Annexes à l'enquête sur les discours de haine sur la plateforme YouTube

En 2020, le médialab avait coordonné une première enquête sur l'antisémitisme sur YouTube dans le cadre du 30e rapport annuel de la CNCDH. Cette enquête avait montré qu'en dépit de la modération active mise en place par la plateforme, il restait environ 0,65 % des commentaires postés qui étaient à caractère antisémite.
En 2022, une nouvelle enquête élargit le champ d'étude à un plus grand nombre de commentaires (35 millions) et à d'autres registres de la haine en ligne : l'hostilité à l'égard des musulmans et de l'islam (HEMI), le racisme, le masculinisme et le complotisme.

Annexe 1 - Pipeline Complet

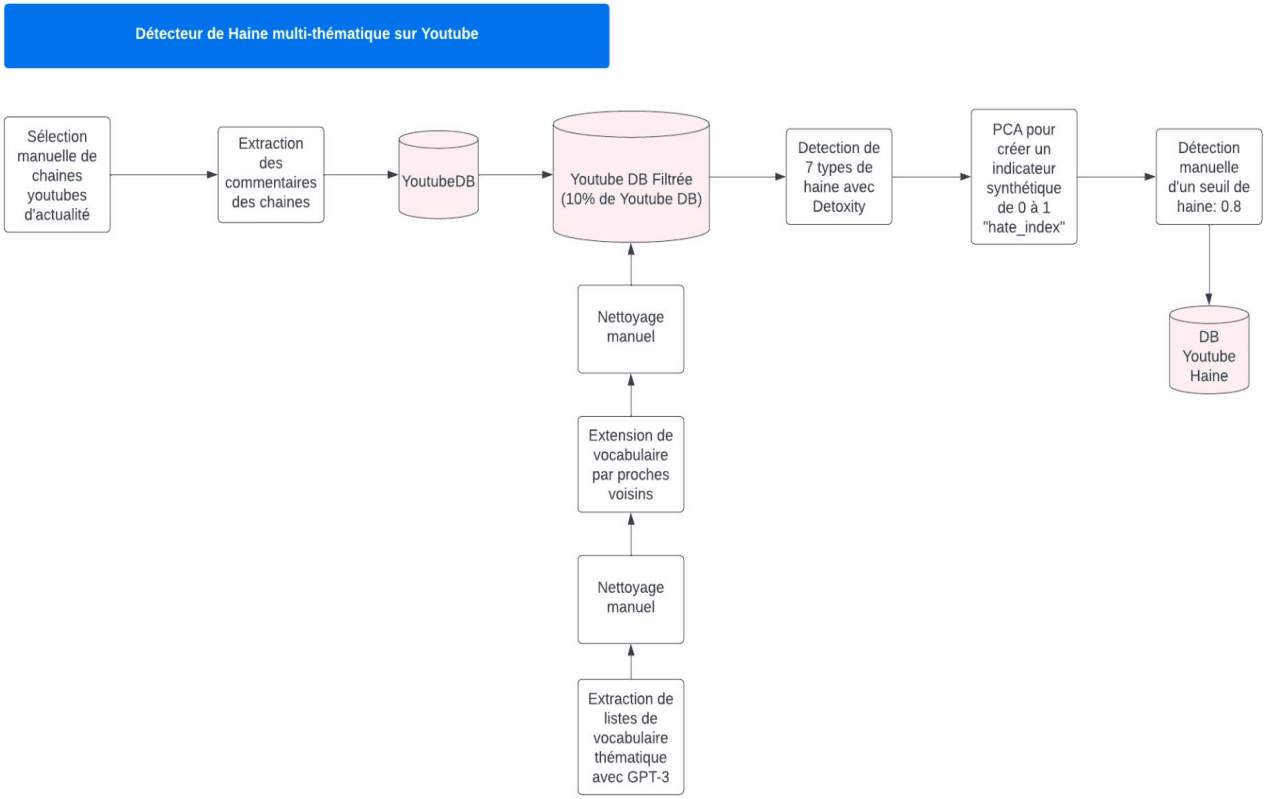

Figure 1: pipeline de description de la création de la base de données de commentaires de haine sur les sujets ciblés du rapport.

Annexe 2 - Création de listes de termes pour isoler des territoires

Tableau 1

Listes des termes d'indexation

Tableau 2

Liste des chaînes en lien avec les territoires d'information

Ce rattachement est, dans une minorité de cas, discutable, car il arrive que certaines chaînes soient commentées massivement par un public qui ne correspond pas forcément à celui ciblé par les créateurs de la chaîne" (Par exemple *Le Crayon* ou *C'est pas sorcier*)

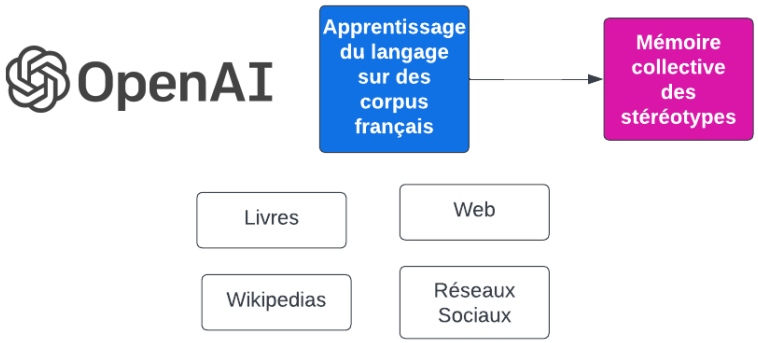

Figure 2 - mémoire paramétrique stéréotypée des GPT (Generative Pretrained Models).

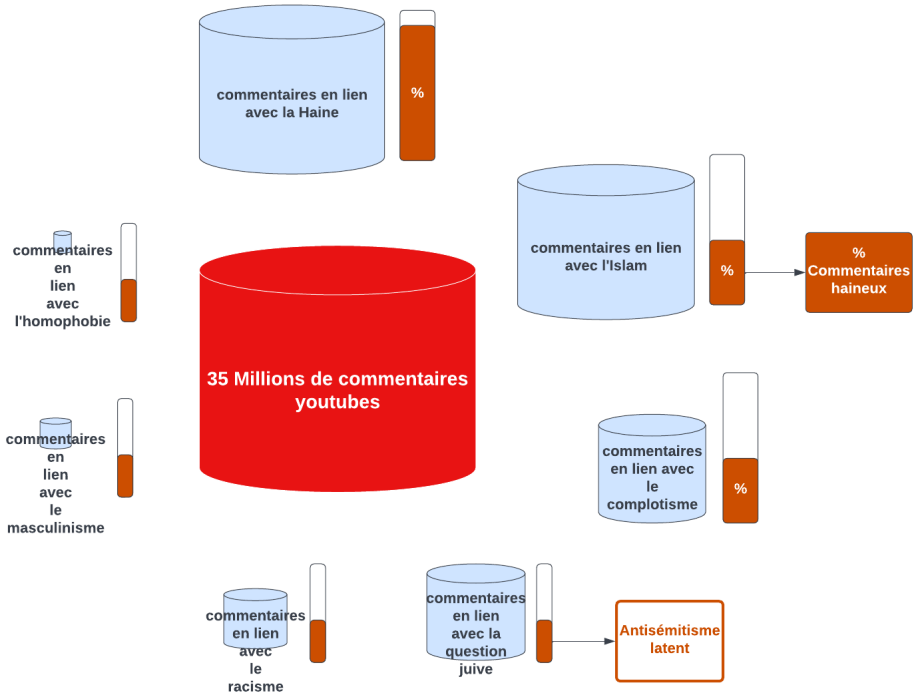

Figure 3 - Le corpus Youtube est subdivisé en sous-corpus d'intérêt. Chaque sous-corpus possède un % de commentaires détectés comme haineux par l'algorithme de détection de haine "Detoxify".

Annexe 3 - Utilisation d'un algorithme de reconnaissance de la haine en ligne

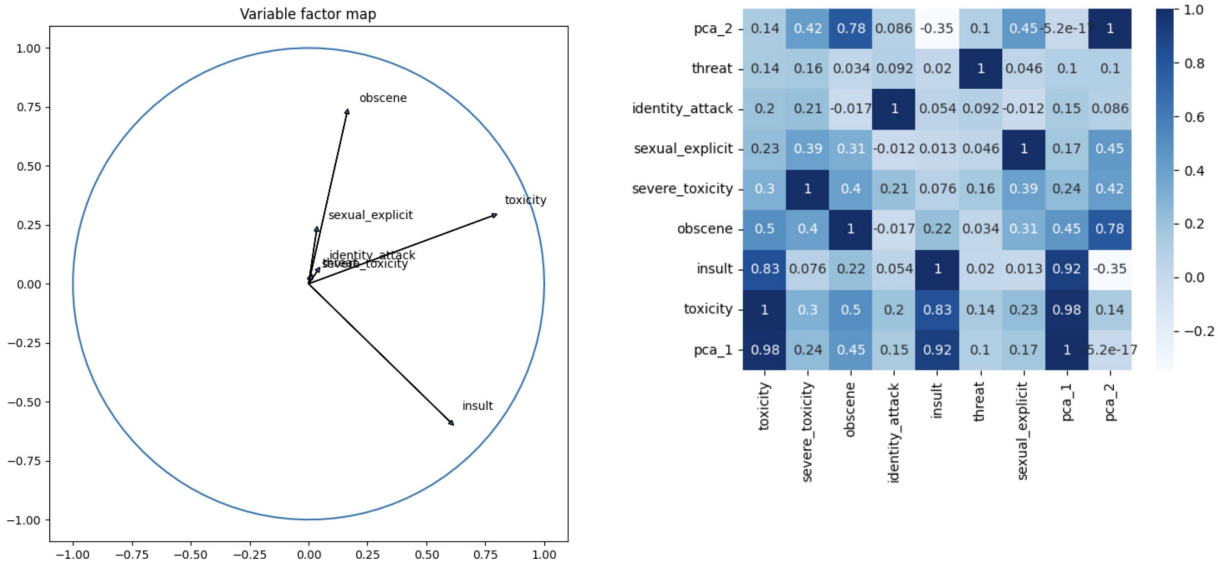

fig 4: carte de corrélation des variables de l'ACP. La première composante de l'ACP corrèle presque parfaitement avec l'indicateur de toxicité *toxicity* (98%) et les *Insultes* (92%). Il corrèle de façon plus modeste avec *l'obscénité (45%) la toxicité sévère (24%) et l'attaque contre l'identité (15%). Enfin, notre indicateur corrèle de façon faible avec les explicitation sexuelle (17%), les menaces (10%) et les attaques contre l'identité (15%).*

Example input	Model type	Language	Toxicity	Severe Toxicity	Obscene	Threat	Insult	Identity attack	Sexually explicit
shut up, you are a liar	original	en	98.38	1.21	25.59	0.29	80.79	0.78	N/A
i am a jewish woman who is blind	unbiased	en	13.52	0.09	0.45	0.35	1.67	22.4	0.14
tais-toi, tu es un menteur	multilingual	fr	99.38	0.79	8.03	0.65	21.74	0.42	2.54
cállate, eres mentirosa	multilingual	es	98.85	0.37	5.9	0.29	29.28	0.35	1.76
kapa çeneni, sen bir yalancısın	multilingual	tr	99.45	1.02	10.13	2.02	29.1	0.53	6.1
stai zitto, tu sei un bugiardo	multilingual	it	99.22	0.32	5.45	0.21	38.71	0.32	0.9
заткнись ты лжец	multilingual	ru	99.24	0.8	10.88	0.61	23.71	0.63	2.22
cala a boca, você é um mentiroso	multilingual	pt	98.96	0.44	6.66	0.67	41.82	0.24	1.53

fig 5: - *extrait du package (https://github.com/unitaryai/detoxify).* Les commentaires en plusieurs langues sont associés à plusieurs catégories de haine. Dans le cas du rapport, nous avons utilisé

le model « multilingual », seul capable de prendre en compte la langue française. La toxicité est un indicateur composite des autres formes de haine. Afin de s'assurer de résilier un détecteur de haine généraliste et applicable dans la plupart des situations, nous avons recréé nous-même un indicateur composite. Pour cela, nous avons passé le modèle sur les données filtrées des dictionnaires et avons réalisé une Analyse en Composantes Principales (ACP) sur les 7 formes de haines différentes : la première composante de l'ACP représente le facteur le plus explicatif de toutes ces formes de haine, en d'autres termes, cette composante représente le dénominateur commun des formes de haine. Les différentes variables corrèlent différemment avec les deux premiers facteurs de l'ACP car chacune des variables exprime une haine différente. La *toxicité* se trouve à mi-chemin des autres formes de haine, ce qui confirme son caractère composite.
Nous avons choisi le premier axe de l'ACP comme notre nouvel indicateur de haine dans le corpus car il expliquait 77% de la variance totale des données et répondait donc à notre enjeu de création d'un détecteur de haine composite.

Annexe 4 - Description de notre algorithme d'analyses de thématiques

La haine que nous avons détectée dans les commentaires peut prendre différentes formes au sein d'un même registre. Il existe par exemple plusieurs formes d'antisémitisme et plusieurs formes de racisme.

Afin de détecter ces différents registres, nous avons utilisé des algorithmes d'analyse thématique. Le *Topic Modeling* est une technique de traitement automatique du langage qui permet de grouper des commentaires en fonction de leur contenu sémantique (mots contenus dans le commentaire, structure du commentaire…). Cette technique est particulièrement utile pour l'analyse de grandes quantités de données textuelles, comme les articles de presse, les blogs, les réseaux sociaux, etc. Les algorithmes de recommandation sur les réseaux sociaux utilisent ce type de stratégie pour comprendre les différents groupes d'intérêt des utilisateurs pour leur recommander du contenu. Les premières approches du *Topic Modeling* étaient principalement orientées statistiquement, telles que l'Allocation Latente de Dirichlet (LDA) (Blei et al., 2003) et la Factorisation de Matrices Non Négatives (NMF) (Févotte et Idier, 2011). Les embeddings ont ensuite été vus comme un moyen de corriger les problèmes ou les limitations des techniques statistiques (Sia et al., 2020). Les nouvelles méthodes telles que Doc2Vec (Mikolov, 2014), Top2Vec (Angelov, 2020) et Bertopic (Grootendorst, 2022) ont été développées récemment. En combinant le *Topic Modeling* et les embeddings, il est possible d'obtenir des résultats de classification de documents plus précis et de visualiser en deux dimensions les résultats du *Topic Modeling*, ce qui facilite l'interprétation des données, indispensable dans ce genre de discipline.

Nous avons fait le choix de présenter les résultats de cette analyse thématique sous la forme de cartes topologiques afin d'aider le lecteur de ce rapport à se faire une idée globale du sujet abordé et d'explorer les résultats.

Afin de créer les topics, nous avons utilisé un modèle multilingue embeddings (Sbert) pour transformer les documents en vecteurs. Nous avons réduit en 2 dimensions avec Umap et avons utilisé une méthode du K-Means pour détecter 20 clusters. Les clusters sont ensuite inspectés manuellement et supprimés ou fusionnés si besoin. Nous avons utilisé une nouvelle méthode, proche de celle développée par Bertopic. Pour la représentation du topic (les quelques termes qui permettent de décrire rapidement le contenu) du topic, nous avons extrait les termes avec le package *spacy*. Cela permet d'avoir des unigrams ou des bigrams précis. Nous avons ensuite filtré les termes par leur occurrence pour être certains d'avoir des termes qui sont souvent utilisés. Nous avons ensuite utilisé une technique du chi2 pour déterminer les termes les plus spécifiques par territoire d'analyse. La spécificité permet d'avoir les documents qui ont une occurrence forte dans un topic mais qui ne sont pas présents dans les autres topics. Cela vient de l'idée qu'on distingue mieux des topics par ce qui les distingue plutôt que par ce qui les lie. La représentation visuelle des topics est indispensable pour aider l'esprit à générer les intuitions et interpréter les résultats. Pour chaque topic, nous avons calculé son centre (le centroid du topic qui correspond

à la moyenne des coordonnées des documents qu'il possède). Sur ces centroids, nous avons placé les termes représentant au mieux le topic (voir paragraphe précédent).

Nous avons utilisé un graphisme qui s'apparente à une véritable carte pour aider l'esprit humain à activer ses intuitions spatiales : deux topic proches sur la carte sont plus similaires sémantiquement que deux topics éloignés. Cela fait écho à la première loi de la géographie de Tobler – *« tout est lié à tout le reste, mais les choses proches sont plus liées que les choses éloignées »*.

Cette approche permet de favoriser le *sensemaking*, qui est définie comme une activité ou un comportement qui implique souvent une étape où la personne chercheuse d'informations crée une représentation de l'espace d'information. Le *sensemaking* est une activité itérative faisant partie du processus de recherche exploratoire. En effet, la personne chercheuse d'informations se déplace dans l'espace d'information et a besoin de donner un sens aux espaces inconnus (Ryen W. White & Resa A. Roth, 2009, p. 32). Pour se déplacer dans l'espace, il est préférable que l'espace ressemble à un espace connu, tel qu'une carte.

Sur la carte, la densité représente le nombre de commentaires en lien avec un topic. Cela permet de se faire une idée des topics qui attirent le plus de commentaires haineux. De plus, le centre de la carte peut être interprété comme la partie structurante du discours de haine: Le centre d'un graphique *d'embeddings* a des propriétés structurelles car il peut fournir des informations sur les relations et les similitudes entre les points du graphique. Comme les points du graphique représentent les commentaires de haine encodés dans un espace à haute dimension, le centre du graphique peut représenter les commentaires les plus courants ou les plus centraux dans le jeu de données. Les topics au centre représentent les topics qui partagent un peu d'information avec les autres topics aux alentours, ce qui leur donne leur propriété structurante: les topics du centre ont plus de relation avec tous les autres topics que les topics situés à des extrémités.

Les méthodes qui consistent à utiliser des réseaux de neurones profonds (« Deep Learning ») possèdent quelques limites bien connues : elles manquent de clarté quant aux choix de leur décision, ce qui interroge sur leur utilité réelle dans le cadre d'études de ce genre. En effet, les algorithmes de Deep Learning ont pu apprendre une notion biaisée des discours de haine en fonction des *datasets* sur lesquels ils ont été entraînés. Par exemple, si l'algorithme trouve le mot « taliban » dans un commentaire, il va associer ce dernier à de la haine car il a appris que les commentaires qui possèdent ce terme sont souvent haineux. Or, il est tout à fait possible de parler des talibans sans développer une rhétorique haineuse. Nous ne détectons pas à proprement parler de la « haine » mais « un discours de haine latent », c'est un dire une « probabilité d'existence » d'un discours haineux dans un commentaire.

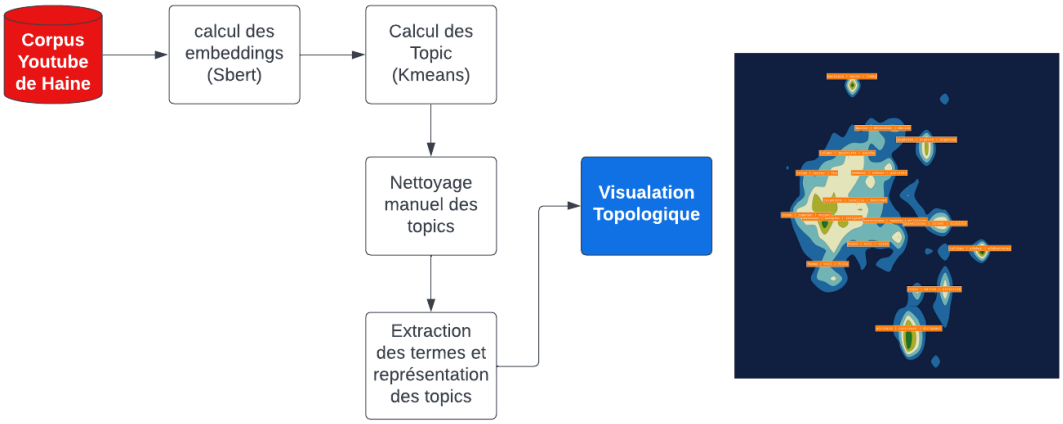

fig 6: description technique de l'algorithme de "Topic Modeling" .

Annexe 5 - Revue de littérature sur les détecteurs de haine

Les détecteurs de haine s'appuient sur les avancées récentes en traitement automatique du langage (TAL) qui permettent de « plonger » du contenu textuel dans des espaces latents. Cette notion de plongement dans des espaces vectoriels connaît un développement important depuis quelques années, en particulier pour analyser le discours de haine (Fortuna, 2018). Ces méthodes se basent par exemple sur la reconnaissance de mots-clés (Dinakar, 2011 ; Dadvar, 2012 ; Isbister et al., 2018), ou sur l'analyse du ton des discours : on parle alors d'analyse de sentiments (Gitari et al. 2015 ; Davidson et al. 2017 ;; Del Vigna et al., 2017). Plus récemment, des méthodes de représentation vectorielle des mots puis de textes entiers fondées sur des réseaux de neurones se sont développées (Al-Makhadmeh et al., 2020). Les algorithmes les plus populaires pour plonger des segments de textes dans ces espaces vectoriels de grandes dimensions (typiquement entre 300 et 700 dimensions) s'appellent doc2vec, paragraph2vec, FastText, Bert ou plus récemment GPT-3 (Djuric, Zhou, Morris, Grbovic, Radosavljevic, Bhamidipati, 2015 ; Schmift, Wiegand, 2010 ; Siegel, 2020). Chaque nouvelle génération de ces modèles de langues neuronaux promet une compréhension toujours plus contextuelle et plus fine de la sémantique et de la syntaxe des contenus textuels et correspond au champ plus large des "embeddings".

Les embeddings sont des représentations vectorielles de mots, phrases ou documents. Ces représentations sont générées par des algorithmes d'apprentissage automatique (Transformers, Bert, Sbert) et permettent de capturer les relations sémantiques entre les mots d'un corpus. Les techniques d'embeddings sont apparues progressivement avec TF-IDF (Joachims, 1996), puis avec Word2Vec, puis Glove et enfin BERT (Devlin et al., 2018) qui est basé sur l'architecture des Transformers (Vaswani et al., 2017). Plus récemment, Sbert (Reimers, 2019) en particulier (Sentence-bert) a atteint des performances d'état-de-l'art sur diverses tâches d'embedding de phrase. Ce modèle est un modèle Transformer basé sur l'architecture de Bert, en particulier celle de Roberta (Liu, 2019). Sbert ajoute une opération de pooling à la sortie de RoBERTa pour dériver un embedding de phrase de taille fixe.