SocArXiv

SocArXiv : 2026-04-04 10.31235/osf.io/vpzmq_v2

YouPol : Une infrastructure collaborative de recherche et une base de données pour le contenu politique sur YouTube et TikTok

Antoine Lemor
Université de Sherbrooke, CIRST, RFICS

Tristan Boursier
Sciences Po Paris & Université du Québec à Montréal

Abstract

Cet article présente YouPol (YouTube and TikTok Political Observatory and Longitudinal database), une infrastructure de recherche continûment mise à jour qui capture ce que les créateurs de contenu politique disent réellement sur les plateformes vidéo. En date d'avril 2026 et en expansion continue, le corpus comprend 25 397 vidéos provenant de 68 chaînes en France et au Québec, avec des transcriptions complètes diarisées par locuteur (645 738 segments, 3,18 millions de phrases annotées) et 7,7 millions de commentaires archivés. L'infrastructure inclut un pipeline de transcription indépendant qui produit des transcriptions de haute qualité indépendamment des sous-titres fournis par les plateformes, et un cadre d'annotation LLM-in-the-loop construit sur la plateforme open-source LLM Tool (Lemor et al., 2025) capable d'entraîner des classifieurs phrastiques pour tout projet de recherche, les projets en cours couvrant la détection du contenu politique, l'idéologie d'extrême droite, la rhétorique genrée et le discours néo-réactionnaire. Pour produire des mises à jour de transcription et de métadonnées en temps réel, YouPol introduit également le Réseau de Calcul Collaboratif YouPol (YCCN), qui permet à tout chercheur collaborateur de contribuer de la capacité de traitement depuis sa propre machine, affranchissant l'observatoire de toute dépendance aux grappes de calcul institutionnelles. YouPol répond à quatre lacunes dans la littérature : (1) la substance idéologique du contenu vidéo politique reste empiriquement inaccessible par les seules métadonnées ; (2) la suppression de contenu et la déplatformation effacent le matériel avant que les chercheurs puissent l'étudier ; (3) les dynamiques longitudinales d'engagement sont sous-exploitées ; et (4) aucune base existante ne préserve les commentaires dans le temps ni ne suit leur suppression. L'observatoire a déjà préservé 2 305 vidéos et trois chaînes entièrement supprimées qui ne sont plus disponibles sur les plateformes. La base et l'API sont accessibles à l'adresse `data.you-pol.com`.

Keywords: sciences sociales computationnelles, discours politique, YouTube, transcription de la parole, TAL, infrastructure de recherche.

Word count : 7223

Introduction

Le 12 mars 2025, la chaîne YouTube d'extrême droite française FDENEWS, qui avait accumulé 1 755 vidéos, 68,3 millions de vues et 103 000 abonnés en près d'une décennie, a été définitivement supprimée par YouTube pour violations répétées de ses règles communautaires. Pour tout chercheur qui n'avait pas anticipé cet événement, les données étaient irrémédiablement perdues. Des milliers d'heures de commentaire politique, des centaines de milliers de commentaires d'utilisateurs et un registre longitudinal de l'un des médias d'extrême droite les plus importants de France avaient simplement cessé d'exister. Ce type d'événement souligne un besoin clair d'infrastructures de recherche capables de collecter, traiter et préserver en continu le contenu vidéo politique *avant* que les plateformes ne le suppriment. L'observatoire présenté dans cet article, YouPol, répond à ce besoin. Parce qu'il surveillait, entre autres chaînes, FDENEWS depuis son inclusion dans le corpus, la suppression n'a eu aucun effet sur le dossier analytique. Chaque vidéo avait déjà été téléchargée et transcrite, ainsi que chaque commentaire et métadonnée. La disparition de la chaîne de YouTube a elle-même été enregistrée comme un point de données.

Cet exemple illustre, en pratique, les difficultés auxquelles sont confrontées les sciences sociales computationnelles dans l'étude du discours politique sur les plateformes (vidéo). Le contenu que les chercheurs ont le plus besoin d'étudier (radical, transgressif, politiquement conséquent) est précisément celui que les plateformes sont les plus susceptibles de supprimer. Sans infrastructure conçue pour une collecte continue et préventive, les chercheurs se retrouvent à n'étudier que ce que les plateformes laissent visible. Cela introduit un biais de survie qui compromet la validité de toute analyse de l'extrémisme politique en ligne. C'est précisément ce à quoi YouPol répond en fonctionnant comme un observatoire permanent qui collecte, transcrit, diarise et annote en continu le contenu vidéo politique sur les plateformes, tout en préservant le matériel que les plateformes suppriment par la suite.

Les plateformes de partage vidéo sont devenues des espaces centraux pour la communication politique, la mobilisation idéologique et la délibération publique (Munger and Phillips 2022). YouTube héberge un écosystème de créateurs de contenu politique dont la production collective rivalise avec les médias traditionnels en portée (Rieder, Coromina, and Matamoros-Fernández 2020). Les affordances algorithmiques de TikTok ont encore transformé la circulation du contenu politique, en particulier parmi les publics plus jeunes (Gerbaudo 2026; Guinaudeau, Munger, and Votta 2022). Pourtant, malgré l'importance politique de ces plateformes, la communauté des sciences sociales computationnelles manque de jeux de données au niveau des transcriptions, maintenus longitudinalement, qui capturent ce qui est réellement *dit* dans les vidéos politiques (Lazer, Pentland, Watts, Aral, Athey, Contractor, et al. 2020). La plupart des études existantes s'appuient sur les titres, les métadonnées ou les schémas de recommandation et passent donc à côté du contenu idéologique substantiel que seules la transcription et l'analyse textuelle fine peuvent révéler.

Cette lacune est particulièrement aiguë hors du contexte anglophone. Alors que le YouTube politique anglophone a reçu une attention soutenue, des audits de parcours de radicalisation (Ribeiro, Ottoni, West, Almeida, and Meira 2020) aux analyses offre-demande de contenu de droite (Munger and Phillips 2022), les écosystèmes politiques francophones restent sous-étudiés. Les travaux antérieurs sur les créateurs de contenu politique français se sont largement appuyés sur les titres et métadonnées des vidéos, négligeant la substance idéologique (Boursier 2025).

L'"ère post-API" de la recherche sur les médias sociaux (Freelon, Monzer, Jeon, Moy, and Williams 2024) a aggravé ce défi. Les restrictions d'accès aux données imposées par les plateformes (Bruns 2019; Tromble 2021), la dépréciation de CrowdTangle et les limites de l'API de recherche de TikTok (Pearson, Silver, Robinson, Azadi, Schillo, and Kreslake 2024) rendent de plus en plus difficile pour les chercheurs de constituer et maintenir des jeux de données complets (Chen, Sherren, Lee, McCay-Peet, Xue, and Smit 2024; Ohme, Araujo, Boeschoten, Freelon, Ram, Reeves, and Robinson 2024). Lorsque les plateformes fournissent des API de recherche, des audits systématiques ont révélé des biais significatifs qui menacent la validité des résultats (Rieder, Padilla, and Coromina 2025; Bai and Gu 2026).

Dans cet article, nous présentons YouPol (YouTube and Tiktok Political Observatory and Longitudinal database), une infrastructure de recherche conçue pour répondre à ces défis. YouPol est un *observatoire permanent* qui collecte, transcrit, diarise et annote en continu le contenu vidéo politique à travers les plateformes. Ses contributions sont triples. Premièrement, YouPol constitue, à notre connaissance, la première base de données de vidéos politiques au niveau des transcriptions de cette ampleur. En date d'avril 2026 et en expansion continue, elle comprend 23 712 vidéos YouTube et 1 685 vidéos TikTok avec diarisation des locuteurs, 7,7 millions de commentaires et 645 738 segments de locuteurs découpés en 3,18 millions de phrases annotées, provenant de 68 chaînes en France et au Québec.

Deuxièmement, l'observatoire fonctionne comme une infrastructure permanente et collaborative plutôt que comme une collecte ponctuelle. Tout chercheur peut rejoindre le YouPol Collaborative Computing Network (YCCN) en installant sur sa machine une application dédiée développée par les auteurs, qui agit ensuite comme un nœud autonome de transcription et de traitement. Le YCCN scanne les chaînes en continu, traite les nouveaux contenus et préserve le matériel que les plateformes suppriment, y compris les vidéos supprimées, les chaînes supprimées et les commentaires modérés. Le traitement initial a été effectué sur l'infrastructure de calcul haute performance de l'Alliance de recherche numérique du Canada ; la production a depuis été transférée au YCCN, rendant l'observatoire pleinement autonome et indépendant des grappes de calcul institutionnelles. Ce qui est particulièrement important dans le contexte actuel de restriction de l'accès aux données, et parfois d'hostilité envers les chercheurs.

Troisièmement, un cadre d'annotation LLM-in-the-loop construit sur la plateforme open-source LLM Tool (Lemor, Dinan, and Gilbert 2025) produit des classifieurs CamemBERT pour

la détection phrastique de l'idéologie politique, de la rhétorique genrée et du discours néo-réactionnaire. Le classifieur de détection politique (`detect_pol`) et le pipeline d'annotation complet sont en production ; l'intégration de TikTok est en cours, avec priorité accordée aux comptes TikTok des créateurs déjà suivis sur YouTube. L'approche suit le paradigme de distillation des connaissances validé par des travaux méthodologiques récents (Pangakis and Wolken 2024; Gilardi, Alizadeh, and Kubli 2023; Alizadeh, Kubli, Samei, Dehghani, Zahedivafa, Bermeo, Korobeynikova, and Gilardi 2025).

1. Revue de littérature

1.1. Bases de données existantes sur le discours politique : couverture et limites

Les sciences sociales computationnelles disposent de plusieurs bases importantes sur le discours politique, mais aucune ne capture le contenu parlé des vidéos politiques sur les réseaux sociaux. Media Cloud (Roberts, Bhargava, Valiukas, Jen, Malik, Bishop, et al. 2021) collecte des articles d'actualité provenant de plus de 60 000 sources avec des mises à jour continues, mais couvre le journalisme textuel, pas les plateformes vidéo. ParlSpeech V2 (Rauh and Schwalbach 2020) fournit 6,3 millions de discours parlementaires de neuf démocraties, et CoCoHD (Hiray, Liu, Song, Shah, and Chava 2024) offre 32 697 transcriptions d'auditions du Congrès américain ; les deux contiennent du texte intégral, mais de procédures institutionnelles plutôt que de communication politique informelle sur les réseaux sociaux. Sur des médias de communication politique plus informels comme YouTube, plusieurs études influentes ont eu tendance à s'appuyer exclusivement sur les métadonnées ou les commentaires. Ribeiro et al. (2020) ont publié des métadonnées et commentaires de plus de 330 000 vidéos, Ledwich and Zaitsev (2020) ont classé environ 800 chaînes politiques, et Rauchfleisch and Kaiser (2020) ont analysé les réseaux d'audience de l'extrême droite allemande. Aucun de ces travaux n'inclut de transcriptions vidéo. Fait intéressant, Lai, Brown, Bisbee, Tucker, Nagler, and Bonneau (2024) ont utilisé les transcriptions d'un sous-ensemble de vidéos YouTube politiques pour estimer l'idéologie des chaînes, mais ont noté une pénurie sévère de données, les transcriptions n'étant disponibles que pour une minorité de leur échantillon. Cette pénurie est particulièrement aiguë dans les contextes non anglophones : aucun corpus de contenu vidéo politique au niveau des transcriptions n'existe pour les écosystèmes francophones ou d'autres langues, où les travaux antérieurs se sont appuyés presque exclusivement sur les titres et métadonnées des vidéos (Gilliotte 2024; Rauchfleisch and Kaiser 2020).

Les quelques études intégrant des transcriptions reposent sur les sous-titres fournis par les plateformes. Sosnovik, Violot, and Humbert (2025) ont collecté plus de 100 000 transcriptions YouTube françaises des élections 2024 via les transcriptions fournies par la plateforme et

appliqué un étiquetage thématique par LLM. Ce corpus ne couvre toutefois qu'une seule période électorale, ne propose pas de diarisation des locuteurs, offre une annotation thématique plutôt que phrastique et constitue un instantané ponctuel. Une limitation fondamentale de cette approche est que les sous-titres automatiques sont souvent de qualité médiocre, ne sont pas disponibles pour toutes les vidéos (en particulier les contenus anciens ou peu populaires) et n'identifient pas les locuteurs individuels. YouPol répond à cette limite par un pipeline de transcription indépendant à l'état de l'art qui combine traitement et nettoyage audio par séparation neuronale des sources (Demucs), reconnaissance vocale de pointe (Whisper large-v3) et diarisation des locuteurs (pyannote.audio). Ce pipeline produit des transcriptions étiquetées par locuteur de haute qualité constante, indépendamment de la disponibilité ou non des sous-titres de la plateforme.

Pinto, Bickham, Salkar, Menezes, Luceri, and Ferrara (2025) ont collecté 3,14 millions d'identifiants vidéo avec des transcriptions générées par IA sur TikTok lors de l'élection américaine de 2024, mais l'ont fait sans annotation TAL ni mise à jour continue. Solovev, Drolsbach, Demirel, and Pröllochs (2026) ont analysé plus de 25 000 vidéos TikTok de l'élection fédérale allemande de 2025 et ont constaté que les appels émotionnels négatifs augmentent significativement l'engagement, bien que leur analyse repose sur un codage computationnel du contenu plutôt que sur une transcription indépendante. Le tableau 1 positionne YouPol par rapport à ces ressources. Aucune base existante ne combine transcriptions vidéo produites par un pipeline de transcription indépendant, diarisation des locuteurs, annotation TAL phrastique, mise à jour continue et préservation du contenu supprimé.

1.2. Quatre lacunes non résolues : transcriptions, préservation du contenu, trajectoires d'engagement et modération des commentaires

La recherche sur le contenu politique des plateformes vidéo a produit des résultats importants

TABLE 1 – YouPol comparé aux bases de données existantes sur le discours politique.

Base de données	Source	Type de contenu	Échelle	MAJ
YouPol	YT + TikTok	Transcriptions + diarisation (français)	25,4K vidéos, 7,7M commentaires	oui
Sosnovik et al. 2025	YouTube	Sous-titres auto (français)	100K vidéos	non
Pinto et al. 2025	TikTok	Transcriptions IA (anglais)	3,14M IDs	non
Ribeiro et al. 2020	YouTube	Métadonnées uniquement (anglais)	330K vidéos, 72M commentaires	non
ParlSpeech V2	Parlement	Transcriptions officielles (9 langues)	6,3M discours	non
Media Cloud	Presse	Articles (20+ langues)	2Mrd articles	oui

mais laisse quatre lacunes fondamentales non résolues.

Premièrement, la dépendance aux métadonnées signifie que la substance idéologique du contenu vidéo politique reste empiriquement inaccessible. Des études influentes sur YouTube ont examiné la migration des utilisateurs vers du contenu extrême (Ribeiro et al. 2020), la concentration de la consommation radicale au sein d'un public restreint (Hosseinmardi, Ghasemian, Clauset, Möbius, Rothschild, and Watts 2021), les dynamiques d'offre et de demande du contenu de droite (Munger and Phillips 2022), les recommandations algorithmiques (Haroon, Wojcieszak, Chhabra, Liu, Mohapatra, and Shafiq 2023), le chevauchement d'audience dans les réseaux d'extrême droite (Rauchfleisch and Kaiser 2020), et la migration entre communautés antiféministes et d'extrême droite à travers 300 millions de commentaires (Mamié, Horta Ribeiro, and West 2021). Sur TikTok, la recherche a abordé l'émergence de "clusters d'intérêt social" formés algorithmiquement (Gerbaudo 2026), le rôle de la viralité par rapport au nombre d'abonnés (Guinaudeau et al. 2022), et l'effet des appels émotionnels négatifs sur l'engagement (Solovev et al. 2026). Si chacune de ces contributions fait avancer notre compréhension des plateformes vidéo politiques, toutes s'appuient sur les métadonnées, les commentaires ou les traces comportementales. Aucune n'a accès à ce que les créateurs disent réellement dans leurs vidéos, ce qui laisse le contenu idéologique du discours politique empiriquement hors de portée.

Deuxièmement, la suppression de contenu et le déplatformage restent invisibles dans les bases existantes. Lorsque YouTube a définitivement supprimé FDENEWS en mars 2025, la plateforme a effacé 1 755 vidéos, 282 000 commentaires et 68,3 millions de vues cumulées. Rauchfleisch and Kaiser (2024) ont analysé la suppression de plus de 11 000 chaînes YouTube entre 2018 et 2019. Ils ont montré que le déplatformage réduit efficacement la portée du contenu d'extrême droite, mais noté que le contenu lui-même est perdu pour les chercheurs. Jhaver, Boylston, Yang, and Bruckman (2021) ont démontré des dynamiques similaires sur Twitter : le déplatformage réduit à la fois l'activité et la toxicité des partisans, mais le contenu supprimé devient définitivement inaccessible pour l'analyse rétrospective. Plus largement, Lakic, Rossetto, and Bernstein (2023) ont constaté que plus de 20 % des URL dans les jeux de données multimédia collectés sur le web ne sont plus accessibles, ce qui compromet la reproductibilité. Rieder et al. (2025) ont montré que l'API de recherche de YouTube elle-même présente une dégradation temporelle sévère et rend les vidéos précédemment indexées progressivement indécouvrables. YouPol préserve tout le contenu de manière préventive. Le corpus comprend actuellement 2 305 vidéos supprimées et trois chaînes entièrement effacées, permettant l'étude de ce que les plateformes choisissent de supprimer et de ses effets sur le discours public.

Troisièmement, les dynamiques longitudinales d'engagement sont sous-exploitées. La plupart des bases capturent les métadonnées à un instant donné, manquant la trajectoire temporelle du contenu politique. Or, l'évolution du nombre de vues, de likes et de la croissance des abonnés révèle comment l'influence politique se développe, comment les audiences réagissent

aux événements extérieurs et comment les interventions des plateformes reconfigurent la visibilité. La recherche sur la viralité du contenu a montré que les vidéos politiques atteignent rarement leur influence par un pic d'exposition unique, mais plutôt par des schémas de diffusion cumulatifs qui se déploient sur des semaines ou des mois (Vosoughi, Roy, and Aral 2018; Gerrand, Ging, Roose, and Flood 2025; Ganesh 2025). De même, la fidélité de l'audience, comprise comme le retour répété des spectateurs vers la chaîne d'un créateur, est un mécanisme central par lequel les communautés idéologiques se consolident au fil du temps (Munger and Phillips 2022). Cette dimension temporelle est particulièrement conséquente pour les stratégies métapolitiques qui, comme noté ci-dessus, opèrent par la normalisation progressive de cadres idéologiques plutôt que par des événements viraux ponctuels (Schilk 2025; Boursier 2026; Norocel 2023). Évaluer si les récits métapolitiques gagnent ou perdent en emprise culturelle exige d'observer le même contenu au fil du temps, et non un instantané unique. YouPol enregistre des instantanés horodatés de métadonnées à chaque observation, produisant des profils longitudinaux d'engagement pour chaque vidéo et chaque chaîne. Combinés à l'analyse de contenu au niveau des transcriptions, cela permet aux chercheurs d'étudier la relation entre ce que disent les créateurs et la manière dont les audiences répondent au fil du temps.

Quatrièmement, bien que l'analyse des commentaires ait produit des résultats importants (Wu and Resnick 2021; Mamié et al. 2021), *aucune base existante ne préserve les commentaires longitudinalement ni ne suit leur suppression.* Wu and Resnick (2021) ont analysé 134 millions de commentaires YouTube et constaté que les conservateurs sont beaucoup plus susceptibles de commenter les vidéos de gauche que les libéraux sur les vidéos de droite, et que les interactions interpartisanes sont plus toxiques que les interactions co-partisanes. Mamié et al. (2021) ont retracé la migration idéologique à travers 300 millions de commentaires. Pourtant, les pratiques de modération des commentaires, l'évolution des sections de commentaires au fil du temps et la suppression systématique du discours des utilisateurs restent invisibles pour la recherche. YouPol extrait et préserve 7,7 millions de commentaires avec provenance complète (auteur, horodatage, likes, nombre de réponses), y compris les commentaires qui ont été ultérieurement supprimés par les créateurs ou les plateformes. Cela ouvre l'étude de la modération des commentaires en tant que stratégie de communication politique, et permet aux chercheurs de croiser le discours des audiences avec l'analyse de contenu produite par le pipeline d'annotation.

L'"ère post-API" (Freelon et al. 2024) a encore exacerbé ces difficultés. Les restrictions d'accès aux données imposées par les plateformes (Bruns 2019; Tromble 2021), la dépréciation de CrowdTangle et les limites de l'API de recherche de TikTok (Pearson et al. 2024) rendent de plus en plus difficile pour les chercheurs de constituer et maintenir des jeux de données complets (Chen et al. 2024; Ohme et al. 2024). Lorsque les plateformes fournissent des API de recherche, des audits systématiques ont révélé des biais significatifs qui menacent la validité des résultats (Rieder et al. 2025; Bai and Gu 2026). YouPol comble ces quatre lacunes simultanément en combinant collecte indépendante au niveau des transcriptions, préservation

préventive du contenu, suivi longitudinal de l'engagement et archivage des commentaires au sein d'une infrastructure unique continuellement mise à jour. Nous introduisons également le Réseau de Calcul Collaboratif YouPol (YCCN), une architecture distribuée qui permet à tout chercheur collaborateur de contribuer de la capacité de traitement depuis sa propre machine, affranchissant l'observatoire de toute dépendance aux grappes de calcul institutionnelles.

1.3. Atteindre le contenu des vidéos politiques : transcription, diarisation et annotation par LLM

Les lacunes identifiées ci-dessus ne peuvent être comblées que si le contenu parlé des vidéos politiques est converti en texte structuré et analysable. Les avancées récentes en traitement de la parole ont rendu cela possible. Proksch, Wratil, and Wäckerle (2019) ont montré que les transcriptions générées par reconnaissance automatique de la parole (RAP) ne biaisent pas systématiquement les mesures en aval telles que l'analyse de sentiment ou le positionnement idéologique. Whisper (Radford, Kim, Xu, Brockman, McLeavey, and Sutskever 2023), entraîné sur 680 000 heures d'audio web faiblement supervisé, a rendu possible une RAP multilingue robuste dans 99 langues sans ajustement spécifique, et atteint un taux d'erreur de mots (WER) d'environ 5,6 % sur le benchmark Fleurs français. WhisperX (Bain, Huh, Han, and Zisserman 2023) a complété ce système par un alignement forcé et une détection d'activité vocale produisant des horodatages au niveau du mot, ce qui est essentiel pour la diarisation des locuteurs, c'est-à-dire l'identification de qui parle à quel moment. Park, Kanda, Dimitriadis, Han, Watanabe, and Narayanan (2022) proposent une revue complète des avancées récentes en diarisation neuronale. YouPol utilise pyannote.audio (Bredin, Yin, Coria, Gelly, Korshunov, Lavechin, Fustes, Titeux, Bouaziz, and Gill 2020; Bredin 2023), qui a atteint des taux d'erreur de diarisation à l'état de l'art grâce à l'entraînement par ensemble de puissance multi-classes (Plaquet and Bredin 2023), et prétraite l'audio par la séparation neuronale des sources Demucs (Défossez, Usunier, Bottou, and Bach 2019; Rouard, Massa, and Défossez 2023) pour isoler la piste vocale de la musique de fond avant la transcription.

Une fois les transcriptions produites, les chercheurs sont confrontés à un choix de stratégie analytique. Les *approches inductives*, comme la modélisation thématique appliquée par Sosnovik et al. (2025) ou le regroupement par réseaux utilisé par Rauchfleisch and Kaiser (2020), laissent émerger les structures à partir des données sans engagement théorique préalable. YouPol suit une *approche déductive* parce que les construits qu'il cible (idéologie d'extrême droite, rhétorique genrée, discours néo-réactionnaire) sont déjà bien définis dans la littérature en science politique, et leur opérationnalisation à travers des livres de codes explicites produit des schémas d'annotation que d'autres chercheurs *peuvent évaluer, contester et reproduire*. Toutefois, l'application de l'annotation déductive à des millions de phrases nécessite des méthodes spécifiques (Grimmer and Stewart 2013). Des travaux récents ont montré que les LLM peuvent égaler ou dépasser les performances des codeurs humains et des annotateurs

experts sur les tâches d'annotation de textes politiques (Gilardi et al. 2023; Törnberg 2025; Ziems, Held, Shaikh, Chen, Zhang, and Yang 2024; Törnberg 2024). La stratégie utilisée dans YouPol, dans laquelle des étiquettes générées par LLM validées par rapport à des annotations humaines entraînent des classifieurs BERT légers pour un déploiement à l'échelle du corpus, a été validée par Pangakis and Wolken (2024) et Heseltine and Clemm von Hohenberg (2024), qui ont montré que des classifieurs entraînés sur des données codées par LLM produisent des résultats comparables à ceux entraînés sur des données codées manuellement. Alizadeh et al. (2025) ont en outre démontré que des LLM open-source affinés peuvent approcher la performance des modèles propriétaires.

YouPol met en œuvre cette approche au moyen de LLM Tool (Lemor et al. 2025), un pipeline hybride open-source dans lequel un LLM guidé par un livre de codes annote un échantillon stratifié, des annotateurs humains codent indépendamment un sous-ensemble de chevauchement pour valider l'accord inter-annotateurs, et les étiquettes validées sont ensuite utilisées pour affiner un classifieur BERT léger pour le déploiement à l'échelle du corpus. La validation empirique de LLM Tool sur un corpus bilingue de 38 451 textes politiques canadiens a montré que des classifieurs XLM-RoBERTa entraînés sur des étiquettes générées par LLM atteignent un Micro F_1 moyen de 66,7 %, la fidélité de l'annotation plutôt que la taille du modèle étant le principal déterminant de la performance en aval. Au sein de YouPol, le premier classifieur déployé (`detect_pol`) atteint un F_1 macro de 92,2 % et une exactitude de 93,5 % sur les données de validation réservées, avec un κ de Light de 0,787 entre les annotateurs humains et le LLM sur l'échantillon de chevauchement de 1 000 phrases. L'architecture hybride procure une accélération de 110 à 1 580× par rapport à l'annotation directe par LLM, ce qui rend le déploiement à l'échelle du corpus réalisable pour des millions de phrases tout en restant fiable avec de faibles coûts humains et monétaires. YouPol applique ce pipeline avec CamemBERTav2 (Antoun, Kulumba, Touchent, Villemonte de la Clergerie, Sagot, and Seddah 2024), un modèle francophone fondé sur l'architecture DeBERTaV3, pour entraîner des classifieurs au niveau de la phrase pour la détection de contenu politique, l'évaluation idéologique et l'analyse de la rhétorique genrée.

2. Conception et portée de l'observatoire

2.1. Principes de conception

Trois principes guident l'architecture de YouPol. *Permanence* : le système surveille et collecte en continu, fonctionnant comme un observatoire vivant ; le contenu est capturé avant que les plateformes ne puissent le supprimer, et les vidéos supprimées ainsi que les chaînes supprimées sont signalées mais préservées intégralement. *Profondeur* : les transcriptions intégrales avec diarisation des locuteurs permettent une analyse au niveau de la phrase de ce que les créateurs

politiques disent réellement, plutôt qu'une inférence à partir des titres ou des métadonnées de la plateforme. *Scalabilité collaborative* : l'infrastructure croît avec sa communauté d'utilisateurs grâce au *Réseau de Calcul Collaboratif YouPol* (YCCN), une architecture distribuée qui élimine la dépendance aux grappes de calcul institutionnelles. Tout chercheur collaborateur peut contribuer de la capacité de traitement en installant une application dédiée sur sa machine, qui fonctionne ensuite comme un nœud de traitement autonome sur le réseau. Le pipeline est agnostique en termes de langue (Whisper pour la reconnaissance vocale, toute variante BERT pour l'annotation) et peut accueillir de nouvelles langues, plateformes et projets d'annotation sans perturber le traitement existant. Le corpus francophone actuel sert de fondation, et l'expansion anglophone est en préparation.

2.2. Portée du corpus

L'observatoire surveille actuellement 68 chaînes à travers deux écosystèmes politiques francophones (France et Québec) et quatre orientations idéologiques. Ce corpus est conçu pour croître à mesure que de nouvelles chaînes sont identifiées et de nouveaux écosystèmes sont intégrés, incluant bientôt l'écosystème anglophone. Les chaînes d'extrême droite promeuvent un discours ethno-nationaliste, anti-immigration ou autoritaire et constituent la catégorie la plus importante. Les chaînes de gauche sont associées à des perspectives progressistes ou de gauche et sont incluses à des fins comparatives. Les chaînes de la manosphère promeuvent des idéologies antiféministes ou "red pill", à l'intersection de la politique de genre et de la radicalisation politique. Les chaînes complotistes sont centrées sur des récits conspirationnistes qui croisent fréquemment le discours d'extrême droite.

La sélection des chaînes a suivi une approche itérative guidée par l'expertise, cohérente avec les pratiques établies dans le domaine (Lewis 2018; Rauchfleisch and Kaiser 2020). Le processus a débuté par une liste de chaînes connues pour leur rôle dans l'écosystème politique francophone, identifiées par les métriques d'audience (vues, abonnés), les travaux antérieurs (Boursier 2025; Gilliotte 2024) et les rapports de veille médiatique. Des chaînes supplémentaires ont été intégrées par échantillonnage en boule de neige via les réseaux de recommandation de YouTube, suivant le protocole itératif jusqu'à saturation décrit par Reveilhac and Nchakga (2024) pour un corpus comparable de 67 chaînes alternatives françaises. Chaque chaîne a été classée selon trois dimensions : orientation idéologique, pays d'origine et genre du créateur. Le corpus résultant couvre tout le spectre d'audience, des grands médias à plus d'un million d'abonnés (Mediapart, BLAST) aux micro-chaînes de moins de 1 000, captant ainsi les influenceurs à forte audience et la longue traîne de la production de contenu politique. La cartographie indépendante de 127 chaînes politiques françaises par Gilliotte (2024) valide la couverture du corpus YouPol. La figure 1 présente la répartition selon les orientations et les pays, et la figure 2 présente la croissance cumulative du corpus au fil du temps.

3. Le pipeline de traitement en sept étapes

La figure 3 présente l'architecture complète du pipeline. L'observatoire fonctionne en boucle continue. Les vidéos nouvellement découvertes alimentent l'étape de collecte, et chaque étape fonctionne en permanence sur le YCCN.

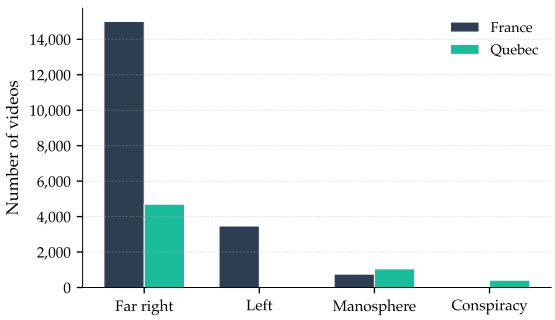

FIGURE 1 – Composition du corpus par orientation idéologique et pays d'origine. Le contenu d'extrême droite constitue la majorité du corpus en France comme au Québec.

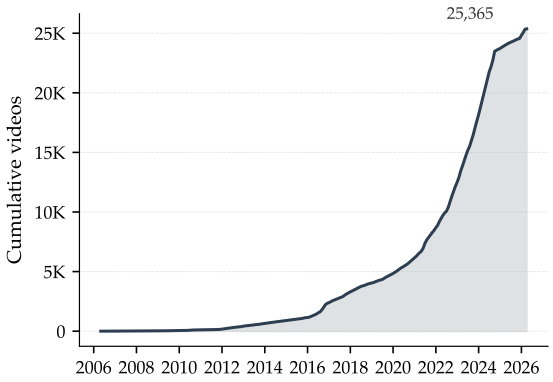

FIGURE 2 – Croissance cumulative du corpus au fil du temps. La production vidéo s'accélère nettement à partir de 2020.

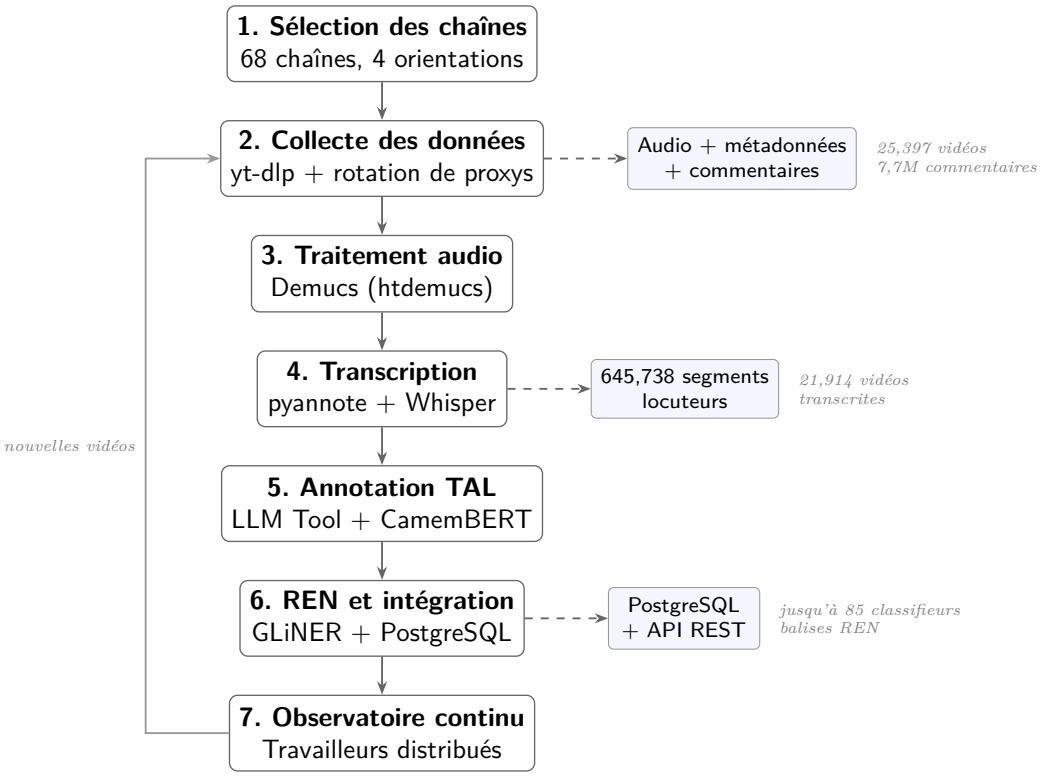

FIGURE 3 – Le pipeline de traitement en sept étapes de YouPol. L'observatoire fonctionne en boucle continue : l'étape 7 réinjecte les vidéos nouvellement découvertes dans l'étape 2.

3.1. Étape 1 : Sélection des chaînes

Les chaînes sont identifiées par une combinaison d'indicateurs d'audience (nombre d'abonnés et de vues), d'analyse de contenu et d'évaluation experte du rôle de chaque chaîne dans l'écosystème de contenu politique francophone. Chaque chaîne est classée selon deux dimensions : l'orientation politique et le pays d'origine. Le corpus couvre actuellement 68 chaînes à travers le spectre politique, du journalisme d'investigation de gauche (par ex., Mediapart, 1,1M d'abonnés) au commentaire d'extrême droite de premier plan (par ex., Frontières · Livre Noir, 565K abonnés). La figure 4 présente les 15 chaînes avec le plus grand nombre cumulé de vues.

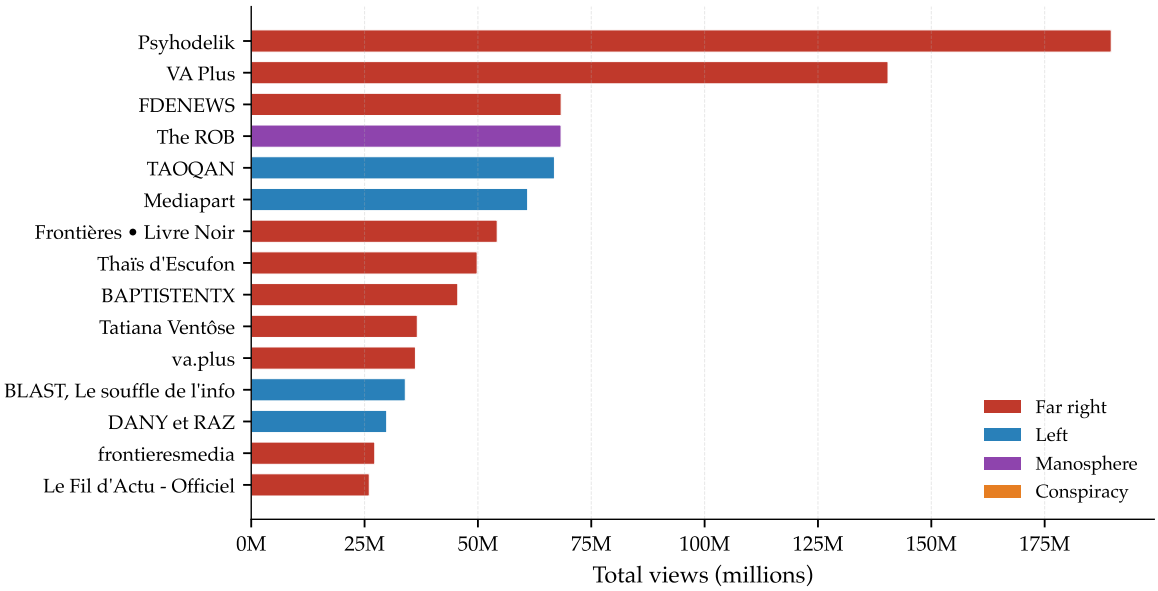

FIGURE 4 – Top 15 des chaînes par vues cumulées, colorées par orientation idéologique. Le corpus capture plus d'un milliard de vues au total (1 156 113 044).

3.2. Étape 2 : Collecte continue des données

Le pipeline scrute en continu toutes les chaînes suivies pour détecter les vidéos nouvellement publiées. [1]

Pour chaque vidéo découverte, le pipeline extrait l'audio au format WAV, les métadonnées (titre, durée, date de mise en ligne, vues, likes, nombre d'abonnés) et les commentaires avec provenance complète (auteur, texte, horodatage, likes, nombre de réponses). Le lot initial de plus de 15 To de données audio a été traité via l'infrastructure de calcul haute performance de l'Alliance de recherche numérique du Canada. Le traitement a depuis été transféré au YCCN

1. Pour YouTube, le scan est effectué via `yt-dlp` avec rotation automatique des cookies. Pour TikTok, un scanner dédié à base de navigateur headless gère l'environnement restrictif de la plateforme par rotation de proxys SOCKS5 (68 serveurs) et mesures anti-détection pour contourner la limitation de débit.

décrit à la section 3.6.

Chaque scan capture un *instantané de métadonnées* horodaté qui enregistre les compteurs de vues, de likes et d'abonnés à chaque observation. Ces instantanés permettent une analyse longitudinale de la dynamique d'engagement. Ils capturent non seulement l'état final mais la trajectoire complète de chaque vidéo, et reproduisent ainsi le tableau de bord et les données d'engagement auxquels les influenceurs politiques ont eux-mêmes accès. Le système suit également le *contenu supprimé* : les vidéos retirées par les créateurs ou les plateformes sont marquées comme `suppressed` plutôt que supprimées de la base de données, et les commentaires supprimés sont également préservés. La section 4.3 détaille l'ampleur du contenu préservé, qui inclut trois chaînes entièrement supprimées et 2 305 retraits individuels de vidéos.

3.3. Étape 3 : Traitement audio

Le contenu vidéo politique contient fréquemment de la musique de fond, des jingles et des effets sonores. Nous abordons ce problème par un traitement audio utilisant Demucs (Défossez et al. 2019; Rouard et al. 2023), spécifiquement le modèle `htdemucs`. Demucs décompose le signal audio en quatre pistes (voix, basse, batterie, autre) ; seule la piste vocale est conservée. Cette étape de prétraitement améliore substantiellement la précision de la transcription, en particulier pour le contenu de commentaire politique où les créateurs utilisent systématiquement la musique pour cadrer les segments.

3.4. Étape 4 : Diarisation des locuteurs et transcription

La transcription procède en trois étapes. Premièrement, pyannote.audio (Bredin et al. 2020; Bredin 2023), avec un seuil de regroupement fixé à 0,75 pour réduire la sur-segmentation, segmente la piste vocale en tours de parole pour identifier les locuteurs distincts et leurs limites temporelles. Cette étape est essentielle pour le contenu politique, qui comporte fréquemment des interviews, des débats et des formats multi-locuteurs. Deuxièmement, Whisper large-v3 (Radford et al. 2023) transcrit chaque segment. Sur le benchmark Fleurs français, Whisper large-v3 atteint un taux d'erreur par mot (WER) d'environ 5,6 %, ce qui le place parmi les meilleurs modèles de reconnaissance vocale open-source pour le français. [2] Troisièmement, chaque segment de locuteur est découpé en phrases individuelles à l'aide de SaT (Segment any Text) (Frohmann, Sterner, Vulić, Minixhofer, and Schedl 2024), un modèle multilingue de détection de limites de phrases (`sat-12l-sm`). Les transcriptions résultantes comprennent 645 738 segments de locuteurs découpés en 3 182 705 phrases individuelles à travers 21 914 vidéos (avril 2026). Chaque phrase est stockée avec son identifiant de locuteur, ses limites temporelles et le contexte du segment original.

2. Fiche du modèle Whisper large-v3, `https://huggingface.co/openai/whisper-large-v3`. Le prétraitement Demucs réduit davantage le WER en isolant la piste vocale du bruit de fond.

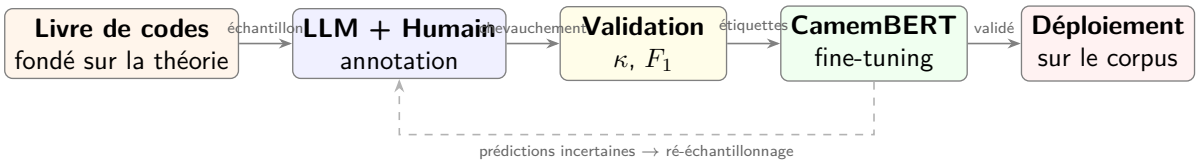

FIGURE 5 – Le cadre d'annotation LLM-in-the-loop. Un livre de codes guide l'annotation par LLM d'échantillons stratifiés. Des annotateurs humains valident un sous-ensemble de chevauchement représentatif. Les étiquettes validées entraînent des classifieurs CamemBERTav2, avec une boucle de raffinement itératif pour les prédictions incertaines.

3.5. Étape 5 : Annotation TAL par LLM-in-the-loop

L'annotation de millions de phrases de transcription pose un défi de scalabilité que les méthodes manuelles ne peuvent relever à un coût raisonnable. YouPol aborde ce problème par un cadre LLM-in-the-loop construit sur la plateforme open-source LLM Tool (Lemor et al. 2025), suivant le paradigme de distillation de connaissances dans lequel des étiquettes générées par LLM, validées par rapport à des annotations humaines, sont utilisées pour entraîner des classifieurs légers et déployables (Pangakis and Wolken 2024; Ziems et al. 2024).

Le cadre, illustré à la figure 5, opère en cinq étapes. Premièrement, pour chaque dimension d'annotation, un *livre de codes* détaillé spécifie la définition du construit, les règles de décision et des exemples annotés, fondés sur la théorie en science politique (le prompt complet du livre de codes pour `detect_pol` est fourni en annexe A). Deuxièmement, un échantillon aléatoire stratifié de typiquement 1 000 phrases est tiré du corpus et assure une représentation équilibrée entre chaînes, orientations idéologiques et types de contenu. Un grand modèle de langage (GPT-5.2) annote cet échantillon selon les instructions du livre de codes, et des annotateurs humains annotent indépendamment les mêmes phrases. L'accord inter-annotateurs est ensuite calculé entre les annotations du LLM et celles des humains. Si l'accord n'atteint pas les seuils acceptables (κ de Light $\geq 0{,}75$, F_1 macro $\geq 0{,}90$), le prompt du livre de codes est révisé et le processus est répété jusqu'à ce qu'un accord satisfaisant soit atteint. Troisièmement, une fois le prompt validé, un échantillon stratifié plus large de plus de 10 000 phrases est tiré et annoté par le LLM seul. Cet ensemble plus large d'étiquettes générées par le LLM sert de données d'entraînement. Quatrièmement, des classifieurs *CamemBERTav2* (Martin, Muller, Ortiz Suárez, Dupont, Romary, de la Clergerie, Seddah, and Sagot 2020) sont fine-tunés sur ces étiquettes. Si la performance du classifieur sur les données de validation est insuffisante, les prédictions les plus incertaines sont ré-échantillonnées, ré-annotées et utilisées pour augmenter l'ensemble d'entraînement. Cinquièmement, les classifieurs validés sont déployés sur l'ensemble du corpus. Chaque projet d'annotation de la base de données YouPol traverse plusieurs itérations de cette boucle avant le déploiement.

Le pipeline d'annotation est structuré en cascade. Le premier classifieur, et le plus fondamental, `detect_pol`, fonctionne comme une porte binaire qui identifie les phrases de transcription

contenant du contenu politique, défini au sens large pour inclure l'actualité, les enjeux sociaux, les acteurs politiques, les rapports de pouvoir et les normes sociales. Cette définition large est fondée sur les conceptions en science politique du "politique" qui s'étendent au-delà du contenu strictement partisan. Le tableau 2 présente les métriques de validation et de performance du classifieur pour `detect_pol`. L'accord inter-annotateurs entre deux annotateurs humains et le LLM sur l'échantillon de chevauchement de 1 000 phrases a donné un κ de Light de 0,787, un F_1 par paires de 89,4 % et un F_1 macro de 92,2 %, ce qui confirme la fiabilité des étiquettes générées par le LLM. Le classifieur CamemBERTav2 entraîné sur ces étiquettes a atteint un F_1 macro de 92,2 % et une exactitude de 93,5 % sur l'ensemble de validation réservé (1 753 phrases). Le classifieur a été déployé sur les 645 738 segments transcrits avec diarisation et produit une étiquette binaire (`1` = politique, `0` = non politique) pour chaque phrase. Seules les phrases classées comme politiques sont ensuite acheminées vers les classifieurs spécifiques aux projets (selon le projet).

La figure 6 présente des résultats préliminaires agrégés par orientation idéologique. La densité politique varie substantiellement : les chaînes de gauche présentent la plus forte proportion de phrases classifiées comme politiques (44,8 %), cohérent avec leur format plus long et orienté analyse. Les chaînes d'extrême droite suivent à 30,1 %, bien qu'elles constituent la plus grande part du corpus. Les chaînes complotistes affichent une densité politique de 24,1 %, tandis que les chaînes de la manosphère présentent la densité la plus faible (6,8 %), résultat de leur format hybride qui mêle commentaire politique, mode de vie et divertissement. La figure 7 présente l'évolution mensuelle de la densité politique moyenne séparément pour la France et le Québec de 2017 à 2024. Les deux courbes montrent que chaque écosystème répond à son propre

TABLE 2 – Performance de validation et du classifieur pour `detect_pol`.

Étape	**Métrique**	**Valeur**
Accord humain–LLM (chevauchement de 1 000 phrases, 2 annotateurs)		
	κ de Light	0,787
	α de Krippendorff	0,787
	F_1 par paires	89,4 %
	F_1 macro	92,2 %
	F_1 pondéré	92,0 %
	Correspondance exacte	88,1 %
	Perte de Hamming	7,7 %
Classifieur CamemBERTav2 (validation, $n = 1\,753$)		
	F_1 macro	92,2 %
	Exactitude	93,5 %
	F_1 (non politique)	94,6 %
	F_1 (politique)	89,3 %
	Époques d'entraînement	11 / 25 (arrêt précoce)

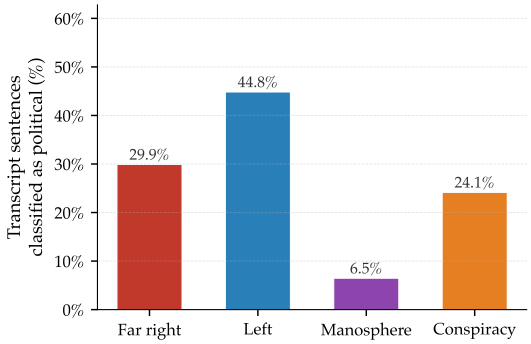

FIGURE 6 – Proportion de phrases de transcription classées comme politiques par `detect_pol`, par orientation idéologique.

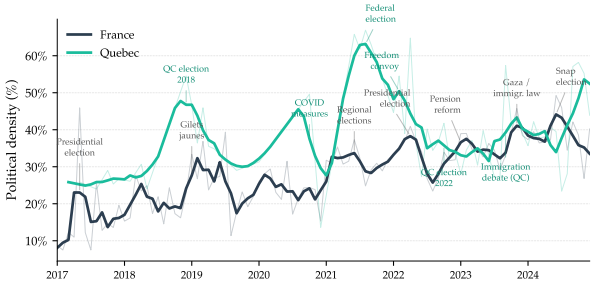

FIGURE 7 – Densité politique moyenne mensuelle par pays (lissage sur 3 mois pour la France, 5 mois pour le Québec en raison de données plus éparses). La France et le Québec répondent à des événements politiques nationaux distincts.

calendrier politique national. Pour la France, des pics clairs apparaissent autour de l'élection présidentielle de 2017, du mouvement des *Gilets jaunes* (pic en janvier 2019), des élections régionales de 2021, de l'élection présidentielle de 2022, des manifestations contre la réforme des retraites (début 2023), de la guerre Israël–Gaza et du débat sur la loi immigration (novembre 2023, pic le plus élevé pré-2024 à 44 %), et de l'élection législative anticipée de 2024 suivant la dissolution de l'Assemblée nationale par Macron (atteignant 49 %, le maximum global). Pour le Québec, un premier pic suit l'élection provinciale de 2018, alors que le commentaire politique tiers émerge dans l'extrême droite québécoise. Les mesures sanitaires liées à la COVID-19 produisent un pic distinct à la fin de 2020, porté par le contenu anti-restrictions. Le pic le plus marqué correspond à la campagne de l'élection fédérale canadienne de 2021 (juin–septembre 2021, dépassant 60 %), suivi du mouvement du Convoi de la liberté (février 2022), de l'élection provinciale de 2022 et du débat sur les seuils d'immigration à l'automne 2023. Entre ces événements nationaux spécifiques, les deux courbes suivent une tendance progressive à la hausse, ce qui suggère une politisation progressive des deux écosystèmes au fil du temps. Cette comparaison transnationale, rendue possible par la profondeur du corpus au niveau des transcriptions, démontre que la densité politique au niveau de la phrase est déterminée par des dynamiques nationales distinctes plutôt que par un calendrier unique partagé.

Trois projets d'annotation opèrent ensuite sur le sous-ensemble de phrases classées comme politiques par `detect_pol`. *Projet 1 : Score idéologique d'extrême droite (SIED).* Fondé sur Boursier and Lemor (2025), le SIED opérationnalise l'idéologie d'extrême droite et néo-réactionnaire à travers un livre de codes couvrant onze macro-catégories : nationalisme, immigration, démocratie, progrès, autorité, tradition, égalité, technologie, libertarianisme, écologie et métaphores fictionnelles (red pill, Le Seigneur des anneaux, Star Wars, Cathédrale). Chaque macro-catégorie contient entre deux et six sous-dimensions, et chaque phrase est annotée sur l'ensemble d'entre elles. Cela produit un profil idéologique multidimensionnel

bien au-delà de la classification binaire. *Projet 2 : Analyse du discours genré (GENRE).* Ce livre de codes classifie la rhétorique genrée selon quatre dimensions : la présence d'un discours genré, sa valence (positive, négative, ambivalente ou nulle), le type de rationalité invoqué (biologique/nature, libéral, empirique ou héroïque/vérité), et la position envers la science. *Projet 3 : Discours néo-réactionnaire technophile (NR).* Ce projet cible l'intersection entre l'enthousiasme technologique et la politique réactionnaire à travers la technologie, le libertarianisme, les métaphores fictionnelles et des dimensions partagées sur l'égalité et la hiérarchie.

Avec `detect_pol` et les dimensions partagées, YouPol est conçu pour déployer jusqu'à 85 classifieurs binaires qui produisent une annotation multidimensionnelle au niveau de la phrase. L'architecture en cascade, où `detect_pol` conditionne toute annotation en aval, assure que les classifieurs fins opèrent sur le sous-ensemble pertinent tout en fournissant une variable de recherche autonome analysable indépendamment.

3.6. Étape 6 : Reconnaissance d'entités nommées et intégration

Toutes les phrases traitées sont soumises à la reconnaissance d'entités nommées à l'aide de GLiNER (Zaratiana, Tomeh, Holat, and Charnois 2024), un modèle de classification de tokens en zero-shot qui identifie les entités sans fine-tuning spécifique à la tâche. Nous utilisons la variante multilingue (`gliner_multi-v2.1`) et extrayons neuf types d'entités adaptés à l'analyse du discours politique : *personne*, *parti politique*, *institution*, *organisation*, *média*, *lieu*, *loi*, *événement* et *idéologie*. Par exemple, la phrase "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") produit {personne : Macron, parti_politique : RN, événement : plan de relance}. Les entités extraites sont stockées aux côtés de chaque phrase dans la base de données et sont pleinement intégrées dans les fonctions de recherche et de filtrage de l'API REST et de l'explorateur web du corpus (voir l'annexe B), permettant aux chercheurs d'interroger le corpus par type d'entité, nom ou cooccurrence. L'ensemble des données est organisé dans un schéma PostgreSQL normalisé comprenant plus de 40 tables (décrit à la section 4.4), indexées pour la recherche, l'agrégation et l'accès programmatique via une API REST fondée sur PostgREST à l'adresse `https://data.you-pol.com/` avec authentification JWT et permissions basées sur les rôles. Une bibliothèque cliente Python est également disponible pour les requêtes programmatiques et l'export de données. [3]

3.7. Étape 7 : Le Réseau de Calcul Collaboratif YouPol (YCCN)

La septième étape n'est pas une étape terminale de traitement mais le mode opérationnel permanent de l'infrastructure. YouPol fonctionne par le YCCN (voir l'annexe B pour une

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

capture d'écran de l'application travailleur), dans lequel des chercheurs collaborateurs contribuent de la capacité de traitement depuis leurs propres machines plutôt que de dépendre d'un calcul institutionnel centralisé. Cette architecture, illustrée à la figure 8, rend l'observatoire entièrement autosuffisant : après la phase initiale de traitement sur l'infrastructure de l'Alliance de recherche numérique du Canada, toute la production a été transférée au YCCN, éliminant les coûts continus et la dépendance aux allocations de calcul institutionnelles.

Un serveur central détenu par l'équipe YouPol crée les tâches de traitement et les place dans une file d'attente partagée adossée à une base PostgreSQL avec verrouillage exclusif, ce qui empêche le traitement en double sans coordination centralisée. Les tâches sont organisées en quatre types indépendants : balayage de chaînes, transcription, extraction de commentaires et actualisation des métadonnées. Chaque machine travailleuse réclame de manière autonome la prochaine tâche disponible, exécute la chaîne de traitement appropriée (téléchargement, traitement audio, transcription, diarisation, annotation) et signale l'achèvement. Si un travailleur devient indisponible, ses tâches inachevées sont automatiquement libérées et réassignées à un autre travailleur.

Le YCCN repose sur un principe de crowdsourcing appliqué aux ressources de calcul. Chaque

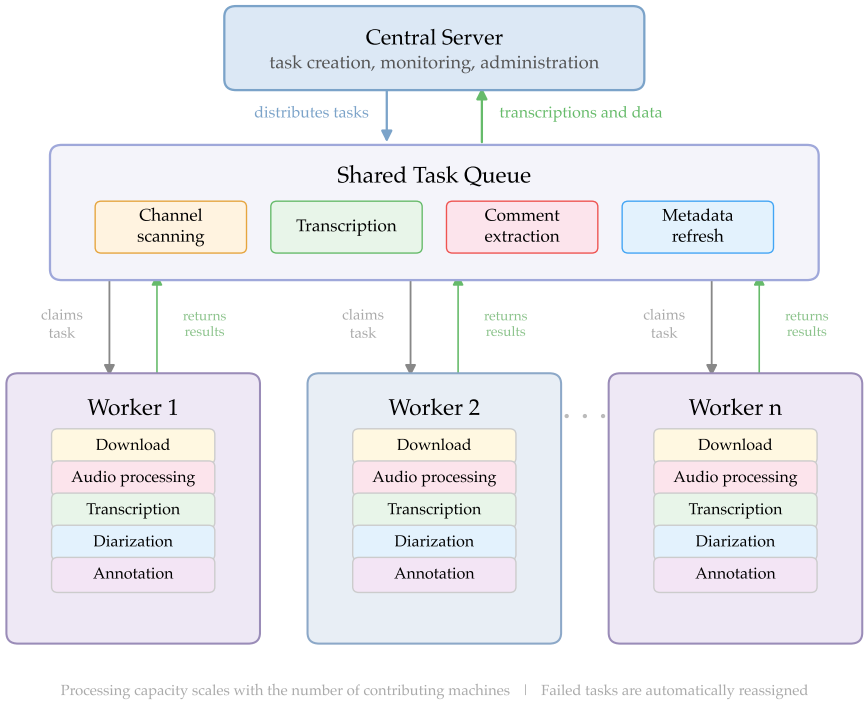

FIGURE 8 – Architecture du Réseau de Calcul Collaboratif YouPol (YCCN). Un serveur central distribue les tâches dans une file d'attente partagée organisée par type. Chaque travailleur réclame indépendamment les tâches et exécute la chaîne de traitement complète. La capacité de traitement croît au fur et à mesure que de nouveaux collaborateurs contribuent des machines.

chercheur collaborateur qui rejoint le projet installe une application dédiée développée par les auteurs sur sa machine. Une fois connectée, la machine commence immédiatement à réclamer et traiter des tâches en tant que nœud autonome. La capacité de traitement de l'observatoire croît ainsi organiquement avec sa communauté. À mesure que le corpus s'étend à de nouveaux écosystèmes, de nouveaux collaborateurs contribuent à la fois des chaînes à surveiller et des ressources de calcul pour les traiter. Ce modèle élimine le goulot d'étranglement traditionnel de l'accès au calcul institutionnel et ses coûts associés, et place l'infrastructure sous le contrôle collectif de la communauté de recherche plutôt que sous les contraintes de fournisseurs de ressources externes. Le système traite actuellement des centaines de nouvelles transcriptions quotidiennement, en parallèle de l'extraction continue des commentaires et des mises à jour de métadonnées.

4. Le jeu de données YouPol

4.1. Statistiques descriptives

Le tableau 3 présente les statistiques clés du jeu de données en date d'avril 2026. La base de données contient 23 712 vidéos YouTube et 1 685 vidéos TikTok, dont 21 914 ont été intégralement transcrites et annotées. 1 798 vidéos YouTube supplémentaires sont en attente de traitement dans la base principale, et le scanner continu a identifié plus de 16 000 vidéos supplémentaires en cours de transcription dans le pipeline. La composante TikTok de l'observatoire, actuellement en expansion active, reproduit le pipeline YouTube complet (collecte, transcription, diarisation, annotation) avec un ensemble dédié de tables et de scanners. La priorité actuelle est d'identifier et d'intégrer les comptes TikTok des créateurs déjà suivis sur YouTube, afin de permettre une comparaison inter-plateformes des mêmes acteurs politiques. À mesure que de nouveaux comptes TikTok sont identifiés, ils entrent dans le même pipeline en sept étapes décrit à la section 3. Une extension à l'écosystème politique anglophone est en préparation. L'architecture du pipeline étant agnostique à la langue (Whisper supporte l'anglais nativement, CamemBERT peut être remplacé par toute variante BERT via LLM Tool), cette extension ne nécessite que l'identification de nouvelles chaînes et des codebooks adaptés, sans changement infrastructurel.

4.2. Croissance temporelle et dynamiques d'engagement

La figure 9 présente la distribution temporelle du corpus. La production vidéo s'est considérablement accélérée depuis 2020, reflétant la croissance de l'écosystème politique sur YouTube. L'année la plus active est 2024 avec 6 012 vidéos, et le contenu d'extrême droite domine sur toutes les périodes, bien que le contenu de gauche et de la manosphère ait connu une croissance substantielle depuis 2021. La figure 10 présente la croissance parallèle de l'engagement

TABLE 3 – Jeu de données YouPol : statistiques descriptives (avril 2026)

Indicateur	**Valeur**
Vidéos YouTube (total)	23 712
Vidéos intégralement transcrites	21 914
Vidéos en attente de traitement	1 798
Vidéos dans le pipeline (découvertes)	>16 000
Vidéos TikTok	1 685
Chaînes suivies	68
Pays	2 (France, Québec)
Orientations idéologiques	4
Période temporelle	mai 2006 – avril 2026
Vues totales	1 156 113 044
Commentaires extraits au total	7 703 663
Likes totaux	14 421 301
Segments diarisés par locuteur	645 738
Phrases individuelles	3 182 705
Classifieurs binaires d'annotation	jusqu'à 85
Données audio initiales traitées	>15 To

des utilisateurs : l'extraction des commentaires a suivi la croissance du corpus, atteignant 1,88 million de commentaires pour les vidéos de 2024 seules, tandis que les vues cumulées dépassent 1,15 milliard.

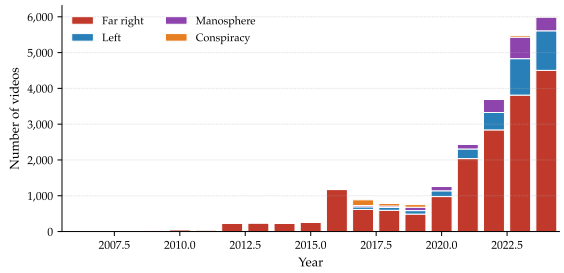

FIGURE 9 – Nombre de vidéos par année et orientation idéologique. La croissance s'accélère à partir de 2020.

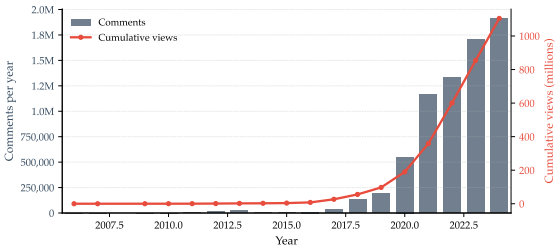

FIGURE 10 – Nombre de commentaires par année (barres) et vues cumulées (ligne).

La figure 11 présente l'évolution du nombre total de vues par orientation idéologique depuis 2012. Si les chaînes d'extrême droite dominent en nombre absolu de vues, les données révèlent des différences structurelles importantes. Les chaînes de gauche atteignent des durées moyennes substantiellement plus longues (48,1 minutes contre 19,1 minutes pour le contenu d'extrême droite ; figure 12), reflétant des formats de contenu différents (journalisme d'investigation contre commentaire). La figure 13 présente la répartition par tranches de vues. La majorité du corpus (10 732 vidéos) se situe dans la tranche 10K–100K, avec 68 vidéos dépassant 1 million de vues, tandis que les 7 053 vidéos en dessous de 1 000 vues représentent la "longue traîne" invisible aux études fondées sur la recommandation algorithmique.

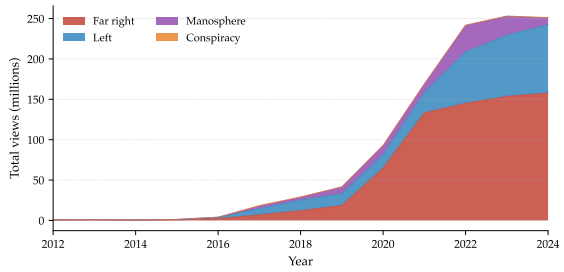

FIGURE 11 – Vues totales par orientation idéologique (2012–2024). Le contenu d'extrême droite domine, avec une accélération après 2020.

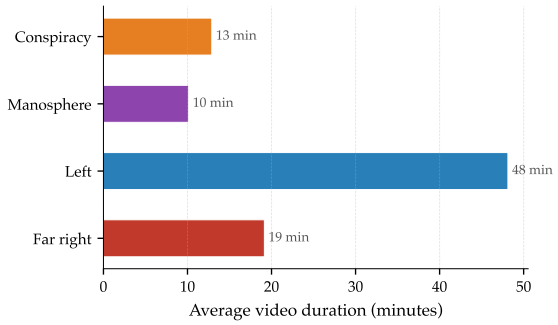

FIGURE 12 – Durée moyenne des vidéos par orientation idéologique.

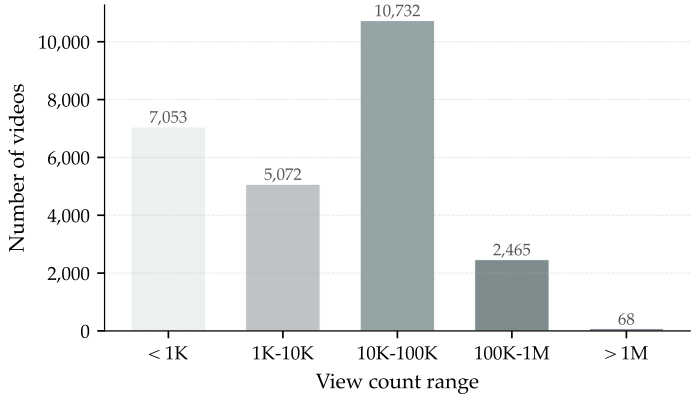

FIGURE 13 – Répartition des vidéos par tranche de vues, du contenu de longue traîne (<1K vues) aux vidéos virales (>1M vues).

4.3. Contenu préservé : vidéos et chaînes supprimées

Une caractéristique distinctive de YouPol est la préservation systématique du contenu que les plateformes suppriment. En avril 2026, la base contient 2 305 vidéos supprimées (9,1 % du corpus) et 1 578 598 commentaires qui ne sont plus accessibles sur la plateforme, dont 312 452 commentaires individuellement supprimés et 1 266 146 commentaires sur des vidéos qui ont été retirées par la suite. Ce matériel correspond à des contenus retirés par YouTube pour violation des règles communautaires, supprimés volontairement par les créateurs, ou devenus indisponibles.

Trois chaînes entières ont été supprimées par YouTube et sont préservées intégralement : FDENEWS (1 755 vidéos, 68,3M vues, extrême droite, France), Virginie Vota (104 vidéos, 6,8M vues, extrême droite, France) et Gabriel Duquette (124 vidéos, 823K vues, manosphère, Québec). Pour ces chaînes, chaque vidéo, transcription, segment, commentaire et instantané de métadonnées reste disponible dans la base. Par ailleurs, 21 chaînes actives ont subi des retraits individuels de vidéos (figure 14). Au total, les vidéos supprimées représentent plus de

110 millions de vues et 6,4 millions de likes qui seraient entièrement perdus pour la recherche sans collecte préventive.

Les schémas de suppression varient selon les orientations idéologiques (figure 15). Les chaînes d'extrême droite affichent le taux de retrait vidéo le plus élevé (11,6 %), suivies par la manosphère (8,5 %), tandis que les chaînes de gauche subissent nettement moins de retraits vidéo (1,6 %) et les chaînes complotistes n'ont aucune vidéo supprimée. La suppression individuelle de commentaires suit un schéma différent : les chaînes de gauche présentent le taux le plus élevé (5,6 %), suivies de l'extrême droite (3,9 %) et de la manosphère (2,9 %). Cette divergence entre les taux de suppression de vidéos et de commentaires selon les orientations constitue en soi un résultat analytique significatif, invisible sans surveillance continue.

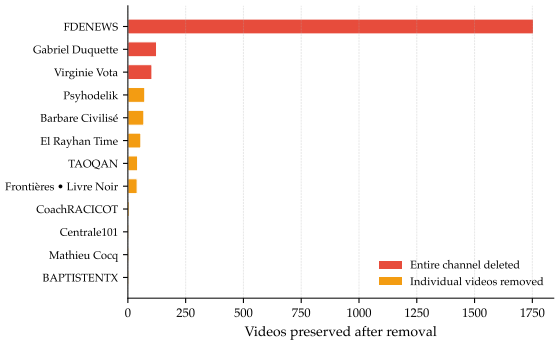

FIGURE 14 – Vidéos préservées après retrait de YouTube, par chaîne. Rouge : chaîne entière supprimée. Orange : retraits individuels depuis des chaînes actives.

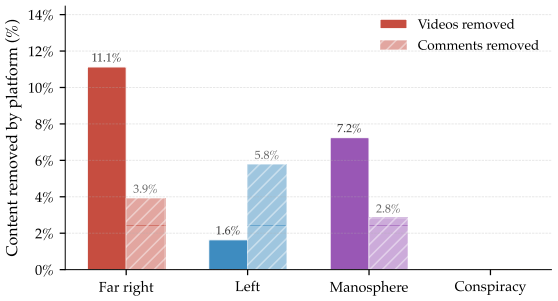

FIGURE 15 – Taux de suppression par orientation idéologique. Plein : vidéos retirées. Hachuré : commentaires retirés. L'extrême droite présente le taux de retrait de vidéos le plus élevé ; les chaînes de gauche le taux de retrait de commentaires le plus élevé.

4.4. Schéma de la base de données et accès aux données

Le jeu de données est organisé dans un schéma PostgreSQL normalisé de plus de 40 tables (voir l'annexe C pour une vue d'ensemble complète). Le schéma comprend cinq groupes fonctionnels : (1) *tables de contenu* stockant les métadonnées et commentaires vidéo pour YouTube et TikTok, avec un suivi des suppressions qui signale le contenu retiré plutôt que de le supprimer ; (2) *tables de transcription* contenant les segments bruts par locuteur (641 613 pour YouTube, 4 125 pour TikTok) et les segments traités avec annotations TAL, étiquettes de détection politique et balises REN (3,18 millions de lignes traitées) ; (3) *vues combinées* agrégeant les deux plateformes ; (4) *tables d'historique de métadonnées* enregistrant des instantanés d'engagement horodatés pour les vidéos (33 746 instantanés) et les chaînes (620 instantanés), permettant une analyse longitudinale ; et (5) *tables de pipeline* pour la coordination du YCCN, les files d'attente de tâches et les journaux d'événements.

Le site du projet à l'adresse `https://you-pol.com/` présente l'observatoire et sa documenta-

tion. L'accès aux données est fourni par la plateforme à l'adresse https://data.you-pol.com/, qui propose trois modes d'interaction (voir l'annexe B pour les captures d'écran). Premièrement, un explorateur de corpus interactif offre une recherche plein texte sur l'ensemble des 645 738 segments de locuteurs et 7,7 millions de commentaires, avec filtrage par chaîne, orientation, pays, période et type d'entité REN. Des tableaux de bord de visualisation affichent les statistiques au niveau du corpus, les comparaisons entre chaînes et les tendances de densité politique. Deuxièmement, une API REST basée sur PostgREST à l'adresse https://data.you-pol.com/ fournit un accès programmatique avec authentification JWT, supportant le filtrage complexe de type SQL, la pagination et les exports en masse CSV/JSON. Troisièmement, une bibliothèque cliente Python [4] encapsule l'API et permet aux chercheurs, y compris ceux menant des travaux qualitatifs, d'interroger le corpus, de récupérer les transcriptions complètes ou les fils de commentaires pour des vidéos spécifiques, et d'exporter les résultats directement dans leurs flux d'analyse. Les trois interfaces appliquent un contrôle d'accès basé sur les rôles avec journalisation des audits.

5. Considérations éthiques et limites

5.1. Éthique

YouPol collecte du contenu publiquement disponible provenant de personnages publics qui ont choisi de participer au discours politique sur des plateformes publiques. Toutes les chaînes suivies sont opérées par des créateurs de contenu disposant d'audiences significatives (de milliers à des centaines de milliers d'abonnés). Comme l'établissent les lignes directrices en éthique de la recherche sur Internet, le contenu produit par des personnages publics dans des forums publics comporte des attentes réduites en matière de vie privée par rapport aux communications privées (Zimmer 2010; Franzke, Bechmann, Ess, and Zimmer 2020). Les commentaires sont collectés à partir de fils publics ; aucune communication privée n'est consultée. Le protocole de recherche a été examiné par le comité d'éthique de l'Université de Sherbrooke, conformément à l'Énoncé de politique des trois conseils sur l'éthique de la recherche avec des êtres humains (Government of Canada 2023).

Cette recherche est menée dans deux juridictions. Dans l'Union européenne, où l'un des auteurs est basé, le Digital Services Act reconnaît explicitement la recherche universitaire vérifiée comme un cas d'usage légitime pour l'accès aux données des plateformes, y compris le contenu des très grandes plateformes en ligne (European Parliament and Council of the European Union 2022). L'infrastructure de collecte indépendante de YouPol est cohérente avec ce cadre. Au Canada et au Québec, des cadres de gouvernance éthique pour la recherche existent, comme en témoigne l'examen éthique mené par l'Université de Sherbrooke, mais

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

aucun mécanisme équivalent ne contraint les plateformes à partager leurs données avec les chercheurs (Government of Canada 2023). L'infrastructure de collecte indépendante de YouPol comble cette lacune tout en demeurant cohérente avec les cadres institutionnels des deux juridictions.

Les identifiants des commentateurs sont stockés en interne à des fins de déduplication et d'analyse de réseaux, mais ne sont pas exposés via l'API publique. L'accès des chercheurs requiert une authentification avec des permissions basées sur les rôles, des limites d'utilisation et une journalisation des audits. Le contenu supprimé par les plateformes est préservé dans la base à des fins de recherche, mais n'est accessible qu'aux chercheurs autorisés dans le cadre d'un accord d'utilisation des données ; il n'est jamais affiché publiquement. La préservation du contenu supprimé soulève des questions éthiques spécifiques ; nous estimons que la valeur de recherche du maintien d'un enregistrement de discours politiquement conséquent, compte tenu notamment du fait que les retraits par les plateformes introduisent des biais systématiques de survivance qui faussent l'étude empirique de l'extrémisme politique (Ohme et al. 2024; Bruns 2019), l'emporte sur le risque d'amplification dans le cadre de ces restrictions d'accès.

5.2. Limites

Plusieurs limites doivent être reconnues. La sélection des chaînes reflète des priorités de recherche délibérées, mais il convient de reconnaître les compromis méthodologiques que cela implique. Deux grandes stratégies coexistent dans la littérature pour la construction de corpus politiques sur les plateformes vidéo. La première, adoptée par YouPol, procède de manière descendante : les chaînes sont identifiées *a priori* par évaluation experte et métriques d'audience, à partir d'une définition conceptuelle préalable de ce qui constitue du contenu politique d'extrême droite (Rauchfleisch and Kaiser 2020; Boursier 2025; Munger and Phillips 2022). La seconde procède de manière ascendante, en laissant le contenu numérique lui-même, certains mots, certains discours, certains schémas de co-engagement, définir l'écosystème d'acteurs *a posteriori* (Ribeiro et al. 2020; Hosseinmardi et al. 2021; Tainturier 2025). Chaque approche comporte des risques épistémologiques distincts. La sélection descendante assure la clarté analytique et la reproductibilité, mais risque le biais de confirmation : le corpus reflète la catégorisation préalable du champ par le chercheur, et peut exclure systématiquement des acteurs dont le positionnement idéologique est ambigu, émergent ou délibérément non marqué. Les approches ascendantes réduisent cette circularité, mais introduisent leurs propres dépendances, envers des mots-clés amorces, des chaînes amorces ou des graphes de recommandation, et peuvent manquer des acteurs qui opèrent aux marges des schémas discursifs indexés. Les deux stratégies font en fin de compte face à ce que Salganik (2019) décrit comme le problème de non-représentativité inhérent à tout corpus numérique intentionnel : ce qui est collecté reflète les choix de collecte, non la population complète du contenu politique.

Le corpus actuel de YouPol surreprésente le contenu d'extrême droite par rapport à l'écosystème politique YouTube dans son ensemble, un choix délibéré qui reflète à la fois la composition du paysage des influenceurs politiques francophones, où les chaînes d'extrême droite représentent une part disproportionnée du nombre total de vues (Munger and Phillips 2022), et les questions de recherche qui motivent la conception de l'observatoire. Toutefois, l'infrastructure que YouPol a construite crée les conditions pour articuler progressivement ces deux paradigmes. Les classifieurs idéologiques au niveau de la phrase, le pipeline de reconnaissance d'entités nommées et les outils d'annotation du discours permettent désormais d'identifier, au sein des transcriptions existantes, les chaînes, figures et médias que les créateurs d'extrême droite citent, approuvent ou avec lesquels ils interagissent activement, des acteurs qui n'auraient pas été captés par la sélection initiale experte. Une expansion du corpus fondée sur le contenu, amorcée non par les métriques d'audience mais par les réseaux discursifs reconstruits à partir des transcriptions vidéo elles-mêmes, permettrait une cartographie plus ascendante de l'écosystème francophone d'extrême droite et fournirait une vérification empirique des frontières tracées par la sélection descendante initiale.

Indépendamment des décisions de construction du corpus, un ensemble de limites techniques contraint ce que YouPol peut actuellement mesurer avec une fiabilité complète. Malgré le prétraitement Demucs, la transcription reste imparfaite pour la parole chevauchée, les accents prononcés ou l'audio dégradé ; les taux d'erreur n'ont pas été systématiquement quantifiés sur l'ensemble du corpus. Le cadre LLM-in-the-loop produit des étiquettes silver-standard plutôt que des annotations humaines gold-standard ; bien que la boucle de raffinement itératif atténue les erreurs systématiques, un certain bruit est inévitable. Malgré l'indépendance infrastructurelle vis-à-vis des API des plateformes, YouPol dépend de la disponibilité des plateformes pour la collecte initiale. Le contenu non encore collecté au moment de sa suppression ne peut être récupéré. Le corpus actuel couvre le contenu francophone, avec une expansion anglophone en préparation. L'extension à des contextes linguistiques supplémentaires nécessiterait l'identification de nouvelles chaînes, le réglage de la RAP spécifique à la langue et des livres de codes culturellement adaptés. Enfin, la collecte de données TikTok fait face à des défis techniques plus importants en raison de mesures anti-scraping plus agressives et reste actuellement moins complète que la couverture YouTube (la collecte des comptes TikTok parallèles aux chaînes YouTube déjà suivies est en cours).

Conclusion

YouPol fournit le premier observatoire de discours politique sur les plateformes vidéo au niveau des transcriptions et continuellement mis à jour. L'infrastructure répond aux quatre lacunes identifiées dans la littérature. Premièrement, elle rend la substance idéologique du contenu vidéo politique empiriquement accessible grâce à un pipeline de transcription indépendant qui

combine la séparation neuronale des sources, la diarisation des locuteurs et la reconnaissance vocale pour produire des transcriptions complètes indépendamment des sous-titres fournis par les plateformes. Deuxièmement, elle préserve le contenu que les plateformes suppriment, y compris 2 305 vidéos supprimées et trois chaînes entièrement effacées, et permet ainsi l'analyse rétrospective de matériel qui serait autrement définitivement perdu pour la recherche. Troisièmement, elle suit les dynamiques d'engagement de manière longitudinale grâce à des instantanés horodatés des métadonnées qui capturent la trajectoire complète de chaque vidéo et chaîne. Quatrièmement, elle archive 7,7 millions de commentaires avec provenance complète, y compris ceux supprimés ultérieurement par les créateurs ou les plateformes, et permet l'étude de la modération des commentaires en tant que stratégie de communication politique.

Un cadre d'annotation LLM-in-the-loop, dans lequel les étiquettes produites par un grand modèle de langage et validées contre des annotations humaines sont distillées en classifieurs légers CamemBERT, fournit une annotation au niveau de la phrase à l'échelle de l'ensemble du corpus. Le premier classifieur déployé, `detect_pol`, attribue une étiquette politique binaire à chacune des 3,18 millions de phrases de la base. La densité politique, définie comme la proportion de phrases de transcription classifiées comme politiques au sein d'une chaîne ou d'une période donnée, varie de 44,8 % pour les chaînes de gauche à 6,8 % pour les chaînes de la manosphère. L'évolution mensuelle de cette mesure en France et au Québec répond à des calendriers politiques nationaux distincts plutôt qu'à une dynamique temporelle partagée. Ces résultats seraient inaccessibles par la seule analyse des métadonnées. Le pipeline d'annotation complet à travers trois projets (idéologie d'extrême droite, rhétorique genrée, discours néo-réactionnaire) étendra cette analyse à des dimensions idéologiques plus fines.

Le YouPol Collaborative Computing Network (YCCN) constitue une contribution distincte à l'infrastructure de recherche. En permettant à tout chercheur collaborateur de contribuer en capacité de calcul depuis sa propre machine, l'observatoire fonctionne indépendamment des grappes de calcul institutionnelles. Le pipeline agnostique à la langue ne requiert que l'identification de nouvelles chaînes et des livres de codes adaptés pour s'étendre à de nouveaux contextes linguistiques et politiques, et une expansion anglophone est en préparation. L'intégration de TikTok est également en cours, avec priorité accordée aux comptes des créateurs déjà suivis sur YouTube afin de permettre une comparaison inter-plateformes des mêmes acteurs politiques.

Une question que la base est bien positionnée pour aborder dans de futurs travaux est de savoir si la densité politique ou l'intensité idéologique d'une vidéo prédit sa suppression ultérieure par la plateforme. Plus largement, YouPol offre un fondement pour la recherche en communication politique qui prend le contenu parlé des vidéos politiques comme objet d'analyse principal.

Le projet est décrit à `https://you-pol.com/`, les données sont accessibles via l'API REST à l'adresse `https://data.you-pol.com/`.

Remerciements

Le traitement initial de plus de 15 To de données audio a été réalisé sur l'infrastructure de l'Alliance de recherche numérique du Canada, dont l'accès a été rendu possible par le Centre interuniversitaire de recherche sur la science et la technologie (CIRST). Nous remercions François Claveau pour son soutien et ses conseils tout au long de ce projet.

Références

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. Prompt d'annotation pour `detect_pol`

Le prompt suivant est utilisé pour instruire le LLM (GPT-5.2) pour la tâche de détection de contenu politique. Le livre de codes définit une conception large du "politique" fondée sur la théorie en science politique, inclut des règles de décision et fournit des exemples annotés pour guider une annotation cohérente. Le prompt est reproduit intégralement ci-dessous.

```
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. Captures d'écran de la plateforme

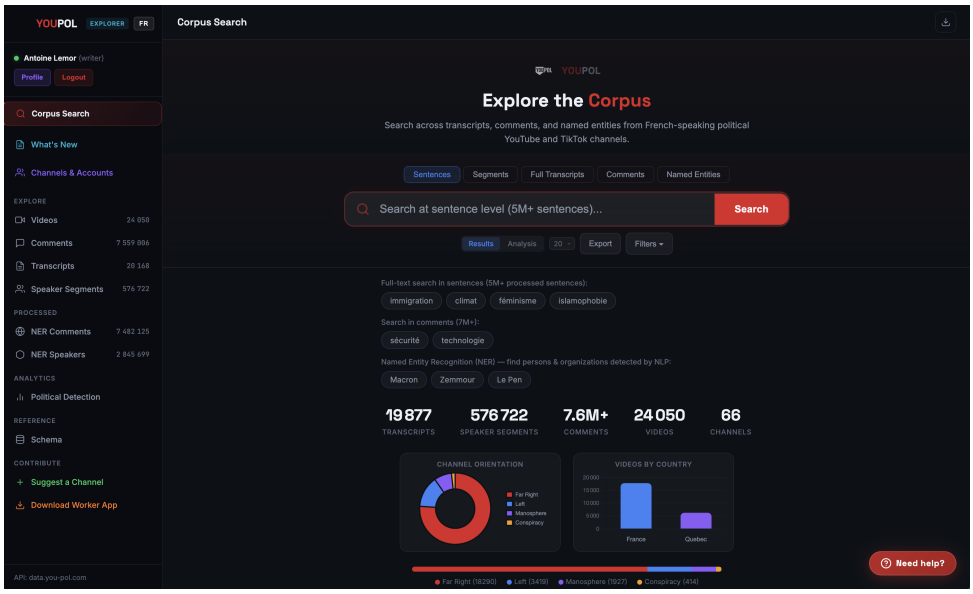

FIGURE 16 – L'API de données YouPol (`data.you-pol.com`). Les chercheurs peuvent parcourir, rechercher, filtrer et exporter les données du corpus. Une bibliothèque cliente Python (`youpol-client`) fournit un accès programmatique.

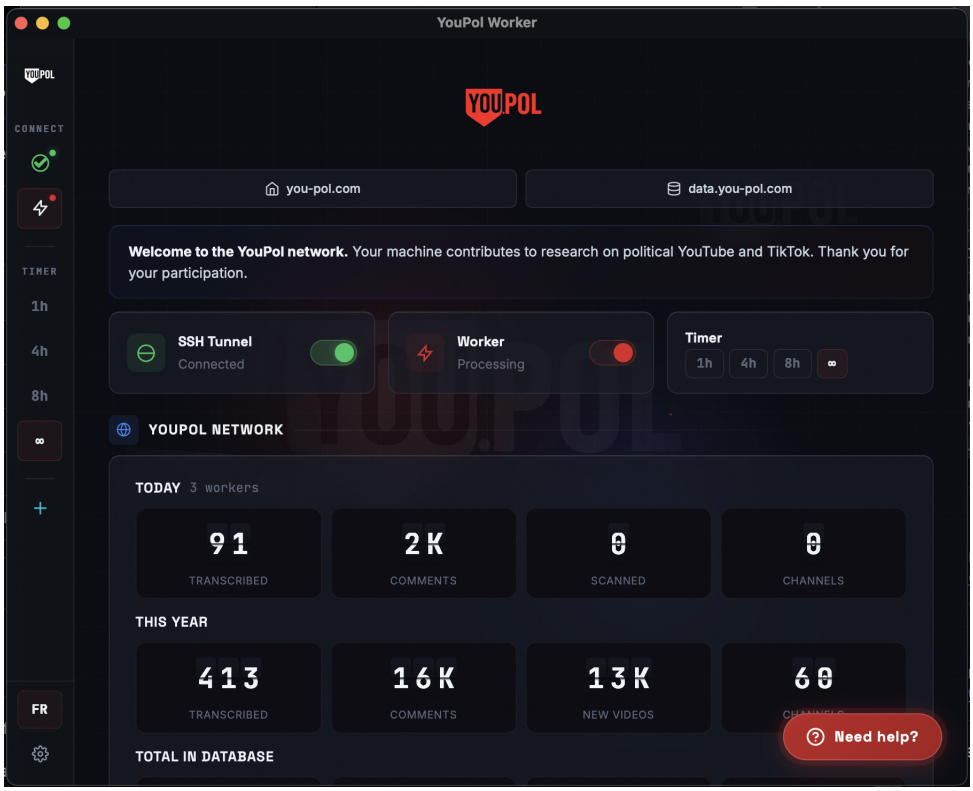

FIGURE 17 – L'application YouPol Worker. Chaque chercheur collaborateur exécute cette application sur sa machine pour participer au YCCN.

C. Vue d'ensemble du schéma de la base de données

Le tableau 4 résume les principaux groupes de tables du schéma PostgreSQL de YouPol, avec les tables représentatives et les nombres de lignes en date d'avril 2026.

TABLE 4 – Vue d'ensemble du schéma de la base de données YouPol (avril 2026).

Groupe	**Tables clés**	**Lignes**
Contenu (YouTube)		
	`youtube_videos`	23 712
	`youtube_comments`	7 585 328
	`channels`	68
Contenu (TikTok)		
	`tiktok_videos`	1 685
	`tiktok_comments`	118 335
Transcriptions (YouTube)		
	`youtube_video_transcripts`	20 234
	`youtube_transcription_speakers`	641 613
	`youtube_transcription_speakers_processed`	3 167 198
Transcriptions (TikTok)		
	`tiktok_video_transcripts`	1 680
	`tiktok_transcription_speakers`	4 125
	`tiktok_transcription_speakers_processed`	15 507
Vues combinées		
	`videos` (toutes plateformes)	25 397
	`transcription_speakers` (toutes)	645 738
	`speakers_with_pol` (toutes)	645 738
Historique des métadonnées		
	`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

La table `transcription_speakers_processed` contient chaque segment de locuteur répliqué

à travers toutes les dimensions d'annotation (détection politique, balises REN), tandis que `speakers_with_pol` fournit une vue dédupliquée avec l'étiquette `detect_pol` pour chaque segment. La table `video_metadata_history` enregistre une ligne par vidéo par cycle d'observation, permettant une analyse longitudinale de l'engagement. Le contenu supprimé est signalé dans la table `videos` (`suppressed = true`) plutôt que supprimé, préservant l'intégralité du registre analytique.

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

SocArXiv 2026 *Submitted:* 2026-04-04
10.31235/osf.io/vpzmq_v2 *Accepted:* 2026-04-05