ISD | Powering solutions to extremism and polarisation

ONLINE CIVIL COURAGE INITIATIVE

Cartographie de la Haine en Ligne

Tour d'horizon du discours haineux en France

Cooper Gatewood, Cécile Guerin,
Jonathan Birdwell, Iris Boyer & Zoé Fourel

À propos de ce rapport

Ce rapport présente les résultats d'un projet de recherche sur l'envergure et la nature des discours haineux en ligne en France.
Il examine différentes catégories de discours haineux, couvrant l'origine ethnique et raciale, le sexe, l'orientation sexuelle, la religion et le handicap.
À l'aide d'outils d'analyse de données sur les réseaux sociaux qui combinent l'apprentissage machine et le traitement du langage naturel à une analyse qualitative, le rapport fournit une analyse détaillée de la dynamique des types de discours haineux les plus répandus sur les plateformes de réseaux sociaux en France.
Ce rapport émet également des recommandations à l'égard des entreprises technologiques, du gouvernement et des organisations de la société civile en matière de mesures pour contrer les discours haineux en ligne. Établi dans le cadre de l'OCCI (Initiative pour le Courage Civique en Ligne), un partenariat stratégique entre l'Institut pour le Dialogue Stratégique (ISD) et Facebook, ce rapport aspire à donner à la société civile et aux décideurs politiques des moyens pour contrer la haine en ligne.

London | Washington DC | Beirut | Toronto

Design par forster.co.uk. Mise en forme par Danny Arter.
Rapport français mis en forme par Eric Bauer.

Les auteurs

Cooper Gatewood
Titulaire d'une maîtrise en affaires internationales de l'Université de Columbia et d'une maîtrise en sécurité internationale de Sciences Po, Cooper travaille dans l'unité de recherche numérique de l'ISD.
Il y est responsable de la recherche quantitative sur la diffusion de discours haineux et clivants en ligne et sur la façon dont ils sont exploités par des acteurs extrémistes.
Cooper élabore par ailleurs des cadres de suivi et d'évaluation pour mesurer l'impact d'un grand nombre de projets d'intervention de l'ISD. Il contribue actuellement aux recherches de l'ISD sur les campagnes de désinformation, en particulier celles qui visent à influencer et à perturber les processus électoraux.
Enfin, il est responsable de l'OCCI en France, coordonnant les activités visant à soutenir la réponse de la société civile à la haine et à l'extrémisme en ligne, et gère l'évaluation quantitative de nombre de programmes de l'ISD, notamment Be Internet Citizens et Young Digital Leaders.

Cécile Guerin
Titulaire d'une maîtrise en histoire internationale de la London School of Economics et d'une maîtrise en anglais de l'École normale supérieure en France, Cécile contribue à de nombreuses publications dont The Guardian, Prospect et The Independent.
Elle est coordinatrice à l'ISD et soutient le travail de développement et d'analyse de l'organisation en Europe. Elle travaille sur l'OCCI, un projet financé par Facebook qui vise à intensifier les efforts de la société civile pour lutter contre les discours haineux et l'extrémisme en ligne.
Elle contribue également aux travaux de recherche et d'élaboration de politiques de l'ISD, en accordant la priorité à l'analyse des réseaux sociaux et à la cartographie des réseaux liés au discours haineux, à l'extrémisme et à la désinformation en ligne.

Les auteurs

Jonathan Birdwell
Titulaire d'une maîtrise (avec distinction) de la London School of Economics and Political Science et d'une licence en sciences politiques et en philosophie de la Tulane University de la Nouvelle-Orléans (Louisiane), Jonathan est directeur adjoint de l'équipe Politique & Recherche à l'ISD.
À ce titre, il est responsable du travail de l'ISD sur les politiques publiques et leurs réseaux d'acteurs, et dirige également les programmes éducatifs.
Il supervise la recherche et les principaux ensembles de données de l'ISD, le suivi et l'évaluation des programmes ainsi que de la validation de tous les documents écrits de l'ISD.
Il œuvre actuellement à l'établissement de partenariats uniques pour l'ISD, à la technologie et aux capacités d'analyse en ligne pour fournir une compréhension à jour de la propagande extrémiste et des tactiques de recrutement, ainsi qu'à l'évaluation des campagnes en ligne et des interventions individuelles en ligne de l'ISD.
Avant de s'engager à l'ISD, Jonathan était chef de programme au groupe de réflexion multipartite britannique Demos pour lequel il a publié plus de 40 rapports de recherche sur des sujets comme l'extrémisme violent islamiste (The Edge of Violence, 2010) et d'extrême droite (The New Face of Digital Populism, 2011).
Ses contributions écrites portent également sur l'éducation (The Forgotten Half, 2011), l'apprentissage social et émotionnel (Character Nation, 2015), l'action sociale des jeunes et leurs attitudes envers la politique (Tune In, Turn Out, 2014), la politique et le marketing numériques (Like, Share, Vote, 2014), la confiance envers le gouvernement (Trust in Practice, 2010) et la religion et l'intégration (Rising to the Top, 2015).

Iris Boyer
Titulaire du diplôme en sciences sociales et humaines décerné par Sciences Po Toulouse et d'une maîtrise internationale en affaires publiques de la Higher School of Economics de Moscou et de la London Metropolitan University, Iris est responsable adjointe de la division Technologie, Communications et Éducation de l'ISD.
Elle supervise nombre de programmes visant à soutenir et amplifier les efforts de la société civile en matière de lutte contre l'extrémisme par le biais de partenariats stratégiques avec des entreprises technologiques et des organisations de terrain.
Elle coordonne également des réseaux multisectoriels regroupant représentants du gouvernement, du monde universitaire, des médias et d´organisations non gouvernementales (ONG).
Elle conseille régulièrement ces différents acteurs sur les dernières tendances de l'extrémisme et les approches les plus efficaces et les plus novatrices pour endiguer sa normalisation.
Elle œuvre également à l'expansion régionale de l'ISD, en particulier en France où elle dirige les activités liées à son développement et son leadership d'opinion.

Zoé Fourel
Titulaire du diplôme en affaires internationales (comprenant des études à la School of Oriental and African Studies à Londres et à Georgetown University à Washington, DC) décerné par Sciences Po Lyon, Zoé est Associée à l'ISD.
En plus d'avoir directement contribué à ce projet de recherche, elle participe activement à à la coordination des activités de l'OCCI en France. De surcroit, elle contribue aux programmes de l'ISD axés sur les réponses de la société civile pour lutter contre la haine et l'extrémisme.

Remerciements

Nous aimerions exprimer notre gratitude à l'équipe de l'ISD, en particulier Henry Tuck et Chloe Colliver, pour leurs commentaires et leurs suggestions judicieuses, sans oublier Hannah Martin qui a coordonné la production du rapport.

Nous tenons à remercier les organisations et les institutions qui nous ont conseillés sur la méthodologie du rapport et nous ont aidés à choisir nos mots-clés, notamment la Ligue Internationale Contre le Racisme et l'Antisémitisme (LICRA), Le Refuge, La Voix des Roms, the American Jewish Committee (AJC) et #JeSuisLà.

Nous aimerions également remercier nos partenaires du Centre d'analyse des réseaux sociaux (CASM), en particulier Carl Miller, Josh Smith, Chris Inskip et Andrew Robertson pour leur soutien dans la formation des classificateurs et l'analyse des données.

Ce rapport n'aurait enfin pas pu être publié sans le soutien financier de Facebook France, qui a subventionné cette recherche. Chez Facebook, nous sommes particulièrement reconnaissants à Clotilde Briend, Hamida Moussaoui et Sarah Yanicostas pour leur soutien.

Toute erreur ou omission est la responsabilité des auteurs.

Sommaire

Résumé

Ces deux dernières années, la haine en ligne est devenue une préoccupation publique majeure en France. À la suite de l'adoption de la loi NetzDG (Loi sur les réseaux sociaux) en Allemagne, l'Assemblée Nationale française a adopté la proposition de Loi Avia contre les contenus haineux sur internet. Cette loi impose aux entreprises de technologie de supprimer les contenus illicites de leurs plateformes. Au moment de la rédaction du présent rapport (novembre 2019), la loi n'a pas encore été adoptée par le Sénat.
Depuis, le 17 Décembre la proposition de loi a été débattue en seconde lecture par les Sénateurs qui ont proposé un certain nombre d'amendements importants y compris sur une de ses mesures phares.

Les entreprises de technologie ont intensifié leurs efforts pour identifier, analyser et quantifier les propos haineux sur leurs plateformes.
Le quatrième rapport du Code de conduite de l'UE sur les discours de haine illégaux en ligne a montré que les entreprises technologiques évaluent à 89 % les contenus signalés dans les 24 heures.[1] En 2018, quatre nouvelles entreprises (Google+, Snapchat, Instagram et Dailymotion) ont adhéré au code de conduite de l'UE, aux côtés des membres fondateurs que sont Facebook, Microsoft, Twitter et YouTube en 2016.[2]

Malgré les changements législatifs amorcés afin de lutter contre les discours de haine en ligne, le rapport parlementaire de la Loi Avia[3] a souligné l'insuffisance de données précises sur l'ampleur du problème.
C'est dans ce contexte que l'OCCI (Initiative pour le courage civique en ligne), un partenariat stratégique entre l'ISD et Facebook, aspire à fournir à la société civile des données de recherche et des ressources de mobilisation afin d'améliorer la réponse civique à donner aux contenus haineux en ligne.

Élément clé du travail de l'OCCI en 2019, ce rapport vient combler le manque d'évidence en offrant une vue d'ensemble et une perspective factuelle des diverses formes de discours haineux en ligne en France sur différentes plateformes de réseaux sociaux, via l'analyse de données accessibles en ligne.
Son objectif est de donner aux décideurs publics, aux entreprises de technologie et aux organisations de la société civile un aperçu de la portée et de la nature des discours haineux en ligne pour éclairer les réponses politiques, technologiques et civiques à mettre en place.
Les résultats de ce rapport s'appuient d'une part, sur une analyse de données réalisée à l'aide d'outils d'écoute des réseaux sociaux et de logiciels de traitement du langage naturel, et d'autre part, sur une analyse qualitative. Pendant cinq mois, nous avons récupéré des données sur différents types de discours haineux à l'aide de mots-clés que nous avons pu identifier en réalisant des recherches sur les milieux extrémistes, en concertation avec des acteurs de la société civile française œuvrant à contrer tous types de manifestations de haine.

Nous avons alors formé des algorithmes pour identifier la proportion de contenus haineux pour les quatre catégories de discours haineux dont les mots-clés renvoyaient le plus grand nombre de messages. Puis, nous avons effectué une analyse qualitative pour les autres types de discours haineux.

Principaux résultats
Ce rapport s'efforce de mettre en évidence les principaux types de discours haineux en ligne en France (tirés de données accessibles publiquement) et ainsi éclairer les réponses civiques, politiques et

> « Les résultats de ce rapport s'appuient d'une part, sur une analyse de données réalisée à l'aide d'outils d'écoute des réseaux sociaux et de logiciels de traitement du langage naturel, et d'autre part, sur une analyse qualitative »

19%

des comptes qui postent le plus de discours haineux avaient un comportement automatisé ou de type bot

13%

des comptes qui postent le plus de discours haineux sont affiliés à des idéologies ou groupes d'extrême droite

technologiques à mettre en œuvre face à ces discours clivants.
Pour ce faire, onze ensembles de données ont été créés à partir d'outils d'écoute des réseaux sociaux en suggérant des termes fréquemment associés à des contenus ciblant des groupes à raison du sexe et de l'orientation sexuelle, de l'origine ethnique et raciale, de la religion et du handicap et dont l'intention était d'inciter à la haine, à la violence ou à la discrimination.

Les quatre plus grands ensembles de données ont été analysés à l'aide d'algorithmes de traitement du langage naturel afin de déterminer l'ampleur du discours haineux. Chacun de ces discours présente ses propres particularités, sachant que certains éléments clés se retrouvent dans l'ensemble des discours haineux. Nous avons constaté ce qui suit :

- **Grâce au machine learning, nous avons pu recenser avec certitude un peu moins de 7 millions de cas de discours haineux en ligne contre les femmes, les personnes de la communauté LGBTQ (lesbiennes, gays, bisexuels, transgenres et queer), les personnes handicapées et les communautés arabes françaises.** Parmi ces cas se trouvaient environ 5,4 millions de discours haineux misogynes, plus d'un million de propos anti-LGBTQ, 265 000 propos discriminatoires fondés sur la capacité physique (capacitistes ou validistes) et 131 000 propos anti-Arabes. Notre capacité à identifier avec certitude et de façon algorithmique les propos haineux pour ces groupes a bénéficié de la très grande envergure des premiers ensembles de données de mots-clés haineux liés à ces groupes, illustrant ainsi l'utilisation répandue des injures et des insultes qui les visaient au départ. Il n'a pas été possible d'identifier de façon algorithmique les propos haineux pour les autres groupes. Une analyse manuelle et qualitative plus poussée a donc été nécessaire. L'ampleur des propos haineux en ligne est vraisemblablement beaucoup plus élevée mais elle ne peut être évaluée en raison du manque d'accès aux communications publiques sur de nombreuses plateformes des réseaux sociaux.

- **La plupart des discours haineux en ligne comprenaient l'utilisation généralisée d'injures et d'insultes basées sur des attaques visant des catégories protégées.** L'utilisation normalisée d'un langage misogyne *(sale pute)*, d'injures homophobes *(pédé)* et d'insultes discriminatoires fondées sur la capacité physique (*mongol*, terme péjoratif pour désigner un trisomique 21) est très répandue dans notre échantillon. Nous avons recensé plus de 4 millions de messages employant un langage misogyne haineux, un peu moins d'un million de messages employant un langage anti-LGBTQ haineux et près de 250 000 messages employant un langage discriminatoire haineux fondé sur la capacité physique. Nous avons souvent constaté des recoupements importants dans les comptes où des propos haineux sont utilisés. Les insultes misogynes, anti-LGBTQ et discriminatoires fondées sur la capacité physique étaient le plus souvent utilisées pour attaquer des figures politiques et des footballeurs.

- **Un faible pourcentage de discours haineux était constitué d'attaques ciblées contre des individus.** Environ 5 % de notre panel de données comprenait des attaques misogynes directes contre des utilisatrices (en raison de leur sexe). Si le machine learning n'a pas permis d'identifier toutes les attaques ciblées pour tous les types de discours, l'analyse qualitative a suggéré que les groupes recevant le plus grand nombre de discours haineux ciblés en ligne étaient les femmes, les Arabes et les musulmans.

- **Parmi les comptes qui postent le plus souvent des propos haineux, 19 % présentaient un comportement automatisé ou de type bot.**[4] Environ 13 % affichaient leur appartenance à un groupe ou à une idéologie d'extrême droite, 4 % étaient associés au mouvement des Gilets Jaunes et 4 % avaient été supprimés au moment de la rédaction du présent rapport. Une petite partie de ces comptes appartenait à des sources d'actualités françaises alternatives.

Sur les dix premiers comptes les plus actifs sur Twitter partageant des mots-clés anti-Musulmans, tous étaient affiliés à l'extrême droite.

• **Des événements, tels que la Journée internationale de la femme ou l'annonce de la nomination de Bilal Hassani à l'Eurovision, ont provoqué des pics de discours haineux.** Parmi les autres événements qui ont provoqué des discours haineux contre les musulmans et les utilisateurs arabes, citons le début du Ramadan et l'attaque de Christchurch.

• **Notre approche basée sur les mots-clés a mis en évidence une faible proportion d'efforts organiques et/ou coordonnés de contre-discours sur les réseaux sociaux.** Ainsi, environ 1 % de l'ensemble de données élaboré à partir de mots-clés misogynes était composé d'utilisateurs qui s'opposaient à l'utilisation du langage et des insultes misogynes. On retrouve également ces tentatives de contre-discours dans les ensembles de données anti-LGBTQ, anti-Noirs et anti-Roms. Ont enfin été constatés des cas de réappropriation des insultes haineuses dans le contre-discours.

• **Un recoupement important existe entre les différents types de discours haineux, ce qui démontre la nécessité d'une analyse « intersectionnelle » des discours de haine en ligne.** Cette constatation est particulièrement vraie pour les discours haineux misogynes et anti-LGBTQ, ainsi que pour les discours haineux anti-Arabes et anti-Musulmans.

• **Notre recherche a démontré le potentiel et les limites que présentent les algorithmes de traitement du langage naturel pour identifier les discours haineux en ligne.** Nos chercheurs ont pu créer les algorithmes pour identifier les propos haineux en ligne avec une précision de l'ordre de 85 % - un niveau élevé pour ce type de recherche.[5] Toutefois, la diversité de la terminologie existante et l'utilisation variée de termes haineux a posé d'importants obstacles, démontrant la nécessité d'une approche globale pour identifier et modérer les propos haineux en ligne.

Recommendations

1. Les opérateurs de plateformes en ligne doivent accroître la transparence du contenu public sur leur plateforme en fournissant un accès ouvert à l'interface de programmation (API) afin de mieux comprendre l'ampleur des discours haineux. Ils doivent fournir un meilleur accès aux données à des organismes de recherche agréés afin de permettre une meilleure compréhension et capacité de réponse rapide et adaptée aux propos haineux en ligne. La démarche actuelle de Twitter fournit un modèle permettant de trouver un équilibre entre la transparence et l'accès à la recherche d'une part et la protection de la vie privée d'autre part.

2. Les organismes de réglementation gouvernementaux et les plateformes en ligne doivent tenir compte des limites des algorithmes de machine learning pour identifier les contenus haineux. L'intelligence artificielle ne doit pas être considérée comme une panacée. Les approches visant à modérer un contenu haineux ou destiné à susciter la haine doivent inclure un examen réalisé par un être humain. Le contenu des images et des vidéos représente un autre défi pour cette approche.

3. Les plateformes en ligne doivent travailler main dans la main avec les acteurs issus des organisations de la société civile pour lutter contre les contenus à caractère haineux demeurant dans le cadre légal, mais néanmoins problématiques et dangereux. La loi Avia imposera aux opérateurs de plateformes en ligne, dès qu'elle sera votée et mise en œuvre, de supprimer les contenus haineux manifestement illicites. Des partenariats avec des organisations de la société civile doivent d'ores et déjà être créés afin de lutter contre le gros volume de contenus haineux qui ne sont pas considérés comme illicites. Ce travail devrait comprendre le contrôle et l'expérimentation d'approches de contre-discours qui font appel à une gamme de techniques d'engagement direct. Ces organisations de la société civile peuvent aider les plateformes en ligne à prendre des décisions concernant les réponses appropriées aux différents types de contenus haineux.

4. Les plateformes en ligne doivent faire preuve d'une transparence accrue sur leurs politiques et approches de modération, ainsi que sur le rôle des algorithmes et des comptes automatisés dans la diffusion de contenus haineux. La transparence

des processus de modération de contenus et une plus grande supervision de la part d'un régulateur gouvernemental, comme indiqué dans le rapport intérimaire de la mission interministérielle,[6] sont nécessaires pour garantir que la modération soit appropriée, précise et dotée de ressources suffisantes. Cette transparence doit porter sur l'ampleur et la nature des plaintes des utilisateurs concernant les contenus haineux et les mesures prises en réponse à ces plaintes. Il est également essentiel d'accroître la transparence sur le rôle des algorithmes dans la diffusion de contenus susceptibles d'être haineux, en particulier lors d'événements susceptibles de générer une augmentation de l'ampleur des contenus haineux tels qu'observés dans notre analyse.

5. Les plateformes en ligne et les organisations de la société civile doivent collaborer à des campagnes efficaces pour lutter contre l'utilisation généralisée et normalisée des insultes dans la société. Ces efforts doivent s'appuyer sur les meilleures pratiques en matière de compréhension des campagnes de changement de comportement, y compris celles qui ont permis de lutter avec succès contre les insultes normalisées ou le racisme ordinaire. Ils doivent se concentrer sur les communautés où ces insultes sont répandues, y compris parmi les supporters de football et la communauté des gamers ou fans de jeux vidéo en ligne. Les campagnes doivent éviter de paraître « politiquement correctes » pour ne pas être contre-productives.

6. Les opérateurs de plateformes en ligne, le gouvernement et les chercheurs doivent davantage tenir compte de la nature intersectionnelle des discours haineux. Ils doivent analyser les expériences des victimes ou des groupes fréquemment soumis à des contenus haineux en ligne. Les plateformes en ligne doivent utiliser ces analyses pour éclairer les politiques de modération et les mises à jour de produits en mettant davantage l'accent sur la protection. La société civile doit s'efforcer de former des coalitions pour lutter plus efficacement contre les aspects intersectionnels de la haine.

7. Les individus qui travaillent dans ou avec les médias, le gouvernement, les autorités locales, la police et les médias en ligne doivent créer un mécanisme coordonné pour réagir aux événements qui ont tendance à provoquer des pics de discours haineux. Cette coordination peut s'inspirer du protocole d'incident de contenu du Global Internet Forum for Countering Terrorism (GIFCT)[7]et s'appuyer sur l'OCCI pour mobiliser les contre-discours dans le contexte de tels événements.

8. Une plus grande attention doit être accordée à la relation entre les contenus haineux en ligne et les crimes ou incidents haineux hors ligne (attentats, attaques, etc.). Ces efforts doivent inclure d'autres recherches sur ces phénomènes pour déterminer s'il existe une corrélation entre eux. Les statistiques de la police française doivent également inclure une catégorie de crimes de haine en ligne.

Glossaire

Discours capacitistes ou validistes Formes de discours haineux qui portent préjudice ou font preuve de discrimination envers les personnes présentant des incapacités et les barrières comportementales et environnementales qui font obstacle à leur pleine et effective participation à la société sur la base de l'égalité avec les autres, selon la définition de la Convention relative aux droits des personnes handicapées (2006).[8]

Discours anti-LGBTQ ou homophobes Le Conseil de l'Europe[9] décrit les discours anti-LGBTQ comme des expressions favorisant ou justifiant l'homophobie ou la transphobie définies comme suit :

> **Homophobie** « la peur irrationnelle et l'intolérance à l'égard de l'homosexualité, des lesbiennes, des hommes gays, des bisexuels, fondées sur des préjugés ».
>
> **Transphobie** « la peur irrationnelle et l'intolérance à l'égard de la non-conformité sexuelle des personnes transgenres fondée sur des préjugés ».

Extrémisme l'ISD définit l'extrémisme comme « la promotion d'un système de croyances qui préconise la supériorité et la domination d'un « endogroupe » sur tous les « exogroupes », propageant un esprit « d'altérité » déshumanisant et contraire à l'application universelle des droits humains. Les groupes extrémistes prônent, par des moyens explicites ou plus subtils, un changement systémique dans la société qui reflète leur vision du monde ».

Harcèlement Propos ou comportements répétés ayant pour objet ou pour effet une dégradation des conditions de vie se traduisant par une altération de la santé physique ou mentale[10]

Discours de haine Définition du Conseil de l'Europe : « toute forme d'expression qui répand, incite, favorise ou justifie la haine raciale, la xénophobie, l'antisémitisme ou toute forme d'intolérence basée sur la haine ».

Discours haineux Au-delà d'une compréhension strictement juridique, le discours haineux se heurte à une utilisation plus normalisée des injures haineuses, d'une part, et à un discours de haine agressif, violent et potentiellement illégal, d'autre part.

Misogynie Ging et Siapera définissent la misogynie comme la haine ou la peur des femmes, « qui peut ne pas impliquer de violence, mais qui implique presque toujours une certaine forme de préjudice ; soit directement sous forme de préjudice psychologique, professionnel, réputationnel ou, dans certains cas, physique ; soit

indirectement, en ce sens qu'elle rend l'Internet moins égal, moins sûr ou moins accessible aux femmes et filles ».[11]

Réappropriation (de termes) Ce mécanisme, la revendication de termes injurieux par des groupes qui en étaient la cible d'origine, vise à donner plus de pouvoir à ceux qui ont été stigmatisés.

Racisme Définition de la Commission européenne contre le racisme et l'intolérance : « le racisme est la croyance qu'un motif tel que la « race », la couleur de peau, la langue, la religion, la nationalité ou l'origine nationale ou ethnique justifie le mépris envers une personne ou un groupe de personnes ».[12]

Discours de haine sexiste Selon le Conseil de l'Europe[13], le discours de haine sexiste est l'une des expressions du sexisme qui peut être définie comme toute supposition, opinion, affirmation, geste ou comportement visant à exprimer du mépris à l'égard d'une personne en raison de son sexe ou de son genre ou de la considérer comme inférieure ou essentiellement réduite à sa dimension sexuelle. Il englobe des expressions qui propagent, incitent à, promeuvent ou justifient la haine fondée sur le sexe. La véritable ampleur de ce phénomène est en partie masquée par le fait qu'un grand nombre des femmes ciblées ne le dénoncent pas.

Injure Définition de l'Anti-Defamation League : « une injure est une marque injurieuse, blessante ou dégradante, souvent fondée sur un groupe identitaire tel que la race, la religion, l'appartenance ethnique, l'identité de genre ou l'orientation sexuelle ».[14]

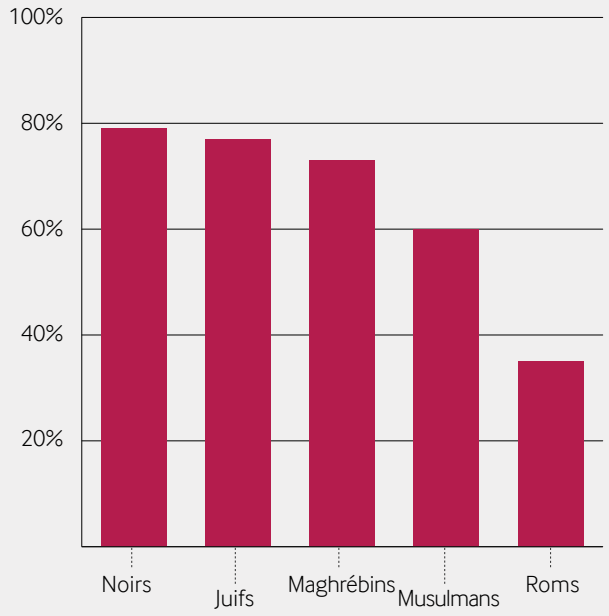
Introduction

La question de la haine en ligne et de ses conséquences dans le monde réel fait fréquemment la une des journaux en France. Deux incidents survenus cette année en sont de parfaits exemples : le 8 février 2019, le quotidien Libération a révélé l'existence de la « Ligue du LOL », un groupe Facebook créé par le journaliste Vincent Glad, dans lequel un certain nombre de journalistes hommes maltraitaient des collègues femmes et coordonnaient des campagnes de harcèlement en ligne contre des femmes journalistes, militantes et politiques. Fin mars, une série d'attaques punitives contre des Roms ont eu lieu en banlieue parisienne. Ces attaques ont été déclenchées par de fausses rumeurs d'enlèvements d'enfants par la population Rom devenues « virales » sur Twitter. Ces évènements ont mis en lumière à la fois la persistance des préjugés contre les Roms et la relation entre la communication en ligne et la haine hors ligne.

Selon un sondage OpinionWay publié en décembre 2018, 59 % des Français ont été victimes d'attaques haineuses sur les réseaux sociaux à un moment ou un autre de leur vie. Les recherches font apparaître une persistance d'attitudes préjudiciables à l'égard des minorités ethniques et religieuses en France. En avril 2019, la Commission Nationale Consultative des Droits de l'Homme (CNCDH) a publié son 28ème rapport sur le racisme, l'antisémitisme et la xénophobie. Ce rapport, qui couvre l'année civile 2018, inclut un indice de tolérance pour cinq groupes minoritaires : les Noirs, les Juifs, les Maghrébins, les Musulmans et les Roms. L'indice de tolérance est basé sur des sondages d'opinion réalisés par Ipsos, où un indice faible indique des niveaux généralement faibles et un indice élevé indique des niveaux généralement élevés de tolérance envers un groupe minoritaire.[15] Comme le montre la Figure 1, les Roms et les Musulmans connaissent les niveaux de tolérance les plus bas de ces cinq groupes minoritaires en France.

Bien que le préjudice ne soit en aucun cas un phénomène nouveau dans la société, les réseaux sociaux offrent une plateforme sur laquelle les discours de haine et les préjugés peuvent être exposés d'une manière visible et mesurable.
Les plateformes de réseaux sociaux offrent également l'occasion à des individus et des groupes

Figure 1
Niveaux de tolérance à l'égard de cinq groupes minoritaires en France, en 2018 [16]



idéologiquement motivés de diffuser des informations erronées et des histoires hautement sensationnalisées afin d'accroître la haine et l'antipathie envers certains secteurs de la société.

On craint de plus en plus que la haine en ligne n'ait des conséquences dans le monde réel. Bien qu'il soit difficile d'établir des liens de causalité entre les attaques haineuses en ligne et les attaques hors ligne, certaines recherches ont établi des corrélations entre elles.
À titre d'exemple, des chercheurs de l'Université de Warwick ont trouvé une association entre le soutien au parti xénophobe « Alternative für Deutschland » et le sentiment anti-réfugiés sur Facebook et les crimes violents contre les réfugiés en Allemagne.[17]

On peut souvent sourcer la violence hors ligne dans des activités en ligne. Une récente série d'attaques perpétrées par l'extrême droite aux États-Unis et en Nouvelle-Zélande a germé sur des plateformes en ligne et a émergé dans un contexte de rhétorique anti-immigrés et anti-Musulmans en ligne. Un rapport d'Amnesty International, qui analyse les messages

sur les réseaux sociaux destinés à des personnalités publiques féminines aux États-Unis et au Royaume-Uni en 2017, a révélé qu'au moins une femme politique britannique ou américaine était victime d'abus en ligne toutes les 30 secondes sur Twitter, quelques mois seulement après l'assassinat de la députée britannique Jo Cox par un extrémiste d'extrême droite.

Les pressions publiques et politiques croissantes en faveur de la lutte contre les discours de haine en ligne ont abouti à l'adoption de nouvelles législations dans plusieurs pays européens.

En juin 2017, le Bundestag allemand a adopté la loi NetzDG (Loi sur les réseaux sociaux).
Cette loi oblige juridiquement les grandes plateformes de réseaux sociaux à supprimer les contenus illégaux dans les 24 heures suivant leur signalement. Entrée en vigueur en octobre 2017, elle permet d'imposer des amendes allant jusqu'à 50 millions d'euros en cas de violations systématiques.

Au Royaume-Uni, le livre blanc du gouvernement sur les préjudices en ligne (Online Harms White Paper) a suscité des préoccupations concernant l'impact des discours de haine en ligne sur les citoyens britanniques et souligné la nécessité d'une plus grande transparence de la part des plateformes de réseaux sociaux sur la modération du contenu.
Ce livre blanc a également proposé une surveillance réglementaire de la façon dont les algorithmes contribuent à la diffusion de contenu nuisible.

De même, le gouvernement français a pris de nouvelles mesures pour réglementer la modération du contenu en ligne notamment avec des mesures portées par la proposition de loi Avia pour lutter contre la haine en ligne adoptée par l'Assemblée le 9 juillet 2019 et examinée par le Sénat en Décembre 2019. La loi doit encore être examinée par le Sénat à l'automne 2019.
À l'instar de son homologue allemande NetzDG, la loi Avia, du nom de sa rapporteure, Laetitia Avia (députée du parti La République en Marche - LREM), vise à réglementer le discours de haine en ligne.
La disposition clé de cette législation, qui n'a finalement pas été adoptée en deuxième lecture par le Sénat, exige que les plateformes suppriment les contenus « manifestement illégaux » (tels que définis par la loi de 2004) dans les 24 heures suivant leur signalement par les utilisateurs *[depuis la rédaction du présent rapport, le projet de loi a été examiné par la commission des lois du Sénat, qui a supprimé l'obligation faites aux plateformes de supprimer les contenus dans un délai de 24 heures]*.
Les opérateurs ne respectant pas cette disposition sont passibles d'amendes pouvant aller jusqu'à 4 % de leur chiffre d'affaires annuel mondial.
Plus généralement, le gouvernement du président Macron a exigé plus de transparence et de collaboration de la part des plateformes en ligne, par exemple avec le renforcement de la coopération judiciaire.[18]

Lacunes en matière de compréhension des discours de haine en ligne : La nécessité de recherches plus approfondies

La Loi Avia a pour objectif de lutter contre la haine en ligne, mais il subsiste de nombreuses lacunes en matière de compréhension de la nature, du volume et de la dynamique de ce phénomène du XXIe siècle.
Le rapport parlementaire de la loi Avia souligne « la prolifération inacceptable de contenus haineux en ligne ».[19] Toutefois, le rapport fait également part des difficultés à parvenir à une compréhension complète de ce phénomène, soulignant par exemple qu'il existe peu d'études rigoureuses sur le sujet.
En conséquence, le rapport qui définit le contexte d'une loi de réglementation en ligne en France s'appuie largement sur des statistiques européennes[20] dans son examen du discours de haine en ligne, ne faisant référence à des statistiques françaises[21] que pour les incidents qui se sont produits hors ligne.

Bien que la loi Avia souligne l'absence d'études systématiques sur la haine en ligne, plusieurs tentatives ont été faites pour combler cette lacune :
En mai 2018, la société de gestion et de modération de contenus Netino a présenté au Secrétaire d'État aux Affaires numériques son Panorama de la haine en ligne. Cette étude analyse des milliers de commentaires Facebook laissés sur les pages des grands médias français. Le rapport montre qu'un commentaire sur dix peut être qualifié de haineux.
Durant toute l'année 2017, le Baromètre mensuel des manifestations de la haine en ligne d'IDPI (Idées, Pratiques, Innovations) a analysé la quantité de discours haineux sur Twitter, en examinant différentes catégories (sexisme, xénophobie, homophobie, antisémitisme et haine musulmane).

La Ligue Internationale Contre le Racisme et l'Antisémitisme (LICRA) surveille régulièrement les discours de haine en ligne.

Toutes ces initiatives s'inscrivent dans la lignée d'études internationales similaires.
En effet, aussi bien en Europe qu'en Amérique du Nord, des organisations à but non lucratif et des instituts de recherche ont tenté d'analyser et de quantifier les discours de haine en ligne. L'organisation Anti-Defamation League (ADL) aux États-Unis, par exemple, a mené une étude d'un an sur les discours de haine antisémites sur Twitter.[22]

Ces tentatives de cartographie des discours de haine en ligne ont apporté des contributions précieuses à la compréhension de cette menace croissante, mais des lacunes subsistent.

Tout d'abord, de nombreux rapports rédigés antérieurement n'incluaient pas de discours misogynes ou capacitistes ou validistes - des catégories souvent exclues des définitions juridiques du discours haineux.[23]
Par ailleurs, d'autres rapports comprenaient des catégories qui ne correspondaient pas aux catégories protégées souvent évoquées dans les définitions du discours de haine, par exemple le « discours agressif », qui ne vise pas une minorité.

Les rapports susmentionnés omettent également d'examiner l'intersectionnalité entre les différentes formes de discours haineux. L'intersectionnalité encourage de fait une analyse de l'identité à plusieurs niveaux (fondée sur la race, le sexe, l'orientation sexuelle, le handicap, etc.), tandis qu'un cadre à cause unique peut limiter la compréhension et l'expérience des victimes de discours haineux.[24]
Une étude récente de la Commission européenne a reconnu l'importance d'une perspective intersectionnelle[25] dans l'examen de la discrimination, reconnaissant néanmoins l'absence de « données ventilées par sexe et par race, et encore moins par d'autres sources de discrimination intersectionnelle, comme l'origine ethnique et le handicap ».[26]

Enfin, les analyses antérieures ont eu tendance à prendre en compte une seule plateforme (généralement Twitter ou Facebook). Ce phénomène est souvent dû à l'accès restreint aux données des plateformes en ligne, ainsi qu'au nombre croissant de plateformes alternatives qui rendent la recherche sur la haine difficile sur le plan méthodologique (étant donné la diversité des sources et le manque de cohérence des données indexées).
Des méthodes plus rigoureuses et plus comparatives sont nécessaires pour quantifier et classifier les discours de haine en ligne sur différentes plateformes, bien que cela soit sérieusement limité par le manque d'accès aux données sur certaines plateformes en ligne.

Notre étude

Ce rapport a pour ambition de compléter la base de recherche décrite plus haut et de fournir une compréhension systématique des discours haineux sur les plateformes en ligne en France qui sont ouvertes à l'analyse quantitative à grande échelle.
Notre objectif est que cette recherche serve de ressource aux organisations de la société civile française qui combattent la haine et l'extrémisme.
Ces organisations sont au cœur de l'OCCI de l'ISD et de Facebook. Elles luttent en première ligne contre les discours haineux et polarisants en ligne et hors ligne. Ce rapport fournira également d'importantes informations aux décideurs politiques pour qu'ils puissent formuler, en toute connaissance de causes, des réponses politiques aux discours de haine et à la polarisation à l'ère du numérique. Ses auteurs espèrent également aider les entreprises technologiques à améliorer les protections contre la haine en ligne pour les communautés ciblées.

Cette étude s'efforce de répondre à deux questions de recherche essentielles :

- Quelle est l'ampleur de la haine dans le discours en ligne en France en considérant un large éventail de groupes de personnes susceptibles d'être victimes de discours haineux, notamment en ce qui concerne l'origine ethnique et raciale, la religion, le sexe, l'orientation sexuelle et le handicap ?
- Quelles sont les caractéristiques des discours haineux en ligne en France, y compris les thèmes clés, les influenceurs et les événements déclencheurs en ligne ou hors ligne ?

53%

des insultes ou commentaires agressifs en ligne sont dirigés contre d'autres utilisateurs*

*Source: Netino

Si notre travail apporte quelques éclaircissements sur l'ampleur, la nature et les cibles des discours haineux en ligne, en plus des recherches existantes, nombre d'obstacles persistent, et entravent une véritable compréhension du discours haineux sur Internet.
Ce manque de compréhension est dû en grande partie au manque d'accès aux données. Twitter reste la plateforme la plus ouverte à la recherche et reste donc la plateforme dominante pour ce type d'analyses. Concernant la recherche et l'analyse des propos haineux sur Facebook et autres plateformes et forums, le manque d'accès API a empêché une analyse complète et exhaustive.

Définition des discours haineux

Le concept de « discours de haine » est incontestablement difficile à définir. En France, le cadre juridique des discours de haine a d'abord été fixé par la loi du 29 juillet 1881 sur la liberté de la presse,[27] complétée par la loi Pleven de 1972,[28] qui condamne « les insultes, la diffamation ou les provocations encourageant la haine et à la violence envers une personne ou un groupe de personnes à raison de leur origine ou de leur appartenance à une ethnie, une nation, une race ou une religion déterminée ».
Cette définition légale a ensuite été modifiée par la loi du 21 juin 2004, qui a ajouté « sexe, orientation sexuelle, identité de genre et handicap ».[29]
La question clé de cette définition d'un point de vue juridique a été la compréhension de termes tels que « insultes »,[30] « diffamation »[31] et « provocations encourageant la discrimination ».
L'interprétation de ces trois termes par les juges a été incohérente[32] et les experts ont souligné les contradictions de la jurisprudence française.[33]
Par exemple, dans une affaire où Michel Houellebecq a été jugé pour avoir déclaré « La religion la plus con, c'est quand même l'Islam »,[34] l'accusé a été déclaré non coupable parce que les tribunaux ont considéré que la formulation était une expression d'opposition à une idéologie et à un système de croyances.
Par ailleurs, l' « humoriste » Dieudonné a été reconnu coupable d'incitation à la haine pour avoir dit : « pour moi les juifs, c'est une secte, une escroquerie »,[35] ce qui a été considéré par les tribunaux comme un préjudice contre une catégorie protégée de personnes en raison de leur origine.

Dans le présent rapport, notre objectif n'a pas été d'identifier les discours de haine illégaux ou de déterminer si certains discours de haine en ligne en France étaient légaux ou illégaux. Notre objectif a plutôt été d'identifier puis d'analyser le large éventail de contextes dans lesquels les « discours haineux » peuvent ou non avoir lieu en ligne : utilisations normalisées d'injures haineuses, d'une part, et discours haineux agressifs, violents et potentiellement illégaux, d'autre part.

Afin de préciser la définition de « discours haineux » utilisée dans notre rapport, nous avons entrepris un examen des recherches et des études existantes dont le but est d'identifier et d'analyser les discours haineux en ligne.

Comme nous l'avons mentionné précédemment, Netino a examiné 15 000 commentaires choisis aléatoirement parmi 15 millions sur les pages publiques Facebook de 25 médias établis afin de cerner les tendances haineuses en ligne.[36] Au premier trimestre de 2019, ils ont identifié quatre principales formes de discours haineux :

- injures ou commentaires agressifs[37] contre

> « En plus des recherches existantes, nombre d'obstacles persistent, et entravent une véritable compréhension du discours haineux sur Internet »

d'autres internautes (comprenant 53,1 % de contenu haineux);
- attaques contre des politiciens (30,1 %);
- attaques contre des personnalités (15.5%);
- attaques contre les médias ou des journalistes (15.1%).

Globalement, 14,3 % des messages sélectionnés aléatoirement se sont révélés agressifs ou haineux.
Le groupe de réflexion IDPI, qui a publié le Baromètre mensuel des manifestations de la haine en ligne cité plus haut, identifie cinq catégories : anti-haine, neutre, racisme ordinaire, propos haineux et « détournement » (Détournement d'un hashtag ou d'un mot pour cibler de manière haineuse une communauté spécifique).

Le Baromètre comprend un échantillon sélectionné aléatoirement de tweets codés manuellement dans ces cinq catégories.[38] Le dernier Baromètre publié, datant de janvier 2018, catégorisait 24 % de l'échantillon comme étant du racisme ordinaire* et 11 % comme explicitement haineux, sur la base d'un échantillon de 2 950 tweets.[39] *défini comme une « Expression traduisant une haine sous-jacente (préjugés, stéréotypes), il correspond à une expression volontairement haineuse ou insultante, tandis que les propos haineux sont une « Expression volontairement haineuse ou insultante ».

Le Baromètre examine en particulier cinq types de haine : homophobie, antisémitisme, sexisme, xénophobie et islamophobie.

Un rapport publié en 2018 par le Conseil Quaker pour les affaires européennes sur le discours de haine contre les migrants s'est concentré sur les aspects les plus délicats des discours de haine.[40] Bien que le rapport ait fondé sa définition des discours de haine sur celle du Conseil de l'Europe, il n'a pas nécessairement tenu compte des « commentaires désagréables, injurieux ou anti-migrants » comme étant des propos haineux, choisissant d'établir les appels à la violence et la déshumanisation comme ses deux critères principaux pour catégoriser le contenu en tant que discours de haine.
À l'autre extrémité du spectre, l'Index de la haine en ligne de l'ADL adopte une définition beaucoup plus large que les définitions juridiques typiques du discours haineux.[41] Elle définit le discours haineux comme suit

« les commentaires contenant des propos dont le but est de terroriser, exprimer des préjugés et du mépris, humilier, dégrader, insulter, menacer, ridiculiser, rabaisser, rabaisser et discriminer en raison de la race, de l'origine ethnique, de la religion, de l'orientation sexuelle, de l'origine nationale, ou du sexe... Y compris également les insultes péjoratives et de groupe, qui comprennent parfois de courts épithètes, généralement négatifs ou de longs narratifs concernant le comportement soi-disant négatif d'un groupe. »[42]

L'approche de l'ADL a également créé plusieurs étiquettes pour catégoriser les messages haineux, notamment l'insulte, le blasphème, la théorie du complot, le sarcasme et la menace.

De même, le groupe de réflexion britannique Demos a créé différentes catégories pour classer la haine misogyne : sérieuse, peu blessante,familière, banale, généralement misogyne, injurieuse et autre (qui comprend des catégories ambiguës comme le contenu subversif et pornographique).[43] Dans une autre étude portant sur le contenu anti-Musulmans sur Twitter, Demos a divisé le contenu haineux en trois catégories : les insultes, les énoncés désobligeants qui lient les Musulmans ou l'islam au terrorisme et les énoncés qui, plus largement, prétendent que les Musulmans détruisent l'Occident sur les plans social et culturel.[44]

Notre approche

Ces différentes approches mettent en évidence la diversité des définitions qui peuvent être adoptées dans l'étude des propos haineux en ligne.
Pour ce rapport, nous avons choisi la définition du Conseil de l'Europe des discours de haine comme « couvrant toutes les formes d'expression qui propagent, incitent, encouragent ou justifient la haine raciale, la xénophobie, l'antisémitisme ou toute autre forme de haine fondée sur l'intolérance ». Toutefois, le classement du contenu en fonction de cette définition a posé certaines difficultés car la frontière entre injures normalisées et propos haineux s'est avérée souvent ardue à définir.

Sachant que les types de discours analysés dans ce rapport n'étaient pas simplement les discours de haine illégaux, mais aussi les discours haineux au sens plus large, cette définition a été interprétée pour inclure les injures normalisées, étant donné leur contribution aux discours de division à la polarisation sociale (au même titre que le « racisme ordinaire » tel que défini par l'IDPI et « familier », « banal » et « généralement misogyne » de la recherche de Demos).
Par conséquent, les utilisations d'injures qui ont été normalisées, mais qui sont fondées sur des attaques de catégories protégées (p. ex., *pute* et *pédé*), ont été incluses dans des sous-ensembles haineux. Bien que les personnes qui utilisent ces termes n'aient pas nécessairement l'intention de propager, d'inciter, de promouvoir ou de justifier la haine, ces termes eux-mêmes sont fondés sur des préjugés qui ont pour effet de normaliser ces types de haine. Dans la mesure du possible, le présent rapport établit une distinction entre ce type de discours haineux normalisé et un discours haineux plus ciblé, caractérisé.

Enfin, tout en précisant l'approche que nous avons adoptée dans ce rapport, il existe une différence importante entre notre approche et celles des études précédentes.
Ces dernières ont permis de développer des catégories plus sophistiquées car elles s'appuyaient en grande partie sur l'analyse qualitative de petits échantillons de contenus de réseaux sociaux. L'échantillon le plus grand parmi les études citées ci-dessus était de 15 000 messages de Twitter dans l'étude Netino. Au contraire, notre objectif a été d'utiliser des algorithmes de traitement du langage naturel et de machine learning pour analyser des ensembles de données beaucoup plus vastes sur le contenu en ligne afin de brosser un tableau plus complet, plus exhaustif du discours haineux en ligne.

Bien que notre analyse ait permis de mieux comprendre les différents types de discours « haineux » identifiés, les étapes suivantes de cette recherche devraient établir dans quelle mesure le traitement du langage naturel et le machine learning peuvent être développés pour analyser les ensembles de données « haineuses » en catégories plus détaillées.A terme, l'objectif serait de construire un système en mesure d'identifier avec certitude les propos haineux, de les analyser en temps réel et de fournir des informations directement aux ONG qui œuvrent pour soutenir et protéger les victimes de discours haineux en ligne en produisant des contre-discours.

> “Les types de discours analysés dans ce rapport n'étaient pas simplement les discours de haine illégaux”

Méthodologie

Comme déjà mentionné, les questions de recherche auxquelles ce rapport a cherché à répondre étaient les suivantes :

- Quelle est l'ampleur de la haine dans le discours en ligne en France en considérant un large éventail de groupes de personnes susceptibles d'être victimes de discours haineux, notamment en ce qui concerne l'origine ethnique et raciale, la religion, le sexe, l'orientation sexuelle et le handicap ?
- Quelles sont les caractéristiques des discours haineux en ligne en France, y compris les thèmes clés, les influenceurs et les événements déclencheurs en ligne ou hors ligne ?

Nous avons de plus cherché à comprendre dans quelle mesure le traitement du langage naturel et le machine learning peuvent être utilisés pour identifier les discours haineux avec certitude, dans les délais et la portée appropriés.

Afin de répondre aux questions de notre recherche, nous avons adopté une approche mixte, combinant l'analyse qualitative de grands ensembles de données avec des techniques de traitement du langage naturel pour identifier et analyser les propos haineux en ligne.

Collecte des données

Nous avons tout d'abord tenté de dresser une liste exhaustive de mots-clés pour saisir les propos haineux classés selon le groupe ou la minorité qu'ils ciblent. La sélection des différentes catégories de discours de haine s'appuie sur les définitions du Conseil de l'Europe, de la Délégation interministérielle à la Lutte contre le Racisme, de l'Antisémitisme et de la Haine anti-LGBTQ (DILCRAH), ainsi que sur les catégories protégées par les termes de service de Facebook et Twitter. La consultation des membres du comité consultatif de l'OCCI en France ainsi que notre souhait d'inclure des catégories souvent sous-représentées dans les études sur les discours de haine en ligne ont documenté certaines des catégories de notre rapport.

Les types de discours et de groupes haineux que nous avons identifiés comprenaient

- discours capacitistes ou validistes
- discours anti-Arabes ou anti-Maghrébins
- discours anti-Asiatiques
- discours anti-Noirs ou anti-Africains
- discours anti-Chrétiens[45]
- discours anti-LGBTQ
- discours misogynes
- discours anti-Musulmans
- discours anti-Roms ou anti-Gitans
- discours anti-Blancs
- discours antisémites[46]

Afin d'identifier des exemples potentiels de ce genre de discours, nous avons dressé des listes de mots-clés pertinents qui sont souvent utilisés de façon haineuse ou en conjonction avec des propos haineux. Ces listes sont basées sur des recherches antérieures de l'ISD ainsi que sur des consultations avec un certain nombre d'organisations de lutte contre la haine œuvrant en France, notamment LICRA, Le Refuge, l'AJC, son initiative #JeSuisLà et La Voix des Roms. La liste complète des mots-clés figure en Annexe 1.

Nous avons interrogé deux outils commerciaux d'écoute des réseaux sociaux en utilisant ces mots-clés pour créer nos premiers ensembles de données. Les messages comprenant au moins un des mots-clés de notre liste ont été ensuite inclus dans nos premiers ensembles de données.
Les outils d'écoute sociale qui ont été utilisés sont :

- **Crimson Hexagon**, outil d'écoute des réseaux sociaux qui rassemble des données provenant d'un certain nombre de sources publiques comme Twitter, la section des commentaires de YouTube, Reddit et autres forums et blogs
- **CrowdTangle**, outil qui rassemble les données de pages et groupes publics Facebook (seulement les messages créés par ces pages et groupes, aucune donnée au niveau de l'utilisateur).

La répartition des sources pour tous les ensembles de données est présentée Tableau 1.
En raison des restrictions d'accès aux données qui varient d'une plateforme à l'autre, l'échantillon est fortement biaisé vers Twitter, dont l'accès est le plus ouvert.
D'autres plateformes limitent l'accès ou sont difficiles à indexer, d'où leur sous-représentation dans le présent rapport. Bien que nous ayons essayé d'inclure autant de plateformes que possible, la restriction des données

constitue toujours un obstacle à ce type de recherche. Voir les sections « Limites » ci-dessous et « Recommandations » ultérieurement dans ce rapport pour plus de précisions sur la question.

Tableau 1 Répartition des sources de données par plateforme

Platform	%
Blogs	1%
Facebook	1%
Forums	4%
Reddit	0%
Tumblr	1%
Twitter	89%
YouTube	4%
Other	1%

Tableau 2 Premiers ensembles de données identifiés à l'aide d'au moins un mot-clé haineux avant l'application des classificateurs de pertinence et de haine

Discours	Nombre de messages
Discours misogynes	7,948,332
Discours anti-Arabes /anti-Maghrébins	1,752,405
Discours anti-LGBTQ	1,695,268
Discours capacitistes ou validistes	565,662
Discours anti-Roms ou anti-Gitans	299,157
Discours anti-Noirs ou anti-Africains	223,483
Discours anti-Musulmans	168,324
Discours anti-Chrétiens	156,047
Discours anti-Blancs	125,654
Discours antisémites	79,289
Discours anti-Asiatiques	45,804

Des outils informatiques commerciaux tels que Crimson Hexagon et CrowdTangle offrent des facultés d'analyse intégrées, y compris des indicateurs de volume dans le temps, ainsi que d'autres termes et contenus fréquemment associés aux mots-clés.
Ils recensent également les utilisateurs qui utilisent les mots-clés le plus souvent. Cependant, ils ne permettent pas de tester ces ensembles de données pour vérifier la pertinence de leurs contenus aux discours haineux. Dans le cas d'un thème aussi délicat que le discours haineux, il est essentiel que les chercheurs puissent tester davantage les données pour vérifier leur pertinence et leur exactitude.
Etant donné la très grande échelle des premiers ensembles de donnés aux discours contenant des mots-clés qui pourraient être misogynes, anti-Arabes ou anti-LGBTQ, ce processus de testing est primordial afin de vérifier que tout le contenu identifié par mots-clés est réellement un discours haineux.

Pour résoudre certaines de ces difficultés, l'ISD s'est associé à CASM pour tirer parti de son outil exclusif de traitement du langage naturel, Method52.
Cet outil nous a permis de créer des algorithmes de machine learning pour reconnaître des types spécifiques de discours ou organiser de très grands ensembles de données pour mettre en évidence des modèles.

Tout d'abord, afin de ne pas exclure de mots-clés importants, nous avons analysé tous nos ensembles de données à l'aide d'un « détecteur d'expressions surprenantes » dans Method52. Nous avons ainsi comparé les ensembles de données recueillies sur les activités haineuses à un corpus de référence de texte général en français, afin d'identifier les mots ou les expressions qui apparaissaient plus fréquemment dans notre échantillon que dans la discussion moyenne en ligne. Ces expressions surprenantes ont été évaluées manuellement pour déterminer qu'aucun ajout à nos listes de mots-clés n'était nécessaire.

En raison de contraintes de temps et de ressources, nous avons sélectionné les quatre ensembles de données les plus importants (discours misogynes, discours anti-Arabes ou anti-Maghrébins, discours anti-LGBTQ et discours capacitistes ou validistes) pour une analyse quantitative plus poussée en utilisant l'outil Method52 afin de confirmer la pertinence du contenu

> Nous avons analysé en détails les quatre plus grands ensembles de données en suivant un processus à quatre étapes

des ensembles de données et d'identifier et analyser spécifiquement un contenu « haineux ».
Étant donné qu'il s'agit des ensembles de données les plus importants, nous avons jugé que les méthodes d'analyse automatisées étaient plus révélatrices que les méthodes manuelles ou par échantillon. Les prochaines itérations de cette recherche devraient se concentrer sur une analyse approfondie similaire pour les autres ensembles de données

Analyse des quatre plus grands ensembles de données

Nous avons analysé en détails les quatre plus grands ensembles de données en suivant un processus à quatre étapes.

Étape 1

Nous avons formé un algorithme de traitement du langage naturel pour séparer les messages pertinents des messages non pertinents dans les quatre ensembles de données. Les messages non pertinents étaient définis comme ceux dans lesquels l'utilisation des mots-clés ne se référait pas à la catégorie protégée concernée. Chacun de ces algorithmes a été formé pour être précis à 85 % au minimum dans l'identification de la pertinence.
NB : Pour en savoir plus sur la formation des algorithmes, voir Annexe 2. Le nombre de messages pertinents des quatre plus grands ensembles de données (provenant de l'algorithme de traitement du langage naturel) est présenté ci-après :

Tableau 3 Nombre de messages utilisés dans l'analyse à la suite du classificateur de pertinence

Ensemble de données	Messages pertinents	%
Anti-femmes ou misogynes	6,669,862	84%
Anti-femmes ou misogynes	1,003,668	57%
Anti-LGBTQ ou homophobes	1,705,196	99%
Capacitistes, validistes ou anti-handicap	344,078	61%

Ces résultats montrent que les listes de mots-clés ont donné des niveaux variables de messages pertinents. Par exemple, les listes de mots-clés anti-LGBTQ et misogynes ont donné une quasi-totalité de messages pertinents.
En revanche, dans l'ensemble de données capacitistes ou validistes, des termes comme *mongolien* sont fréquemment apparus dans les messages relatifs à la Mongolie, qui ont été classés comme non pertinents. Dans l'ensemble de données anti-Arabes, la majorité des messages non pertinents étaient liés à la cuisine, étant donné l'utilisation du terme *beur* comme mot-clé (avec des messages incluant le mot beurre). Dans la section LGBTQ, le mot-clé *pédale* a donné du contenu non pertinent. Dans la section sur la misogynie, les exemples de messages non pertinents comprenaient l'utilisation de termes dérivés d'injures misogynes mais qui ne font pas référence à des personnes, comme le terme saloperie (saleté ou ordure).

Étape 2

Nous avons formé un deuxième algorithme pour identifier les propos haineux dans les données restantes. Pour déterminer ce qui constitue un discours haineux, nous nous sommes référés à la définition du discours de haine du Conseil de l'Europe.
Comme mentionné précédemment, cette définition est interprétée pour inclure des injures que l'Anti-Defamation League définit comme, « une remarque injurieuse, blessante ou dégradante, souvent fondée sur un groupe identitaire tel que la race, l'origine ethnique, la religion, l'appartenance ethnique, l'identité de genre ou l'orientation sexuelle ». Tous ces algorithmes ont été formés pour être précis à 85 % au minimum[47] pour identifier les messages haineux (Tableau 4).

Tableau 4 Nombre de messages provenant des ensembles de données pertinents qui étaient identifiés comme étant des « discours haineux »

Ensemble de données	Messages haineux	%*
Anti-femmes ou misogynes	5,453,603	82%
Anti-Arabes ou anti-Maghrébins	131,731	13%
Anti-LGBTQ ou homophobes	1,007,034	59%
Capacitistes, validistes ou anti-handicap	265,126	77%

*Pourcentage d'ensemble de données « pertinents »

Étape 3

Nous avons ensuite formé un troisième algorithme pour fournir une plus grande précision à chacun des sous-ensembles de messages haineux pour être à même de séparer les propos haineux en fonction des catégories suivantes :

- **injures haineuses généralisées**, messages dans lesquels des injures sont utilisées pour rabaisser ou dévaloriser une personne ou un groupe de personnes (y compris l'utilisation normalisée de termes comme *pute* et *pédé*)
- **propos haineux ciblés**, messages dans lesquels un utilisateur qui semblait appartenir à une catégorie protégée a été directement attaqué sur la base de cette catégorie dans un message personnalisé ou dans lesquels un langage haineux a été employé pour attaquer une personne ou un groupe de personnes sur la base de son appartenance (perçue) à cette catégorie
- **autres discours haineux**, y compris des propos haineux dans des contextes vagues (p. ex. l'emploi d'un langage haineux dans des comptes pornographiques et des citations de propos haineux)
- **contre-discours**, y compris les discours qui s'inscrivaient à la fois dans le cadre de campagnes coordonnées et une riposte organique contre l'emploi de termes de haine ou haineux.

Étape 4

Nous avons effectué une analyse fondée sur les communautés des quatre plus grands ensembles de données afin d'identifier des réseaux distincts d'utilisateurs, de pages ou de groupes qui emploient des propos haineux (selon la classification de l'Étape 2). L'analyse algorithmique a procédé à un groupement pour identifier les communautés qui emploient des discours haineux précis, fondés sur les termes les plus fréquemment utilisés dans les ensembles de données. Ce processus a permis d'obtenir une analyse plus précise des types de discours haineux présents en ligne et des différents centres de discussion. Les résultats de cette analyse sont présentés sous forme de graphiques en réseau de termes-clés dans les sections ci-dessous.

Analyse de comptes qui sont apparus dans plus d'un ensemble de données

L'outil Method52 a également été mis à contribution pour recenser les comptes qui figuraient dans plus d'un de nos ensembles de données, à savoir des comptes qui utilisaient plusieurs types de propos haineux.
Ce recensement nous a permis d'examiner les cas où plusieurs types de discours haineux ont été utilisés par les mêmes comptes et où différents types de discours haineux se recoupent.

Pour en savoir plus sur Method52 et le processus de classification, voir l'Annexe 2.

Analyse de plus petits ensembles de données

Les plus petits ensembles de données ont été analysés à l'aide de Crimson Hexagon et CrowdTangle. Bien que ces outils présentent des limites, ils nous ont néanmoins permis de faire quelques observations intéressantes.[48]
Trois chercheurs de l'ISD ont analysé des échantillons de données grâce aux différentes fonctionnalités offertes par ces outils. Pour chaque catégorie de discours, les chercheurs de l'ISD ont identifié des pics d'activité et examiné le contenu et les mots-clés les plus souvent mentionnés au cours de la période pertinente afin de déterminer quel événement avait entraîné une intensification de la conversation.
L'utilisation des fonctionnalités pour identifier les éléments de contenu les plus partagés nous a permis d'identifier les thèmes et les tendances clés de l'échantillon de données.
Il nous a enfin été possible, à l'aide de ces outils, de recenser les comptes les plus actifs qui mentionnent les mots-clés de nos listes. Ce recensement nous a permis d'évaluer dans quelle mesure les mots-clés étaient principalement employés dans un contexte haineux par des utilisateurs actifs.

Limites

Bien que notre rapport vise à être aussi complet que possible, ce type d'étude comporte un certain nombre de limites.

La première est, encore et toujours, la question de l'accès aux données. Nos ensembles de données proviennent de données accessibles au public provenant des principales plateformes de réseaux sociaux (Facebook, Twitter, YouTube) et d'autres sources indexées par les outils d'écoute des réseaux sociaux avec lesquels nous travaillons.
Cependant, il existe de nombreux canaux privés, fermés et cryptés auxquels nous n'avons pas eu un accès systématique ou informatique. Par conséquent, le présent rapport, ses conclusions et ses recommandations ont été limitées aux données accessibles au public et uniformément indexées. Il ne traite donc pas de l'écosystème plus large des canaux privés et des petits sites qui peuvent aussi héberger des contenus haineux.

La seconde tient à notre choix d'une approche fondée sur des mots-clés pour recueillir les données. Étant donné la nature dynamique du langage et du langage en ligne en particulier, il est possible que nos ensembles de données aient manqué des propos haineux qui ne contiennent pas les termes de nos listes de mots-clés. Inversement, comme le démontre l'analyse de la pertinence montrée plus haut, les approches fondées sur des mots-clés peuvent parfois tirer des données non pertinentes. De plus, le discours haineux ciblant des groupes spécifiques peut varier considérablement, non seulement sur le plan de la terminologie, mais aussi sur le plan du nombre de termes et/ou d'injures qui existent dans la société pour désigner un groupe particulier. Ce phénomène peut entraîner une iniquité dans le nombre de mots-clés employés pour recueillir des données sur chaque type de discours, ce qui peut influencer le nombre de résultats donnés et la profondeur des données.
Bien que nous ayons pris de nombreuses mesures pour minimiser l'impact de ces limites (en consultant des collègues d'autres organisations expertes et en employant une analyse de détection d'expressions surprenantes), certaines omissions ont pu persister.

La troisième concerne les outils d'écoute des réseaux sociaux observés qui retirent de leurs bases de données les messages qui ont été supprimés des plateformes en ligne pour violation des standards de la communauté, des termes de service ou des lois locales.
Par conséquent, nos ensembles de données n'incluent probablement pas certains des messages les plus odieux qui ont pu être supprimés par les plateformes avant ou pendant la collecte des données. Cela peut expliquer, par exemple, la quantité relativement faible de contenus antisémites dans nos ensembles de données, malgré un pic des crimes haineux hors ligne en France, selon les statistiques officielles.[49]

> « Nos algorithmes ont été formés pour être précis à 85 % minimum pour identifier les messages haineux »

Enfin, nos algorithmes ont été formés pour être précis à 85 % minimum[50] pour identifier les messages haineux. Le traitement du langage naturel et le machine learning n'étant pas des sciences exactes (les humains non plus), il y a toujours un certain degré d'erreur dans les décisions prises par les algorithmes. Une précision globale de 85 % est globalement conforme aux performances équivalentes présentées dans les travaux récents évalués par nos pairs dans le même domaine.[51]
Cependant, comme pour toute expérience avec du machine learning, il est impossible d'être précis à 100 % dans la reconnaissance des nuances multiples qui existent dans le langage naturel, en particulier dans un domaine aussi lourd et spécifique au contexte que le discours haineux.[52] En effet, même l'examen humain n'est pas précis à 100 %. Il s'agit là d'une autre limite et d'un autre défi pour cette étude et pour la modération automatisée du contenu en général, comme nous le verrons plus loin. Néanmoins, les scores de confiance produits par Method52 confèrent à ce rapport une plus grande transparence et une plus grande précision que d'autres algorithmes de machine learning plus opaques (voir Annexe 2).

Analyse des Discours

Ce chapitre présente les analyses effectuées sur les 11 différents discours et ensembles de données décrits plus haut. Ces derniers ensembles sont divisés en quatre catégories principales :

- sexe et orientation sexuelle,
- origine ethnique et raciale,
- religion et
- handicap.

L'analyse de chacun de ces discours comprend les principaux résultats de notre analyse, notamment :

- l'ampleur des contenus ou des discours « haineux » comprenant une « terminologie haineuse »,
- les événements ou déclencheurs de pics dans la conversation,
- les thèmes clés et
- l'analyse des dix comptes les plus actif.

Certaines catégories de discours haineux ont donné des résultats étonnamment faibles, ce qui pourrait être le résultat d'un certain nombre de facteurs.
En effet, ces résultats peuvent en partie s'expliquer par la modération du contenu réalisée en amont par les plateformes de réseaux sociaux, par les limites ou biais dans les mots-clés employés et par l'accès limité aux données sur les différentes plateformes, tel que souligné dans la section méthodologie.

Sexe et orientation sexuelle
Ce chapitre s'intéresse aux discours misogynes et anti-LGBTQ.
Ces deux ensembles de données se caractérisent par leur volume élevé et l'usage normalisé d'injures misogynes et homophobes dans divers contextes, en particulier l'emploi des termes comme *pute* et *pédé*. Les propos haineux de ces deux catégories se recoupent aussi beaucoup, ce qui indique une confusion générale entre le sexe et l'orientation sexuelle.

Discours misogynes

Principaux résultats

- Sur les 6,6 millions de messages pertinents identifiés, notre algorithme a placé **5,5 millions de messages dans la catégorie des discours misogynes haineux** (Figure 2).

- Ces messages incluent notamment **4 millions de messages contenant des insultes misogynes généralisées et environ 300 000 messages étant des attaques haineuses ciblées** contre des utilisateurs manifestement de sexe féminin. Plus d'un million de messages dans la catégorie « autres » de discours haineux concernaient principalement des comptes qui partageaient du contenu pornographique.

- Un algorithme formé pour reconnaître le contre-discours a permis d'identifier **100 000 messages (seulement 1 % des messages pertinents) contenant à la fois des contre-discours organiques et coordonnés**.

- Les discours misogynes haineux étaient caractérisés par **leur volume constant dans la durée** par rapport aux autres discours analysés dans ce présent rapport. **Le volume des discours misogynes haineux dépassait presque 200 000 posts, chaque semaine**.

- **La Journée internationale de la Femme** en mars a correspondu à la plus forte hausse de discours haineux généralisés. Les deux pics les plus faibles de discours haineux ciblés ont été causés par des attaques misogynes sur des individus, supprimées par les plateformes.

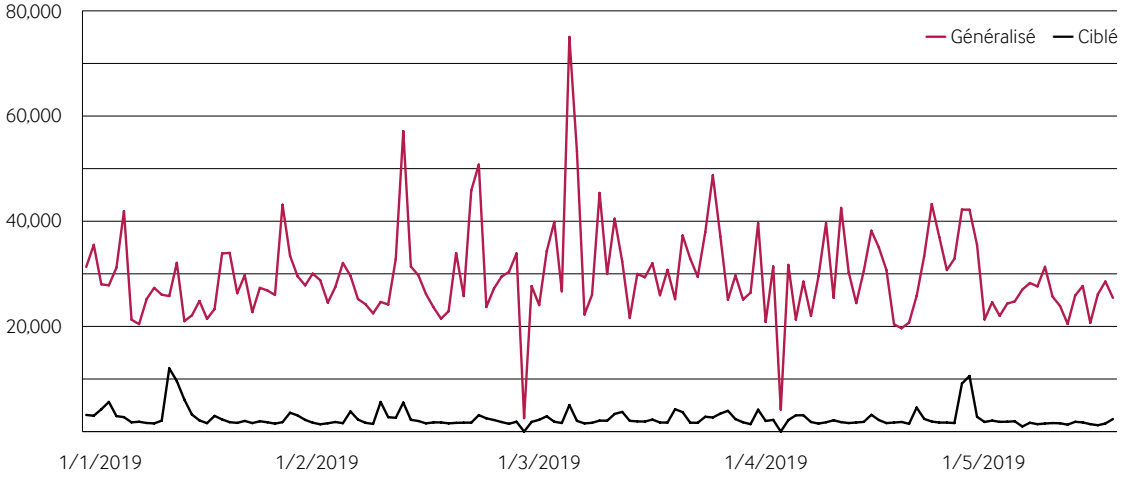
- **Des injures misogynes généralisées** ont été souvent utilisées pour exprimer des opinions antigouvernementales, par exemple attaquer le Président Emmanuel Macron en employant des termes comme fils de pute. Les attaques contre les femmes politiques et les femmes d'influence représentent une tendance clé. C'est notamment le cas de Marlène Schiappa, secrétaire d'Etat française à l'Égalité, qui a fait l'objet de nombreuses agressions en ligne.

- **Les comptes et les pages les plus actifs** en matière de discours misogynes haineux comprennent cinq comptes bot ou semi-automatisés avec des intérêts généraux ou ponctuels, comme le football, trois comptes en faveur des Gilets Jaunes et quelques utilisateurs individuels.

- **Les comptes les plus fréquemment mentionnés** dans les discours misogynes haineux appartiennent à des personnalités politiques et sportives, dont beaucoup sont des hommes, même si certains sont ceux de particuliers susceptibles d'avoir été victimes de harcèlement.

- Le contenu misogyne affiche des recoupements avec d'autres types de contenus haineux, notamment le contenu anti-Arabes et le contenu anti-LGBTQ.

Figure 2
Volume de messages contenant des discours misogynes haineux entre le 1er janvier 2019 et le 31 mai 2019



Discours misogynes
Exemples de discours haineux

[redacted]

Journée de la femme ?humhum je dirais plutôt journée de la pute.

Figure 3 Contenu haineux posté à l'occasion de la Journée internationale de la femme (source : Twitter)

[redacted]

Putain mais faites lui fermer sa gueule à cette grosse chienne twitter.com/NathannAbraham...

Figure 4 Utilisation généralisée d'injures misogynes (source : Twitter)

[redacted]

A partir du moment ou Macron nous traite de "foule haineuse", est-ce qu'on a le droit de lui dire que c'est un enculé de fils de pute de bourgeois au service des milliardaires??
Non?
Ben on le dit tout de même....

Figure 5 Langage misogyne contre Emmanuel Macron (source : Twitter)

[redacted]

Eh sale pute on te parle @AudreyPulvar

Figure 6 Injures misogynes ciblant la journaliste Audrey Pulvar (source : Twitter)

MarleneSchiappa
@MarleneSchiappa

«Je salue la démarche du @CreditMutuel qui lance une enquête interne sur la #cagnotte auprès de sa filliale #leetchi Cette cagnotte est contraire aux valeurs de la République. Elle est indécente. »

[redacted]

On donne de l'argent à qui on veut grosse =pute, par contre quand tu vas recevoir une balle dans la tête ça serra contraire aux valeurs de la république ou pas ?????

Figure 7 Un message ciblant Marlène Schiappa, qui inclut des menaces violentes (source : Twitter)

[redacted]

@AudreyPulvar sale vieille poufiasse hypocrite de merde , t'ose ouvrir ta gueule pour le steak à Ribéry mais faut rien dire sur le fait que tes lunette de salope font 10 fois le prix de son steak , va faire des dons aussi au lieu d'ouvrir ta vieille chatte sombre pute

Figure 8 Injures misogynes ciblant Audrey Pulvar après qu'elle a demandé au footballeur Frank Ribéry de donner son argent à des organisations caritatives (source : Twitter)

Discours anti-LGBTQ
Principaux résultats

- Sur le 1,7 million de messages pertinents identifiés, notre algorithme a placé **1 million de messages dans la catégorie des discours anti-LGBTQ haineux** (Figure 9).

- **Presque tous les propos haineux anti-LGBTQ sont utilisés de façon généralisée.** Il n'a pas été possible de former un algorithme pour identifier avec certitude les attaques haineuses ciblées en raison du volume relativement faible de l'échantillon.

- **Les discours homophobes et transphobes haineux sont devenus « normaux » dans de nombreux contextes.** La forme abrégée de pédé (pd), elle-même une forme abrégée de pédéraste (se référant à l'homosexualité et dérivée de la même racine que la pédérastie), se trouve associée à la plus forte proportion de discours haineux.

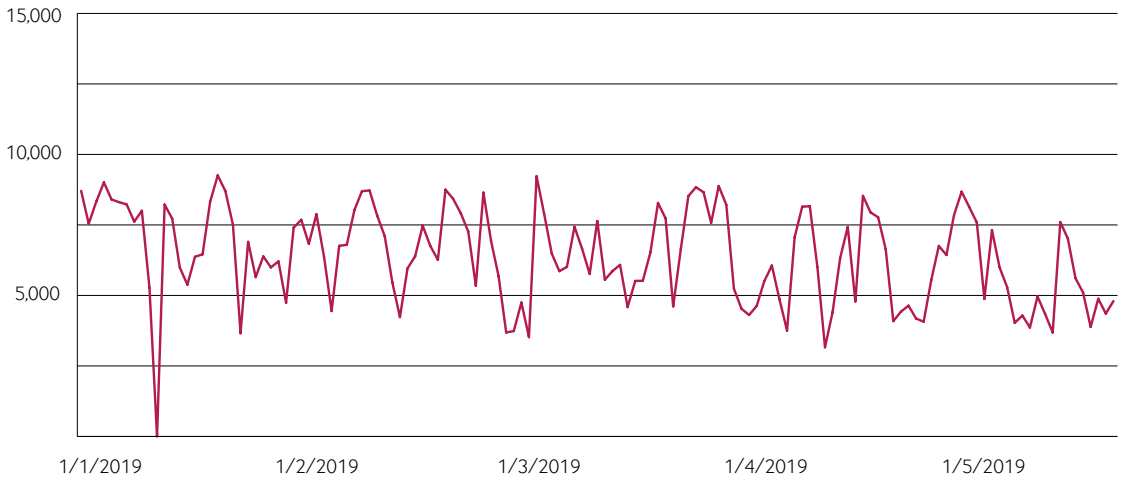
- Nous avons identifié un **petit sous-ensemble de comptes qui se réapproprient des injures homophobes,** ce qui peut être considéré comme une forme de contre-discours organique.

- **Bilal Hassani, le Français qui a représenté la France à l'Eurovision**, a été l'une des principales cibles des discours haineux contre les LGBTQ au cours de cette période.

- **Les comptes et les pages les plus actifs** comprenaient six noms affichant des comportements de bot. Quatre de ces comptes avaient le terme « bot » dans leur nom d'utilisateur et affichaient des volumes élevés de contenus faisant référence à des jeux vidéo et utilisant des injures anti-LGBTQ dans leurs messages. Les quatre suivants étaient des comptes d'utilisateurs individuels, dont deux avaient été interrompus. Les pages les plus actives sur Facebook semblaient toutes être soit des comptes officiels de plateformes médiatiques ou de journaux, soit des pages d'intérêt général qui ne semblaient pas utiliser de mots-clés anti-LGBTQ de façon haineuse. Ces constatations soulignent la difficulté d'identifier les pages haineuses à l'aide d'une approche basée sur des mots-clés dans les logiciels d'analyse disponibles sur le marché.

- **Les personnalités politiques françaises** ont été les cibles les plus fréquentes des discours haineux contre les LGBTQ.

Figure 9
Volume de messages contenant des discours haineux anti-LGBTQ entre le 1er janvier 2019 et le 31 mai 2019



Discours anti-LGBTQ
Exemples de discours haineux

Je vois beaucoup de personnes qui le critiquent en disant qu'il est à l'Eurovision car il est homosexuel et pas car il est talentueux... Ben vous avez totalement raison, il casse les couilles cet enculé twitter.com/AkumaBoucher/s...

Figure 10 **Un tweet haineux ciblant Bilal Hassani (source : Twitter)**

PREZYDENT VESKAYE
7 months ago · 264,573 likes

SVP dites moi, aujourd'hui journée mondiale de la femme, est ce que les PD sont concernés?

Figure 11 **Message démontrant la « normalisation » des injures anti-LGBTQ dans les conversations en ligne ainsi que l'intersectionnalité des discours anti-LGBTQ et misogynes (source : Facebook)**

Apres les drag-queens sous extasies du 14 juillet a l'Elysee , un travelo represente la France a l'eurovision ! Le travail d'humiliation du peuple suit sa feuille de route !

Figure 12 **Propos haineux ciblant Bilal Hassani (source : Twitter)**

#BilalHassani nargue ouvertement les Français. Sa robe de princesse et sa chanson débile sont une insulte à la France et à notre culture. vielle de 2000 ans. Il est le chansonnier des #LGBT et ... le chouchou de notre Ministre de la Culture !? Quelle honte !!! #Eurovision 2019

Figure 13 **Un tweet ciblant Bilal Hassani (source : Twitter) Eurovision2019**

Discours anti-LGBTQ
Exemples de contre-discours

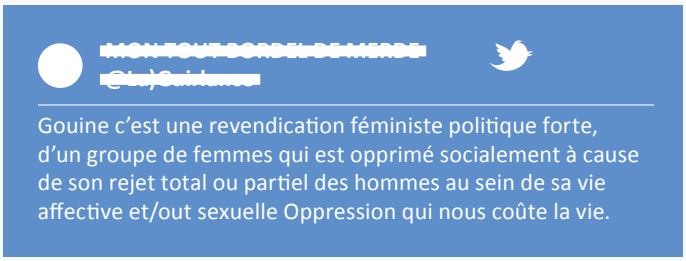


Figure 14 **Tweet se réappropriant le terme *gouine* (source : Twitter)**

Figure 15 **Tweet condamnant un premier message sur Twitter affirmant que « être gay est à la mode » et partageant une photo de l'émission Touche pas à mon poste ! présentée par Cyril Hanouna, connu pour ses déclarations controversées sur la communauté LGBTQ**

Figure 16 **Message de contre-discours de SOS Homophobie, tweeté suite à une attaque transphobe à Paris le 2 avril 2019 (source : Twitter)**

Étude de Cas

[redacted]

Depuis 2014, en 5 évènements, la seule personne ayant gagné pour sa musique date de 2015. LE HASARD On l'aime pas, non pas parce qu'il est gay, mais parce qu'il se comporte comme une pute Oublie pas de me bloquer car tu n'as pas d'arguments

[redacted]

mdr "queen Hassani" quel fiotte

Figure 17 Un tweet injurieux ciblant Bilal Hassani (source : Twitter)

[redacted]

Le pd d'hassani est une humiliation pour notre peuple

[redacted]

Une pensée pour tout les fdp de lgbt ,la maladie ne triomphera jamais #Eurovision

Figure 18 Un tweet injurieux ciblant Bilal Hassani (source : Twitter)

Attaques dirigées contre le chanteur de l'Eurovision Bilal Hassani

Les agressions subies par Bilal Hassani, chanteur à l'Eurovision 2019, sont apparues comme un thème clé dans notre analyse, avec une quantité considérable de propos haineux à son encontre.
Deux petits pics de discussions haineuses apparus au cours du premier semestre de l'année sont liés à lui : l'annonce de sa nomination au concours en janvier et l'annonce de son nouvel album en mars. La vidéo de la chanson de la compétition, Roi, mettait l'accent sur l'acceptation et la démarginalisation des personnes homosexuelles et aux identités de genre variées, ce qui a provoqué de nombreuses insultes.

Le compte Twitter d'Hassani figurait parmi les dix comptes les plus cités dans les messages, y compris notre liste d'injures anti-LGBTQ. Les tweets présentés ci-dessous montrent les types de harcèlement que le chanteur a subi, y compris des injures comme *enculé* (dérivé de l'argot pour sodomiser), *travelo* (argot pour travesti) et *fiotte* (terme injurieux pour homosexuel). Son nom était également présent de manière significative dans l'ensemble de données haineuses anti-LGBTQ, comme on peut le voir dans les Figures 10, 12, 13, 17 et 18.

Origine ethnique et raciale

Ce chapitre comprend une analyse des discours anti-Arabes, anti-Roms, anti-Noirs, anti-Blancs et anti-Asiatiques. Nous avons utilisé le traitement du langage naturel pour analyser l'ensemble de données anti-Arabes en raison de la taille de l'ensemble initial de données généré par mots-clés. Cet ensemble de données montre un recoupement important avec le discours anti-Musulmans, comme nous le verrons dans les sections suivantes.
Les ensembles de données anti-Roms et anti-Noirs comprenaient une partie importante de comptes de lutte contre les stéréotypes et les propos haineux.
L'ensemble de données anti-Blancs a été dominé par le débat sur l'existence du racisme anti-Blancs.
Enfin, les mots-clés anti-Asiatiques ont donné lieu à un débat sur le racisme anti-Asiatiques, bien qu'avec peu de propos haineux.

Discours anti-Arabes

Principaux résultats

- Sur le million de messages pertinents identifiés, **notre algorithme a placé 132 000 messages dans la catégorie des discours anti-Arabes haineux** pendant la période étudiée (Figure 19). Les diverses façons dont les mots-clés anti-Arabes ont été employés ont néanmoins rendu difficile la formation d'un algorithme précis. Ce sous-ensemble n'est certainement pas représentatif de l'ampleur du contenu haineux de l'ensemble de données. En effet, la taille de l'ensemble de données n'a pas permis de mieux classer ces messages haineux pour en déduire une distinction claire entre propos haineux ciblés et propos haineux généralisés.

- Les manières dont les internautes emploient des mots-clés anti-Arabes sont diverses. Parfois, les titulaires de comptes utilisent des insultes génériques pour attaquer les Arabes, tandis que d'autres usent d'insultes anti-Arabes pour cibler d'autres groupes. Cette dernière situation est constatée avec le terme *racaille*, utilisé depuis longtemps par les groupes d'extrême droite et nationalistes pour se référer aux personnes d'origine arabe. Toutefois, pendant la période de l'étude, ce terme a été souvent appliqué aux Gilets Jaunes et à la police.

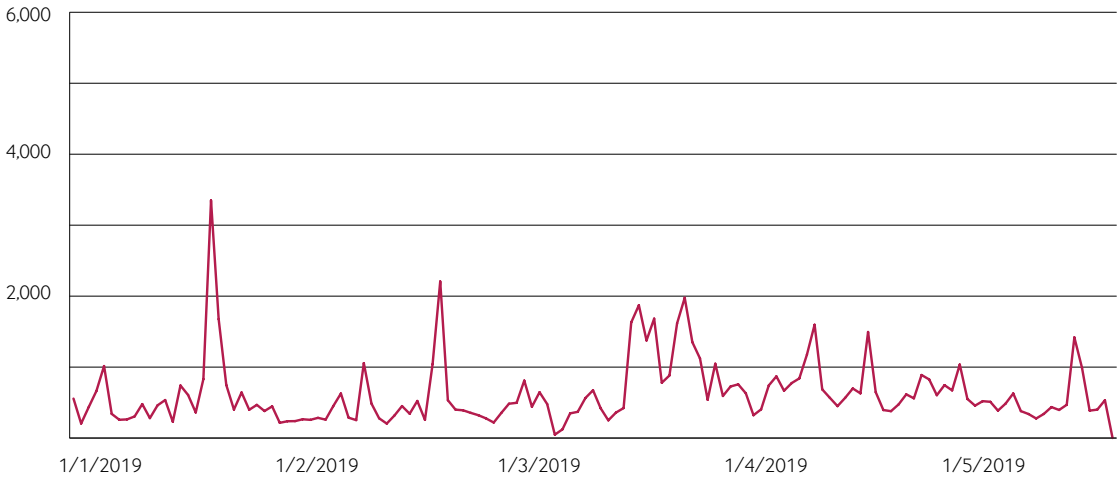
- **Les événements hors ligne, y compris l'attaque de Christchurch, ont conduit à une intensification des discours anti-Arabes haineux, en particulier par les messages mentionnant la théorie du Grand Remplacement ;** cependant, les messages non haineux contenant des mots-clés anti-Arabes ont également augmenté après l'attaque, dont beaucoup se sont opposés à l'idée d'un « grand remplacement » et ceux qui propagent ces opinions. Notre algorithme a classé comme haineuses 6 % des discussions au cours de la semaine suivant l'attaque de Christchurch.

- **Parmi les comptes les plus actifs dans la propagation des discours haineux anti-Arabes, certains proviennent de l'extrême droite, d'autres de groupes identitaires connus.** La théorie du Grand Remplacement et le sentiment anti-migrant sont apparus comme des thèmes récurrents dans l'ensemble de données haineuses.

- Les comptes les plus souvent cités appartiennent également à des personnalités d'extrême droite fréquemment mentionnées dans des messages contenant des propos haineux anti-Arabes.

- Nous avons constaté un recoupement important avec le langage misogyne, évident dans l'utilisation du terme *beurette*, la forme féminine de *beur* (un terme argot utilisé pour désigner les personnes d'origine nord-africaine, ou les Arabes en général). Ce terme a été employé dans environ 9 % des messages de l'ensemble de données haineuses anti-Arabes. Nous avons également relevé des efforts de réappropriation du terme.

Figure 19
Volume de messages contenant des discours anti-Arabes haineux entre le 1er janvier 2019 et le 31 mai 2019



Discours anti-Arabes
Exemples de discours haineux

[redacted]

mdr encore une beurette qui le critique alors que sa suce sur viva street , allez fout moi le camps

Figure 20 **Un tweet associant des insultes anti-Arabes et misogynes (source : Twitter)**

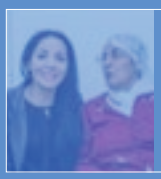
[redacted]

Pendant que l'UNEF appelle au massacre de blance, que les banlieues sont tenues par la racaille islamiste , le pouvoir en place dissout en association dont l'objet était de venir en aide aux SDF francais. #BastionSocial

Figure 21 **Un tweet utilisant le terme *racaille* (source : Twitter)**

Riposte Laique
@1RiposteLaique

La beurette Bouchera Azzouz, ou l'islamiste dévoilée



La beurette Bouchera Azzouz, ou l'isla
Ce ne'st pas le bon sens qui, comme le croyait a tort Descartes, est la chose du monde la mieux partagee, mais l'ingratitude

Figure 22 **Tweet d'une plateforme d'extrême droite associant les femmes arabes et les islamistes (source : Twitter);**

Discours anti-Arabes
Exemples de contre-discours

[redacted]

Beurette est un terme raciste, misogyne et rabaissant pour les femmes maghrébines. Et si je dois le tweeter tout les jours pour que ça rentre dans vons petits cerveau de moineau, je le farais.

Figure 23 **condamnant l'utilisation humiliante du terme *beurette* (source : Twitter)**

Edwy Plenel
@edwyplenel

Le «grand replacement» théorisé par Renaud Camus, mis en fiction par Michel Houellebecq et promu médiatiquement par Éric Zemmour n'est pas une opinion dont il faudrait débattre mais une idéologie meurtrière. Mon analyse via @Mediapart

L'ideologie meurtriere promue par Zemmour
Dans le sillage de l'ecrivain Renaud Camus, Éric Zemmous prétend que le people français subit un...

Figure 24 **Tweet condamnant l'attentat de Christchurch et la théorie du Grand Remplacement (source : Twitter)**

Discours anti-Arabes
Exemples de contre-discours

Benoît Hamon
@benoithamon

C'est donc au nom de le théorie du grand remplacement que le terroriste d'extrême droite et islamophobe a tué 49 personnes. Le massacre d'innocents, voilà où mène la haine de ceux qui colportent en France comme en Nouvelle Zélande, cetta idéologie odieuse. #christchurch

Figure 25 Tweet de Benoit Hamon condamnant l'attentat de Christchurch et la théorie du Grand Remplacement (source : Twitter)

[redacted]

Et quand tu snap le tableau de bord de tes copines du coup t'es une quoi ?vs rendez fou et juste le mot beurette est grv dégradant jsp si il s'en rends compte

[redacted]

déjà beurette ça veux dire une personne né sur le territoire français dont les parents sont issue de pays d'Afrique du Nord mdr y'a rien de dégradant le mot a juste été détourné par des racaille comme Osman le revenant.

Figure 26 Tweet condamnant l'utilisation humiliante du terme *beurette* et la réponse qui nie le sens dégradant du terme (source : Twitter)

Étude de Cas

Discussions concernant le Grand remplacement après l'attentat de Christchurch

L'un des principaux thèmes qui s'est dégagé de cet ensemble de données a été la théorie du Grand Remplacement.[53] L'ISD a formé des algorithmes pour identifier les cas dans lesquels le terme « grand remplacement » a été utilisé d'une manière visant à provoquer ou à inciter à la haine contre des individus et des communautés arabes.
Ces messages ont représenté environ 4 % de l'échantillon global et plus de la moitié (55 %) du sous-ensemble haineux. Les racines de cette théorie remontent à Renaud Camus[54] dont l'ouvrage Le Grand Remplacement, a été publié en 2011. La théorie mentionnée dans ce livre affirme que les populations européennes blanches seront délibérément remplacées au niveau ethnique et culturel par la migration et la croissance des communautés minoritaires.

Dans l'ensemble de données haineuses, dans la semaine suivant l'attentat de Christchurch,

Renaud Camus
@RenaudCamus

Grand Remplacement, Islamisation, Déculturation : résister au péril de civilisation youtu.be/pxQ_WwBI30Y via @YouTube

Figure 27 Tweet de Renaud Camus plaidant pour la lutte contre le Grand Remplacement (source : Twitter)

le nombre de mentions du grand remplacement a augmenté, comme l'illustre la Figure 28.
Les événements récents, en particulier l'attentat de Christchurch, soulignent le rôle joué par la rhétorique du grand remplacement dans les activités d'extrême droite et les attaques violentes.
Rappel : le dernier rapport de l'ISD sur le grand remplacement soulignait : « Il est clair que la théorie préconise une action radicale contre les communautés minoritaires, y compris le nettoyage ethnique, la violence et le terrorisme ».[55]

La promotion de cette idéologie par des personnalités publiques a de plus contribué à la normalisation de cette rhétorique et peut être attribuée à la diffusion de contenus haineux au niveau communautaire.
La carte du réseau de notre ensemble de données haineuses démontre la centralité de l'idée du grand remplacement pour cette communauté en ligne. Les utilisateurs qui évoquent cette idée font référence aux Arabes en utilisant des injures comme rats, porc et arabe de service.

← La première ministre néo-zélandaise a visité le centre social de Phillipstown le 16 mars 2019, moins de 24 heures après l'attaque contre deux mosquées à Christchurch

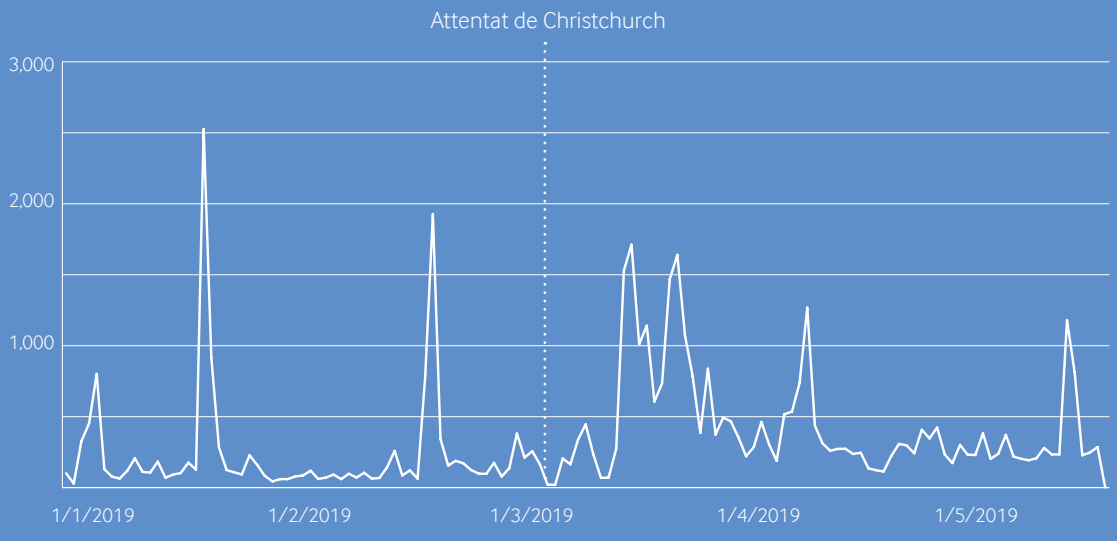


Figure 28
Mentions du terme « Grand Remplacement » dans l'ensemble de données haineuses anti-Arabes entre le 1er janvier 2019 et le 31 mai 2019

Discours anti-Roms ou anti-Gitans

Principaux résultats

- Les mots-clés anti-Roms ou anti-Gitans sélectionnés pour ce rapport ont donné 299 157 résultats (Figure 29).

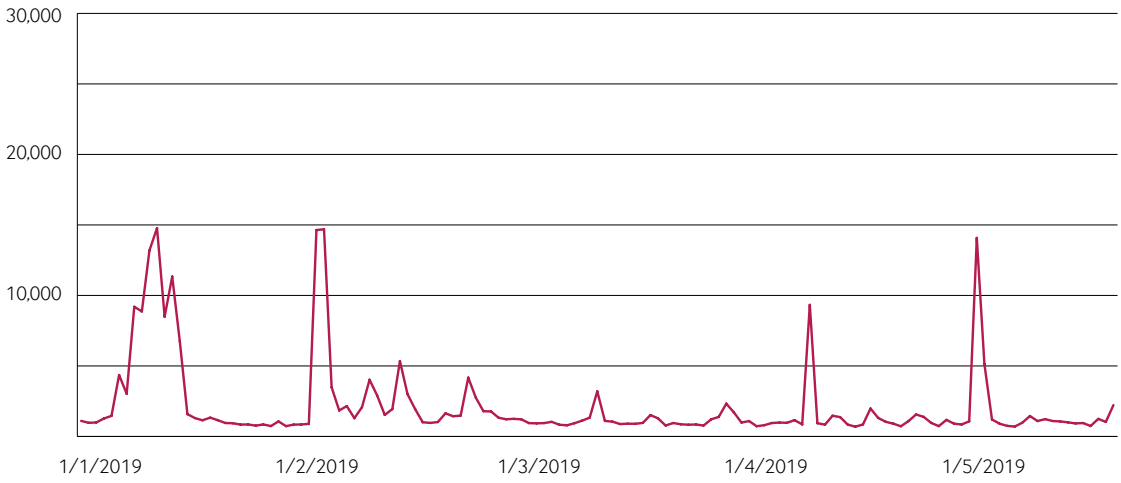
- Les discussions sur les mots-clés anti-Roms et anti-Gitans ont connu un regain d'intérêt à la suite de certains événements d'actualité, notamment **l'arrestation de Christophe Dettinger, militant Gilets Jaunes** en janvier et **les commentaires controversés de l'ancienne ministre des Affaires européennes Nathalie Loiseau sur son passage à l'ENA**. À la suite de l'incident impliquant Christophe Dettinger, **les comptes liés au Mouvement des Gilets Jaunes ont soutenu l'inclusion des Roms dans le mouvement**. Ces types de messages constituaient la majorité de l'ensemble de données.

- **Les messages haineux étaient largement axés sur les stéréotypes qui associent les Roms à la criminalité et les qualifient de fléaux pour la société,** comme en témoignent les messages présentés ci-après.

- **Les comptes les plus actifs associés à notre liste de mots-clés appartenaient en grande partie aux membres de la communauté Rom ou gitane,** ce qui suggère que la communauté s'est appropriée ces termes.

- Bien que le niveau de discours haineux semble faible dans l'échantillon, une série d'attaques contre les Roms et les Gitans fin mars 2019 a mis en lumière la manière dont la désinformation peut être utilisée en ligne pour inciter à la haine et aux attaques violentes. Cette étude de cas illustre également les difficultés de l'identification de contenus haineux en ligne.

Figure 29
Volume de messages contenant des mots-clés anti-Roms entre le 1er janvier 2019 et le 31 mai 2019



Discours anti-Roms ou anti-Gitans

Exemples de discours contenant des mots-clés anti-Roms ou anti-Gitans

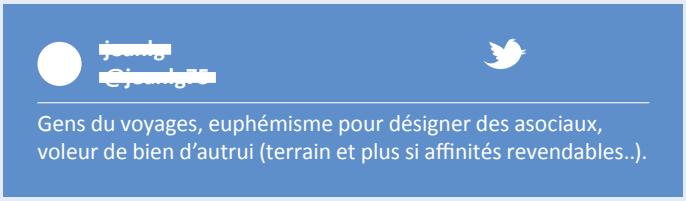


Figure 30 Un tweet employant le stéréotype de voleurs pour les Roms (source : Twitter)

Figure 31 Message Facebook en réponse aux commentaires d'Emmanuel Macron dans le cadre de l'affaire Dettinger (source : Facebook)

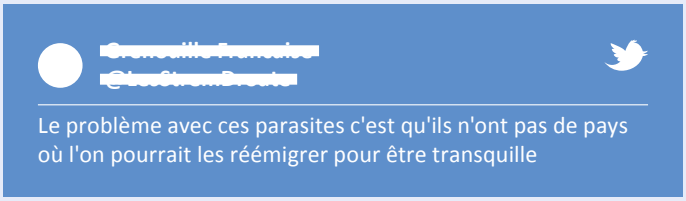


Figure 32 Tweet haineux anti-Roms (source : Twitter)

Étude de Cas

La campagne de désinformation sur les enlèvements perpétrés par des Roms.

Dans la nuit du 26 mars 2019, une série d'agressions contre des Roms ont eu lieu dans la banlieue parisienne. Ces attaques, qui ont fait l'objet d'une large couverture dans les médias et sur les plateformes d'extrême droite, ont été déclenchées par de fausses rumeurs sur les réseaux sociaux selon lesquelles des Roms seraient responsables d'enlèvements dans cette région.

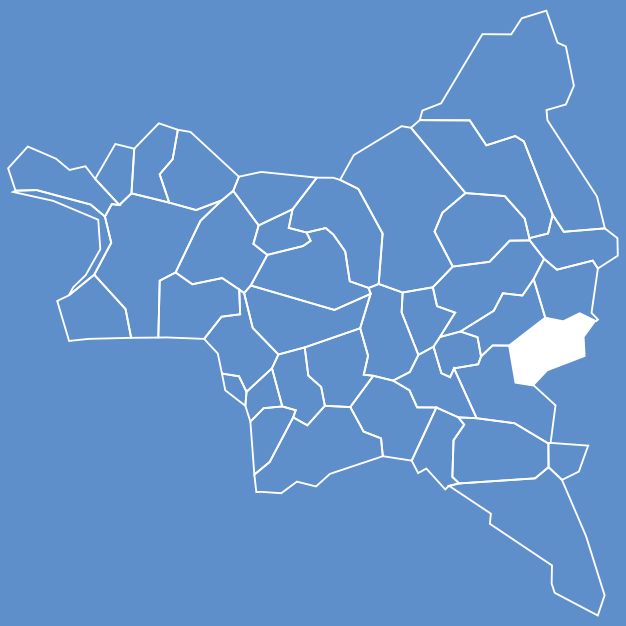


Figure 33 Carte de la région Seine-Saint-Denis

La rumeur est devenue virale avec un tweet, posté le 24 mars, faisant état d'une tentative d'enlèvement présumée par un réseau Rom de trafic d'organes à Montfermeil. Ce message, retweeté plus de 12000 fois, a reçu plus de 8000 « likes ».

Une analyse par l'ISD des messages de Facebook sur les Roms dans les 24 heures qui ont suivi les agressions a montré que

Étude de Cas

La campagne de désinformation sur les enlèvements perpétrés par des Roms.

les contenus « super performants » étaient principalement ceux diffusés des médias grand public, ainsi que des messages de l'ONG antiraciste SOS Racisme condamnant les attentats, preuves d'un engagement rapide du contre-discours.

Les médias d'État russes, dont RT France et Sputnik, ont été particulièrement actifs dans la couverture des événements sur Facebook. Des commentaires d'utilisateurs sous un certain nombre de ces articles affirmaient que les rumeurs d'enlèvements et de trafic d'organes par des Roms étaient en fait fondées (Figure 34). Certains exprimaient des opinions anti-Roms.

La rumeur qui a entraîné la série d'attaques en mars a commencé quelques jours avant les événements quand des clubs de football locaux ont posté des messages avertissant les parents contre des risques d'enlèvements (Figure 36). Le magazine de football Panamefoot a publié un article sur l'enlèvement d'enfants le 23 mars, partageant un message du Club de Football Aulnaysien. Les clubs de football ont ainsi commencé à envoyer aux parents des sms pour les alerter des pseudos « risques d'enlèvements ».

La rumeur a débuté quand le 24 mars, la vidéo d'une camionnette blanche censée appartenir aux kidnappeurs est apparue sur Snapchat. Les rumeurs ont ensuite commencé à se propager sur Facebook et Twitter. Un tweet du 24 mars a largement amplifié la désinformation, déclenchant des attaques haineuses contre les Roms. Le message est d'abord apparu sur un compte Twitter, dont le titulaire se décrit comme un étudiant turc de 17 ans luttant contre la désinformation. Le 25 mars, un autre message Snapchat diffusant de fausses informations sur

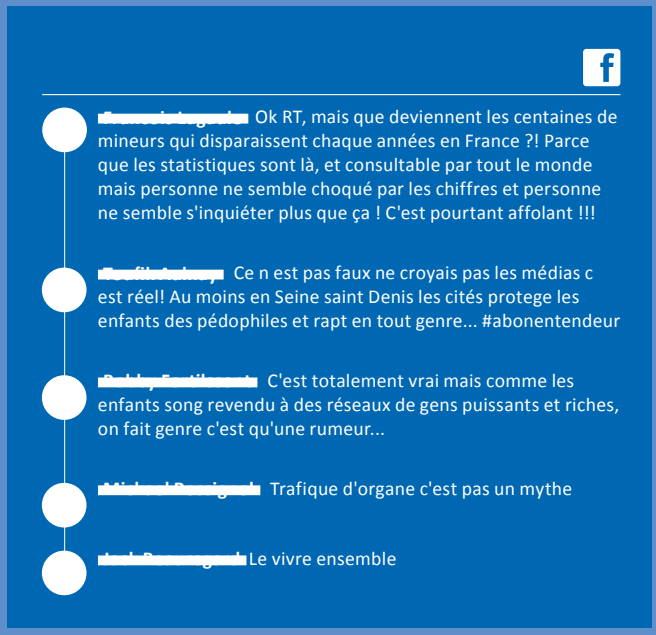


Figure 34 Commentaires sur Facebook partageant des théories de complot à la suite d'attaques anti-Roms (source : Facebook)

des enlèvements commis par des Roms a donné lieu à plusieurs attaques contre la communauté Rom.

Dans les jours qui ont suivi les agressions, une page Facebook est apparue sur laquelle étaient partagés des témoignages de parents dont les enfants étaient censés avoir été enlevés. Des articles ont également été largement diffusés sur des sites Web consacrés à la lutte contre la pédophilie, certains d'entre eux présentant des théories du complot concernant une dissimulation médiatique d'enlèvements par des pédophiles. Un article, paru le 26 mars, partageait ce même point de vue. Il a été posté par le site Web de lutte contre la pédophilie « Wanted Pedo », connu pour partager des fausses informations et dont la page Facebook a été fermée en 2016. Malgré la couverture médiatique des agressions et les tentatives de dénonciation des fausses rumeurs, la désinformation a continué de se répandre.

L'équipe de fact-checking du quotidien Libération Checknews a retracé la rumeur jusqu'en 2012 lorsque trois Roms ont été accusés d'avoir kidnappé un enfant. À l'époque, la justice avait rejeté l'affaire car les tests ADN réalisés n'ont pas confirmé les accusations. Depuis, plusieurs rumeurs, toutes démenties par les autorités locales, sont apparues sur plusieurs plateformes (Facebook et WhatsApp). En France, en 2018, la police a enquêté sur deux tentatives d'enlèvement par des hommes conduisant une camionnette blanche.

Curieusement, nous n'avons constaté aucune augmentation significative du discours haineux anti-Roms pendant les attaques. Plusieurs facteurs peuvent expliquer cette constatation. Snapchat a joué un rôle déterminant dans la

Figure 35 Message envoyé par un club de football pour alerter du risque d'enlèvements d'enfants

diffusion de la campagne de désinformation et des récits haineux sur la communauté Rom. La nature de la plateforme rend le contenu de Snapchat extrêmement difficile à analyser, les vidéos étant supprimées dans les 24 heures suivant leur publication. De plus, notre choix d'une approche par mots-clés ne nous a pas permis d'analyser le contenu vidéo.

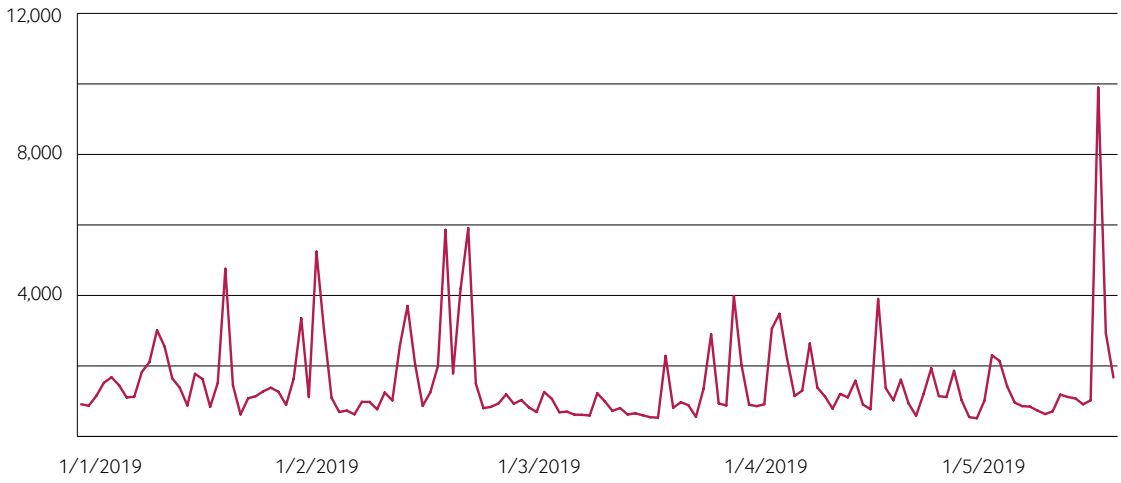
Sur des plateformes telles que Twitter et Facebook, le contenu lié à ces attaques n'a pas utilisé les mots-clés sélectionnés pour notre recherche. Les utilisateurs employaient régulièrement les termes kidnappeur, fdp et Roumain. Bien que les termes kidnappeur et Roumain aient été sélectionnés au départ, ils ont dû être abandonnés car générateurs de trop de faux positifs.

Discours anti-Noirs ou anti-Africains

Principaux résultats

- Les mots-clés anti-Noirs ou anti-Africains sélectionnés pour ce rapport ont donné 223 483 résultats (Figure 36).
- **Les messages comprenant des mots-clés anti-Noirs ou anti-Africains ont été favorisés par des événements spécifiques,** notamment des attaques racistes contre le comédien Donel Jack'sman, la mort d'Ange Dibenesha pendant sa garde à vue et des controverses liées à la représentation des Noirs ou des Africains dans la société française.
- **Le pic le plus important de mots-clés anti-Noirs ou anti-Africains a eu lieu entre le 12 et le 16 mai,** en réponse à l'introduction d'un tableau à l'Assemblée nationale, pour célébrer l'anniversaire de la première abolition de l'esclavage en France en 1794. Ce tableau a déclenché un débat autour de son racisme présumé et de sa représentation d'un visage noir.
- **Nous avons également constaté des pics de conversations du 17 au 20 février, à la suite de la commercialisation controversée par des boulangeries françaises de pâtisseries** nommées « tête de nègre » et « bamboula » (terme péjoratif pour décrire une personne de couleur). La réaction du public contre les pâtisseries a entraîné leur retrait des étagères de certains magasins.[56] Au cours de notre étude, nous avons identifié plusieurs comptes qui ont profité de cet incident pour déclarer que l'usage de termes anti-Blancs n'entraînerait pas le même degré d'indignation.
- Les conversations sur le racisme institutionnalisé sont apparues comme un thème clé parmi les mots-clés choisis.
- Les comptes des réseaux sociaux qui employaient des mots-clés anti-Noirs et anti-Africains montrent le plus souvent un clivage entre les comptes explicitement anti-immigration et les comptes tenus par les membres des communautés noires ou africaines en France.

Figure 36
Volume de messages contenant des mots-clés anti-Noirs ou anti-Africains entre le 1er janvier 2019 et le 31 mai 2019



Discours anti-Noirs ou anti-Africains

Exemples de discours contenant des mots-clés anti-Noirs et anti-Africains

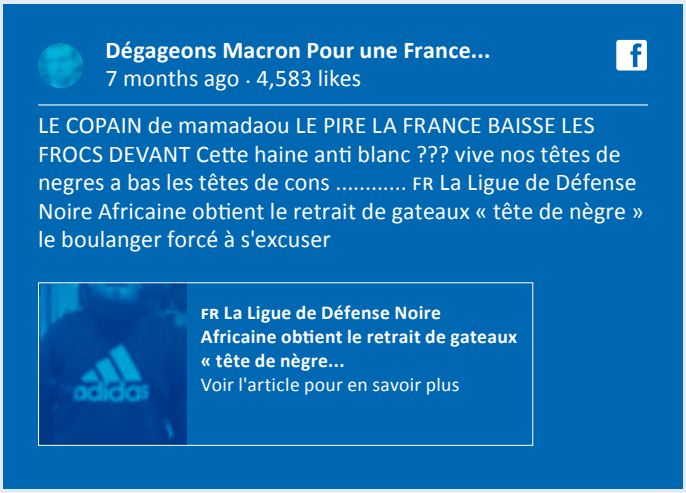


Figure 37 **Message suivant la commercialisation de pâtisseries nommées « tête de nègre » (source : Facebook)**

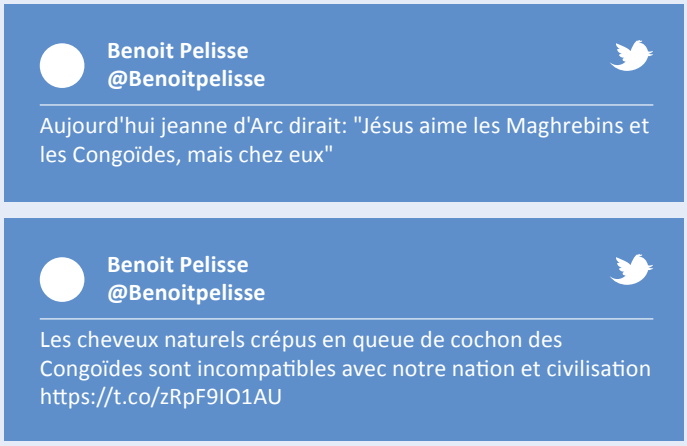


Figure 38 **Tweets de Benoit Pélisse, le second compte en terme d'activité employant des mots-clés associés à des discours haineux anti-Noirs (ce compte est aujourd'hui suspendu). Benoit Pélisse est l'ancien président de Fraternité & Réémigration, une organisation qui promeut le retour de populations non-européennes (source : Crimson Hexagon)**

> Les conversations sur le racisme institutionnalisé sont apparues comme un thème clé parmi les mots-clés choisis

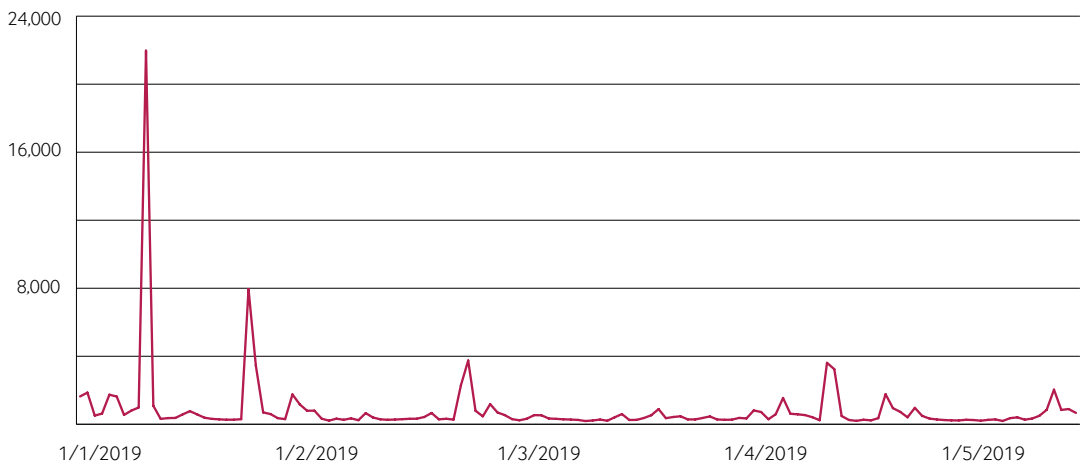
Discours anti-Blancs
Principaux résultats

- Les mots-clés anti-Blancs sélectionnés pour ce rapport ont donné 125 654 résultats (Figure 39).[57]
- **Nous n'avons pas identifié de messages employant les termes babtou ou toubab avec une intention haineuse ou dans le but de maltraiter les Blancs.** Les messages générés par les mots-clés anti-Blancs se concentrent sur la discussion des stéréotypes concernant les Blancs et ont suscité un débat sur l'existence du racisme anti-Blancs.
- **Un pic majeur dans la conversation a eu lieu entre le 20 et le 26 janvier,** causé par un message faisant référence à des stéréotypes de personnes noires et blanches les unes contre les autres (utilisant le terme babtou), sans être ouvertement haineux.
- **Les comptes les plus actifs portaient sur des discussions sur l'antiracisme et mettaient en doute l'existence d'un racisme anti-Blancs.** Quatre comptes utilisent des mots-clés anti-Blancs pour parler de l'oppression des Noirs par les Blancs et de la colonisation française. Les pages les plus actives sur Facebook ne partagent pas de termes haineux, mais semblent plutôt être de faux positifs puisque *babtou* et *toubab* proviennent d'un mot wolof signifiant blanc.

Discours anti-Blancs
Exemples de conversations comportant des mots-clés anti-Blancs

Figure 40 **Tweet utilisant le terme *toubab* (source : Twitter)**

Figure 39
Volume de messages contenant des mots-clés anti-Blancs entre le 1er janvier 2019 et le 31 mai 2019



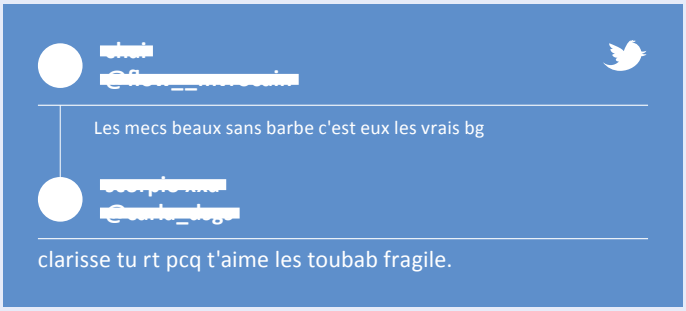


Figure 41 **Tweet utilisant le terme *toubab* (source : Twitter)**

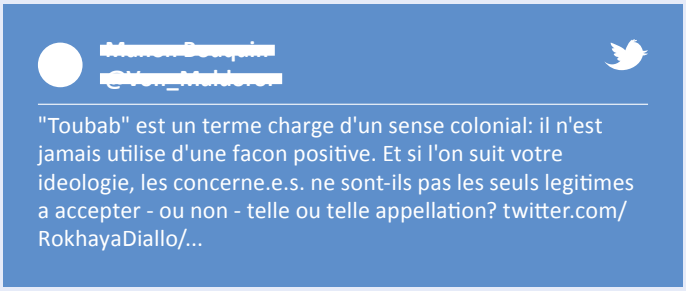


Figure 42 **Tweet affirmant que le terme *toubab* a des connotations négatives (source : Twitter)**

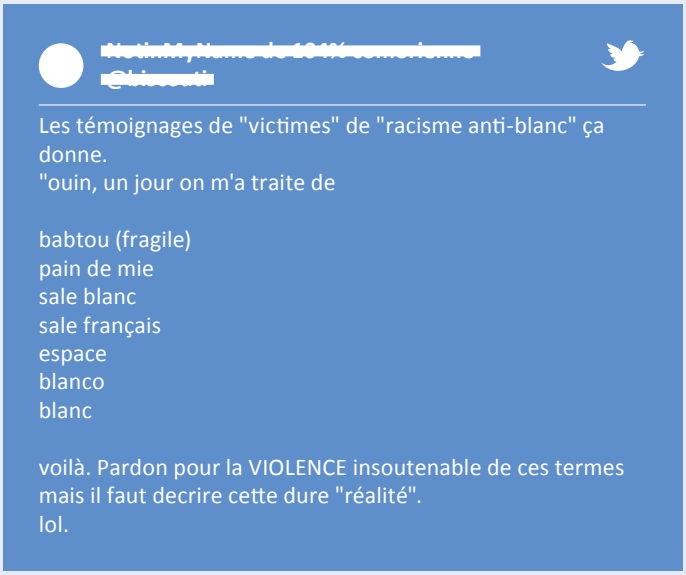


Figure 43 **Tweet affirmant que le racisme anti-Blancs n'est comparable à aucune autre forme de racisme (source : Twitter)**

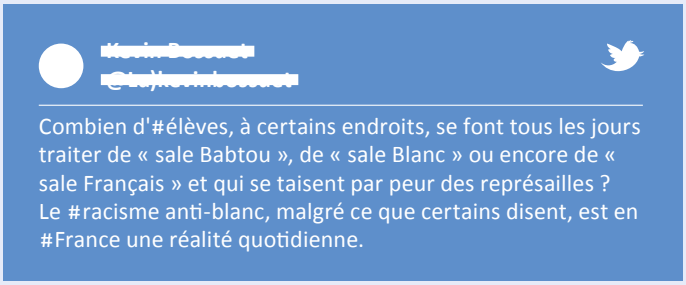


Figure 44 **Tweet affirmant que le racisme anti-Blancs est réel (source : Twitter)**

Discours anti-Asiatiques
Principaux résultats

- Les mots-clés anti-Asiatiques sélectionnés pour ce rapport ont donné 45 804 résultats (Figure 45).
- **Les mots-clés anti-Asiatiques ont été déployés dans divers contextes, y compris dans des expressions françaises** bien établies comme « c'est du chinois »**. Nos chercheurs n'ont pas trouvé d'exemples précis de contenu haineux ciblé.**
- L'analyse qualitative d'un échantillon de messages utilisant l'expression *bridés* a montré qu'elle était utilisée dans une variété de contextes sans thème unificateur clair, la majorité des messages contenant l'expression yeux *bridés*. Ces messages allaient de la beauté et du maquillage à des questions portant sur les yeux *bridés* (exemples ci-après).

Discours anti-Asiatiques
Exemples de contenus en ligne comportant des mots-clés anti-Asiatiques

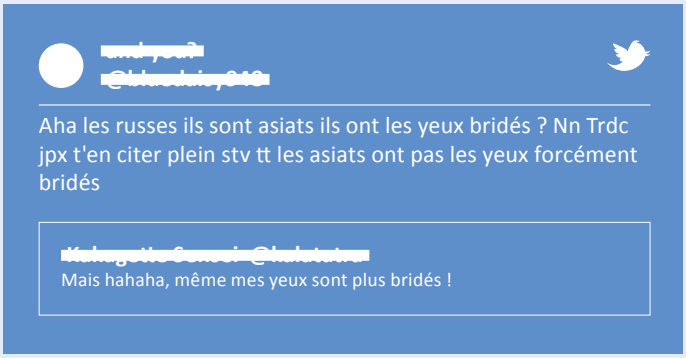


Figure 46 **Tweet au sujet des yeux *bridés* (source : Twitter)**

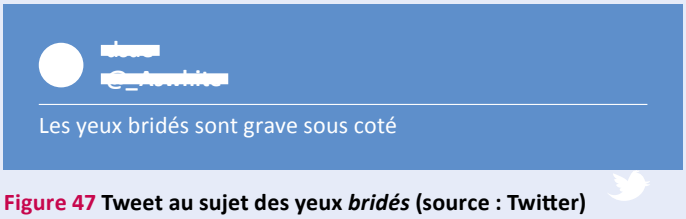
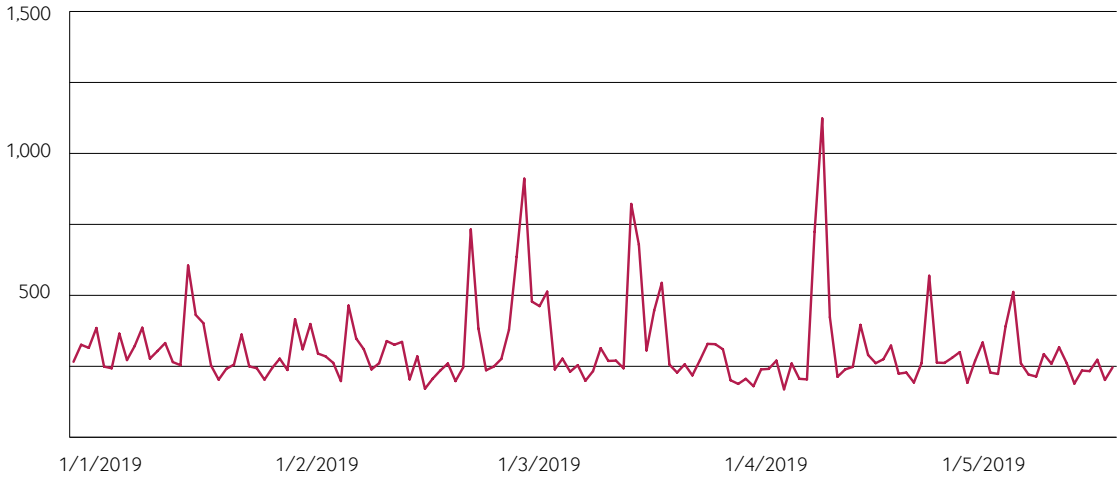


Figure 47 **Tweet au sujet des yeux *bridés* (source : Twitter)**

Figure 45
Volume de messages contenant des mots-clés anti-Asiatiques entre le 1er janvier 2019 et le 31 mai 2019



Discours anti-Asiatiques
Exemples de contre-discours

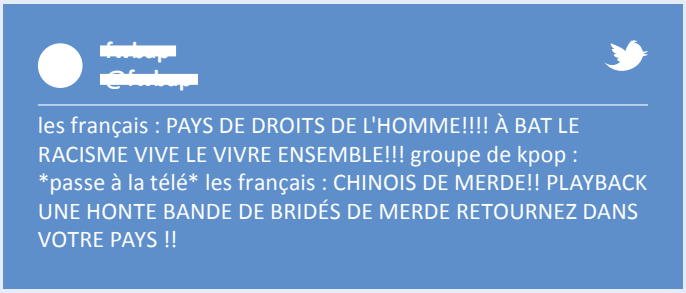


Figure 48 **Tweet condamnant la haine anti-Asiatiques (source : Crimson Hexagon)**

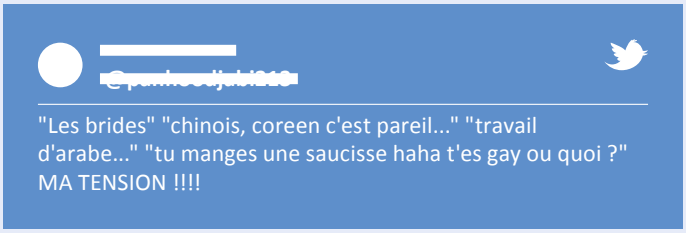


Figure 49 **Tweet se moquant des préjugés et préconceptions, notamment au sujet des Asiatiques (source : Twitter)**

Religion

Cette section comprend les discours anti-Musulmans, antisémites et anti-Chrétiens. L'ensemble des données recueillies à partir de mots-clés anti-Musulmans a révélé une forte proportion de propos haineux, indiquant dans certains cas des efforts coordonnés. L'ensemble de données antisémites contenait un sous-ensemble important de contenu lié à l'agression d'Alain Finkielkraut lors d'un rassemblement des Gilets Jaunes, ainsi que des critiques à l'égard d'Israël. Les discussions sur les mots-clés anti-Chrétiens étaient surtout axées sur les débats publics concernant la place de l'Islam dans la société française, avec peu de haine anti-Chrétiens évidente.

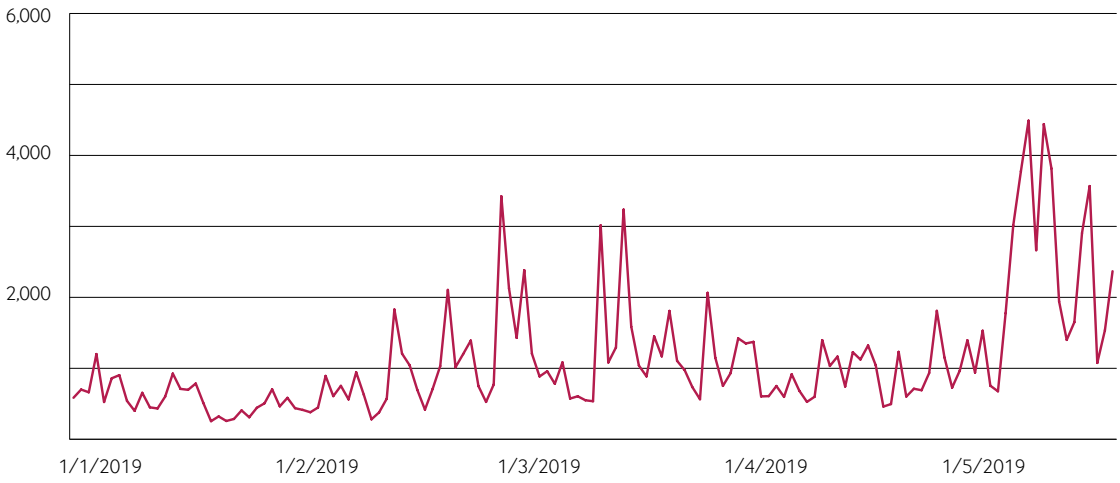
Discours anti-Musulmans

Principaux résultats

- Les mots-clés anti-Musulmans sélectionnés pour ce rapport ont donné 168 324 résultats (Figure 50).
- **L'augmentation des conversations utilisant des mots-clés anti-Musulmans est étroitement liée à des événements spécifiques, notamment le début du Ramadan et la controverse sociétale liée aux signes extérieurs de la foi musulmane, comme l'affaire de discrimination Etam.**
- L'analyse qualitative suggère que les discours employant des mots-clés anti-Musulmans étaient principalement haineux et montraient une mobilisation anti-Musulmans concertée autour d'événements hors ligne.
- **Les discussions sur l'islamisation présumée de la France et l'invasion musulmane ont dominé l'échantillon,** avec divers exemples anecdotiques présentés par des utilisateurs qui semblaient avoir des opinions anti-Musulmans.
- **Les titulaires de comptes anti-Musulmans clairement identifiables ont été les plus actifs dans l'utilisation de mots-clés anti-Musulmans, ce qui suggère qu'il existe des efforts concertés pour cibler la communauté musulmane en ligne.** Certains comptes prétendaient être basés au Canada, soulignant la nature internationale des discours haineux en ligne.

> L'analyse qualitative suggère que les discours employant des mots-clés anti-Musulmans étaient principalement haineux

Figure 50
Volume de messages contenant des mots-clés anti-Musulmans entre le 1er janvier 2019 et le 31 mai 2019



Discours anti-Musulmans
Exemples de contenus en ligne comportant des mots-clés anti-Musulmans

Figure 51 Message Facebook affirmant que le burkini est un signe de l'islamisation de la France (source : Facebook)

Figure 52 Message sur l'islamisation provenant d'une page d'extrême droite (source : Facebook)

Discours anti-Musulmans
Exemple de contre-discours

Figure 53 Tweet condamnant un article du quotidien Daily Mail sur l'islamisation présumée de la ville de Saint Denis (source : Twitter)

Étude de Cas

Pétition Damoclès « Non à l'islamisation de la France » début du Ramadan

Au cours de la période étudiée, le volume de messages anti-Musulmans a atteint un pic marqué au début du Ramadan en France. De ce fait, les pics les plus importants ont eu lieu la semaine du 5 mai (plus de 20 000 tweets). Ce niveau élevé d'activité a été déclenché par une pétition réclamant la fin de l'islamisation de la France, contenu le plus partagé par rapport aux mots-clés choisis. La pétition a été lancée par Damoclès, une association française dont l'objectif est de réduire l'immigration musulmane en France. Cette pétition a été largement partagée sur Twitter et Facebook.

Figure 54 Facebook partageant la pétition de Damoclès (source : Facebook)

Figure 55 Message de Damoclès soutenant que les jeunes femmes musulmanes obtiennent de fausses autorisations médicales de leur médecin pour éviter d'aller à la piscine (source : Facebook)

Discours antisémites
Principaux résultats

- Les mots-clés antisémites sélectionnés pour ce rapport ont donné 79 289 résultats (Figure 56).
- **L'intensification des conversations employant des mots-clés antisémites a connu un pic (plus de 30 000 messages) après l'attaque contre Alain Finkielkraut,** lors d'une manifestation des Gilets Jaunes en février 2019. Les manifestants ont utilisé le terme « sale juif » pendant cette période, en très grande majorité pour signaler l'incident plutôt que comme une attaque directe (très peu de discours haineux antisémites).
- La France a constaté une augmentation de 74 % des agressions antisémites en 2018. L'écart entre le faible volume de discours haineux que nous avons trouvés et les preuves de haine antisémite hors ligne pourraient être le résultat de notre choix de mots-clés, d'un accès limité à l'intégralité des données de toutes les plateformes ou du fait que le contenu haineux antisémite ne se limite pas à l'utilisation de certains mots-clés.
- **Les critiques de la politique israélienne dans les territoires palestiniens sont apparues comme un autre thème clé de cet ensemble de données,** bien que la plupart des messages ne soient pas explicitement haineux (voir Figure 58).
- Les comptes les plus actifs recouvrent une combinaison de comptes explicitement antisémites (dont la plupart ont été suspendus ou supprimés par les plateformes), de comptes des Gilets Jaunes (principalement liés à l'incident contre Finkielkraut), de comptes de type bot (messages sur divers thèmes) et de comptes progressistes (qui critiquent les politiques israéliennes).

Discours antisémites
Exemple de contenu en ligne comportant des mots-clés antisémites

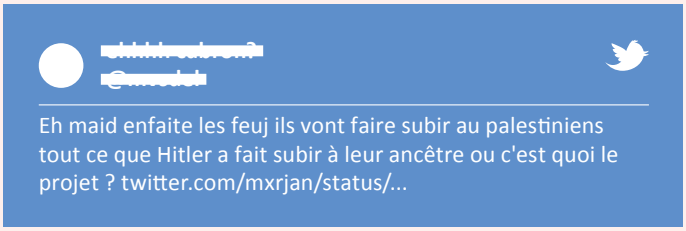
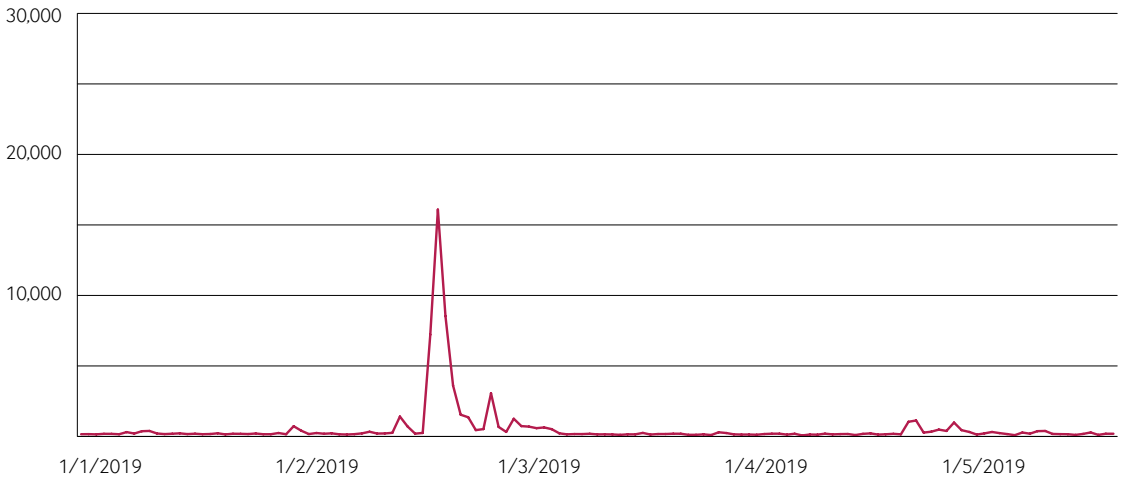


Figure 57 Tweet affirmant que les Juifs traitent les Palestiniens de la même manière que les Nazis traitaient les Juifs pendant la Seconde Guerre mondiale (source : Twitter)

Figure 56 Volume de messages contenant des mots-clés antisémites entre le 1er janvier 2019 et le 31 mai 2019



Discours anti-Chrétiens

Principaux résultats

- Les mots-clés anti-Chrétiens sélectionnés pour ce rapport ont donné 156 047 résultats (Figure 58).

- **Emploi de mots-clés anti-Chrétiens en corrélation avec des controverses spécifiques liées à l'Islam :** Discussions concernant le tweet viral d'un jeune internaute prénommé Hugo qui comparait l'image d'une foule rassemblée pour la prière à la Mecque à une scène faisant référence à In ze boîte, un jeu télévisé pour enfants diffusé sur Gulli. La religion du jeune homme et le soutien qu'il a reçu ont suscité des discussions sur les préjugés « pro-Musulmans », opposant l'Islam au Christianisme.

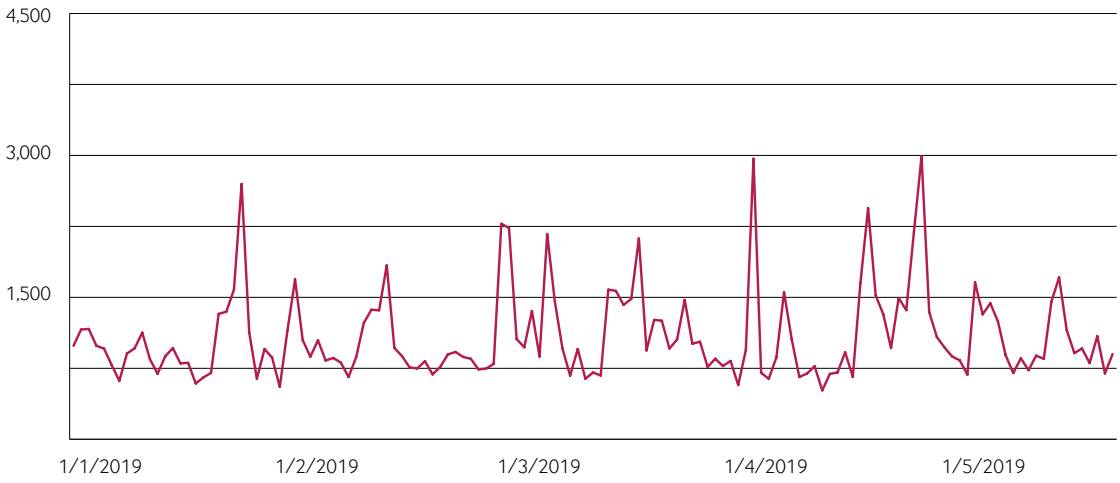
- La principale conclusion de l'ISD est que **les mots-clés anti-Chrétiens sont massivement employés non pas pour cibler les Chrétiens, mais pour critiquer l'Islam et pour conduire des conversations anti-Musulmans.** Par exemple, de nombreux titulaires de comptes d'extrême droite dans l'ensemble de données ont affirmé que l'on accorde moins d'attention à la haine anti-Chrétiens qu'à la haine anti-Musulmans afin d'apaiser les minorités religieuses.

- **Le terme *mécréant* était utilisé par les personnalités politiques et les influenceurs qui condamnent son utilisation par les islamistes (voir Figure 59).** Ce phénomène a entraîné des pics de contenus anti-immigration. Dans d'autres cas, le terme mécréant était utilisé pour citer des sourates du Coran. Un certain nombre d'influenceurs d'extrême droite employaient le terme mécréant dans notre échantillon pour exprimer implicitement un sentiment anti-Musulmans.

- **Les termes désignant le sectarisme (*bigot*) figuraient dans 7 % des messages.** Ils étaient fréquemment utilisés pour attaquer des positions socialement conservatrices concernant les droits des LGBTQ et l'avortement.

- La plupart des comptes actifs étaient composés de comptes bots et de comptes détenus par des utilisateurs engagés politiquement, dont la plupart pouvaient être décrits comme libéraux.

Figure 58
Volume de messages contenant des mots-clés anti-Chrétiens entre le 1er janvier 2019 et le 31 mai 2019



Discours anti-Chrétiens

Exemples de contenus en ligne comportant des mots-clés catégorisés comme anti-Chrétiens

Figure 59 Message de Marine Le Pen utilisant le terme *mécréant* et affirmant que les Islamistes bafouent les valeurs françaises (source : Twitter)

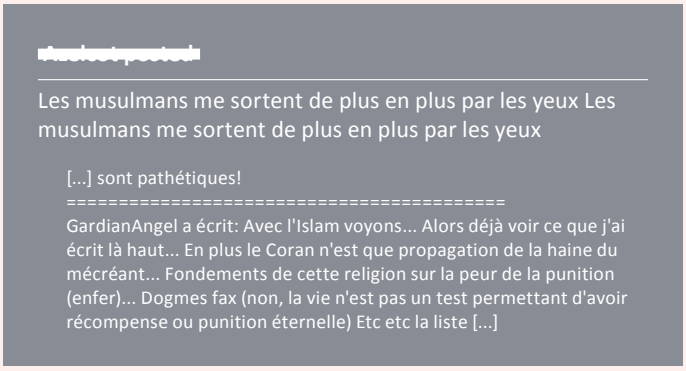


Figure 60 Message anti-Musulmans qui emploie le mot *mécréant* (source : Doctissimo)

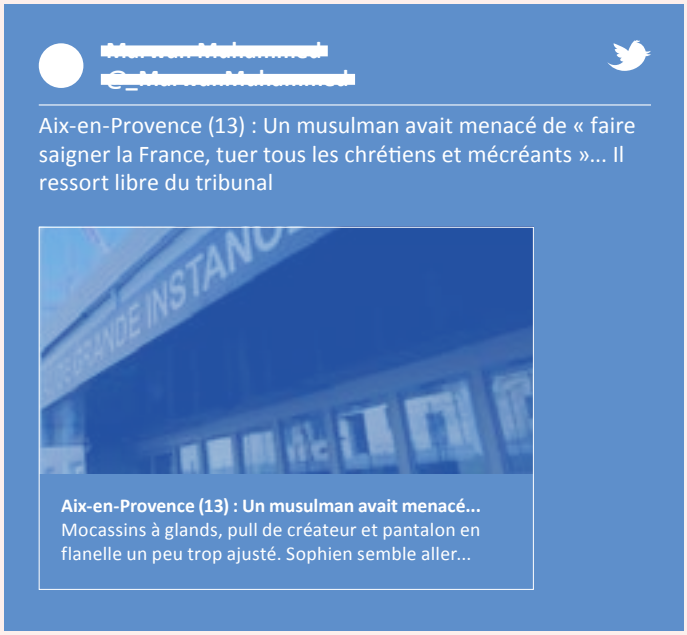


Figure 61 Tweet sur la libération d'un homme qui avait été jugé pour avoir menacé de « tuer tous les Chrétiens » (source : Twitter)

Figure 62 Message critique de l'influenceur du mouvement identitaire Damien Rieu qui reprend la citation d'une sourate par une mosquée contenant le terme *mécréant* (source : Twitter)

Discours anti-Chrétiens

Exemples de contenus en ligne comportant des mots-clés catégorisés comme anti-Chrétiens

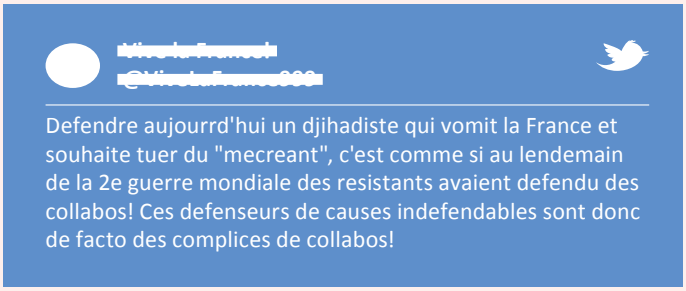


Figure 63 **Message affirmant que la France est trop laxiste à l'égard des djihadistes (source : Twitter)**

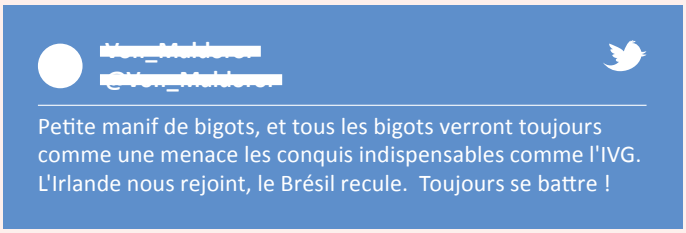


Figure 64 **Message pro-avortement qui emploie le terme *fanatique* (source : Twitter)**

Handicap

Discours capacitistes ou validistes

Principaux résultats

- Sur les 344 000 messages pertinents identifiés, notre algorithme a placé environ 77 %, soit 265 000 messages, dans la catégorie des discours haineux capacitistes ou validistes pendant la période de l'étude (Figure 65).

- **Presque tous ces messages haineux étaient des insultes ou des injures fondées sur le capacitisme, le tout normalisé dans le discours en ligne**.

- **Le pic le plus marqué dans le discours haineux capacitiste s'est retrouvé dans des messages qui se moquaient du rappeur Koba LaD, le qualifiant de *gogole*** (terme péjoratif pour désigner un trisomique 21). Le tweet a été retweeté près de 8 000 fois.

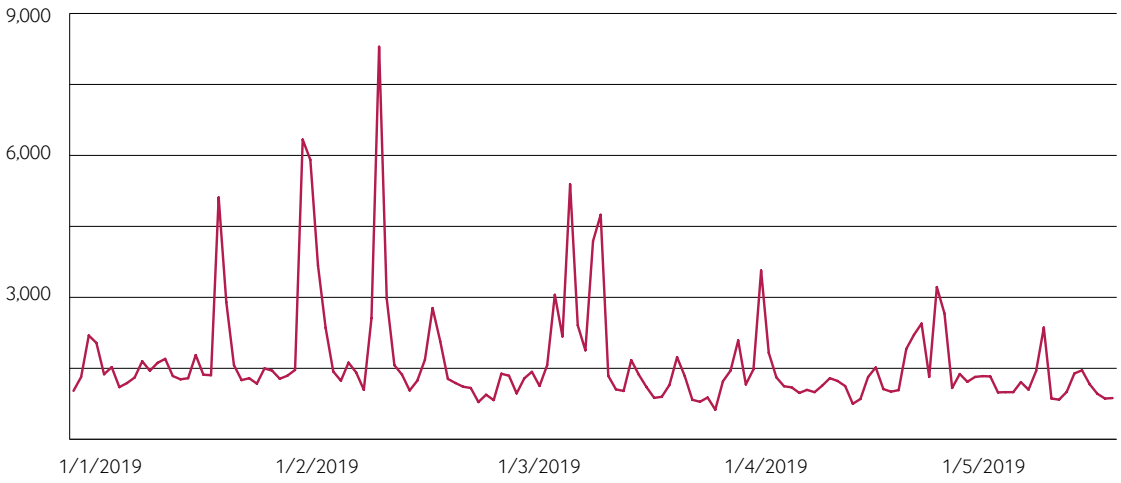
- **Les discours haineux capacitistes sont largement dominés par un vocabulaire axé sur la trisomie.** Ce vocabulaire a été utilisé sans discernement pour insulter l'intelligence des homologues internautes. ***Gogole, mongole* et *attardé*, termes argotiques pour désigner les trisomiques 21, font tous partie de l'échantillon et sont utilisés dans les conversations courantes pour faire référence au manque d'aptitudes mentales d'une personne, indiquant ainsi une généralisation de propos haineux capacitistes en ligne.**

- Les comptes les plus actifs en matière de partage de contenu haineux capacitiste appartiennent à des individus manifestant en grande partie un goût des jeux (gamers).

- Comme pour beaucoup d'autres discours, les comptes les plus fréquemment mentionnés en conjonction avec les discours haineux étaient les politiciens, les comptes sportifs et les médias d'information.

- La Journée mondiale de la trisomie, qui a eu lieu le 21 mars 2019, a provoqué un pic dans l'utilisation des mots-clés capacitistes et validistes dans l'ensemble, comme l'a fait un tweet de la mère d'une jeune trisomique (Figure 70). Cependant, les messages haineux n'ont pas connu de pics à ces occasions, ce qui suggère que les messages de contre-discours ont peut-être atteint leur objectif.

Figure 65
Volume de messages contenant des mots-clés capacitistes ou validistes entre le 1er janvier 2019 et le 31 mai 2019



Discours capacitistes ou validistes
Exemples de discours haineux

Figure 66 Exemples de discours haineux (source : Facebook)

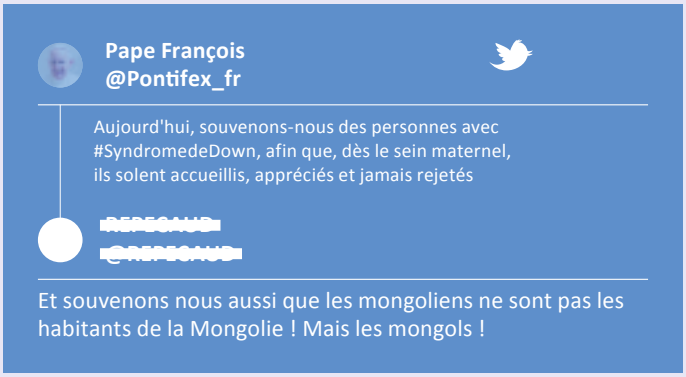


Figure 67 Tweet employant des mots-clés capacitistes (source : Twitter)

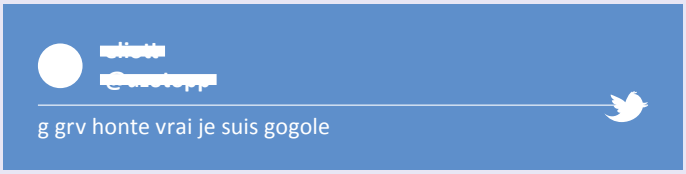


Figure 68 Tweet dans lequel l'auteur emploie un terme capacitiste pour se rapporter à lui-même (source : Twitter)

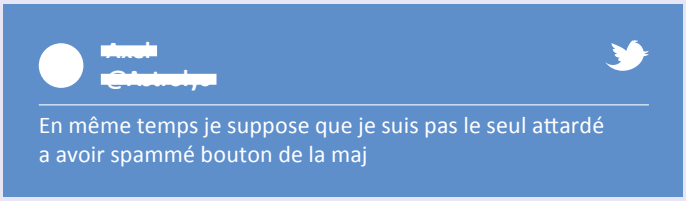


Figure 69 Tweet dans lequel l'auteur emploie un terme capacitiste pour se rapporter à lui-même (source : Twitter)

Discours capacitistes ou validistes
Exemple de contre-discours

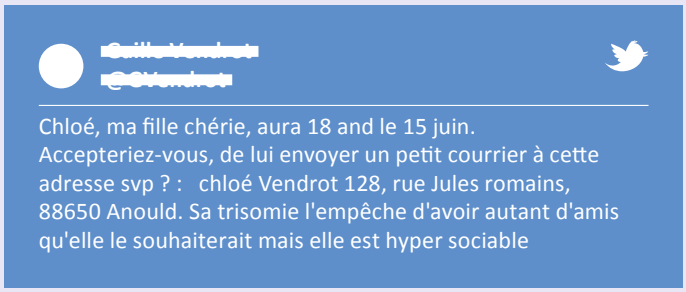


Figure 70 Message envoyé par la mère d'une adolescente trisomique (source : Twitter)

Intersectionnalité des discours haineux

Dans l'analyse des différents types de discours haineux, il est essentiel de se souvenir que les individus peuvent être la cible d'un certain nombre d'injures haineuses différentes en raison de différents aspects de leur identité.
À titres d'exemples, les Musulmans arabes sont souvent ciblés à la fois en raison de leur origine ethnique et de leur religion ; les homosexuels de couleur sont ciblés en raison de leur orientation sexuelle mais aussi de leur origine ethnique et raciale.
Quant aux femmes, elles sont victimes de différents types de haine en raison de leur sexe et d'autres catégories d'identité. Elles sont de fait particulièrement susceptibles d'être ciblées par la violence en ligne. Un rapport de novembre 2017 du Haut Conseil à l'Égalité français a révélé que 73 % des femmes avaient été victimes de violence en ligne.[58]

La façon dont des individus ou des groupes sont la cible d'attaques ou de préjugés fondés sur leurs identités multiples est communément appelée intersectionnalité,[59] ce qui fait référence aux intersections d'identités différentes.

Ce concept est particulièrement important pour comprendre comment les personnes ciblées par des propos haineux en ligne peuvent être victimes d'agressions à multiples facettes.
Il est en outre important de ne pas oublier que des internautes isolés peuvent être responsables de la diffusion d'un certain nombre de discours haineux ou polarisants.
Une analyse à plusieurs niveaux est nécessaire pour bien comprendre les tendances des discours haineux qui se manifestent en ligne puis pour trouver des solutions efficaces.

Ce chapitre présente notre analyse des intersections entre les contenus haineux ciblant différents groupes. Pour ce faire, nous avons identifié les comptes qui employaient de multiples types de discours haineux recensés par nos algorithmes de traitement du langage naturel, ce qui nous a permis d'identifier les types de discours haineux qui se recoupent fréquemment.

Le tableau suivant présente le pourcentage de recoupement dans les comptes utilisant des propos haineux de chaque discours analysé grâce au machine learning. Les chiffres indiquent le pourcentage d'utilisateurs des deux ensembles de données, en pourcentage de chaque colonne.

Tableau 5 Pourcentage de recoupement entre les utilisateurs de différents groupes utilisant des propos haineux

	Capacitistses ou validistes	Anti-Arabes	Anti-LGBTQ	Miso-gynie
Capacitistes ou validistes	100%	33%	31%	22%
Anti-Arabes	11%	100%	10%	7%
Anti-LGBTQ	33%	32%	100%	24%
Misogynie	60%	56%	63%	100%

Étant donné que les discours misogynes généralisés et banalisés sont en ligne, il n'est pas surprenant de constater que ceux qui utilisent un langage misogyne haineux représentent également plus de 50 % des comptes utilisant d'autres types de discours haineux. Les utilisateurs qui emploient un langage haineux anti-LGBTQ constituent, de la même façon, environ un tiers des messages haineux capacitistes ou validistes et anti-Arabes, soulignant une fois de plus son usage répandu.

Nous avons aussi examiné dans quelle mesure les mots-clés des quatre ensembles de données haineuses se recoupaient.
Les cartes du réseau de mots-clés illustrées par les figures 71 à 74 présentent les résultats de cette analyse. Chaque point sur les cartes représente un compte, et chaque compte est connecté à un terme que le titulaire du compte a utilisé dans un message haineux.

Les recoupements les plus importants ont été constatés entre le contenu misogyne, le contenu anti-Arabes et le contenu anti-LGBTQ

- Cinq mots-clés misogynes sont apparus dans la carte du langage anti-LGBTQ (Figure 71).
- 13 mots-clés anti-LGBTQ figuraient dans la carte misogyne (Figure 72), dont *tante, fiotte* et *pédé,* termes péjoratifs pour désigner les homosexuels. Bilal Hassani s'est avéré être un personnage central des discours haineux misogynes et anti-LGBTQ, démontrant à quel point ces types de discours haineux se recoupent et à quel point le sexe et l'orientation sexuelle sont souvent confondus, particulièrement dans les discours haineux.
- Le terme *pute* apparaît fréquemment dans les messages anti-LGBTQ et dans les messages haineux capacitistes ou validistes.
- Le terme *pd,* au centre des discours haineux anti-LGBTQ, figurait dans les quatre ensembles de données haineuses.
- Le terme *beurette,* féminin de *beur* (terme argotique qui désigne les personnes d'ascendance nord-africaine, ou les Arabes en général), était utilisé dans environ 9 % des messages de l'ensemble de données haineuses anti-Arabes (Figure 74).
- Le terme *chienne,* au centre des discours misogynes haineux, était également associé au discours haineux anti-Arabes.
- Bien que le langage haineux capacitiste ou validiste ait démontré un recoupement plus limité avec d'autres types de discours haineux, des mots comme *pédé* et *pute* sont apparus comme des termes significatifs dans cet ensemble de données. Cela démontre, de nouveau, comment ces injures sont utilisées de manière généralisée, à l'instar des injures capacitistes ou validistes comme *mongol* et *attardé.*

Notre analyse a enfin mis en évidence un amalgame entre Arabes et Musulmans, ainsi qu'entre Arabes et Islamistes dans les discours haineux. Les messages haineux ciblant les Musulmans étaient omniprésents dans l'analyse de l'ensemble de données anti-Arabes, et de nombreux messages contenaient à la fois une rhétorique anti-Arabes et anti-Musulmans.

Les figures 71 à 74 comprennent les mots-clés de nos listes initiales et d'autres termes qui peuvent être utilisés à des fins haineuses. Ces termes ont été inclus pour comprendre dans quelle mesure ils apparaissaient dans les discours haineux identifiés par nos algorithmes.

> “Les recoupements les plus importants ont été constatés entre le contenu misogyne, le contenu anti-Arabes et le contenu anti-LGBTQ”

Figure 71 Carte de réseau des mots-clés dans l'ensemble de données haineuses anti-LGBTQ

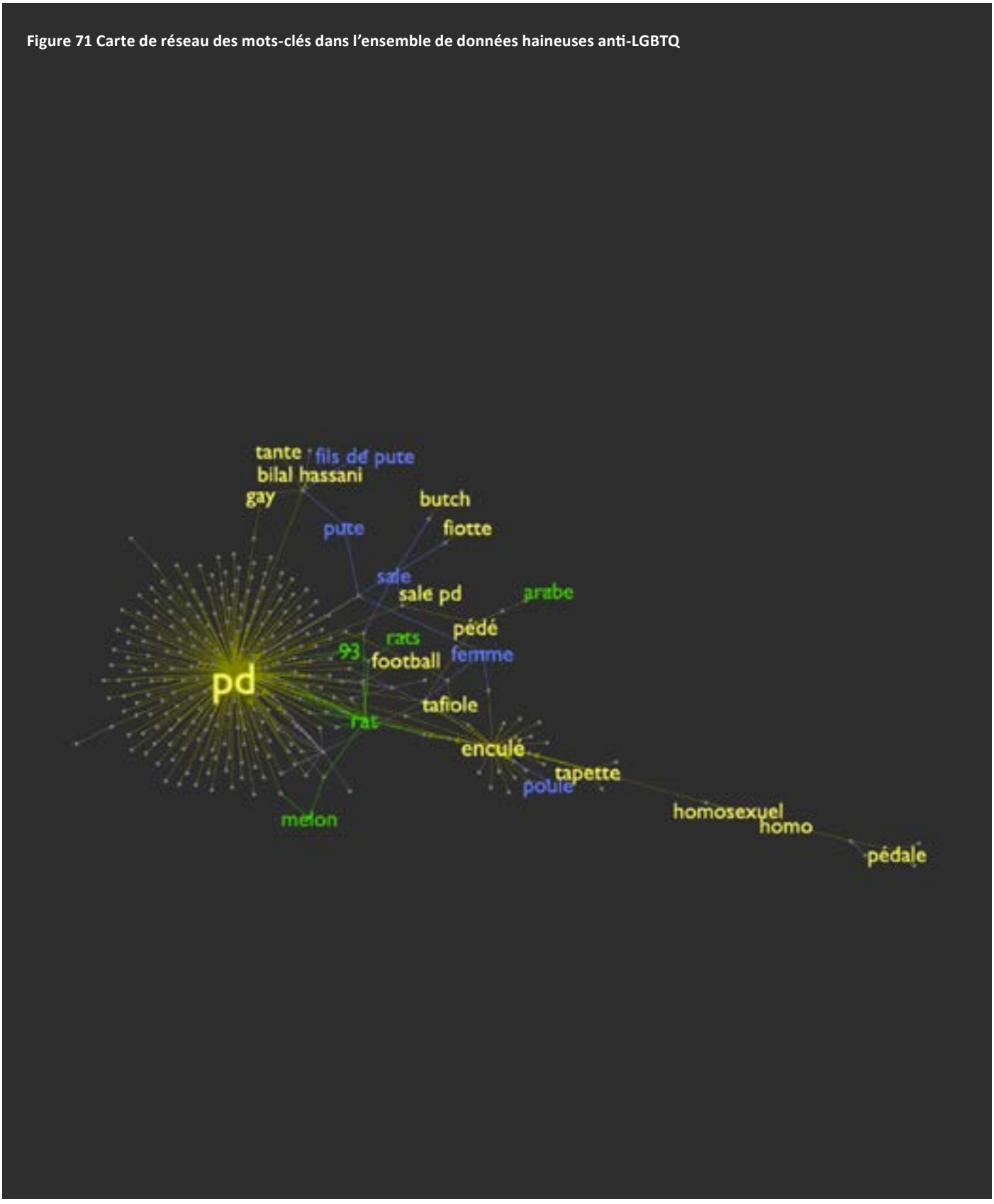


Figure 72 Carte de réseau des mots-clés dans l'ensemble de données haineuses misogynes

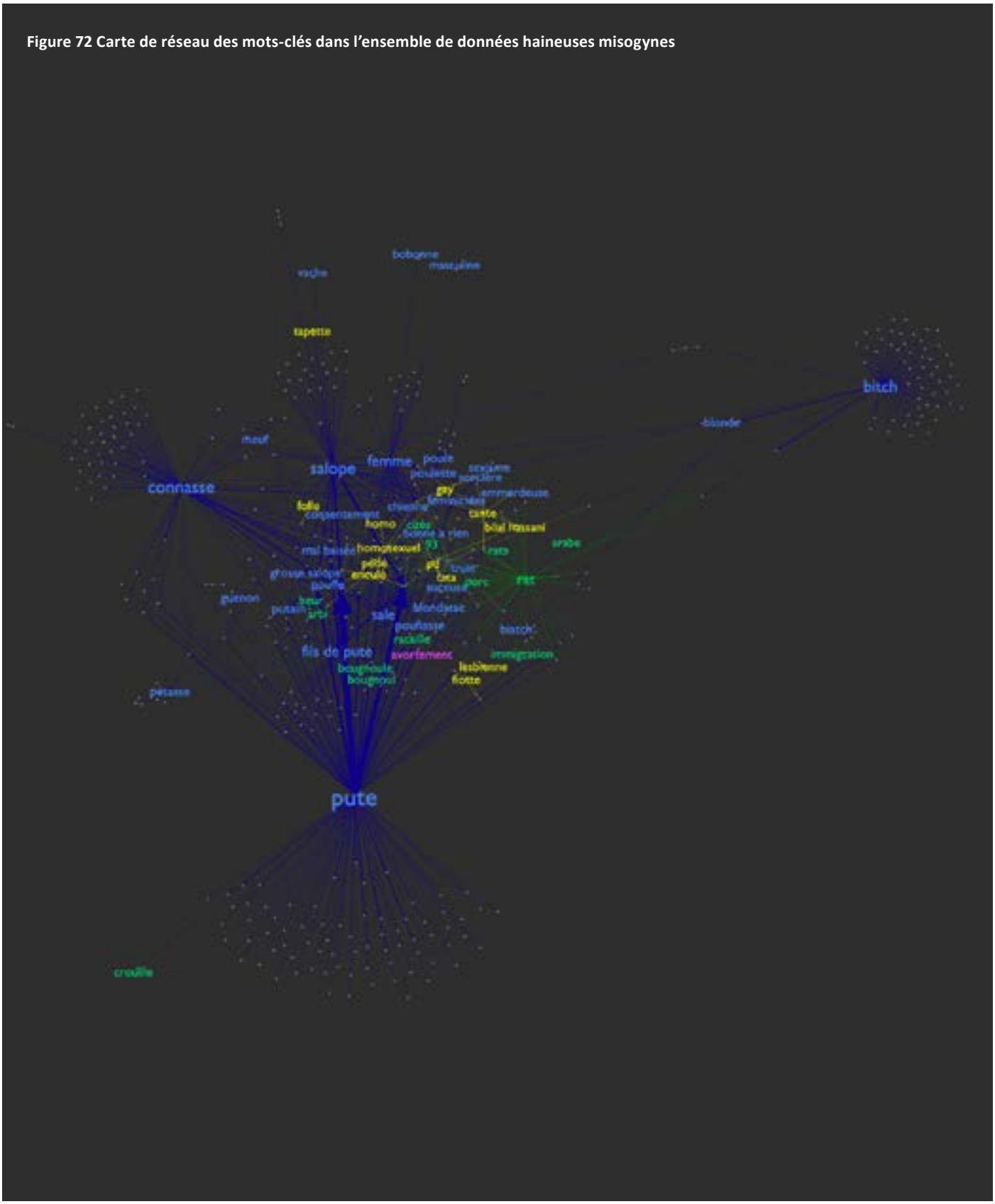


Figure 73 Carte de réseau des mots-clés dans l'ensemble de données haineuses capacitistes ou validistes

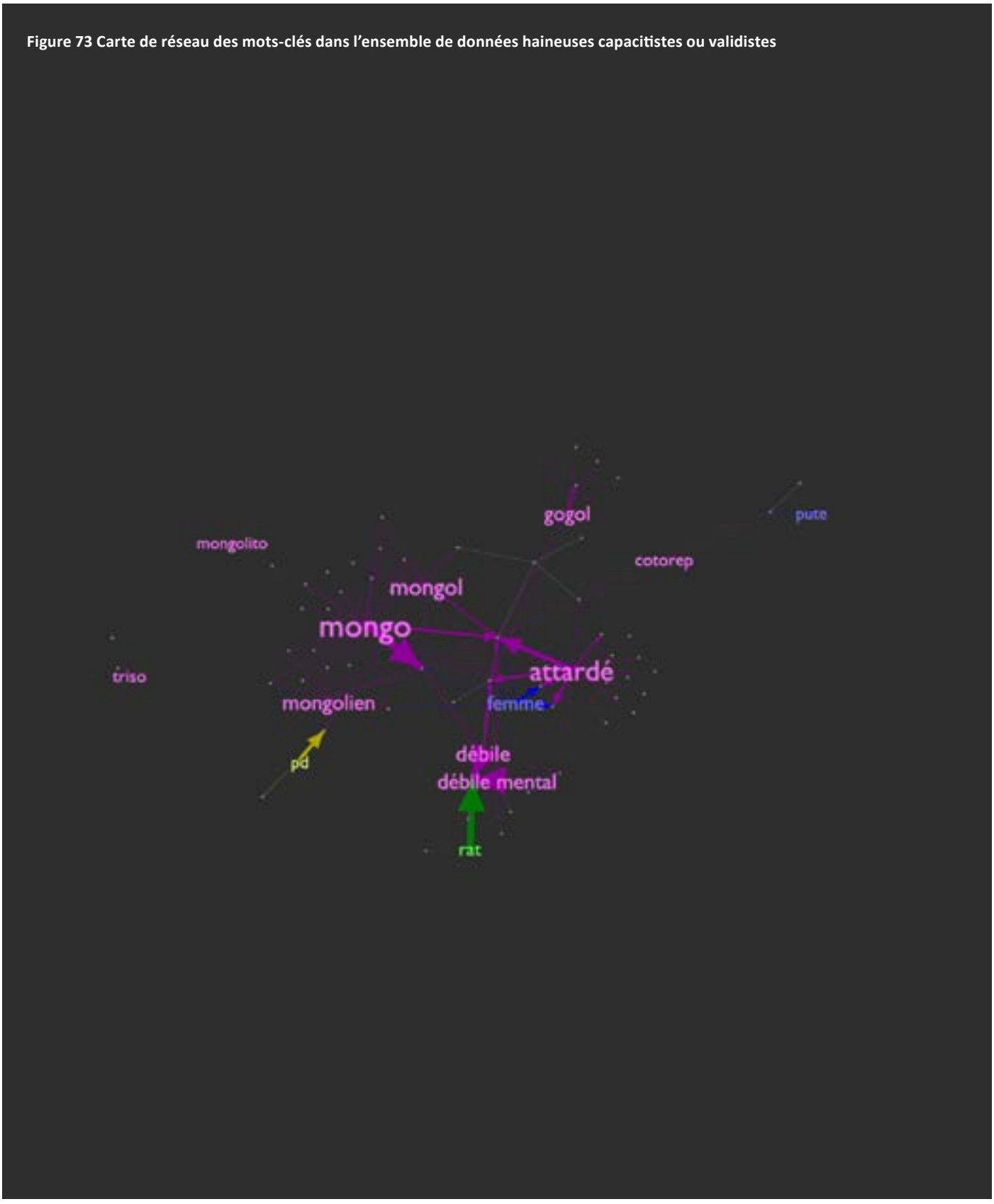
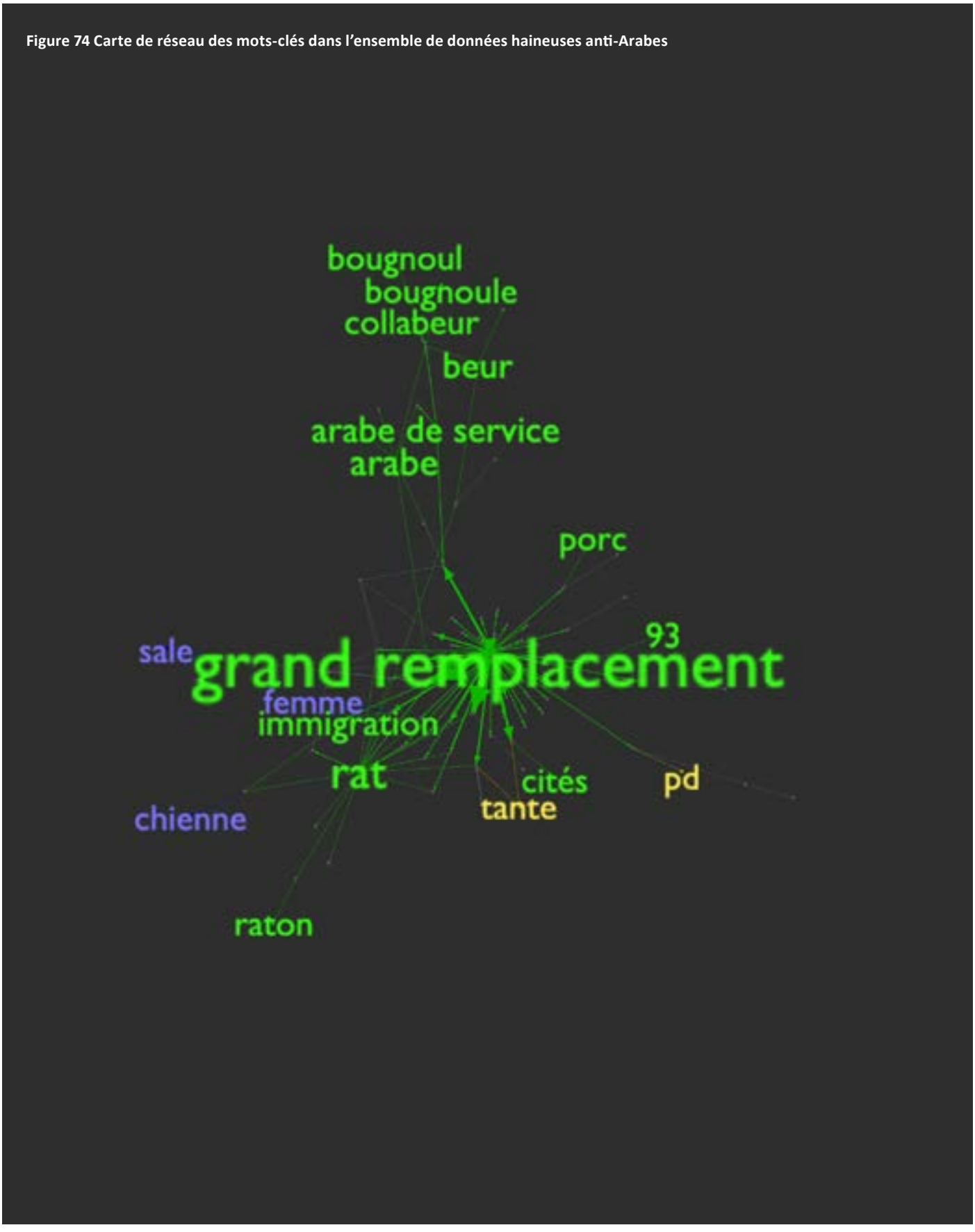


Figure 74 Carte de réseau des mots-clés dans l'ensemble de données haineuses anti-Arabes



Recommandations

Dans ce rapport, nous avons tenté de fournir une vue d'ensemble complète des discours haineux en ligne en France afin que le plus grand nombre puisse s'y référer. Nous avons également cherché à envisager tous les groupes dont les membres subissent ou affirment subir des agressions haineuses afin de comprendre la portée, la nature et les facteurs qui déclenchent ces conversations. Notre but a été d'examiner les plateformes de réseaux sociaux, dans la mesure où l'accès aux données sur ces plateformes était fourni par le biais de leurs API et d'outils informatiques commerciaux. Nous avons enfin cherché à confronter l'application du traitement du langage naturel et du machine learning à la difficulté d'identifier et d'analyser le discours haineux à grande échelle.
Bien que les plateformes elles-mêmes utilisent ces outils et ces moyens techniques, notre projet de recherche a été la première tentative de ce type dans le domaine public en France.

L'étude a révélé un certain nombre de perspectives pour le gouvernement, les plateformes en ligne, les organisations de la société civile et les chercheurs qui s'efforcent de comprendre la portée et la nature des discours haineux en ligne pour les combattre.

Notre recherche intervient à un moment charnière du contexte politique français.
En effet, comme mentionné dans l'introduction du présent rapport, la loi Avia, si elle est adoptée en ces termes en seconde lecture, exigera bientôt que les plateformes en ligne « à fort trafic » suppriment les « contenus manifestement illégaux » dans les 24 heures suivant leur notification (comme mentionné page 24 depuis la rédaction du présent rapport, le projet de loi a été examiné par la commission des lois du Sénat, qui a supprimé l'obligation faites aux plateformes de supprimer les contenus dans un délai de 24 heures). Ces contenus comprennent notamment des propos illégaux au regard du droit français, y compris la « provocation publique à la haine » et les « abus, diffamation et incitation à la discrimination, à la haine ou à la violence à l'égard d'une personne ou d'un groupe en raison de son origine, de son appartenance ou non à un groupe ethnique, national, racial ou religieux ainsi que de son orientation sexuelle ou son handicap ». La loi vise également à faciliter le signalement par les utilisateurs des contenus qui leur semblent illégaux.

> « Nos recommandations tiennent pleinement compte des efforts positifs que les entreprises de réseaux sociaux ont déployés jusqu'à maintenant »

La loi Avia représente un pilier essentiel de la démarche du gouvernement français, sans en être l'élément unique. Le Plan national contre le racisme et l'antisémitisme (2018-2020) propose un certain nombre de mesures pour « lutter contre les contenus haineux, racistes et antisémites », notamment les activités au niveau européen. En outre, l'équipe de mission interministérielle du gouvernement français, qui comprend la DILCRAH (Délégation Interministérielle à la Lutte Contre le Racisme, l'Antisémitisme et la Haine anti-LGBT - https://www.dilcrah.fr/) ainsi que sept experts de haut niveau et trois rapporteurs permanents de divers ministères,[60] a présenté ses premières réflexions sur l'établissement d'un cadre général de réglementation des réseaux sociaux, à commencer par la question des propos haineux en ligne, comme le décrit le rapport de mission intérimaire intitulé Créer un cadre français pour rendre les plateformes de réseaux sociaux responsables.[61] Les conclusions que nous tirons de la recherche présentée dans ce rapport et les recommandations que nous formulons ont été examinées en tenant compte de ce contexte politique.

Nos recommandations tiennent pleinement compte des efforts positifs que les entreprises de réseaux sociaux ont déployés jusqu'à maintenant.
Par exemple, tel que cité dans le rapport interministériel, Facebook a fourni plus de transparence sur le contenu de ses « standards de la communauté », accru les ressources destinées à la modération des contenus par des êtres humains et des machines, créé un programme de « signalement de confiance » et fourni des rapports

de transparence.
Ces efforts ont permis aux chercheurs et aux organisations de la société civile de mieux comprendre l'ampleur de la haine, ainsi que les difficultés des entreprises et les mesures qu'elles prennent.
Toutefois, pour aborder l'existence et l'impact des contenus haineux en ligne, il faudra prendre d'autres mesures, notamment placer la société civile au cœur de la réponse.

1. Les plateformes en ligne doivent accroître la transparence des communications et des contenus publics sur leur plateforme en fournissant un accès ouvert à l'interface de programmation (API) afin de mieux comprendre l'ampleur des discours haineux.
Comme indiqué dans le document intérimaire de la mission interministérielle, il existe « une asymétrie extrême des informations » entre les réseaux sociaux et la société civile, ce qui ébranle la confiance dans l'approche autorégulatrice des entreprises face aux contenus à caractère haineux. Le niveau actuel d'accès limite la capacité des chercheurs à comprendre le problème, ce qui rend difficile l'élaboration de réponses appropriées pour les gouvernements et la société civile.

Les plateformes en ligne doivent examiner comment fournir un meilleur accès aux données de communications publiques par le biais de leur API afin d'aider les chercheurs et la société civile à mieux comprendre et à combattre les propos haineux en ligne. Toutefois, il est essentiel de prendre en considération toutes les précautions relatives à la protection de la vie privée. Ainsi, il est crucial que seuls les instituts de recherche légitimes aient accès à ces données et seulement dans les cas où la recherche répond à certains critères de valeur publique et respecte les droits des utilisateurs. Plus l'accessibilité est grande, plus les risques que ces données soient utilisées par des acteurs malveillants à des fins de profilage et de ciblage en vue de semer la haine et la division, de recruter ou de calomnier, sont élevés. Les récentes modifications apportées aux conditions d'accès à l'API de Twitter peuvent servir de modèle équilibré favorisant le développement d'importants travaux de recherche d'une part, et garantissant le respect du droit à la vie privée des utilisateurs d'autre part.

En sus de l'accès à l'API, une plus grande transparence pourrait inclure des entreprises de réseaux sociaux fournissant une répartition détaillée du discours haineux qui se trouve sur leur plateforme, triée en fonction des groupes ciblés par un tel discours ainsi que des types de discours identifiés (par exemple, selon qu'il s'agisse du niveau 1, 2 ou 3 si on se réfère à la politique de Facebook sur le contenu haineux).
Dans la mesure du possible, ces informations devraient être fournies rapidement et à grande échelle aux organisations de la société civile qui luttent contre la haine.

Si ces préconisations ne sont pas possibles, alors les opérateurs de plateformes en ligne doivent développer des outils personnalisés et des algorithmes pour analyser les discours haineux de façon transparente et accessible.
Une des principales ambitions de l'OCCI est d'aider les plateformes en ligne à communiquer ces données à des organisations locales et régionales de la société civile de façon accessible et exploitable et de conseiller la société civile en matière de stratégies de réponse efficaces.

2. Les organismes de réglementation gouvernementaux et les plateformes en ligne doivent tenir compte des limites des algorithmes de *machine learning* pour identifier les contenus haineux.

> « Une plus grande transparence pourrait inclure des entreprises de réseaux sociaux fournissant une répartition détaillée du discours haineux qui se trouve sur leur plateforme »

Cette étude a mis en évidence les limites du traitement du langage naturel et du *machine learning* - en particulier des logiciels commerciaux - pour fournir une identification fiable du contenu haineux pertinent et une répartition précise des différents types de contenu haineux. Cela a des répercussions sur la capacité d'exécuter ce travail rapidement et à grande échelle.

Les images et les contenus vidéo posent tous deux d'autres défis à cette approche. L'une des études de cas les plus significatives et les plus inquiétantes de la recherche a été la campagne de désinformation en ligne qui prétendait que les communautés Rom enlevaient des enfants. Des vidéos ont commencé au niveau local pour se diffuser ensuite sur Snapchat.
Ce type de contenu vidéo n'est pas aisément analysable par mots-clés ou par approches basées sur le langage. Il est primordial d'investir davantage d'efforts dans l'analyse algorithmique de ces contenus visuels tout en garantissant l'intégrité de ce processus d'analyse en matière de protection de la vie privée. Par exemple, certains mécanismes pourraient permettre l'analyse de contenus vidéo associés à des légendes appelant à la violence ou utilisant un contenu dégradant, ou des injures contre des catégories protégées, plutôt que l'analyse de tout le contenu vidéo téléchargé, ce qui n'est ni pratique ni fiable et pose des problèmes de confidentialité.

Alors que les gouvernements commencent à réglementer les opérateurs de plateformes en ligne et que ces derniers s'appuient de plus en plus sur des algorithmes pour identifier et modérer les propos haineux, toutes les parties doivent reconnaître les limites de cette approche et l'importance que le contexte joue dans la modération du contenu haineux. L'intelligence artificielle ne devrait pas être considérée comme une panacée et les méthodes de modération du discours haineux se doivent d'adopter des visions et approches globales pour être efficaces.

Les efforts doivent se concentrer sur l'application de modèles de modération plus centrés sur l'être humain, donner la priorité à l'expérience personnelle des utilisateurs en matière de bien-être et de capacité à s'exprimer en toute sécurité, offrir davantage de possibilités de personnalisation de la modération, dans les limites des cadres juridiques existants.

Bien que les modèles de modération des opérateurs de plateformes en ligne comportent déjà des niveaux d'examens réalisés par des machines et des êtres humains, il reste possible d'introduire des nouveaux processus d'apprentissage et de classification du langage humain bien plus robustes et approfondis.
Ces processus améliorés doivent être combinés à des niveaux supplémentaires d'analyse de l'intelligence artificielle s'appuyant sur une gamme plus complète de facteurs à observer systématiquement, y compris une meilleure prise en compte de l'interprétation des utilisateurs (catégorisation) et une observation humaine et algorithmique de la nature des interactions entre les utilisateurs impliqués.
Ce point est particulièrement primordial dans le contexte du harcèlement et permettra aux utilisateurs en ligne de contrôler le contenu des zones grises d'un point de vue juridique, un contenu pouvant être pourtant très perturbant et indésirable d'un point de vue personnel, avec les risques potentiels associés, allant de leur impact mental à la possibilité que l'utilisateur s'autocensure ou réplique des comportements malveillants ou extrémistes, etc.

3. Les plateformes en ligne doivent travailler main dans la main avec les organisations de la société civile pour lutter contre les contenus à caractère haineux qui sont légaux, mais néanmoins problématiques et dangereux.
La loi Avia imposera aux plateformes en ligne l'obligation de supprimer les contenus haineux manifestement illicites. D'autres partenariats avec les organisations de la société civile doivent être créés en parallèle afin de lutter contre le gros volume de contenus haineux qui ne sont pas considérés comme illicites.

Dans ce contexte, les réseaux sociaux doivent entreprendre des recherches sur l'expérience et les points de vue de groupes qui font souvent l'objet de propos haineux en ligne ou qui sont à la base d'injures généralisées. Ces études aideront les réseaux sociaux à adopter une approche axée sur l'utilisateur pour comprendre les types de contenus haineux les plus nuisibles afin d'aider à prendre des décisions quant à la priorisation de la modération de contenu. Elles peuvent aussi aider les entreprises à développer et à appliquer une gamme de solutions correspondant à la sévérité des sanctions, que ce soit la suppression ou la minimisation de la portée, des rappels ciblés des termes

de service, des initiatives éducatives, des contre-discours ou un soutien aux victimes. Les organisations de la société civile peuvent aider les entreprises à prendre des décisions sur les réponses appropriées aux différents types de contenus haineux. Cela contribuera, nous en sommes persuadés, à renforcer la confiance et les partenariats entre les entreprises de réseaux sociaux et les groupes de la société civile.

Les plateformes en ligne doivent également travailler avec les organisations de la société civile pour expérimenter et tester une série de techniques d'engagement direct qui vont au-delà de la simple création et diffusion de campagnes de contre-discours. Nos recherches démontrent que le changement d'attitude et le recrutement idéologique se font moins par le contenu que par un engagement en ligne sur des forums, dans des groupes ou par la messagerie directe. Ainsi, le programme de l'ISD Counter Conversations a démontré qu'un engagement direct peut mener à une conversation soutenue avec une personne qui partage publiquement un contenu haineux.[62] La prochaine génération de contre-discours efficaces et d'interventions en ligne se déroulera dans les « espaces itératifs » en ligne où se réunissent ceux qui sont attirés par les idéologies extrémistes : dans les fils de commentaires et les forums aux contenus de « zone grise » pour lesquels il n'existe pas de bases solides pour les supprimer.

4. Les plateformes en ligne doivent faire preuve d'une transparence accrue sur leurs politiques et approches de modération, ainsi que sur le rôle des algorithmes et des comptes automatisés dans la diffusion de contenus haineux.

Il est difficile, pour les personnes externes aux entreprises, d'évaluer l'efficacité du signalement de contenus haineux et du processus d'évaluation et de suppression de ces contenus, bien que des efforts aient été déployés dans le cadre des exercices de surveillance du code de conduite de l'UE sur la lutte contre les propos haineux illégaux en ligne.[63]

En raison de ce manque de données, il est difficile de déterminer si les différences identifiées dans l'ampleur du discours haineux ciblant certains groupes sont attribuables à des politiques de modération plus efficaces. Le manque de transparence expose les entreprises à la critique potentielle que leurs politiques de modération de contenu protègent efficacement certains groupes plus que d'autres.

Pour répondre à ces questions, les entreprises technologiques doivent lever le voile sur leurs pratiques de modération et amener les chercheurs et les organisations de la société civile à comprendre leurs approches et certains des défis auxquels ils sont confrontés.

En réponse au *Livre blanc sur les préjudices en ligne du Royaume-Uni*, l'ISD a présenté une série de recommandations et d'arguments en faveur d'une transparence accrue de la part des réseaux sociaux dans quatre catégories : contenu et communications, publicité, plaintes et recours et algorithmes.

La transparence des processus de modération des contenus et une surveillance accrue de la part d'un organisme de réglementation sont nécessaires pour s'assurer que les processus sont appropriés, exacts et dotés de ressources suffisantes, comme l'indique le rapport intérimaire de la mission interministérielle. Cette transparence doit porter sur l'ampleur et la nature des plaintes des utilisateurs concernant les contenus haineux et les mesures prises en réponse à ces plaintes. Les entreprises doivent veiller à avoir une expertise interne au sujet de ces questions et un dialogue systématique avec les groupes fréquemment ciblés. Il est également important pour les entreprises technologiques d'assurer une plus grande transparence sur les conditions de travail et le soutien moral apporté aux modérateurs de contenu.

Il est enfin essentiel d'accroître la transparence sur le rôle des algorithmes dans l'amplification des contenus susceptibles d'être haineux, en particulier lors des événements où l'on a constaté une augmentation de l'ampleur des contenus à caractère haineux.

Un organisme de réglementation pourrait ainsi exiger cette transparence des plateformes en ligne. Dans le même ordre d'idées, il est nécessaire de faire preuve d'une plus grande transparence dans la compréhension du rôle des comptes automatisés dans la diffusion de contenus haineux.

Notre recherche a révélé qu'un grand nombre de comptes de type bot diffusaient du contenu haineux, y compris de nombreux comptes qui semblaient avoir été supprimés par des plateformes de réseaux sociaux à une date ultérieure.

De fait, une meilleure compréhension du phénomène

et de l'impact des contenus haineux en ligne exige une transparence accrue sur ces questions.

5. Les plateformes en ligne et les organisations de la société civile doivent collaborer à des campagnes efficaces pour lutter contre l'utilisation généralisée et normalisée des insultes dans la société.
Bien que Facebook et Twitter abordent le sujet des injures qui « ciblent négativement » ou qui sont « non consensuelles » dans leurs standards de la communauté, l'utilisation répandue de ces termes, telle que présentée dans ce rapport, peut rendre extrêmement difficile l'identification des cas à grande échelle. La situation est compliquée par le fait que certains groupes se sont réappropriés ces termes.
Un effort visant à éliminer de telles injures à grande échelle entraînerait une réaction très négative et serait excessivement draconien.
En même temps, comme le démontre notre recherche, il est possible d'identifier des utilisations haineuses et ciblées de ces injures qui pourraient aider à éclairer les campagnes visant à les combattre.

Les campagnes menées par la société civile pour dénoncer les injures haineuses ou ciblées doivent inclure des campagnes de récupération ou de réappropriation des termes utilisés comme injures, des programmes éducatifs pour aborder leur utilisation dans les jeunes communautés et des programmes de communication visant à souligner l'impact négatif que des injures normalisées peuvent avoir sur certaines communautés.

Les personnes qui lancent ces efforts doivent pouvoir s'appuyer sur les meilleures pratiques en matière de compréhension des campagnes de changement de comportement, y compris celles qui ont permis de lutter avec succès contre les insultes normalisées ou le racisme ordinaire.
Elles doivent se concentrer sur les communautés où ces insultes sont répandues, y compris parmi les supporters de football et la communauté des *gamers* ou fans de jeux vidéo en ligne.

Les praticiens doivent également être conscients du risque de réaction négative : une campagne mal conçue et mal ciblée pourrait se révéler contre-productive et finir par réduire à néant l'utilisation de telles injures par certaines personnes dans le cadre d'une culture internet subversive qui s'oppose au « politiquement correct »

6. Les opérateurs de plateformes en ligne, le gouvernement et les chercheurs doivent tenir compte davantage de la nature intersectionnelle des discours de haine.
Des recherches plus approfondies doivent être menées sur l'aspect intersectionnel des discours de haine.
Jusqu'à présent, très peu d'études ont examiné la nature intersectionnelle du discours haineux. Nos recherches ne donnent qu'un aperçu de ces tendances.
De plus gros efforts et investissements sont nécessaires pour soutenir des recherches supplémentaires qui s'avèrent nécessaires.
L'accent mis sur le discours haineux intersectionnel doit s'étendre au personnel des entreprises de réseaux sociaux et aux militants de la société civile.
Ces catégories d'acteurs doivent être conscientes des aspects intersectionnels du discours haineux et de son impact sur différentes populations.
Les campagnes ne doivent pas se concentrer exclusivement sur un type de discours haineux car les différents types de discours haineux ne sont pas clairement délimités.
Les militants de la société civile doivent tenir compte des différents types de discours haineux utilisés pour cibler les communautés. Cela s'applique non seulement

> « Les campagnes menées par la société civile pour dénoncer les injures haineuses ou ciblées doivent inclure des campagnes de récupération ou de réappropriation des termes utilisés comme injures »

aux campagnes de communication, mais aussi à d'autres types d'intervention.
Nous appelons ces mêmes militants à former des coalitions pour aborder plus efficacement les aspects intersectionnels de la haine, en faisant fi des lignes idéologiques traditionnelles.

Les plateformes de réseaux sociaux doivent enfin tenir compte de l'aspect intersectionnel du discours haineux lorsqu'elles créent et mettent en œuvre des politiques de modération de contenu.
Si les comptes sont identifiés comme étant des récidivistes, ils doivent faire l'objet de recherches plus approfondies pour savoir s'ils propagent ou non d'autres types de haine.

7. Les médias, le gouvernement, les autorités locales, la police et les plateformes en ligne doivent créer un mécanisme coordonné pour réagir aux événements qui ont tendance à provoquer des pics de discours haineux.
Ce rapport a démontré comment des événements particuliers peuvent susciter des discussions haineuses. Les entreprises technologiques doivent être préparées à ce genre de pics.
Les militants de la société civile doivent se préparer aux pics de discours haineux qui peuvent survenir à la suite d'événements d'actualité importants. Ils pourraient s'inspirer, pour ce faire, du Protocole sur les incidents liés au contenu du GIFCT pour prévenir le partage de contenu à caractère violent et terroriste et tirer parti de l'Initiative de lutte contre le crime organisé de l'Ontario.

Par exemple, les entreprises technologiques pourraient fournir aux organisations de la société civile davantage de données sur l'ampleur et la nature du discours haineux qui a lieu à la suite d'événements particuliers, fournir gratuitement des crédits publicitaires et des conseils aux organisations de la société civile qui produisent des campagnes de contre-discours après leur survenue.

8. Une plus grande attention doit être accordée à la relation entre les contenus haineux en ligne et les crimes ou incidents haineux hors ligne (tels que les attaques).
Ces efforts doivent inclure d'autres travaux de recherche sur ces phénomènes pour déterminer s'il existe une corrélation entre eux.

L'ISD œuvre actuellement à la mise au point de capacités de cartographie de la haine en ligne par géolocalisation afin d'aider la société civile locale et les gouvernements infranationaux à mieux comprendre l'ampleur et la nature du contenu haineux en ligne provenant de leur région.
Cette capacité a d'abord été développée par CASM Consulting LLP, partenaire technologique de l'ISD, en partenariat avec le bureau du maire de Londres pour la police et la criminalité et la Metropolitan Police : en plus de pouvoir identifier et géolocaliser avec certitude les discours haineux en ligne au niveau des arrondissements londoniens, l'étude a démontré une corrélation entre les statistiques sur les crimes haineux en ligne et hors ligne.
Cette fonctionnalité peut aider le personnel des administrations locales et régionales à prévoir quand et où certaines communautés risquent davantage d'être la cible d'attaques haineuses.
Grâce à cette fonctionnalité, les autorités peuvent canaliser plus efficacement les ressources nécessaires pour prévenir de telles attaques, par exemple en mettant en place une police préventive de proximité, en sensibilisant la population au signalement d'attaques haineuses et en soutenant les victimes.

Les statistiques officielles de la police française sur les crimes motivés par la haine fournissent déjà une « carte sensible » montrant la répartition des discours de haine en France.
Des efforts doivent être faits pour examiner comment la répartition géographique des crimes haineux en France correspond à celle des discours de haine en ligne.
La police française doit envisager d'inclure une nouvelle catégorie de crimes haineux en ligne dans ses statistiques officielles. En effet, au Royaume-Uni, il est, depuis 2015, obligatoire d'inclure de telles statistiques.

Si la police française était tenue de signaler les crimes de haine en ligne, les données et les statistiques accumulées au fil du temps pourraient fournir une base pour comprendre la relation entre les contenus haineux en ligne et les attaques motivées par la haine hors ligne.

Annexes

Annexe 1 : Listes de mots-clés

Les listes de mots-clés créés pour ce rapport d'étude figurent ci-dessous, comme saisies dans les deux outils d'écoute sociale mentionnés dans la section sur la méthodologie (Crimson Hexagon et CrowdTangle).

Mots-clés anti-Arabes ou anti-Maghrébins

Crimson Hexagon
beur* OU beurette* OU rabza* OU beureu* OU bicot* OU raton* OU bounioul* OU bougnoule* OU bougnoul* OU racaille* OU crouille* OU blédard* OU métèque* OU « arabe de service » OU « sale arabe » OU « arabe voleur » OU « arabe pas français » OU « arabe étrangers » OU « grand remplacement »
CrowdTangle
beur, beurette, rabza, beureu, bicot, raton, bounioul, bougnoule, bougnoul, racaille, crouille, blédard, métèque, arabe de service, sale arabe, arabe voleur, arabe pas français, arabe étrangers, grand remplacement

Mots-clés anti-Noirs ou anti-Africains

Crimson Hexagon
« sale noir » OR négroide* OR congoide* OR bamboula* OR « face de pygmée » OR macaque* OR métèque* OR banania* OR nègre* OR négresse* OR « noir de service » OR « nègre de maison » OR « race inférieur » OR babouin*
CrowdTangle
sale noir, négroide, congoide, bamboula, face de pygmée, macaque, métèque, banania, nègre, négresse, noir de service, nègre de maison, race inférieur, babouin

Mots-clés anti-Roms ou anti-Gitans

Crimson Hexagon
tsigane* OR tzigane* OR romanichel* OR romanichelle* OR sinté* OR sintée* OR manouche* OR gitan* OR gitane* OR bohémien* OR bohémienne* OR caraque* OR « noi » OR manouche* OR manouches* OR yéniches* OR « race nomade » OR « sale roumaine » OR « sale roumain »
CrowdTangle
tsigane, tzigane, romanichel, romanichelle, sinté, sintée, manouche, gitan, gitane, bohémien, bohémienne, caraque, noi, manouche, manouches, yéniches, race nomade, sale roumaine, sale roumain

Mots-clés anti-Asiatiques

Crimson Hexagon
niakoué* OR niakouée* OR niaqué* OR niaquée* OR niaquoué* OR niaquouée* OR chinetoque* OR chinetoc* OR chinetok* OR noich* OR noichi* OR « face de citron » OR bridé* OR « sale asiat » OR « sale chinois »
CrowdTangle
niakoué, niakouée, niaqué, niaquée, niaquoué, niaquouée, chinetoque, chinetoc, chinetok, noich, noichi, face de citron, bridé, sale asiat, sale chinois

Mots-clés anti-Blancs

Crimson Hexagon
babtou OR toubab
CrowdTangle
babtou, toubab

Mots-clés anti-Musulmans

Crimson Hexagon
muzz OR « envahisseur musulman » OR « immigration islamique » OR islamisation OR « musulman terroriste » OR « islam terrorisme » OR « danger islam » OR « menace islamiste » OR « musulman délinquant »
CrowdTangle
muzz, envahisseur musulman, immigration islamique, islamisation, musulman terroriste, islam terrorisme, danger islam, menace islamiste, musulman délinquant

Mots-clés antisémites

Crimson Hexagon
youpin* OR youpine* OR feuj* OR « sale juif » OR « sale juive » OR juiverie OR « complot juif » OR « mafia juives » OR hoananas OR « pornographie mémorielle » OR shoax OR « propagande sioniste » OR « juif voleur » OR « juif rothschild » OR « juif franc-macon » OR « juif traitre » OR « sous race de juif'
CrowdTangle
youpin, youpine, feuj, sale juif, sale juive, juiverie, complot juif, mafia juives, hoananas, pornographie mémorielle, shoax, propagande sioniste, juif voleur, juif rothschild, juif franc-macon, juif traitre, sous race de juif.

Mots-clés anti-Chrétiens

Crimson Hexagon
mécréant* OR kouffar* OR kâffir* OR bigot* OR bigotte* OR « grenouille de bénitier » OR « sale catho »
CrowdTangle
mécréant, kouffar, kâffir, bigot, bigotte, grenouille de bénitier, sale catho

Mots-clés misogynes

Crimson Hexagon
« féministe hystérique » OR salope* OR pute* OR biatch* OR bitch* OR poufiasse* OR connasse* OR emmerdeuse* OR pétasse* OR salasse* OR grognasse* OR guenon* OR « mal baisée » OR blondasse* OR « femme hystérique » OR « féministe de service » OR poulette* OR tepu* OR pouffe* OR « bonne à rien » OR chaudasse* OR « meuf vulgaire » OR « féministe totalitaire » OR « grosse chienne » OR gonzesse* OR « sale meuf » OR « chienne féministe » OR bobonne*
CrowdTangle
féministe hystérique, salope, pute, biatch, bitch, poufiasse, connasse, emmerdeuse, pétasse, salasse, grognasse, guenon, mal baisée, blondasse, femme hystérique, féministe de service, poulette, tepu, pouffe, bonne à rien, chaudasse, meuf vulgaire, féministe totalitaire, grosse chienne, gonzesse, sale meuf, chienne féministe, bobonne

Mots-clés LGBTQ

Crimson Hexagon
pédé* OR pd OR pédale* OR tapette* OR tarlouze* OR tantouze* OR tafiole* OR fiotte* OR gouine* OR gouinasse* OR camionneuse* OR butch* OR goudou* OR follasse* OR travelo* OR enculé* OR pédéraste* OR lopette* OR « broutte minou » OR « sale pd » OR « sale gay »
CrowdTangle
pédé, pd, pédale, tapette, tarlouze, tantouze, tafiole, fiotte, gouine, gouinasse, camionneuse, butch, goudou, follasse, travelo, enculé, pédéraste, lopette, broutte minou, sale pd, sale gay

Mots-clés capacitistes ou validistes

Crimson Hexagon
mongo* OR mongol* OR mongolien* OR mongolito* OR gogol* OR attardé* OR « espèce d'handicapé » OR cotorep* OR triso* OR « débile mental »
CrowdTangle
mongo, mongol, mongolien, mongolito, gogol, attardé, espèce d'handicapé, cotorep, triso, débile mental

Annexe 2 : Method52

Technologie

La technologie de base utilisée dans ce projet s'appelle Method52. Elle a été conçue pour permettre aux personnes qui n'ont pas de formation formelle en sciences des données de collecter, d'analyser et de visualiser des ensembles de données très volumineux et non structurés. C'est particulièrement le cas pour les grands ensembles de données composés de textes, notamment ceux tirés des réseaux sociaux, mais aussi pour les ensembles de données composés d'emails, de données provenant de forums et de données internes et exclusives détenues par de grandes organisations.

Le principe de conception de Method52 est de créer un environnement de développement via une interface utilisateur graphique. Les utilisateurs sélectionnent, configurent et connectent un certain nombre d'éléments pour créer une base de données de travail sur mesure dans laquelle les données circulent. Chacune de ce bases de données est conçue pour accomplir une tâche particulière. Souvent, un certain nombre de bases de données sont elles-mêmes reliés entre elles pour remplir une fonction particulière axée sur la recherche. Method52 comprend 82 éléments dont beaucoup d'entre eux peuvent être configurés pour effectuer un certain nombre de tâches différentes.

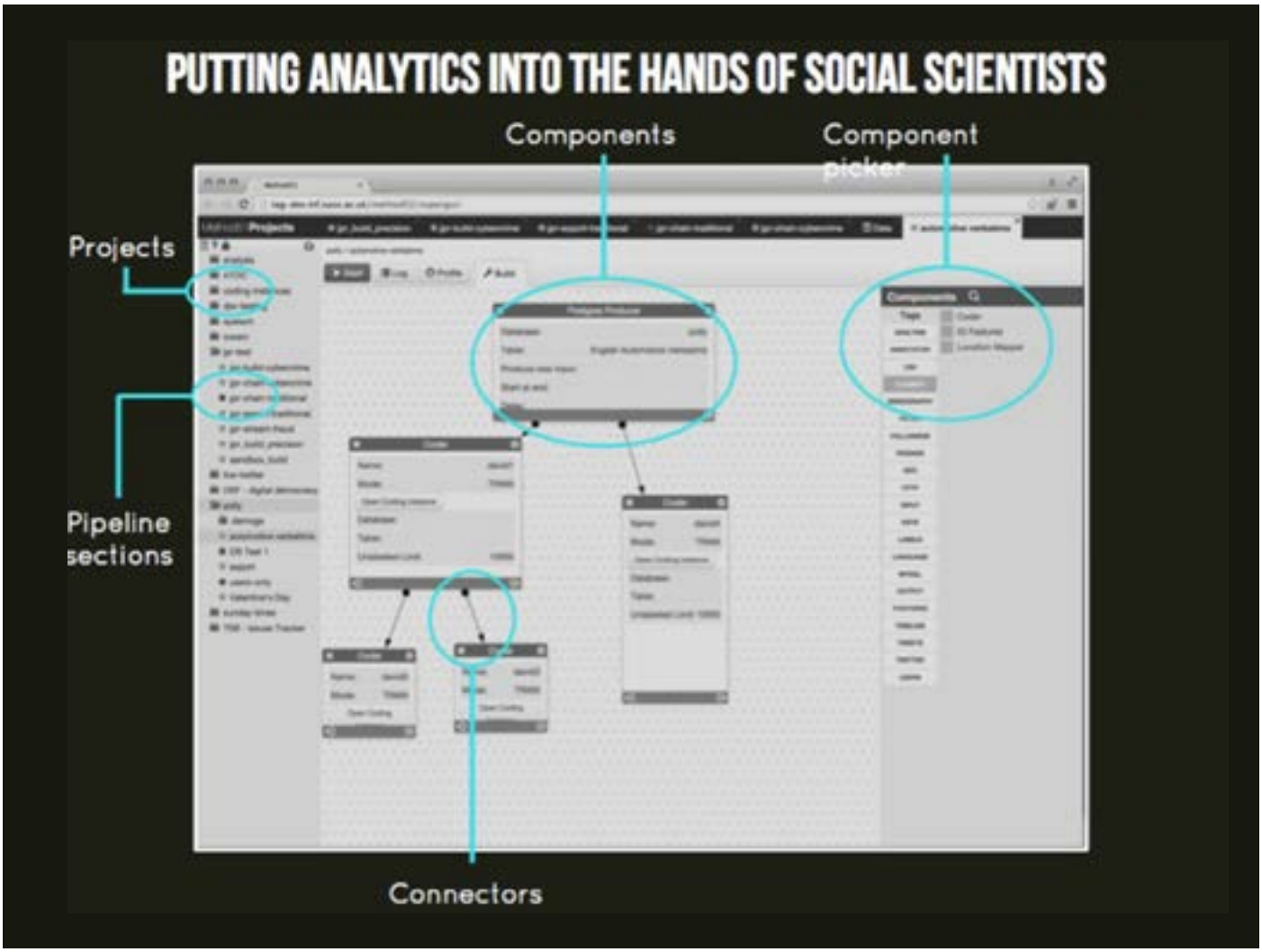


Figure 77 Exemple de pipeline Method52

Classification des thèmes

Method52 permet aux chercheurs de former des algorithmes pour diviser (« classer ») les documents en catégories, selon la signification des documents et sur la base du texte qu'ils contiennent.
Dans ce but, Method52 utilise une technologie appelée traitement du langage naturel, une branche de la recherche en intelligence artificielle. Elle combine des approches développées dans les domaines de l'informatique, des mathématiques appliquées et de la linguistique.

Tout au long de ce projet, les chercheurs de l'ISD ont annoté les messages selon qu'ils les considéraient comme faisant partie ou non des catégories que chaque algorithme a été formé à distinguer. Cette méthode « enseigne » à l'algorithme à repérer les tendances dans l'utilisation de la langue associée à chaque catégorie choisie.
L'algorithme recherche alors des corrélations statistiques entre le langage utilisé et les catégories attribuées pour déterminer dans quelle mesure les mots et les digrammes entrent dans les catégories prédéfinies.

Pour mesurer l'exactitude des algorithmes dans les catégories choisies par l'analyste, nous avons utilisé une approche dite « gold standard » (référence absolue) : pour chaque algorithme, environ 100 documents ont été sélectionnés aléatoirement à partir de l'ensemble de données pertinent pour former un ensemble de test « gold standard » pour chaque classificateur.
Ceux-ci sont ensuite codés manuellement dans l'une des catégories définies ci-dessus. Les 100 documents sont ensuite supprimés de l'ensemble de données principal et ne sont donc pas utilisés pour former le classificateur.
À mesure que chaque classificateur est formé, le logiciel rend compte de la précision avec laquelle le classificateur classe le « gold standard » par rapport à la décision de l'analyste.

Il existe plusieurs façons de mesurer la précision des classificateurs. Nous avons choisi d'utiliser celle dite de l'exactitude globale : le pourcentage de probabilité qu'un document sélectionné aléatoirement dans l'ensemble de données soit placé dans la catégorie appropriée par l'algorithme.

Processus de formation

La création d'algorithmes pour catégoriser et séparer les messages a joué un rôle important dans la méthode de recherche utilisée dans le présent rapport.
Cette méthode répond à un défi général de la recherche sur les réseaux sociaux, les données produites et recueillies régulièrement étant trop volumineuses pour être traitées manuellement.

Les classificateurs de traitement du langage naturel fournissent une fenêtre d'analyse sur ces types d'ensembles de données. Ils sont formés par des analystes sur un ensemble spécifique de données pour reconnaître la différence linguistique entre différents types de données. Cette formation a été dispensée à l'aide de Method52.

Ainsi, chaque classificateur a été créé à l'aide de l'interface utilisateur Web de Method52 pour passer par les huit phases décrites ci-dessous.

Phase 1 : Définition des catégories

Les critères officiels expliquant comment les messages doivent être annotés sont élaborés. En pratique, cela signifie qu'un petit nombre de catégories (entre deux et cinq) sont définies. Ce seront les catégories dans lesquelles le classificateur tentera de placer chaque message. La définition exacte des catégories se développe tout au long de l'interaction précoce des données. Ces catégories ne sont pas établies a priori, mais de façon itérative, en fonction de l'interaction du chercheur avec les données. L'idée que le chercheur se fait d'une catégorie est souvent contestée par les données elles-mêmes, ce qui entraîne une redéfinition de cette catégorie.
Ce processus permet de s'assurer que les catégories reflètent les données probantes plutôt que les idées préconçues ou les attentes de l'analyste.
Cette idée est conforme à une méthode sociologique bien connue nommée « théorie ancrée ».

Phase 2 : Création d'un ensemble de données test « Gold standard »

Cette phase fournit une source de vérité par rapport à laquelle la performance du classificateur est testée.
Un certain nombre de messages (généralement 100, plus si l'ensemble de données est très volumineux) sont sélectionnés aléatoirement pour former un ensemble de test « Gold standard ». Ceux-ci sont codés

manuellement dans les catégories définies au cours de la Phase 1. Les messages constituant ce « Gold standard » sont ensuite retirés de l'ensemble de données principal et ne sont pas utilisés pour former le classificateur.

Phase 3 : Formation
Cette phase décrit le processus par lequel les données de formation sont introduites dans le modèle statistique, nommé « balisage ». Dans le cadre d'un processus appelé « apprentissage actif », chaque message non étiqueté de l'ensemble de données est évalué par le classificateur en fonction du niveau de confiance relatif à la présence du message dans la bonne catégorie. Le classificateur sélectionne le message ayant le score de confiance le plus faible et le présente à l'analyste humain via une interface utilisateur de Method52.
L'analyste lit alors chaque message et décide à quelle catégorie pré affectée (voir Phase 1) il doit appartenir.
Un petit groupe d'entre eux (généralement une dizaine) sont soumis comme données de formation et le modèle de traitement du langage naturel est recalculé.
L'algorithme de traitement du langage naturel recherche ensuite des corrélations statistiques entre le langage utilisé et le sens exprimé pour arriver à une série de critères fondés sur des règles et présente au chercheur un nouvel ensemble de messages pour lesquels il a un faible niveau de certitude dans le modèle recalculé.

Phase 4 : Examen de la performance et modification
Le classificateur mis à jour est ensuite utilisé pour classer chaque message dans l'ensemble de test « Gold standard ». Les décisions prises par le classificateur sont comparées aux décisions prises (en Phase 2) par l'analyste humain. Sur la base de cette comparaison, des statistiques de performance des classificateurs - « rappel », « précision » et « global » - sont créées et évaluées par un analyste humain.

Phase 5 : Nouvelle formation
Les Phases 3 et 4 sont répétées jusqu'à ce que le rendement du classificateur cesse d'augmenter. Cet état est appelé « plateau » et, lorsqu'il est atteint, il est considéré comme la performance optimale pratique qu'un classificateur puisse raisonnablement atteindre. Le plateau se produit généralement dès 200-300 messages annotés, bien que cela dépende du scénario : plus la tâche est complexe, plus les données de formation sont nécessaires.

Phase 6 : Traitement
Lorsque la performance du classificateur a atteint le « plateau », le modèle de traitement du langage naturel est utilisé pour traiter tous les messages restants de l'ensemble de données dans les catégories définies pendant la Phase 1, en utilisant des règles déduites des données sur lesquelles l'algorithme a été formé.
Le traitement crée une série de nouveaux ensembles de données - un pour chaque catégorie de signification - contenant chacun les messages considérés par le modèle comme relevant le plus probablement de cette catégorie.

Phase 7 : Création d'un nouveau classificateur (Phase 1) ou d'une analyse post-traitement (Phase 8)
En pratique, les classificateurs sont conçus pour fonctionner ensemble : chacun est capable d'effectuer une tâche assez simple à très grande échelle (filtrer les messages pertinents des messages non pertinents, trier les messages en catégories générales de sens ou séparer les messages contenant un type de message clé de ceux en contenant un autre).
Lorsque les classificateurs fonctionnent ensemble, on les appelle « cascade ».
Des cascades de classificateurs ont été utilisées dans cette étude. Une fois la phase 7 terminée, il faut décider s'il faut retourner à la phase 1 pour construire le prochain classificateur au sein de la cascade ou, si la cascade est terminée, pour passer à la phase finale, à savoir l'analyse post-traitement.

Phase 8 : Analyse post-traitement
Une fois que les messages ont été traités, les nouveaux ensembles de données sont systématiquement analysés et évalués à l'aide d'autres techniques.

Notes sur la formation

Langue étrangère
La portée de cette recherche est géographiquement localisée en France, dans le but d'identifier les tendances haineuses dans la langue française, de sorte que tout contenu qui était entièrement en langue étrangère a été classé comme non pertinent pour cette recherche. L'exemple le plus parlant est celui de l'ensemble de données anti-LGBTQ, car un grand parti politique italien utilise le sigle PD (Partito Democratico ou Parti démocratique).

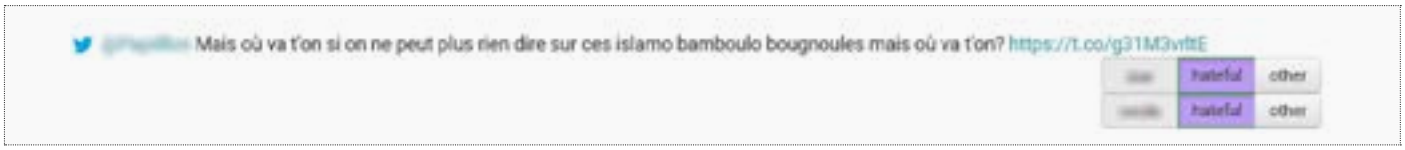
Certains messages combinaient le français et une autre langue étrangère (principalement l'anglais). Les messages ont été jugés non pertinents s'il y avait moins de 4 mots en français.

Informations et actualités
Dans certaines circonstances, les ensembles de données comprenaient des messages ou du contenu provenant de diffuseurs d'informations.

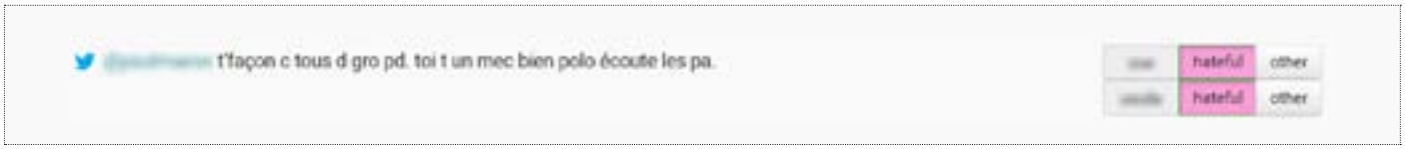
Le présent rapport se concentrant principalement sur les discussions organiques en ligne et non sur les retweets de titres, tous les messages des diffuseurs d'informations établis ont été déclarés non pertinents.

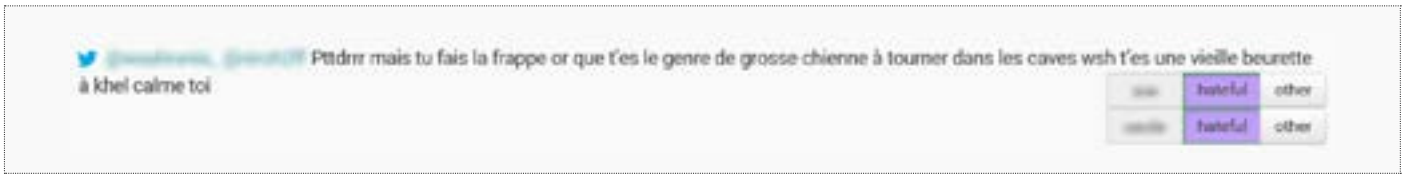
Exemples de classification
Pour les quatre discours analysés par Method52, nous avons inclus les types de messages suivants dans la catégorie « haineux » :

- Messages contenant des injures ayant clairement été utilisées pour insulter :

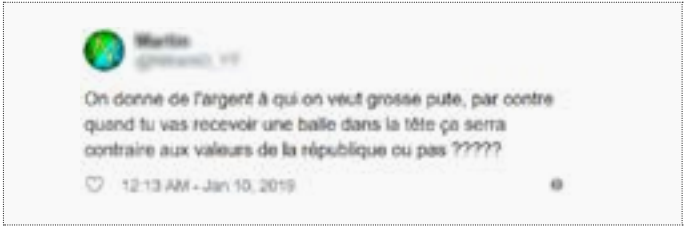


- des messages qui déshumanisaient et dépréciaient des individus en raison de leur appartenance à une catégorie protégée :





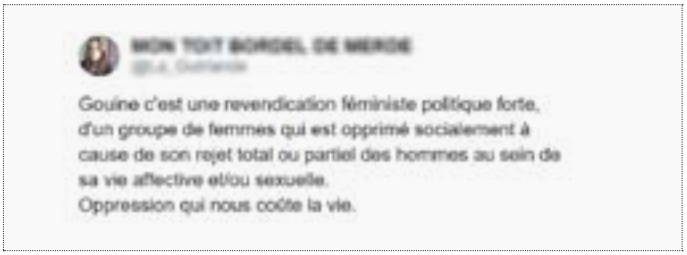
• Messages contenant des menaces ou des appels à la violence (par exemple Marlène Schiappa, mentionnée dans ce rapport) :



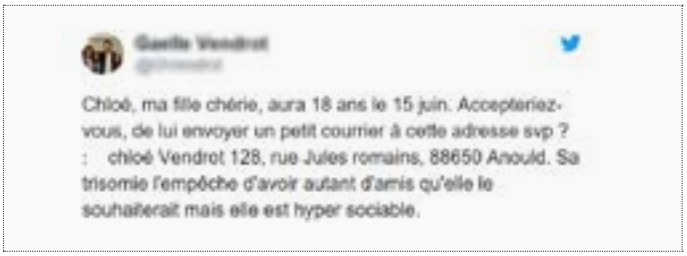
Catégorie « non haineux » :

• Messages où des insultes étaient employées pour nommer d'autres utilisateurs (l'un des signes était l'utilisation de guillemets) :

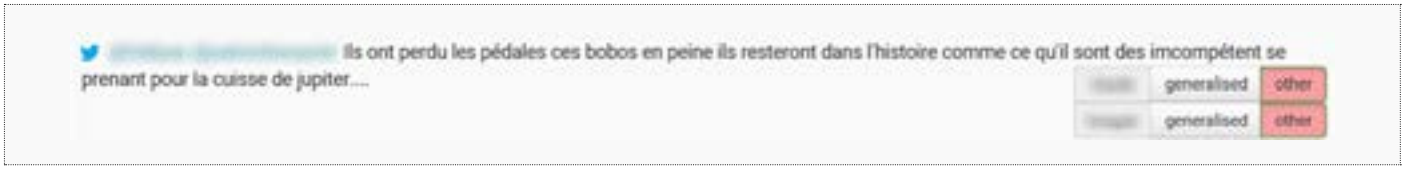
• Messages qui se réappropriaient des termes haineux et des insultes : par exemple *gouine* (exemple de *gouine* dans le rapport dans la section anti-LGBTQ) :



• Messages employant des termes haineux dans le cadre du partage de l'information : *triso* a souvent mentionné des contenus qui incluaient le terme trisomique (quelqu'un atteint de trisomie) dans des articles sur la recherche et l'information sur la trisomie, ainsi que pour sensibiliser la population :



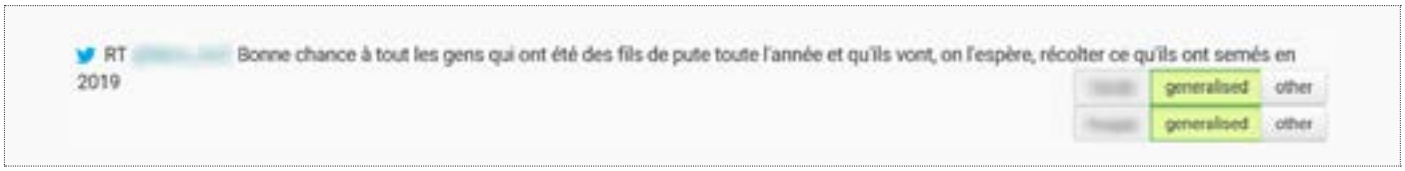
• Messages dont le contenu non pertinent qui avait été omis par le classificateur de pertinence initial, par exemple *beurre* ou *beur* dans l'ensemble de données anti-Arabes et *pédale* dans l'ensemble de données anti-LGBTQ :



Classification généralisée et précise du discours misogyne
L'ensemble de données misogynes était suffisamment vaste pour permettre une classification plus précise des messages.
Tout d'abord, un classificateur a été formé pour identifier l'utilisation généralisée de termes misogynes.
Un second classificateur a ensuite été formé pour identifier les propos haineux ciblés ainsi que les contre-discours figurant dans l'ensemble de données.

Dans la catégorie généralisée « haineux », nous avons inclus :

• le contenu qui ne visait pas la communauté ou le groupe visé par le sens original du mot (p. ex. fils de *pute*, qui a des racines misogynes, s'applique à un large éventail de cibles) :



• les messages où une personne en particulier ne peut être identifiée comme étant directement visée par des insultes (p. ex. les messages qui ne contiennent que le mot *salope* par opposition à une phrase clairement ciblée comme tu es une *salope*) et où le message ne contient pas de menace directe :

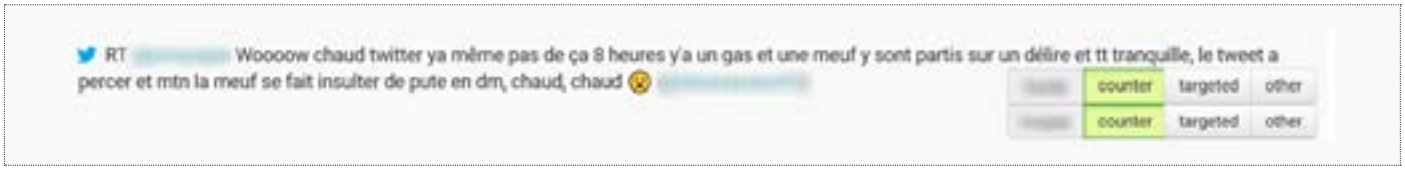
• les messages où des injures étaient utilisées contre des personnages imaginaires ou fictifs :

Tous les autres messages ont été inclus dans la catégorie « autres ».

Nous avons ensuite pris la catégorie « autres » et nous l'avons séparée au moyen d'un classificateur à trois voies, qui séparait ce contenu en contre-discours, discours haineux ciblés et autres (peu clairs ou non pertinents).

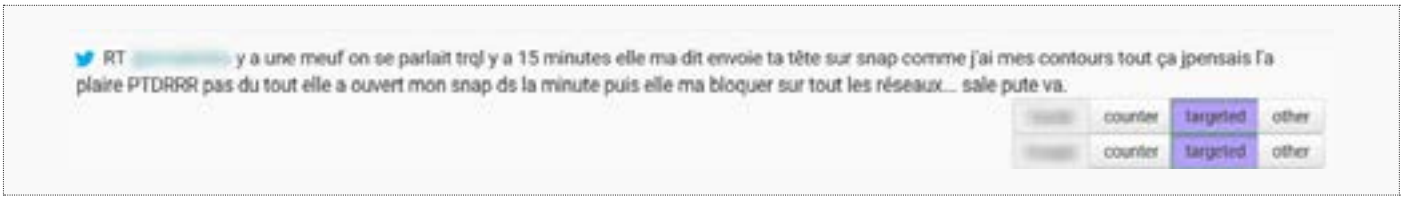
Dans la catégorie « contre-discours », nous avons inclus :

• les condamnations claires de l'utilisation d'injures et d'une rhétorique haineuse



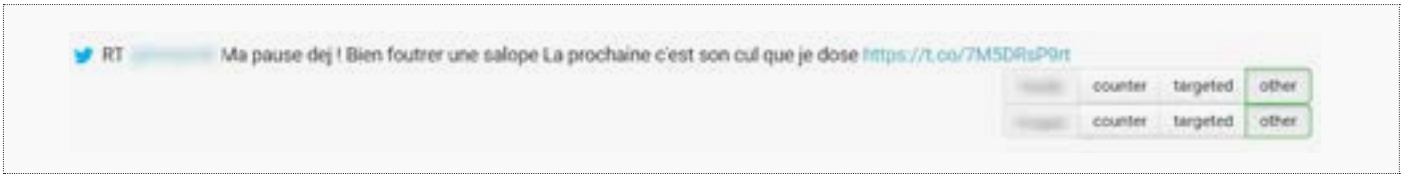
Dans la catégorie ciblée, nous avons inclus :

- les menaces directes (violence, menaces de mort ou de viol)
- l'utilisation d'injures et d'insultes visant directement une personne de la catégorie protégée (p. ex., « tu es une grosse pute » lorsque l'utilisateur ciblé peut raisonnablement être identifié comme une femme)



Dans la catégorie « autres », nous avons inclus :

- le contenu pornographique contenant un langage misogyne
- tous les messages non pertinents restants



Annexe 3 : Statistiques d'exactitude des classificateurs

L'exactitude globale de chaque classificateur est la probabilité qu'un message sélectionné aléatoirement dans l'ensemble de données appartienne à la catégorie dans laquelle il a été classé.

Misogyne

Pertinence du classificateur

Étiquette	Précision	Rappel	FB1	Étiqueté
Pertinent	0.973	0.917	0.944	136
Non pertinent	0.586	0.820	0.683	104
Exactitude globale	0.905			

Classificateur haineux

Étiquette	Précision	Rappel	FB1	Étiqueté
Haineux	0.831	0.888	0.858	86
Autres	0.500	0.382	0.433	52
Exactitude globale	0.773			

Classification généralisée « haineux »

Étiquette	Précision	Rappel	FB1	Étiqueté
Généralisée	0.796	0.750	0.772	368
Autres	0.500	0.565	0.531	257
Exactitude globale	0.693			

Classification précise

Étiquette	Précision	Rappel	FB1	Étiqueté
Contre	0.889	0.762	0.821	29
Ciblée	0.426	0.606	0.500	107
Autres	0.867	0.801	0.833	182
Exactitude globale	0.765			

Anti-Arabes

Pertinence du classificateur

Étiquette	Précision	Rappel	FB1	Étiqueté
Pertinent	0.910	0.966	0.937	70
Non pertinent	0.889	0.741	0.808	70
Exactitude globale	0.905			

Classificateur haineux 1

Étiquette	Précision	Rappel	FB1	Étiqueté
Haineux	0.462	0.750	0.571	143
Autres	0.882	0.682	0.769	175
Exactitude globale	0.700			

Classificateur haineux 2

Étiquette	Précision	Rappel	FB1	Étiqueté
Généralisée	0.670	0.678	0.674	324
Autres	0.750	0.743	0.747	326
Exactitude globale	0.715			

Anti-LGBTQ

Pertinence du classificateur

Étiquette	Précision	Rappel	FB1	Étiqueté
Pertinent	0.884	0.985	0.932	113
Non pertinent	0.962	0.750	0.843	129
Exactitude globale	0.905			

Classificateur haineux

Étiquette	Précision	Rappel	FB1	Étiqueté
Haineux	0.875	0.824	0.848	76
Autres	0.447	0.548	0.493	67
Exactitude globale	0.767			

Capacitistes ou validistes

Pertinence du classificateur

Étiquette	Précision	Rappel	FB1	Étiqueté
Pertinent	0.968	0.772	0.859	71
Non pertinent	0.514	0.905	0.655	50
Exactitude globale	0.800			

Classificateur haineux

Étiquette	Précision	Rappel	FB1	Étiqueté
Haineux	0.928	0.772	0.847	62
Autres	0.184	0.500	0.269	41
Exactitude globale	0.747			

Notes de fin

01 *https://europa.eu/rapid/press-release_IP-19-805_en.htm*

02 *https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/countering-illegal-hate-speech-online_en.*

03 *« Rapport fait au nom de la commission des lois constitutionnelles, de la législation et de l'administration générale de la république sur la proposition de loi, après engagement de la procédure générale de la république sur la proposition de loi, après engagement de la procédure accélérée, visant à lutter contre la haine sur internet (n° 1785) », 19 juin 2019, ihttp://www.assemblee-nationale.fr/2019/rapports/r2062.asp.*

04 *Notre méthodologie de recensement des comptes présentant une activité inorganique et potentiellement robotique s'inspire des lignes directrices établies par le Digital Forensic Research Lab (DFR Lab) du Conseil de l'Atlantique et de l'Oxford Internet Institute (OII) : https://medium.com/dfrlab/botspot-twelve-ways-to-spot-a-bot-aedc7d9c110c.*

05 *Voir par exemple : Pete Burnap et Matthew L. Williams, 'Cyber Hate Speech on Twitter : An Application of Machine Classification and Statistical Modeling for Policy and Decision Making', Policy & Internet, Wiley Periodicals. 2015.*

06 *http://www.iicom.org/images/iic/themes/news/Reports/French-social-media-framework---May-2019.pdf*

07 *Le GIFCT a été créé en 2017 par Facebook, Microsoft, Twitter et YouTube pour aider à formaliser la coopération dans l'industrie afin de freiner la propagation du terrorisme et de l'extrémisme violent en ligne. En mai 2019, le premier ministre néo-zélandais Jacinda Ardern et le président français Emmanuel Macron ont annoncé l'Appel à l'action de Christchurch, suite à l'attentat terroriste perpétré à Christchurch en mars 2019. En réponse, le GIFCT a créé le Content Incident Protocol (Protocole en cas d'incident de contenu) , qui est un « système triagé visant à minimiser la diffusion en ligne de contenus terroristes ou extrémistes violents résultant d'une attaque réelle contre des civils / innocents sans défense ». Pour en savoir plus sur le GIFCT et le protocole d'incident de contenu, cliquez ici : https://gifct.org/transparency/.*

08 *https://www.un.org/disabilities/convention/pdfs/factsheet.pdf*

09 *https://rm.coe.int/hate-crimes-against-lgbti/168073dd37*

10 *Définition basée sur l'article 222-33-2 du Code pénal français), https://www.legifrance.gouv.fr/affichCodeArticle.do?idArticle=LEGIARTI000029336939&cidTexte=LEGITEXT000006070719&dateTexte=20140806.*

11 *Debbie Ging et Eugenia Siapera, « Introduction », Feminist Media Studies, Vol. 18, Issue 4, numéro spécial sur la misogynie en ligne, https://www.tandfonline.com/doi/full/10.1080/14680777.2018.1447345.*

12 *https://rm.coe.int/ecri-glossary/1680934974.*

13 *https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=0900001680651592.*

14 *https://www.adl.org/education/resources/tools-and-strategies/slurs-and-biased-language.*

15 *For more on the methodology of this index, see: https://www.cncdh.fr/sites/default/files/cncdh_rapport_2017_bat_basse_definition.pdf.*

16 *Source : CNCDH, Rapport sur la lutte contre le racisme, l'antisémitisme et la xénophobie, 2018.*

17 *Karsten Müller et Carlo Schwarz, Fanning the Flames of Hate: Social Media and Hate Crime, 30 novembre 2018, 30 November 2018, https://warwick.ac.uk/fac/soc/economics/staff/crschwarz/fanning-flames-hate.pdf.*

18 *En juin 2019, Facebook a accepté de fournir aux autorités françaises des adresses IP pour identifier les personnes qui publient des contenus illégaux sur leur plateforme. Cette mesure constitue un changement majeur puisque jusqu'à présent, Facebook ne partageait des informations privées que pour des contenus liés au terrorisme ou à la pédopornographie. Pour accéder à des contenus illégaux, les autorités françaises ont dû se soumettre à un processus long et complexe du système juridique américain. En savoir plus : https://www.nouvelobs.com/societe/20190625.OBS14918/facebook-fournira-desormais-a-la-justice-francaise-les-adresses-ip-des-auteurs-de-propos-haineux.html.*

19 *« Rapport fait au nom de la commission des lois constitutionnelles, de la législation et de l'administration générale de la république sur la proposition de loi, après engagement de la procédure générale de la république sur la proposition de loi, après engagement de la procédure accélérée, visant à lutter contre la haine sur internet (n° 1785) », 19 juin 2019, www.assemblee-nationale.fr/15/rapports/r2062.asp.*

20 *Voir le Code de conduite de la Commission européenne sur la lutte contre les discours de haine illégaux en ligne, février 2019, https://ec.europa.eu/info/sites/info/files/code_of_conduct_factsheet_7_web.pdf*

21 *Pour cette section, le rapport fait référence à la Commission nationale consultative des droits de l'homme, La lutte contre le racisme, l'antisémitisme et la xénophobie (année 2017), Mai 2018, pp. 38–50, www.cncdh.fr/fr/publications/rapport-2017-sur-la-lutte-contre-le-racisme-lantisemitisme-et-la-xenophobie.*

22 *Anti-Defamation League, Quantifying Hate: A Year of Anti-Semitism on Twitter, 2018, https://www.adl.org/resources/reports/quantifying-hate-a-year-of-anti-semitism-on-twitter.*

23 *Le Code de conduite de la Commission européenne sur la lutte contre les discours de haine illégaux en ligne, février 2019, qui est mentionné dans le rapport contextuel de la Loi Avia pour identifier les tendances en ligne des discours de haine, ne mentionne pas les contenus haineux qui visent les femmes et les personnes handicapées. Le code de conduite s'appuie sur la décision-cadre 2008/913/JHA du 28 novembre 2008, qui inclut « l'incitation publique à la violence ou à la haine visant un groupe de personnes ou un membre d'un tel groupe, défini par référence à la race, la couleur, la religion, l'ascendance, l'origine nationale ou ethnique ».*

24 *Kimberlé Crenshaw, « Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics', University of Chicago Legal Forum, Vol. 1989, Numéro 1, pp. 139–67, https://chicagounbound.uchicago.edu/cgi/viewcontent.cgi?article=1052&context=uclf.*

25 *« L'intersectionnalité met en évidence les failles des lois qui, en matière de discrimination, se concentrent sur un motif unique. Premièrement, cette focalisation sur un seul motif omet le fait que toute personne a un âge, un genre, une orientation sexuelle, des convictions et une origine ethnique; et que beaucoup peuvent avoir ou acquérir une religion ou un handicap. L'intersectionnalité vise à saisir et à réparer les torts subis par ceux qui se trouvent à la croisée de ces différentes relations. » Pour en savoir plus, consultez le rapport Commission européenne, Intersectional Discrimination in EU Gender Equality and Non-discrimination Law, Mai 2016, http://ohrh.law.ox.ac.uk/wordpress/wp-content/uploads/2016/07/Intersectional-discrimination-in-EU-gender-equality-and-non-discrimination-law.pdf.*

26 *Voir Commission européenne, Intersectional Discrimination in EU Gender Equality and Non-discrimination Law, May 2016, http://ohrh.law.ox.ac.uk/wordpress/wp-content/uploads/2016/07/Intersectional-discrimination-in-EU-gender-equality-and-non-discrimination-law.pdf.*

27 *Voir Loi du 29 juillet 1881 sur la liberté de la presse, notamment les articles 24, 29, 32, 33, https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=LEGITEXT000006070722.*

28 *Ajout apporté en 1972 : la diffamation et l'injure sont punies plus sévèrement lorsqu'elles sont commises « à raison de leur origine ou de leur appartenance ou de leur non appartenance à une ethnie, une nation, une race ou une religion déterminée ».*

29 *Ces principes ont été repris en droit pénal par les articles R 625-7 et R 625-8 du Code pénal français. Ceci est important car la loi de 1881 sur la liberté de la presse ne s'applique que dans les espaces publics (à l'ère d'Internet, toute publication sur les réseaux sociaux qui n'est pas publique ne relèverait pas de cette loi) ; avec le code pénal, le contenu partagé dans les espaces privés relève également de cette réglementation.*

30 *Le terme Injures est compris dans l' article 29 de la loi du 29 juillet 1881 comme « toute expression outrageante, termes de mépris ou invective qui ne renferme l'imputation d'aucun fait »).*
www.legifrance.gouv.fr/affichTexte.do?cidTexte=LEGITEXT000006070722.

31 *La diffamation est définie à l'article 29 de la loi 29 juillet 1881 sur la liberté de la presse comme « toute allégation ou imputation d'un fait qui porte atteinte à l'honneur ou à la considération de la personne ou du corps auquel le fait est imputé ».*
https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=LEGITEXT000006070722.

32 *Pour en savoir plus, lire l'article de Nathalie Droin intitulé « L'appréhension des discours de haine par les juridictions françaises : entre travail d'orfèvre et numéro d'équilibriste », 2018, https://journals.openedition.org/revdh/4302*

33 *Pour en savoir plus, merci de consulter l'article de Nathalie Droin intitulé « L'appréhension des discours de haine par les juridictions françaises : entre travail d'orfèvre et numéro d'équilibriste », 2018. https://journals.openedition.org/revdh/4302.*

34 *« La religion la plus con, c'est quand même l'Islam », pour en savoir plus, lire TGI Paris, 17ème ch., 22 octobre 2002, Légipresse, 2003, nos. 198-I, p. 12.*

35 *« pour moi les juifs, c'est une secte, une escroquerie »', pour en savoir plus, lire Cass. Crim. 15 mars 2005, n° 04-84.463, Bull. crim., no. 90; Dr. pén., 2005, Comm. no. 85, note M. Véron.*

36 *Cette recherche est basée sur l'analyse de 15 000 commentaires choisis aléatoirement, parmi les 15 millions de commentaires extraits de 25 pages Facebook. Pour en savoir plus sur l'ensemble de la méthodologie, veuillez consulter la section méthodologie (page 1) sur https://netino.fr/panorama-de-la-haine-en-ligne-2019/.*

37 *La définition de l'injure utilisée dans le présent rapport va au-delà des propos haineux dirigés contre des catégories protégées et inclut des insultes généralisées telles que « va te faire foutre ».*

38 *Dans cette recherche, parmi les 2 964 271 tweets recueillis au cours du mois de janvier 2017, 2 950 tweets ont été sélectionnés et classés aléatoirement ; pour en savoir plus à ce sujet, veuillez consulter la section méthodologie ici : http://www.idpi.fr/wp-content/uploads/2018/02/2018-02-Baromètre-IDPI-Janvier-2017.pdf.*

39 *Voir Baromètre mensuel des manifestations de la haine en ligne – Janvier 2018, IDPI, pour en savoir plus sur la méthodologie et les définitions. https://www.idpi.fr/actualites/barometre-mensuel-des-manifestations-de-la-haine-en-ligne-janvier-2018/*

40 *Sylvain Mossou et Andrew Lane, Anti-migrant Hate Speech, Conseil Quaker pour les affaires européennes, 2018, http://www.qcea.org/*

41 *Anti-Defamation League, The Online Hate Index, 2018. Cette étude était basée sur l'analyse de 9 000 commentaires Reddit choisis aléatoirement et codés parmi 80 000 commentaires tirés (codés manuellement dans les catégories « haineux » ou « non haineux »). https://www.adl.org/resources/reports/the-online-hate-index.*

42 *Encyclopedia of Political Communication, Sage, 2008.*

43 *Voir la section de la méthodologie de Jamie Bartlett, Misogyny on Twitter, Demos, 2014, https://demos.co.uk/project/misogyny-on-twitter/. Deux études ont été réalisées : la première a extrait des tweets géographiquement situés au Royaume-Uni qui mentionnaient le terme « viol ». Parmi les 138 662 tweets, il en est resté 108 044 après le classificateur de pertinence ; parmi ces tweets, 500 tweets ont été sélectionnés aléatoirement pour l'analyse. Dans la deuxième étude, les mots-clés utilisés étaient « salope » et « pute » avec 161 744 tweets recensés au Royaume-Uni ; il en est resté 131 711 après le classificateur de pertinence, dont 500 ont ensuite été sélectionnés aléatoirement pour l'analyse.*

44 *Carl Miller et Josh Smith, Anti-Islamic Content on Twitter, Demos, 2017 ; étude fondée sur 143 920 tweets réunis et identifiés comme « dérogatoires à l'égard des musulmans ». https://demos.co.uk/project/anti-islamic-content-on-twitter/*

45 *Comme pour le discours anti-Blancs, dans les sociétés occidentales, le concept de discours anti-Chrétiens est souvent évoqué par des groupes marginaux pour légitimer leurs agendas anti-immigrés, et en cherchant ce discours, nous avons cherché à déterminer s'il s'agissait réellement d'une partie importante du discours haineux en ligne.*

46 *Le concept de discours haineux anti-Blancs est souvent utilisé par l'extrême droite comme preuve du sentiment anti-Blancs des groupes minoritaires et, dans certains cas, comme preuve de nettoyage ethnique ou de « génocide blanc ». Dans le cadre de cette étude préliminaire, nous nous sommes efforcés de comprendre dans quelle mesure ce type de discours existe en ligne et de démontrer son importance (ou son insignifiance) relative en ligne par rapport à d'autres types de discours haineux.*

47 *À l'exception de l'algorithme anti-Arabes. Voir « Discours anti-Arabes » pour en savoir plus.*

48 *Les principales limitations sont les suivantes : le contenu potentiellement non pertinent ne peut être séparé de l'échantillon et*

une analyse linguistique détaillée ne peut être effectuée pour distinguer les utilisations haineuses et non haineuses de certains mots-clés.

49 *https://www.lemonde.fr/societe/article/2019/02/11/la-justice-saisie-a-cause-de-plusieurs-tags-antisemites-a-paris_5422154_3224.html.*

50 *À l'exception de l'algorithme anti-Arabes. Voir « Discours anti-Arabes » pour en savoir plus.*

51 *Burnap et Williams, « Cyber Hate Speech on Twitter », 2015.*

52 *Pour en savoir plus sur l'exactitude des algorithmes de traitement du langage naturel et leur application aux réseaux sociaux, voir Jamie Bartlett, Vox Digitas, Demos, 2014, https://demos.co.uk/project/vox-digitas/*

53 *La théorie du Grand Remplacement est la croyance que « les populations européennes blanches sont délibérément remplacées au niveau ethnique et culturel par la migration et la croissance des communautés minoritaires. Cette propagation s'appuie souvent sur des projections démographiques pour mettre en évidence les changements démographiques en Occident et la possibilité que les populations d'origine ethnique blanche deviennent des groupes minoritaires ». Voir le rapport de l'ISD Great Replacement Theory report. https://www.isdglobal.org/isd-publications/the-great-replacement-the-violent-consequences-of-mainstreamed-extremism/*

54 *Pour en savoir plus sur Renaud Camus et la théorie du Grand remplacement, voir https://urlz.fr/bwKQ.*

55 *Jacob Davey et Julia Ebner, 'The Great Replacement': The Violent Consequences of Mainstreamed Extremism, ISD, 2019, https://www.isdglobal.org/isd-publications/the-great-replacement-the-violent-consequences-of-mainstreamed-extremism/.*

56 *https://www.bfmtv.com/societe/a-grasse-une-boulangerie-soupconnee-de-vendre-des-gateaux-racistes-867194.html.*

57 *Les mots-clés anti-Blancs ont été inclus dans ce rapport étant donné que le racisme anti-Blancs est souvent cité par des extrémistes pour justifier leurs préjugés à l'égard des immigrés. L'inclusion de mots-clés anti-Blancs dans ce rapport de recherche visait en partie à comprendre dans quelle mesure, le cas échéant, ce type de racisme est exprimé en ligne.*

58 *En finir avec l'impunité des violences faites aux femmes en ligne : une urgence pour les victimes, http://haut-conseil-egalite.gouv.fr/IMG/pdf/hce_rapport_violences_faites_aux_femmes_en_ligne_2018_02_07-3.pdf*

59 *Kimberlé Crenshaw, « Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics', University of Chicago Legal Forum, Vol. 1989, Numéro 1, pp. 139–67, https://chicagounbound.uchicago.edu/cgi/viewcontent.cgi?article=1052&context=uclf.*

60 *Cette équipe de mission interministérielle est composée de sept experts de haut niveau et de trois rapporteurs permanents issus de différents ministères (Culture, Intérieur, Justice, Économie, Services du Premier ministre) ; DILCRAH (Délégation interministérielle à la lutte contre le racisme, l'antisémitisme et la haine anti-LGBTQ) ; DINSIC (Délégation interministérielle du numérique et des systèmes d'information et de communication) ; et autorités administratives indépendantes ARCEP (Autorité des communications électroniques et des Postes) et CSA (Conseil supérieur de l'audiovisuel). Cette équipe a travaillé avec Facebook en janvier et en février pour explorer comment un nouveau système de régulation des réseaux sociaux pourrait être mis en place pour compléter les instruments existants et mieux atteindre les objectifs de politique publique relatifs à la conciliation des libertés publiques et à la sauvegarde de l'ordre public sur les réseaux sociaux.*

61 *http://www.iicom.org/images/iic/themes/news/Reports/French-social-media-framework---May-2019.pdf.*

62 *Jacob Davey, Jonathan Birdwell et Rebecca Skellett, Counter Conversations, ISD, 2018, https://www.isdglobal.org/isd-publications/counter-conversations-a-model-for-direct-engagement-with-individuals-showing-signs-of-radicalisation-online/.*

63 *https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/countering-illegal-hate-speech-online_en.*

ISD London | Washington DC | Beirut | Toronto
Association à but non lucratif enregistrée : 1141069

www.isdglobal.org

ISD | Powering solutions to extremism and polarisation

PO Box 7814, London
United Kingdom, W1C 1YZ
www.isdglobal.org