227-5.

A. Annotation Prompt for `detect_pol`

The following prompt is used to instruct the LLM (GPT-5.2) for the political content detection task. The codebook defines a broad conception of "the political" grounded in political science theory, includes decision rules, and provides annotated examples to guide consistent annotation. The prompt is reproduced in full below.

```
You are a text annotator specialized in the analysis
of political discourse.

This annotation task is part of a scientific research
project that measures the presence of political content
in social media discourse (transcriptions of YouTube
videos).

The objective of this step is to determine whether a
sentence is political or non-political, according to a
broad definition of politics.

This filtering step precedes ideological annotation.
Precision and restraint are therefore essential.

## General Instructions

- Output a single JSON object containing one key
  ("political") every time.
- The value is either "yes" or "no".
- The classification is binary.
- When in doubt, prefer "no".
- Do not infer intentions beyond what is expressed.
- A sentence may be political even if it is expressed
  emotionally, polemically, humorously, or ironically,
  does not use explicit political vocabulary, or
  expresses a personal opinion about a collective issue.
- Output the JSON only, with no comments.

## Definition of Politics (Broad Definition)

A sentence is considered political if it refers,
explicitly or implicitly, to at least one of the
following:
- Current affairs, public debates, or media
  controversies
- Social issues (e.g. immigration, security, gender,
```

```
  ecology, education, religion, identity, inequality,
  technology, public health, etc.)
- Political actors, institutions, or collective rules
  (state, government, parliament, justice system,
  elections, laws, parties, public policies, media
  institutions, etc.)
- Power relations, collective conflicts, ideological
  oppositions
- Social norms or collective values presented as
  desirable, threatened, declining, or needing reform

## Not Political

A sentence is non-political ("no") if it:
- Refers only to personal life, private anecdotes, or
  individual experiences without broader collective
  implications
- Is purely narrative, descriptive, technical, or
  conversational without societal relevance
- Concerns entertainment, lifestyle, or storytelling
  without reference to collective issues

## Annotated Examples

Example 1:
"Le gouvernement doit agir immediatement pour reformer
le systeme de sante, les hopitaux sont a bout."
-> {"political": "yes"}

Example 2:
"Hier j'ai fait des crepes avec ma fille, c'etait
super sympa."
-> {"political": "no"}

Example 3:
"On vit dans une societe ou les riches deviennent de
plus en plus riches et les pauvres de plus en plus
pauvres, c'est quand meme hallucinant."
-> {"political": "yes"}

Example 4:
"Macron a encore fait un discours vide, comme
d'habitude. Ce mec ne represente personne."
```

```
-> {"political": "yes"}

Example 5:
"J'ai regarde la derniere saison de cette serie,
franchement les acteurs sont incroyables."
-> {"political": "no"}

Example 6:
"Le probleme c'est que personne veut parler de
l'insecurite dans les quartiers, c'est tabou."
-> {"political": "yes"}

Example 7:
"Mon pote il a ouvert un resto a Lyon, ca marche
super bien apparemment."
-> {"political": "no"}

Example 8:
"L'ecole publique est en train de mourir, et tout
le monde s'en fout."
-> {"political": "yes"}

Example 9:
"On nous impose un modele de societe qui ne
correspond pas a ce que veulent les gens."
-> {"political": "yes"}

Example 10:
"C'est vrai que le cafe en grain c'est quand meme
bien meilleur que les capsules."
-> {"political": "no"}

Expected JSON Keys: {"political": ""}
```

B. Platform Screenshots

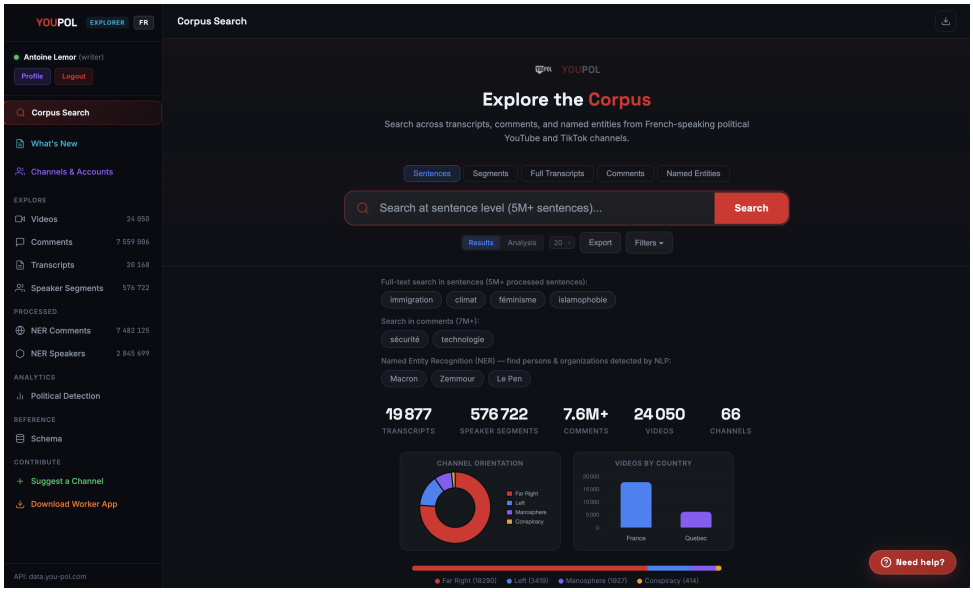

Figure 16: The YouPol data API (`data.you-pol.com`). Researchers can browse, search, filter, and export corpus data. A Python client library (`youpol-client`) provides programmatic access.

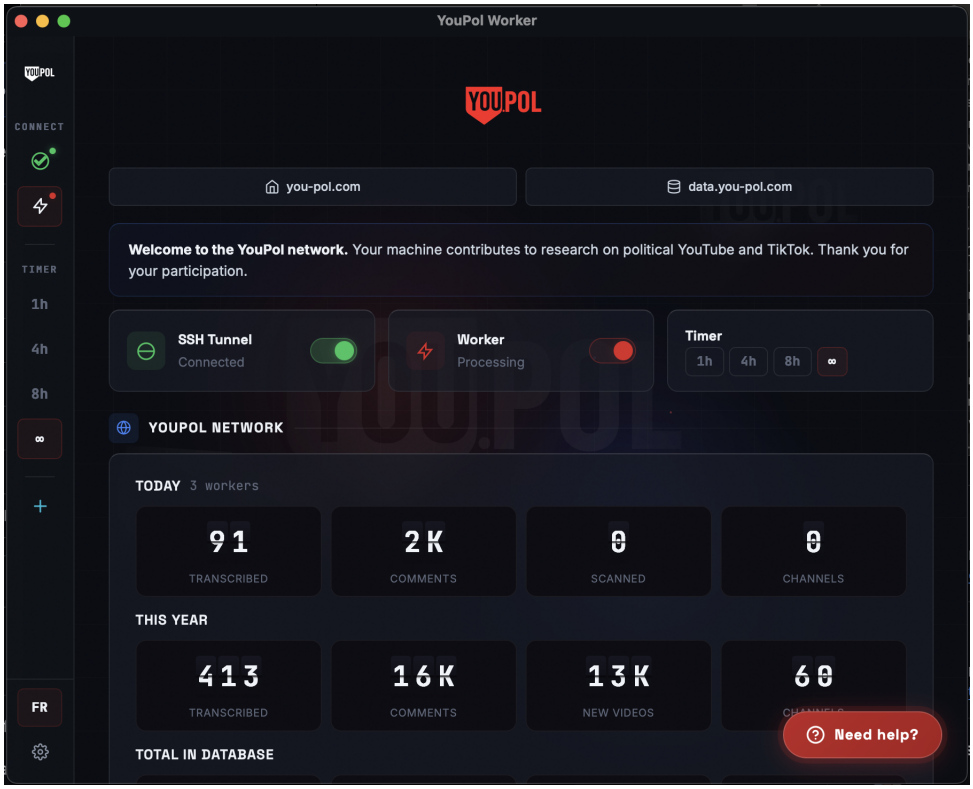

Figure 17: The YouPol Worker application. Each collaborating researcher runs this application on their machine to participate in the YCCN.

C. Database Schema Overview

Table 4 summarizes the main table groups in the YouPol PostgreSQL schema, with representative tables and row counts as of April 2026.

Table 4: YouPol database schema overview (April 2026).

Group	Key tables	Rows
Content (YouTube)		
	`youtube_videos`	23,712
	`youtube_comments`	7,585,328
	`channels`	68
Content (TikTok)		
	`tiktok_videos`	1,685
	`tiktok_comments`	118,335
Transcripts (YouTube)		
	`youtube_video_transcripts`	20,234
	`youtube_transcription_speakers`	641,613
	`youtube_transcription_speakers_processed`	3,167,198
Transcripts (TikTok)		
	`tiktok_video_transcripts`	1,680
	`tiktok_transcription_speakers`	4,125
	`tiktok_transcription_speakers_processed`	15,507
Combined views		
	`videos` (all platforms)	25,397
	`transcription_speakers` (all)	645,738
	`speakers_with_pol` (all)	645,738
Metadata history		
	`video_metadata_history`	33,746
	`channel_metadata_history`	620
Pipeline (YCCN)		
	`pipeline_videos`	16,167
	`pipeline_new_videos`	14,608
	`pipeline_events`	23,488
	`pipeline_workers`	5

The `transcription_speakers_processed` table contains each speaker segment replicated

across all annotation dimensions (political detection, NER tags), while `speakers_with_pol` provides a deduplicated view with the `detect_pol` label for each segment. The `video_metadata_history` table records one row per video per observation cycle, enabling longitudinal engagement analysis. Suppressed content is flagged in the `videos` table (`suppressed = true`) rather than deleted, preserving the full analytical record.

Affiliation:

Antoine Lemor
Université de Sherbrooke, CIRST, RFICS
Sherbrooke, QC, Canada
E-mail: antoine.lemor@usherbrooke.ca
URL: https://antoinelemor.github.io/

Tristan Boursier
Sciences Po Paris & Université du Québec à Montréal
E-mail: tristan.boursier@sciencespo.fr

SocArXiv Website https://socopen.org/
SocArXiv Preprints https://osf.io/preprints/socarxiv

Preprint 2026 *Submitted:* 2026-04-01
10.31235/osf.io/vpzmq$_v$2 ***Accepted:***