YOUPOL : UNE BASE DE DONNÉES TEXTUELLES DE PLUS DE 20 000 VIDÉOS D'INFLUENCEURS POLITIQUES FRANCOPHONES SUR YOUTUBE (2006–2024)

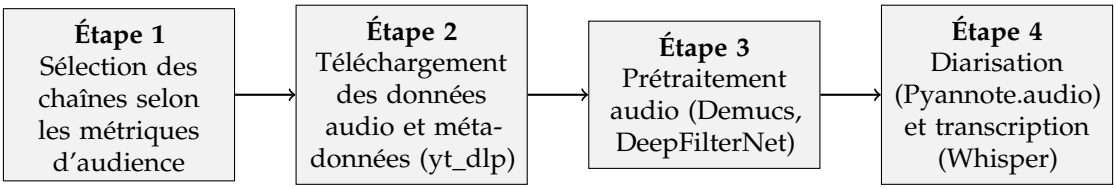
Antoine Lemor[1] & Tristan Boursier[2]

Nous présentons une base de données compilant les transcriptions de plus de 20 000 vidéos d'influenceurs politiques francophones (France-Québec) sur YouTube depuis 2006. Conçue pour être directement utilisable pour des analyses de traitement automatique du langage naturel (TALN), cette base de données inclut également toutes les métadonnées de chaque vidéo, notamment plus de 7 millions de commentaires. Le corpus cible spécifiquement les principaux créateurs de contenu politique (Finlayson, 2022) couvrant l'ensemble du spectre politique français et québécois, de l'extrême gauche à l'extrême droite (Riedl et al., 2021). Le corpus se distingue par son envergure, sa granularité (incluant la diarisation des locuteurs) et surtout sa capacité à analyser longitudinalement et computationnellement le contenu des vidéos—là où les études précédentes se concentraient uniquement sur les titres. Cette base de données permet ainsi l'analyse longitudinale de la diffusion des idées politiques sur YouTube dans le temps et à travers l'ensemble du spectre politique en France et au Québec—ce qui, à la connaissance des auteurs, n'avait jamais été accompli auparavant.

Méthodologie, pipeline et logiciels

Le pipeline utilisé pour construire la base de données comprenait quatre étapes principales, représentées dans la Figure 1 ci-dessous. Tout le code a été écrit en Python. (Étape 1) Les vidéos ont été collectées à partir de liens de chaînes YouTube spécifiquement identifiées pour leur contenu politique francophone et leur rôle dans l'écosystème des créateurs de contenu politique sur YouTube (p. ex., nombre de vues et d'abonnés). (Étape 2) Les métadonnées (p. ex., nom de la chaîne, nombre de vues, commentaires, abonnés, etc.) et les fichiers audio des vidéos ont ensuite été extraits à l'aide de la bibliothèque yt_dlp. (Étape 3) Les pistes audio des vidéos ont été prétraitées pour réduire le bruit à l'aide de Demucs et DeepFilterNet. (Étape 4) La segmentation des locuteurs a été effectuée à l'aide de pyannote.audio pour la diarisation, permettant l'identification et la délimitation des interventions des différents locuteurs au sein de chaque vidéo. Whisper, le modèle de transcription open source d'OpenAI, a ensuite été utilisé pour générer les transcriptions des pistes audio, chaque segment étant réattribué à son locuteur respectif identifié lors de la diarisation. Les transcriptions ont ensuite été structurées avec des horodatages et des identifiants de locuteurs, assurant l'alignement entre le texte, la diarisation et l'audio. Toutes les données ont été stockées dans une base de données SQL divisée en trois tables (métadonnées, commentaires et transcriptions).

Figure 1. Pipeline de construction de la base de données YOUPOL.



Implémentation et matériel

L'implémentation du pipeline a fait face à plusieurs défis techniques, avec deux problèmes principaux : (1) la taille des pistes audio téléchargées (>15 To) ; et (2) la puissance de calcul requise pour exécuter le pipeline. Pour gérer le volume de données audio, les pistes audio ont été téléchargées en lots successifs de 1 To sur une machine personnelle. Chaque lot a ensuite été téléversé sur les serveurs de l'Alliance de recherche numérique du Canada, auxquels le premier auteur a accès. Le prétraitement, la diarisation et la transcription ont ensuite été effectués pour chaque lot sur les grappes de calcul de l'Alliance. Les transcriptions finales ont ensuite été agrégées dans la base de données SQL finale, stockée sur un serveur privé.

État de la base de données et validité

Le prétraitement audio et la transcription sur les grappes de calcul sont toujours en cours—40% ont été complétés au moment de la rédaction—mais devraient être terminés d'ici janvier 2025. Bien que Whisper offre une qualité de transcription particulièrement élevée et que des tests préliminaires aient été effectués avant le lancement du pipeline pour assurer la qualité des transcriptions et de la diarisation, un test sera effectué pour valider la qualité des transcriptions complétées une fois la base de données finalisée. Un échantillon représentatif de phrases sera sélectionné et annoté manuellement pour valider la qualité des transcriptions. La base de données sera définitivement complétée début février 2025, et son contenu sera migré vers un serveur public, avec accès fourni sur demande avant la publication finale.

Originalité et intérêt scientifique

À la connaissance des auteurs, YOUPOL est la première base de données francophone permettant l'analyse longitudinale du discours politique sur YouTube (Forchtner, 2020 ; Mirrlees, 2018 ; Stephan, 2024) au niveau du contenu tout en incluant toutes les méta-données de chaque vidéo. Ce faisant, la base de données facilitera une grande variété d'analyses sur la diffusion des idées politiques—particulièrement celles de l'extrême droite (Carter, 2018)—et leur contenu dans le temps et à travers le spectre politique. La base de données permet également des analyses novatrices se concentrant sur les déterminants du contenu des vidéos. Il sera ainsi possible d'identifier les stratégies discursives employées par les créateurs de contenu pour générer plus de vues (Boursier, 2022, 2024), avec une variation du contenu potentiellement liée aux variations d'audience pour une chaîne donnée. La base de données est déjà liée à deux projets de recherche en cours. Le premier vise à étudier les déterminants des commentaires haineux (Voirol & Martini, 2023) sous les vidéos YouTube, en fonction du contenu des vidéos et de la diffusion des idées d'extrême droite. Le second vise à examiner l'impact et l'utilisation des arguments scientifiques sur ces mêmes commentaires, selon l'orientation politique des chaînes YouTube. Plus important encore, cette base de données permet un large éventail d'études basées sur le traitement automatique du langage naturel (TALN) et l'annotation du contenu vidéo, ce qui, à la connaissance des auteurs, n'avait jamais été accompli auparavant.

Références

Boursier, T. (2022). White Supremacism on YouTube : How to Rewrite History from a Racist Point of View. Dans C. Kaiser, O. Khiari, & V. S. Lühr (Éds.), *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. (Éd.). (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] Chercheur postdoctoral, 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] Chercheur postdoctoral, Sciences Po Paris & Université du Québec en Outaouais