



Hackathon 2026 - Master Humanités numériques Ecole nationale des chartes - PSL

L'avortement à la télévision française (1960-2024)

Sujet :

Exploration d'un corpus de fiches documentaires et de transcriptions de sujets de JT et magazine

Données :

Fiches documentaires et transcriptions (majoritairement via Whisper) de sujets de JT et de magazines relatifs à l'avortement (1960-2024)

Méthodes de collecte (par INA) :

Matching des termes "Avortement", "ivg" et " interruption volontaire de grossesse" dans les notices documentaires

Github :

https://github.com/Phineaax/Hackathon_INA

Plan compte rendu :

Partie 1 : Traitement des thématiques

Partie 2 : Embeddings et extraction

PARTIE 1 : Traitement des thématiques

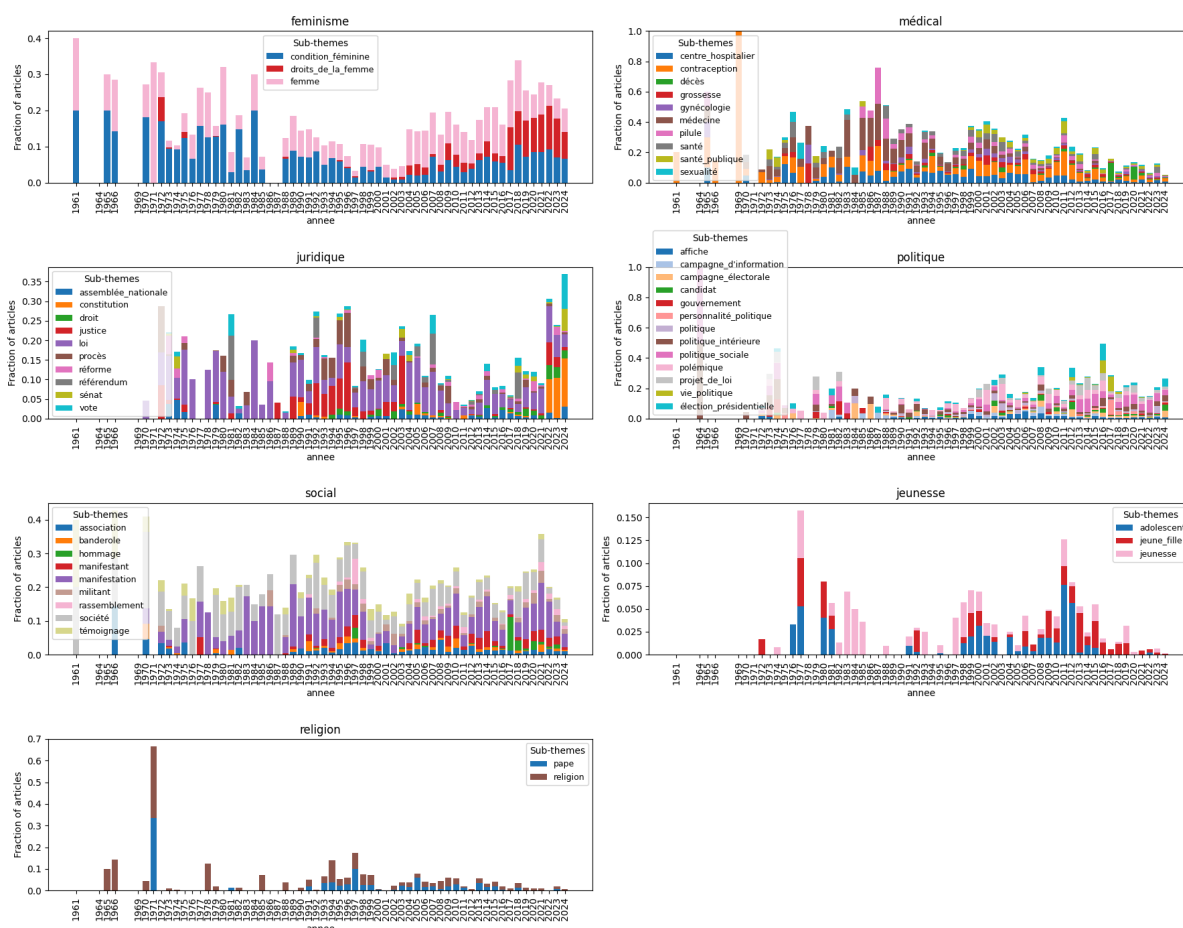
Github notebook PARTIE 1 :

https://github.com/Phineaax/Hackathon_INA/blob/main/Notebook%20traitement%20des%20the%CC%81matiques.ipynb

Identification des thèmes principaux

- > Identification de thèmes dans les documents traités: construction de grandes catégories à partir des thèmes les plus communs dans la colonne “thes”, représentation de la fréquence d’apparition des thèmes pour chaque année proportionnellement au nombre de documents publiés et au nombre de thèmes traités par an.
- > Même tâche relativement à une sélection de chaînes de diffusion.

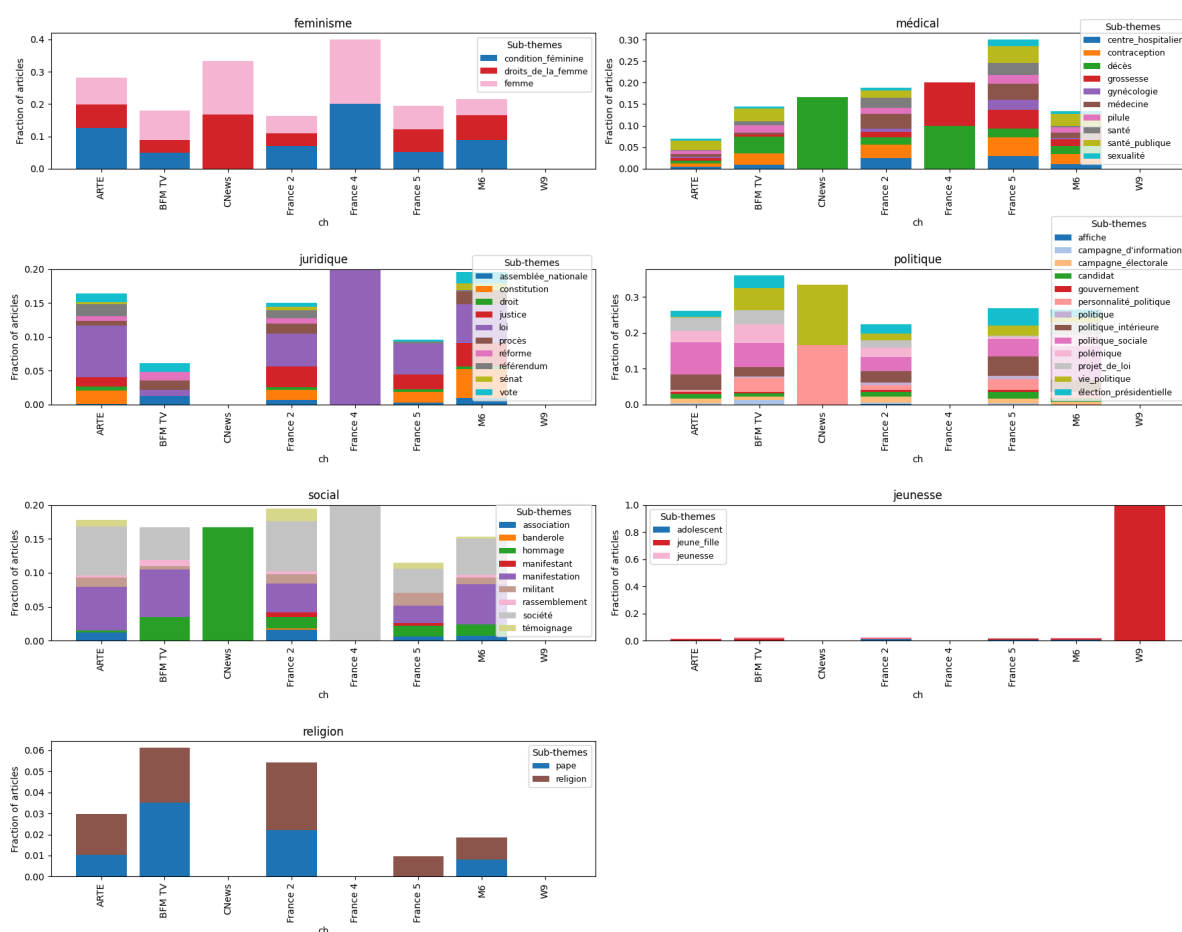
Thèmes principaux selon les années



Observations :

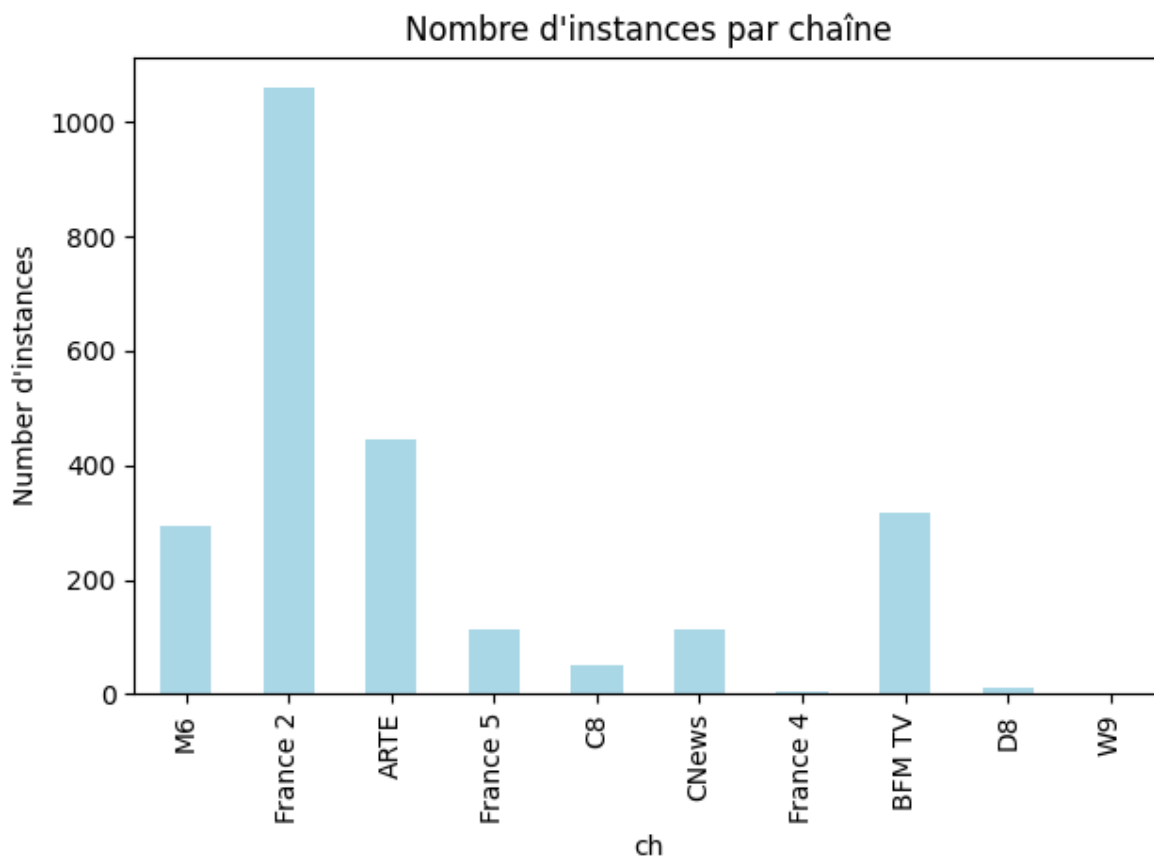
- Apparition du topic des droits de la femme dans les années 2000 puis augmentation de sa pertinence
- Baisse des thèmes médicaux dans le traitement de l'IVG depuis les années 2000.
- Pic de topics juridiques dans les années 2022 à 2024. Creux dans les années 2008 à 2021.
- Pic politique en 2013, vraisemblablement à l'occasion de la loi Taubira.

Thèmes principaux par chaînes



Observations

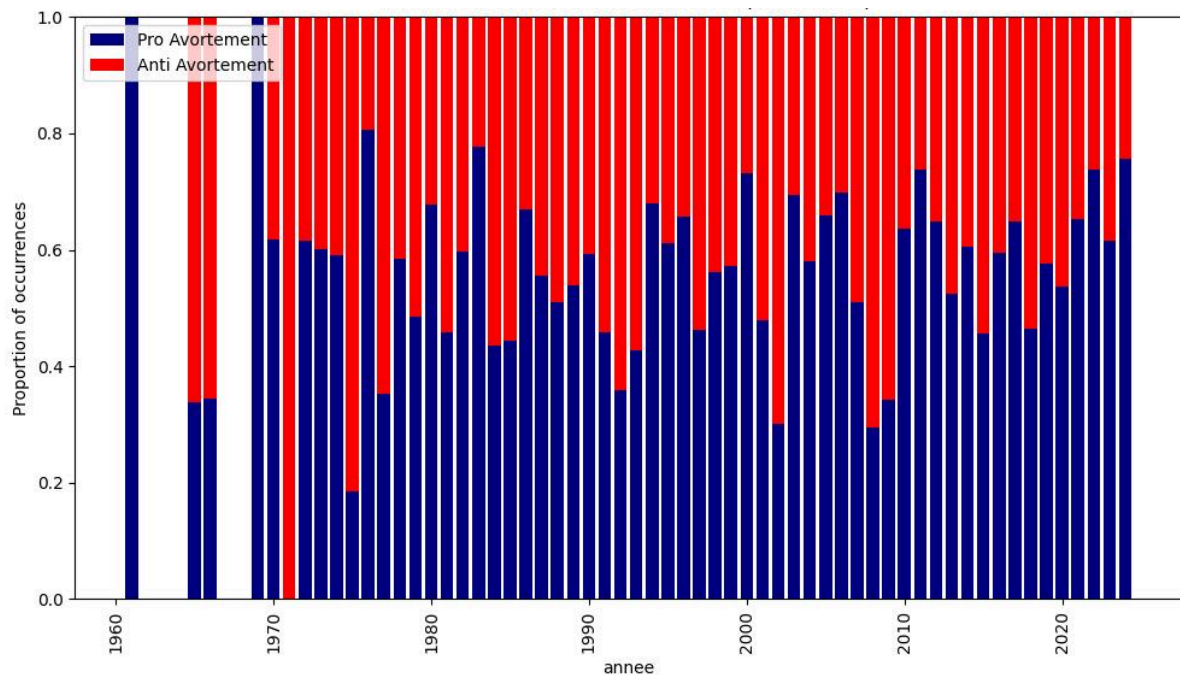
- Certaines chaînes traitent de thèmes peu divers au sein des catégories : CNews ne parle que d'hommage pour le social et de décès pour le médical. Grande diversité sur d'autres



Identification de la tendance pro ou anti avortement

- > Construction de listes de mots connotant un discours pro ou anti avortement (voir annexe). Calcul de la fréquence de textes pro ou anti avortement pour chaque année.
- > De même pour chaque chaîne.

Fréquence relative de discours pro ou anti avortement en fonction du temps



Observations

- Globalement, les chaînes ne sont pas clivantes.
- Quelques marginaux très orientés :
 - Boulogne-Sur-Mer, F3 Lyon locale, Rennes locale, F3 Marseille locale, F3 Périgueux locale, Rennes-Haute Bretagne.
- La première chaîne nationale la plus anti-avortement est CNews.

Pro Avortement vs Anti Avortement (normalized)

Proportion of occurrences

Pro Avortement

Anti Avortement

Media Outlet	Pro Avortement (Proportion)	Anti Avortement (Proportion)
M6 Marseille	0.00	1.00
Télé Wallis et Futuna	0.00	1.00
Télé Toulouse	0.00	1.00
F3 Marseille Locale	0.00	1.00
Rennes-Haute Bretagne	0.00	1.00
Public Sénat	0.40	0.60
Monte Carlo TMC	0.45	0.55
2ème chaîne	0.50	0.50
France 3 Pays de la Loire	0.50	0.50
Télé Nouvelle Calédonie	0.50	0.50
France 3 Rhône	0.50	0.50
France 3 Ile de France	0.50	0.50
France 3 Centre	0.50	0.50
Assemblée Nationale	0.50	0.50
France 3 Occitanie	0.50	0.50
France 3 Bretagne	0.50	0.50
BFM TV	0.55	0.45
TF1	0.60	0.40
Canal 5	0.60	0.40
La Chaîne Info	0.60	0.40
France 3 Lorraine	0.60	0.40
France 3 Alsace	0.60	0.40
Nantes 3	0.60	0.40
France 3	0.60	0.40
France 5	0.60	0.40
La Cinquième	0.60	0.40
France 3 Basse Normandie	0.60	0.40
France 3 Normandie	0.60	0.40
France 3 Centre	0.60	0.40
France 3 Midi Pyrénées	0.60	0.40
France 3	0.60	0.40
France 3 Provence Alpes	0.60	0.40
France 3	0.60	0.40
France 3	0.60	0.40
Télé Martinique	0.60	0.40
France 3 Bourgogne	0.60	0.40
France 3 Auvergne	0.60	0.40
Télé RFO Paris	0.60	0.40
France 3 Champagne Ardenne	0.60	0.40
France 3 Alpes	0.60	0.40
France 3 Pas de Calais	0.60	0.40
France 3 Languedoc Roussillon	0.60	0.40
France 3 Haute Normandie	0.60	0.40
France 3 Poitou Charente	0.60	0.40
France 3	0.60	0.40
Télé Guadeloupe	0.60	0.40
Télé Polynésie Française	0.60	0.40
France 3 Auvergne	0.60	0.40
Télé Saint-Pierre Martin	0.60	0.40
Télé Saint-Pierre Martin	0.60	0.40
Toulouze Bastide	0.60	0.40
Bayonne	0.60	0.40
Boulogne-Sud	0.60	0.40
Bastide	0.60	0.40
D8	0.60	0.40
Télé Guyane	0.60	0.40
Via Stella	0.60	0.40
F3 Lyon	0.60	0.40
Grénoble Locale	0.60	0.40
F3 Montpellier	0.60	0.40
F3 Nice	0.60	0.40
F3 Orléans	0.60	0.40
F3 Reims	0.60	0.40
F3 Périgueux	0.60	0.40
F3 Saint-Etienne	0.60	0.40
La Rochelle	0.60	0.40
Meiz	0.60	0.40
La Rochelle	0.60	0.40
TV5 France Belgique Suisse	0.60	0.40

- Globalement les chaînes ne sont pas clivantes. Quelques marginaux très orientés : Boulogne-Sur-Mer, F3 Lyon locale, Rennes locale, F3 Marseille locale, F3 Périgueux locale, Rennes-Haute Bretagne. On note que la première chaîne nationale la plus anti-avortement est CNews.

```
pro_avortement = [
    "choix", "liberte", "autonomie", "sante", "reproductive",
    "legal", "securisee", "femme", "egalite", "contraception",
    "droits", "protection", "planification", "maternite",
    "decision", "acces", "soutien", "prevention", "controle", "securite"
]

anti_avortement = [
    "vie", "foetus", "enfant", "naissance", "conception",
    "sacre", "bebe", "provie", "interdiction", "lutte",
    "defense", "moralite", "famille", "responsabilite",
    "respect", "opposition", "antiivg", "innocent", "naturel",
    "precieux",
    "meurtre", "assassinat", "suppression", "elimination", "destruction", "homicide" ]
```

PARTIE 2 : Embeddings et extraction

Github notebook PARTIE 2 :

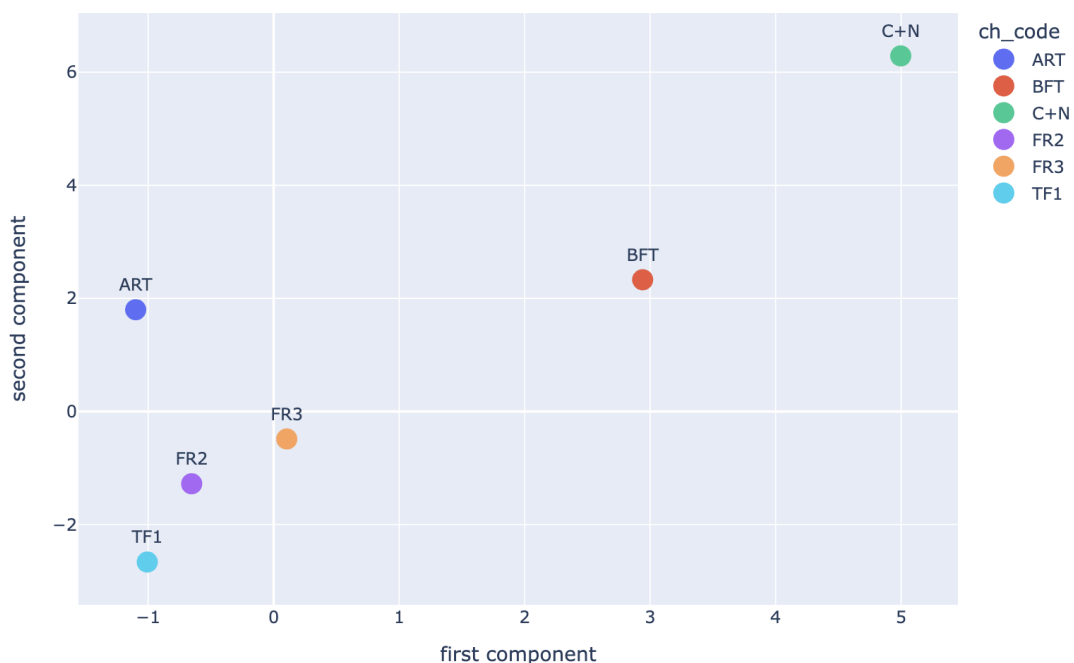
https://github.com/Phineaax/Hackathon_INA/blob/main/Embeddings_et_extraction.ipynb

But : Etudier la proximité des discours et les positionnements idéologiques par chaînes et dans le temps par l'utilisation de différents types d'embeddings

1) Embeddings (Fasttext et All-mini-LM) et visualisation

- Vectorisation des transcriptions via Fasttext (vecteurs de mots)
 - 300 dimensions
- Vectorisation des transcriptions via embedding de textes (All-mini-LM) (vecteurs de textes)
 - Avantage : fenêtre contextuelle non disponible dans les modèles classiques de word embedding
 - Embedding de dimension 384

→ Réalisation d'une visualisation 2D avec réduction de dimensionnalité TSNE du positionnement relatif du discours de chaque chaîne



Visualisation des embeddings (All-mini-LM) moyens par chaînes

Observations :

- L'axe 1 (abscisses) peut être interprété comme un axe d'orientation politique. Cnews et BFMTV sont les plus à droite, ARTE et TF1 les plus à gauches

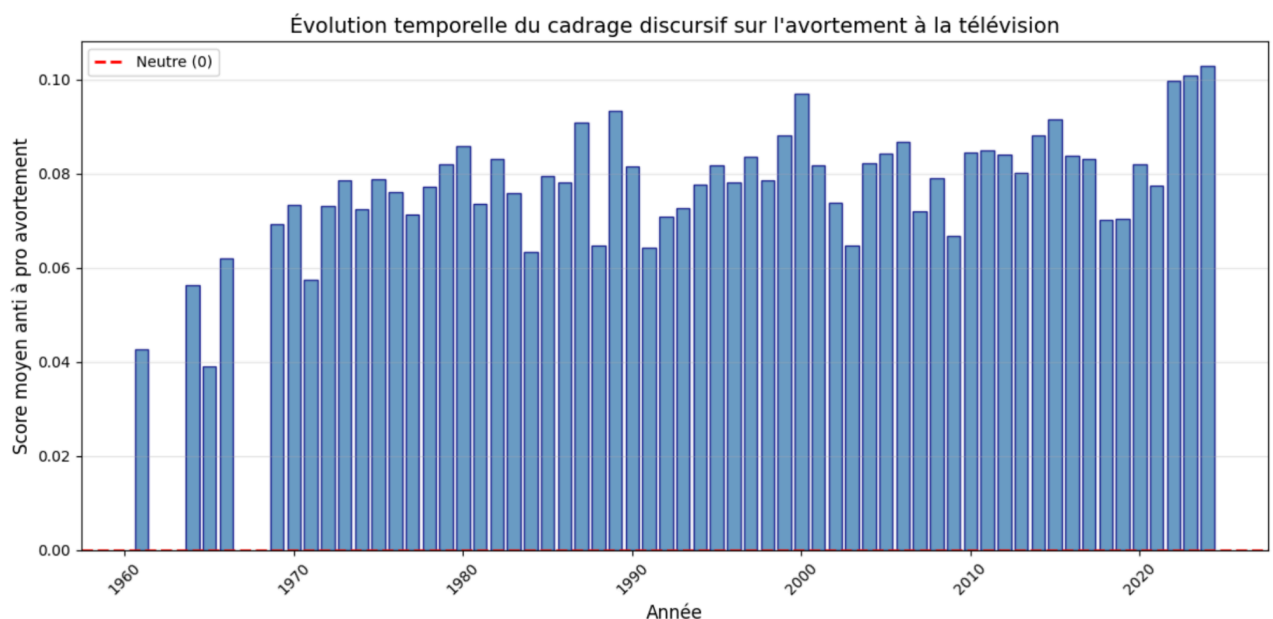
France 2 et France 3 sont assez proches entre eux

- Axe 2 difficilement interprétable

2) Création d'un axe anti-avortement à pro-avortement à partir d'embeddings de mots caractéristiques

- Liste de mots vecteur pro-avortement (protection, soin) :
"choix", "liberte", "santé", "contraception", "droit", "protection",
"soutien", "prévention"
- Liste de mots vecteur anti-avortement (vision criminalisante) :
"interdiction", "meurtre", "assassinat", "suppression", "elimination",
"destruction", "homicide", "immoral", "honte", "délit"

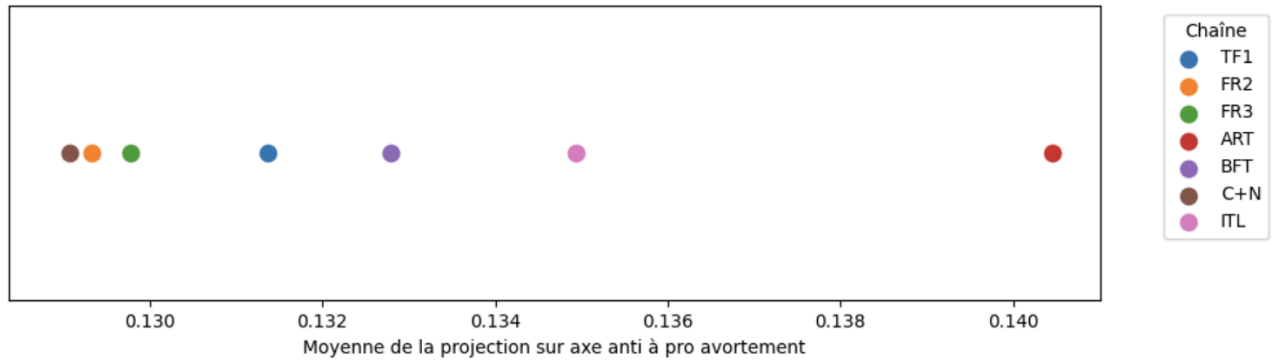
→ Visualisation de l'évolution de l'orientation moyenne (sur l'axe anti-avortement à pro-avortement) des sujets de JT et magazines dans le temps et par chaînes



On observe un glissement vers un cadrage plus pro-avortement au cours du temps mais pas d'évolution linéaire. On distingue un pic sur les années 2022, 2023 et 2024 qui pourrait être expliqué par l'évènement médiatique de la constitutionnalisation de l'IVG en mars 2024.

3) Attribution de scores à chaque Sujet de JT et magazines

→ Position moyenne des chaînes sur l'axe anti-avortement à pro-avortement (visualisation des centroïdes par chaînes)



Arte se différencie des autres chaînes en étant une chaîne particulièrement orientée pro-avortement. Les chaînes CNews et France 2 apparaissent être les plus anti-avortement. Cependant, il est important de noter que les variations entre chaînes sont très minimales.

4) Extraction

Nous avons cherché à résoudre un problème inhérent aux données de l'INA concernant les transcriptions des sujets de JT et magazine. Les transcriptions sont produites à partir des informations de diffusion associées aux notices (chaîne et horaires de diffusion). Or, les horaires de diffusion reposent souvent sur des grilles de programmes prévisionnelles, qui peuvent différer des horaires de diffusion réelle. Cette approximation entraîne fréquemment des décalages dans les transcriptions. Ainsi, certaines transcriptions peuvent inclure des segments qui ne relèvent pas du programme étudié. Les sujets de JT et magazines comportent aussi régulièrement ce que nous pourrions considérer comme du bruit dans notre approche, c'est-à-dire du contenu qui n'est pas en lien avec l'avortement tels que des contenus d'information sur d'autres thématiques ou bien des publicités et programmes météo. Pour résoudre cela, nous employons une approche à base de règle. Nous récupérons toutes les phrases qui contiennent les mots "ivg", "avortement" et "interruption volontaire de grossesse" des transcriptions de paire avec la phrase précédent l'extrait et les deux phrases suivantes. Nous avons extrait manuellement les parties discutant d'avortement sur un échantillon de 50 transcriptions (25 prises au hasard avant les années 2000, 25 prises au hasard après les années 2000) pour comparer avec les performances de l'algorithme.

Résultats extraction :

- Précision moyenne : 0.775
- Rappel moyen : 0.608
- F1-score moyen : 0.681

Limites extraction :

- Ne prend pas en compte les variations dans la quantité de contexte donné

- Match le mot “avortement” même quand ce dernier ne fait pas référence à une ivg. Exemple : “l’avortement du projet...”.

Poursuites :

Utilisation (en local) ou sur un serveur protégé d’un LLM tel que Mistral afin d’extraire les extraits pertinents de manière plus précise et contextuelle.