YOUPOL: A TEXTUAL DATABASE OF OVER 20,000 VIDEOS BY FRANCOPHONE POLITICAL INFLUENCERS ON YOUTUBE (2006–2024)

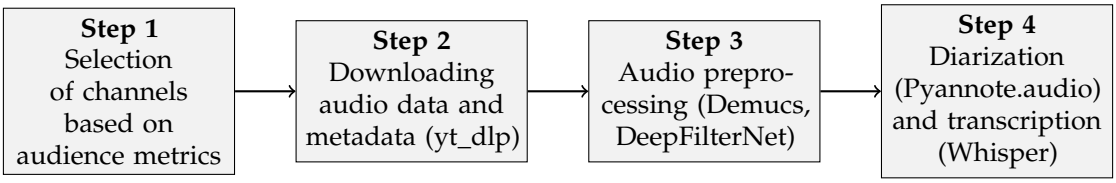
Antoine Lemor[1] & Tristan Boursier[2]

We introduce a database compiling the transcripts of over 20,000 videos by francophone political influencers (France-Quebec) on YouTube since 2006. Designed to be directly usable for Natural Language Processing (NLP) analyses, this database also includes all metadata for each video, notably more than 7 million comments. The corpus specifically targets major political content creators (Finlayson, 2022) spanning the entire French and Quebec political spectrum, from the far left to the far right (Riedl et al., 2021). The corpus is distinctive due to its scale, granularity (including speaker diarization), and especially its capacity to longitudinally and computationally analyze video content—where previous studies focused only on titles. This database thus enables the longitudinal analysis of the dissemination of political ideas on YouTube over time and across the entire political spectrum in both France and Quebec—something that, to the authors' knowledge, has not been previously accomplished.

Methodology, Pipeline, and Software

The pipeline used to construct the database consisted of four main stages, represented in Figure 1 below. All code was written in Python. (Step 1) Videos were collected from YouTube channel links specifically identified for their francophone political content and their role in the ecosystem of political content creators on YouTube (e.g., number of views and subscribers). (Step 2) Metadata (e.g., channel name, number of views, comments, subscribers, etc.) and video audio files were then extracted using the yt_dlp library. (Step 3) Audio tracks from the videos were preprocessed to reduce noise using Demucs and DeepFilterNet. (Step 4) Speaker segmentation was performed using pyannote.audio for diarization, enabling the identification and delimitation of different speakers' interventions within each video. Whisper, OpenAI's open-source transcription model, was then used to generate transcripts of the audio tracks, with each segment reassigned to its respective speaker identified during diarization. The transcripts were subsequently structured with timestamps and speaker identifiers, ensuring alignment between text, diarization, and audio. All data was stored in an SQL database divided into three tables (metadata, comments, and transcripts).

Figure 1. YOUPOL Database Construction Pipeline.



Implementation & Hardware

The implementation of the pipeline faced several technical challenges, with two main issues: (1) the size of the downloaded audio tracks (>15 TB); and (2) the computational power required to execute the pipeline. To manage the volume of audio data, the audio

tracks were downloaded in successive 1 TB batches on a personal machine. Each batch was then uploaded to the servers of the Digital Research Alliance of Canada, to which the first author has access. Preprocessing, diarization, and transcription were subsequently performed for each batch on the Alliance's compute clusters. The final transcriptions were then aggregated into the final SQL database, stored on a private server.

Database Status and Validity

Audio preprocessing and transcription on the compute clusters are still ongoing—40% have been completed at the time of writing—but are expected to be finished by January 2025. Although Whisper offers particularly high transcription quality and preliminary tests were conducted before the pipeline was launched to ensure the quality of both the transcriptions and diarization, a test will be carried out to validate the quality of the completed transcriptions once the database is finalized. A representative sample of sentences will be selected and manually annotated to validate transcription quality. The database will be definitively completed by early February 2025, and its contents will be migrated to a public server, with access provided upon request before final publication.

Originality and Scientific Interest

To the authors' knowledge, YOUPOL is the first francophone database enabling longitudinal analysis of political discourse on YouTube (Forchtner, 2020; Mirrlees, 2018; Stephan, 2024) at the content level while including all metadata for each video. By doing so, the database will facilitate a wide variety of analyses on the dissemination of political ideas—particularly those of the far right (Carter, 2018)—and their content over time and across the political spectrum. The database also allows for novel analyses focusing on the determinants of video content. It will thus be possible to identify the discursive strategies employed by content creators to generate more views (Boursier, 2022, 2024), with content variation potentially linked to audience variations for a given channel. The database is already tied to two ongoing research projects. The first aims to study the determinants of hate comments (Voirol & Martini, 2023) under YouTube videos, based on the content of the videos and the dissemination of far-right ideas. The second aims to examine the impact and use of scientific arguments on these same comments, depending on the political orientation of YouTube channels. Most importantly, this database enables a wide range of studies based on Natural Language Processing (NLP) and video content annotation, something that, to the authors' knowledge, has not been accomplished before.

References

Boursier, T. (2022). White Supremacism on YouTube: How to Rewrite History from a Racist Point of View. In C. Kaiser, O. Khiari, & V. S. Lühr (Eds.), *Temporalities of Diversity / Temporalités de la diversité / Zeitlichkeiten der Vielfalt* (p. 65-87). Waxmann.

Boursier, T. (2024). La banalisation du suprémacisme blanc sur YouTube : Analyse des convergences et influences idéologiques au sein de l'extrême droite française. *Politique et sociétés*, 42(1).

Carter, E. (2018). Right-wing extremism/radicalism: Reconstructing the concept. *Journal of Political Ideologies*, 23(2), 157-182.

Finlayson, A. (2022). YouTube and Political Ideologies: Technology, Populism and Rhetorical Form. *Political Studies*, 70(1), 62-80.

Forchtner, B. (Ed.). (2020). *The far right and the environment: Politics, discourse and communication*. Routledge, Taylor & Francis Group.

Mirrlees, T. (2018). The Alt-Right's Discourse on 'Cultural Marxism': A Political Instrument of Intersectional Hate. *Atlantis: Critical Studies in Gender, Culture & Social Justice*, 39(1), 49.

Riedl, M., Schwemmer, C., Ziewiecki, S., & Ross, L. M. (2021). The Rise of Political Influencers—Perspectives on a Trend Towards Meaningful Content. *Frontiers in Communication*, 6, 1-7.

Stephan, G. (2024). Faire carrière dans les médias de « réinformation » : Les dynamiques d'engagement dans les mobilisations informationnelles d'extrême droite en France (2007-2022). *Politiques de communication*, N° 22(1), 91-122.

Voirol, O., & Martini, É. (2023). La fabrique discursive de la haine : Affects, agitation fasciste et « politique du ressentiment ». *Réseaux*, N° 241(5), 39-77.

---

[1] Postdoctoral researcher, Université de Montréal & Université de Sherbrooke, Centre interuniversitaire de recherche sur la science et la technologie (CIRST), Réseau francophone international en conseil scientifique (RFICS). antoine.lemor@umontreal.ca

[2] Postdoctoral researcher, Sciences Po Paris & Université du Québec en Outaouais