ENJEUX MÉDIATISÉS : INTERACTION ENTRE HAINE ET POLARISATION

François-Xavier BOIVIN

Marc-Antoine DE HENAU

Université du Québec à Montréal

Résumé

Les discours de haine sont, par nature, difficiles à détecter, puisqu'ils nécessitent une grande connaissance du contexte, de la communauté visée et des termes et expressions utilisés pour dissimuler la haine aux yeux du grand public. Toutefois, la détection et surtout l'amoindrissement des discours haineux sont des objectifs souhaitables pour toute plateforme sociale numérique, surtout en temps de conflit. Notre projet de recherche s'attaque donc à ce sujet sur deux plans ; d'abord, la façon dont l'information en ligne est diffusée, spécifiquement sur le point de la polarisation, définie ici comme la prise de position pour un camp plutôt qu'un autre dans un contexte qui fait polémique. Ensuite, nous analyserons la capacité des outils lexicaux, utilisés principalement pour leur aisance à traiter rapidement des quantités énormes de données. Pour ce faire, nous avons analysé les réponses présentes sous les publications qui touchent les récents développements du conflit israélo-palestinien chez deux sources médiatiques : Radio-Canada, une source perçue comme étant plus neutre, et le Tsahal, une source considérée comme plus polarisée. Nous avons trouvé dans ce projet que la polarisation d'un média ne semble pas liée à une hausse de la proportion de discours haineux, toutefois il semble y avoir un volume beaucoup plus grand de contenu total, ce qui génère par conséquent plus de discours haineux en nombres bruts, malgré une portée de plateforme qui devrait sensiblement être la même. De plus, les outils lexicaux que nous avons utilisés pour notre analyse nous ont montré des lacunes importantes dans leur capacité à détecter de façon fiable et efficace les discours de haine.

Mots-clés : discours de haine, médiatisation, enjeux polémiques, polarisation, discours en ligne

1. INTRODUCTION

Dans l'ère numérique, la facilité de communication entraîne naturellement une facilité à partager des discours de haine, particulièrement sur des réseaux sociaux comme Twitter, maintenant X, qui a vu une augmentation drastique des discours de haine depuis l'acquisition de la plateforme par Elon Musk en octobre 2022 (Hickey et al, 2025). Le but d'amoindrir les discours de haine peut s'accomplir de deux façons, idéalement en conjonction. Il faut d'abord tenter de prévenir la création de discours de haine et ensuite être capable de les détecter efficacement pour les supprimer s'ils sont publiés. Dans notre recherche, nous nous pencherons sur ces deux stratégies en nous demandant si la façon dont les sources médiatiques traitent l'actualité a un effet sur la création de discours haineux, puis en étudiant les capacités d'outils lexicaux, souvent utilisés pour leur simplicité et leur rapidité de traitement d'information, un atout nécessaire pour un outil qui surveille une plateforme de presque 600 millions d'utilisateurs en février 2025 (Statista, 2025).

Dans la deuxième section de notre travail, nous introduirons les assises théoriques sur lesquelles cet article se base ; nous présenterons entre autres le champ d'études d'analyse de discours assisté de corpus, nous justifierons le choix des sujets traités et nous présenterons le logiciel d'analyse 4CAT. La troisième section nous permettra d'introduire et d'expliquer les questions de recherche du présent article ainsi que nos hypothèses. Nous chercherons à savoir si la couverture d'enjeux polémiques par un média polarisé génère plus de discours haineux dans les commentaires que la couverture d'un média plus neutre sur le même sujet. Nous chercherons également à tester les limites de l'approche lexicale dans la détection des discours de haine. Dans la quatrième section, nous discuterons des modules HateBase analysis et Find co-words du logiciel 4CAT, les modules utilisés dans notre analyse. De plus, nous expliquerons la méthode utilisée pour monter nos corpus. Nous présenterons les résultats obtenus et discuterons de nos conclusions dans les sections éponymes.

2. CONTEXTE SOCIOPOLITIQUE

Puisque ce travail a comme sujet le traitement médiatique d'un sujet d'actualité polémique, cette mise en contexte contiendra deux volets ; le premier portera sur cet aspect médiatique. Avant de pouvoir affirmer qu'une source journalistique est neutre, il faut savoir comment nous définissons la neutralité en journalisme. Pour choisir une source qui a le plus de chance d'être neutre, afin de la placer en contraste avec la source polarisée, nous nous fions au jugement de Media Bias/Fact Check (MBFC), une organisation indépendante qui analyse le biais des sources journalistiques. En cherchant les sources disponibles, nous avons décidé de prendre le compte X du média Radio-Canada (CBC) comme source neutre, qui selon MBFC, est un média avec un niveau de reportage factuel Élevé (High) et avec un biais faible-modéré pour la gauche libérale (Left-Center). C'est également une bonne source à choisir puisque, comme la source polarisée que nous avons choisie, CBC est financée au moins partiellement par son gouvernement respectif. La source polarisée choisie est le compte X du Tsahal (IDF), @IDF s'occupe des communications pour les forces armées israéliennes, il s'agit donc du média d'un parti directement impliqué dans le conflit dont il est question dans cet article. Pour ces raisons, pour les besoins du projet et malgré une absence d'analyse par MBFC, nous évaluons IDF comme une source biaisée. Un addendum important à faire est que la neutralité médiatique n'implique pas un respect égal accordé aux deux camps du conflit ; nous ne présentons pas IDF comme une source sioniste et CBC comme une source antisioniste. Notre seule considération repose sur la présentation des deux sources et leur comportement général sur le sujet.

Le choix du conflit israélo-palestinien comme sujet principal n'est pas un choix hasardeux. Il a été choisi pour son impact mondial, le grand intérêt médiatique qui y est porté et les opinions hautement divisées qu'il suscite (Silver L. et al, 2020). Ce projet ne cherche pas à affirmer ou infirmer un côté du conflit ou un autre. Il est choisi purement parce que c'est un sujet suffisamment vaste et polémique pour permettre une étude pertinente sur les discours haineux.

Les définitions des discours haineux sont multiples. Dans le cadre de notre travail, nous nous baserons sur la définition de discours de haine telle que décrite dans *Encyclopedia of the American Constitution* (Nockleby, 2000), qui définit le discours de haine comme « toute communication qui dénigre une personne ou un groupe sur la base de caractéristique tels l'ethnicité, le genre, l'orientation sexuelle, la nationalité, la religion ou une autre caractéristique.

Par contre, comme nous travaillons avec plusieurs outils, il est bon de remarquer que chacun utilise une définition qui lui est propre. Par exemple, les Conditions d'utilisations de la plateforme Twitter stipulent qu'un utilisateur ne peut pas « promouvoir la violence contre, menacer ou harceler une personne sur la base de l'ethnie, la nationalité, la classe sociale, l'orientation sexuelle, le genre, la religion, l'âge ou le handicap. » Pour les besoins de ce projet, nous utiliserons ces définitions pour détecter par exemple les faux positifs ou les faux négatifs de l'analyse HateBase.

Cette publication est originale puisqu'elle explore le lien potentiel entre la polarisation d'une source médiatique et les discours de haine qu'elle génère. De plus, l'utilisation du logiciel 4CAT dans l'analyse des discours haineux est encore une nouvelle occurrence et nous espérons par notre recherche approfondir la connaissance sur les capacités de cet outil dans le domaine. La relation entre les discours haineux et la polarisation des médias est pertinente à étudier puisqu'elle permet de fournir une description qualitative d'un comportement néfaste des utilisateurs qui peut potentiellement être dissuadé par un changement de l'organisation médiatique, assurant ainsi un meilleur espace de discussion public qui contient moins de discours haineux. Ce portrait comportemental peut ensuite être utilisé par le public, entreprise ou institution gouvernementale afin de prendre des décisions éclairées par rapport à la plateforme utilisée afin d'avoir un effet désirable sur le public, par exemple.

3. QUESTIONS DE RECHERCHE ET HYPOTHÈSE

Dans le cadre de ce travail, nous tenterons de répondre à deux questions de recherche. La première porte sur la polarisation de la source médiatique et la présence de haine ; la deuxième est une question qui ne porte pas sur les résultats, mais sur les outils utilisés, et elle sera alimentée par nos impressions après l'utilisation de ces outils. La première question est :

(1) Est-ce que la polarisation d'une source médiatique sur un sujet polémique entraînera plus de discours haineux chez les utilisateurs qui y répondent?

Notre hypothèse est que oui ; si une source médiatique prend position dans un conflit, ceci pourrait causer un effet d'entraînement qui poussera les internautes à parler de façon plus haineuse. La deuxième question est :

(2) Est-ce que les outils lexicaux sont suffisants pour la détection de discours haineux?

Comme les outils sont fondamentalement différents, la méthodologie pour répondre à cette question le sera également. Pour HateBase, qui donne un résultat positif ou négatif, nous considérerons l'outil comme un échec si le nombre de faux positifs est supérieur à 50 % ; nous ne pourrons pas traiter les faux négatifs, puisque la taille du corpus ne nous permettrait pas de faire une annotation manuelle dans un délai raisonnable. Pour les bigrammes, qui sont un outil de traitement plus qualitatif que quantitatif, nous ne pourrons pas avoir de décision objective ; nous allons plutôt observer quelles lacunes nous avons remarquées s'il y a lieu. Notre hypothèse est que ces outils ne seront pas bien adaptés à la détection de discours haineux, donc que HateBase aura un taux de faux positifs > 50 % et que les bigrammes auront un défaut évident qui se manifestera lors du traitement.

4. MÉTHODOLOGIE

Nous travaillerons avec les commentaires qui se trouvent sous les publications et non les publications elles-mêmes pour deux raisons. D'abord, le fait de traiter des commentaires nous offrira un échantillon beaucoup plus grand pour la construction du corpus. Ensuite, il nous semble raisonnable de penser que le taux de discours haineux serait plus important chez un groupe d'internautes sans prestance sociale à maintenir que chez une organisation reconnue. Comme nos outils vont principalement trouver la présence de termes haineux, il semble peu probable que ceux-ci soient utilisés par CBC ou IDF. Nous avons choisi d'observer la communauté qui interagit avec les publications du compte Twitter de l'IDF (@IDF) puisque leur couverture du conflit est polarisée par leur implication directe dans ce dernier. Comme nous l'avons déjà mentionné plus haut, nous cherchons à mesurer la relation entre la prise de position médiatique et le degré de discours de haine. Nous avons choisi d'observer les interactions des utilisateurs avec les publications du compte Twitter de la CBC (@CBCNews) qui couvrent le conflit. Ceux-ci serviront de contrôle. Nous trouvons un autre avantage à ce choix, car les deux pages ont un nombre sensiblement similaire d'abonnés, soit 3,7M pour CBC et 2,9M pour IDF. Comme le nombre d'abonnés est similaire, nous nous attendons à ce que les publications aient une portée similaire.

Critères de sélection La publication-mère doit…	**Raisonnement**
Avoir > 20 réponses	Se concentrer sur les publications volumineuses, pour faciliter la récolte
Contenir UN des mots-clés suivants: *israel, palestine, gaza, netanyahu, hamas*	Obtenir des données pertinentes à notre sujet de recherche
Avoir été publiée entre le 1er octobre 2023 et le 30 janvier 2024 OU le 1er octobre 2024 et le 30 janvier 2025	Réduire le champ de recherche et fournir des résultats qui montreront l'évolution dans le temps

Tableau 1. Critères de sélection des publications.

Les publications dont nous noterons les commentaires (publications mères) seront choisies selon trois critères, la première étant que la publication doit avoir plus de 20 réponses. Ceci est non seulement par souci d'efficacité, mais également parce que les échanges, et par le fait même les désaccords, sont plus susceptibles de contenir le genre de données qui pourraient nous intéresser. Ensuite, pour s'assurer que la publication mère soit pertinente à notre étude, elle devra contenir au moins un des cinq mots-clés suivants : *israel*, *palestine*, *gaza*, *netanyahu*, *hamas*. Finalement, puisque le conflit est déjà actif depuis plus d'un an, nous ne choisirons que les publications faites pendant une des deux périodes visées, soit d'octobre 2023 à janvier 2024, les quatre premiers mois du conflit, ou octobre 2024 à janvier 2025, les quatre mois précédant le moment où nous avons effectué notre tâche de recherche. Ces périodes nous permettront de voir comment les discours ont évolué un an après le 7 octobre 2023.

La création du corpus sera faite grâce à l'outil Zeeschuimer, une extension du fureteur Firefox qui permet de racler les données d'une publication de média social simplement en faisant défiler la page. Ceci créera un fichier avec l'énoncé des publications, leur date de publication, le nom du

compte associé, etc. Toutes les consultations sur Twitter seront faites à partir d'un compte vierge pour éviter toute influence possible de l'algorithme de la plateforme sur le contenu qui est visionné. Comme les données sont publiquement accessibles, nous ne ferons pas d'efforts d'anonymisation dans le corpus lui-même ; toutefois, les exemples cités plus loin ne comporteront pas plus d'information que la citation directe. Pour nous assurer que les énoncés analysés sont uniquement des réponses, nous éliminerons toute publication recueillie si elle a été publiée par @IDF ou @CBCNews ou s'il ne s'agit pas d'une réponse. La colonne « is_reply » de Zeeschuimer permet de faire la distinction. Seul le contenu en anglais sera traité ; cette séparation sera faite par l'outil 4CAT lui-même, qui ne prend pas en compte les énoncés qu'il ne peut pas traiter.

Une fois le corpus monté, nous avons utilisé l'outil 4CAT pour faire l'analyse des données. 4CAT est un outil modulaire qui permet de faire plusieurs manipulations statistiques différentes, et ce, sans nécessiter un travail de programmation ou une expertise en informatique. Les deux modules que nous avons exploités sont « HateBase analysis » et « Find co-words ». Le premier est un module qui se base sur le lexique de HateBase (HB). HB est un lexique fondé en 2013 et fermé en 2022 qui comporte 3984 termes haineux dans 198 langues différentes. Ces termes sont issus des recherches du Sentinel Project, une organisation non gouvernementale canadienne qui agit contre les actes violents commis internationalement. Pendant le temps d'activité de HB, les internautes pouvaient aussi suggérer des additions au lexique, offrir des définitions supplémentaires ou fournir des exemples d'utilisation des termes haineux répertoriés. Depuis 2022, HB est devenu le Weaponized Word et est vendu comme un grand lexique à utiliser pour détecter la présence de discours haineux sur des plateformes sociales, pour des organisations gouvernementales, corporatives ou autres. De son côté, HB est devenu librement consultable pour le public et c'est comme ça que le lexique fait partie de l'outil 4CAT. Lorsque le module HateBase analysis est lancé, les énoncés du corpus sont analysés pour détecter toute présence d'un des termes contenus dans le lexique. Si c'est le cas, le terme utilisé sera répertorié et classifié selon les deux critères de classification de HB : l'ambiguïté et le niveau d'offense. Un terme est ambigu s'il a plus d'une définition possible et qu'au moins une de ces définitions est haineuse. Ensuite, le terme reçoit un score d'offense sur 100 qui correspond au grade d'offense qui se trouve sur la ressource en ligne. Le tableau 1 explique le système de grade et de score. Les mots soulignés sont des exemples de mots non ambigus, selon HB.

Grade	Score	Exemples
Extrêmement offensant	85 - 100	*property, chief, slave*
Hautement offensant	65 - 84	*gay, hillbilly, hoe*
Modérément offensant	25 - 64	*trash, egg*
Légèrement offensant	1 - 24	*idiot, bird, anglo*

Tableau 2. Gradation selon HateBase.

Une fois l'analyse de HB faite, nous réviserons manuellement les résultats positifs pour décider si, selon la définition fournie par HB, le terme haineux mis en évidence est utilisé comme sa définition haineuse. S'il n'est pas utilisé dans un contexte haineux, nous le compterons comme un faux positif. Un taux de faux positif >50% sera considéré comme un échec de l'outil d'analyse HateBase.

Le deuxième module utilisé est « Find co-words », le module de 4CAT qui permet la détection de bigrammes. Le module permet de faire des bigrammes ou des trigrammes ; nous avons décidé de travailler avec les premiers. Pour nous assurer de trouver des bigrammes significatifs à notre question, nous avons choisi quatre mots-noyaux, et ce sont les mots précédents ou suivant ces quatre mots qui sont à l'étude. Comme notre projet porte sur les discours de haine et que ceux-ci sont définis par de la violence verbale envers des groupes démographiques, nos mots-noyaux seront des termes démographiques. Nous avons choisi deux termes ethnoreligieux et deux termes de nationalité, chacun se plaçant d'un bord ou de l'autre du conflit. Les termes de nationalité sont *israeli* et *palestinian* ; les termes ethnoreligieux sont *jewish* et *muslim*. Nous choisirons parmi la liste des bigrammes les cinq termes les plus fréquents pour fins d'analyse. Les occurrences uniques ne seront pas comptabilisées. Après avoir trouvé les cinq occurrences les plus communes, nous utiliserons le lexique de valence NRC-VAD v.2.1 (Mohammad, 2025) qui qualifie pour plus de 20 000 mots anglais la valence, soit l'attitude positive ou négative d'un locuteur typique pour le mot. Cette catégorisation va de -1 à 1. Une valence inférieure à zéro indique une attitude négative chez un locuteur typique, -1 étant le plus négatif. Une valence supérieure à zéro indique une attitude positive, 1 étant le plus positif. Ces scores seront une façon numérique de quantifier l'attitude générale envers chaque bigramme et seront un appui à l'analyse des mots-noyaux.

5. RÉSULTATS

Dans le cadre de notre travail, nous avons récupéré un total de 99458 publications uniques (tweets). Les tweets récupérés sont séparés dans le tableau 2 selon leur source et la date de publication du contenu mère dont ils sont les commentaires.

	@IDF	@CBCNews	Total
Oct. 23 - Jan. 24	69 749	5 331	75 080
Oct. 24 - Jan. 25	22 753	1 585	24 338
Total brut	92 502	6 956	99 458
Nombre d'items ignorés par 4CAT*	2 583	1 231	3 814
Total net	89 919	5 725	95 644

Tableau 3. Description du corpus.

L'outil 4CAT est fait d'une telle façon que les énoncés qui ne peuvent pas être traités sont automatiquement supprimés lorsque le corpus est enregistré. Il reste donc 95 644 tweets après l'élimination des 3 814 énoncés intraitables.

Nous avons remarqué la différence importante entre la taille du corpus CBC et IDF, ce dernier était plus de 15 fois plus grand que le premier. Nous avons considéré cette différence et en avons tiré deux conclusions. La première est que nos tests analytiques, soit HateBase analysis et Find co-words ne seront pas affectés, puisque nous travaillons avec des proportions et non des nombres bruts, donc la différence de volume ne devrait pas avoir d'impact statistique sur nos résultats. Il est

tout de même important de noter que cette différence existe. La deuxième conclusion que nous avons tirée est que, dans l'esprit de la première question de recherche, un volume plus important est quand même problématique, puisque les discours de haine posent un risque informationnel[1] pour les personnes qui pourraient le lire. La section qui suit touchera plus sur ce sujet.

Les résultats de HateBase amènent des conclusions intéressantes. On peut remarquer que CBC semble avoir une proportion plus importante d'énoncés comportant des termes haineux. Par contre, IDF a dix fois plus d'énoncés contenant de tels termes au total. De plus, en regardant la moyenne du score d'offense reçu pour chaque énoncé considéré haineux par HateBase, même si le corpus IDF compte proportionnellement moins d'énoncés haineux, ils sont généralement plus offensants que la moyenne des termes utilisés dans le corpus CBC.

	IDF	CBC
Nombre total de tweets	89 919	5 725
Nombre de tweets contenant des termes haineux (*proportion*)	1 240 (*1,38%*)	123 (*2,15%*)
Valeur moyenne d'offense	58,18%	47,20%
Valeur médiane d'offense	68%	44%

Tableau 4. Résultats du module HateBase analysis.

Les résultats du module Find co-word se trouvent en Annexe 1. Ces tableaux montrent les cinq mots les plus fréquents qui suivent ou précèdent les quatre mots-noyaux, et ce pour chacun des deux corpus. Pour chaque mot, le nombre entre parenthèses représente le nombre d'occurrences total dans son corpus, alors que le nombre entre crochets représente la valence. Un mot n'a pas de valence s'il n'est pas présent dans le lexique NRC-VAD. Le numéro de valence qui se trouve sous chaque mot-noyau est la valence moyenne de ses bigrammes et non la valence du mot-noyau lui-même. La moyenne n'est pas pondérée selon la fréquence relative de chaque bigramme.

Nous remarquons que toutes les moyennes de valence des mots-noyaux sont négatives, sauf pour le terme jewish, qui a dans les deux corpus une valence positive. Le nombre important de mots négatifs n'est pas le fruit du hasard ; le contexte du conflit armé implique l'utilisation de mots comme *soldier*, *army*, *kill* et *hostages*. Nous remarquons toutefois la présence du mot *terrorist* dans les bigrammes de *muslim* pour les deux corpus et *genocidal* dans les bigrammes de *jewish*, celui-là uniquement présent dans le corpus de CBC. La présence de ces mots suggère une accusation faite envers ces groupes ethnoreligieux. Ces résultats seront importants pour la prochaine section.

6. DISCUSSION

Avant de considérer les résultats de HB, il faut qualifier la fiabilité de l'outil. Selon notre inspection manuelle des résultats positifs émis par HB, seulement 36,76% des positifs sont des vrais positifs,

[1] Les risques informationnels, *cognitohazard* en anglais, sont des idées ou des énoncés qui peuvent causer de la détresse psychologique par le simple fait de les observer ou d'en prendre conscience.

c'est-à-dire que le terme haineux répertorié était réellement utilisé dans un contexte haineux. Ceci tombe sous la barre des 50% qui aurait marqué HB comme un outil fiable. Il nous faudra donc considérer que HateBase est un outil non fiable pour la détection des discours haineux, ce qui répond partiellement à notre deuxième question de recherche. La plupart des faux positifs étaient comme (3), des termes haineux ambigus utilisés dans un contexte non-haineux. Le mot *chief* souligné est celui que HateBase a détecté comme étant haineux, puisqu'il peut être utilisé de façon moqueuse pour parler d'une personne autochtone, ce qui n'est pas le cas ici.

> (3) This holiday is not truly complete without the return of our hostages. "For the full remarks from LTG Herzi Halevi, Chief of the General Staff"

Malgré cette conclusion fâcheuse, les résultats de HB ne sont pas impertinents à considérer. Même en retirant environ 65% des énoncés haineux, le nombre d'énoncés dans le corpus IDF reste significativement plus volumineux que le corpus CBC, et le volume de discours haineux également. Comme des recherches précédentes l'ont démontré, la couverture de sujets polémiques et politiquement chargés augmente de façon significative la présence de discours haineux (Zannettou et al, 2020). Nos résultats suggèrent qu'une couverture plus polarisée d'un sujet polémique encourage plus d'interactions de la part des utilisateurs, amenant par le fait même plus de discours haineux. Même si cet effet n'est pas proportionnellement significatif, il est très significatif sur le volume brut d'énoncés, haineux ou non. Comme on retrouve plus d'énoncés haineux, il faut se souvenir du risque informationnel que posent les discours de haine, qui peuvent causer de la détresse par le simple acte de les lire et considérer que même si le résultat n'est pas proportionnel, la présence d'un plus grand nombre brut d'éléments haineux constitue plus de haine, ainsi l'esprit de notre première question de recherche reste intact ; la polarisation d'une source médiatique semble avoir un effet sur le volume d'énoncés qui y répondent, ce qui engendre plus de discours haineux.

Les résultats des bigrammes permettent eux aussi de tirer des conclusions, même si elles sont plus qualitatives que quantitatives. D'abord, la seule valence moyenne positive se trouvant avec le mot-noyau *jewish* est particulièrement intéressante puisque c'est un résultat unique qui se reflète dans les deux corpus. La valence ne semble donc pas être affectée par la polarisation. Comme mentionné plus tôt, la présence de mots à valence négative est compréhensible dans le contexte d'un conflit armé, ce qui peut expliquer l'attitude généralement négative que l'on trouve dans les bigrammes. Pour nous, les résultats les plus intéressants sont les bigrammes de *muslim* qui comportent *terrorist* dans les deux corpus et le bigramme de *jewish* qui contient, dans le corpus CBC, le mot *genocidal*. Il est intéressant pour nous de noter que ces deux résultats différents ont des parallèles très forts. Dans les deux cas, il s'agit d'un groupe ethnoreligieux qui se voit accusé et insulté pour les actions violentes d'une minorité qui, bien qu'elle revendique le même trait identitaire, n'a pas de lien intrinsèque ou direct avec la population entière. La connexion médiatique entre *muslim* et *terrorist* est bien connue, prenant racine dans le début des années 2000 avec les attentats, entre autres celui du 11 septembre 2001, et l'intervention militaire des États-Unis au Moyen-Orient. Depuis, le lien fâcheux entre l'identité musulmane et le terrorisme s'est implanté dans l'imaginaire commun, si bien que 56,18% des personnes musulmanes disent sentir de la détresse psychologique causée par cette association (Naeem, 2022). De l'autre côté, l'association entre *genocidal* et *jewish* est plus nouvelle, puisque les personnes juives sont plus souvent considérées comme victimes de violence que coupables. Cette nouvelle tendance prend ses racines dans le même courant que l'association *muslim*/*terrorist*, puisqu'on blâme ici la population juive entière pour les actions violentes de l'État

israélien. Même si la connexion n'est pas aussi forte que la précédente, n'apparaissant que dans un de nos deux corpus, elle représente quand même un effet néfaste pour la population visée alors que 75% des personnes juives en Europe disent se sentir injustement blâmées pour les actions du gouvernement israélien (EUAFR, 2024). Les bigrammes permettent donc, en ajoutant un élément contextuel, de voir se dessiner des tendances et de détecter des narratifs haineux mieux que HB est capable de le faire.

Nous notons la présence de bigrammes *genocide* et *palestinian* que nous ne traitons pas comme *genocidal*/*jewish*. D'abord, nous considérons que la présence de ce mot n'est pas une erreur, puisque les actions de l'État israélien à Gaza sont considérées par Amnesty International et le Universtity Network for Human Rights Association comme un génocide. De plus, comme le mot utilisé avec *palestinian* est le nom *genocide*, alors qu'avec *jewish* on trouve l'adjectif *genocidal*, il y a dans le deuxième cas une accusation qui ne figure pas dans le premier. Toutefois, il est important pour nous de montrer que nous avons considéré la présence de ce bigramme et qu'il a volontairement été écarté de la question, puisqu'il ne s'agit pas d'un discours haineux pertinent à notre recherche.

Il faut tout de même considérer la fiabilité de l'outil et de son analyse. Pour notre considération, l'analyse de bigrammes semble avoir émis des résultats réels, puisque c'est un travail quantitatif. Comme les résultats ne sont pas réfutables comme le sont ceux de HB, il faut plutôt considérer la qualité de l'information recueillie. Nous considérons qu'une analyse par bigrammes peut démontrer des tendances dans les corpus, mais que ce n'est pas un bon outil pour la détection des discours de haine, puisque les résultats peu fréquents ne sont pas suffisamment bien répertoriés pour les détecter, que ce soit pour les éliminer dans le cas de la modération d'un réseau social ou pour des besoins de recherche.

Notre méthodologie contient plusieurs limites que d'autres recherches pourront exploiter afin de trouver de nouvelles avenues et arriver à des conclusions différentes des nôtres. D'abord, la limite de temps et d'expérience a nécessité l'utilisation d'outils peu fiables. D'autres chercheurs pourraient utiliser un outil de détection des discours haineux comme BERT-HateXplain, qui permet de détecter les discours haineux de façon plus fiables et de justifier ces choix. De plus, il serait intéressant de trouver un corpus neutre avec un nombre équivalent d'énoncés, puisque le volume disproportionné de nos deux corpus a été une entrave à la recherche.

7. CONCLUSION

Comme réponse à nos questions de recherche, il semblerait qu'une source médiatique plus polarisée entraînerait, lors de discussions sur un sujet polémique, plus de réactions et de réponses qu'une couverture médiatique neutre. Par ce fait, une couverture médiatique créera par volume plus de discours haineux, même si le taux proportionnel reste sensiblement le même.

Nous considérons que les outils d'analyse lexicaux, comme HB ou les bigrammes, peuvent faire paraître des tendances, mais qu'ils sont généralement mal adaptés pour la détection de discours haineux.

Plus de recherches sur le sujet pourraient grandement approfondir la question, en utilisant différents types d'outils, par exemple les outils LLM (Large Language Model) tels que Chat-GPT pourraient

effectuer des tâches plus variées et précises en travaillant non pas avec un outil créé pour un seul but, mais en faisant plutôt effectuer au modèle une tâche personnalisée au contexte de recherche. 4CAT dispose justement d'un module qui permet de faire traiter un corpus par une instance de LLM, ce qui ouvre une avenue intéressante pour le futur de la recherche en linguistique informatique. D'autres avenues de recherche pour approfondir notre sujet pourraient également voir si d'autres sujets, d'autres plateformes sociales ou d'autres sources médiatiques génèrent les mêmes résultats.

ANNEXE 1 : Résultats du module d'analyse Find co-words et valence, corpus IDF

Mot précédent	Mot-noyau	Mot suivant
killed (49) [-0.896] israel (21) innocent (19) [0.458] idf (18) praying (18) [0.292]	israeli (2202) [-0.06]	soldiers (109) [-0.216] army (109) [-0.166] people (96) [0.208] civilians (94) [-0.020] government (73) [-0.146]
israel (19) stand (18) [0.200] praying (5) [0.292]	jewish (724) [0.318]	people (85) [0.208] cousins (25) [0.520] israel (10) community (9) [0.500] nation (8) [0.188]
israel (4) arab (3) [0.000] palestine (3)	muslim (173) [-0.211]	countries (18) [0.000] terrorist.s (14) [-0.981] country (7) [0.396] stand (7) [0.200] invasions (5) [-0.458]
innocent (48) [0.458] killing (32) [-0.840] hamas (30) genocide (27) [-0.960] occupied (24) [-0.326]	Palestinian (1252) [-0.142]	people (176) [0.208] children (157) [0.714] civilians (127) [-0.020] land (80) [0.164] hostages (46) [-0.674]

ANNEXE 2 : Résultats du module d'analyse Find co-words et valence, corpus CBC

Mot précédent	Mot-noyau	Mot suivant
killed (5) [-0.896] released (3) [0.688]	israeli (168) [-0.173]	hostages (14) [-0.674] government (11) [-0.146] military (8) [-0.208] civilians (7) [-0.020] propaganda (7) [0.042]
israel (3) genocidal (2) [-1.000]	jewish (143) [0.192]	people (8) [0.208] leaders (7) [0.666] community (4) [0.500] supremacy (3) [0.588]
government (2) [-0.146]	muslim (42) [-0.142]	countries (6) [0.000] country (5) [0.396] terrorist (3) [-0.981] votes (3) [0.164]
hamas (7) genocide (4) [-0.960] innocent (4) [0.458] pro (3) [0.334] killing (3) [-0.840]	palestinian (182) [-0.059]	officials (21) [0.292] people (17) [0.208] hostages (14) [-0.674] children (9) [0.714]

RÉFÉRENCES

Amnesty International. (2024). ‘You Feel Like You Are Subhuman’ Israel’s Genocide Against Palestinians In Gaza. Index : MDE 15/8744/2024. https://www.amnesty.org/en/documents/mde15/8668/2024/en/

European Union Agency for Fundamental Rights. (2024). Jewish People’s Experiences and Perceptions of Antisemitism. EU Survey of Jewish People. https://fra.europa.eu/en/publication/2024/experiences-and-perceptions-antisemitism-third-survey

Hickey D., Fessler D. M. T., Lerman K., Burghardt K. (2025) X under Musk’s leadership: Substantial hate and no reduction in inauthentic activity. PLoS ONE 20(2): e0313293. https://doi.org/10.1371/journal.pone.0313293

Mathew, B., Saha, P., Yimam, S. M., Biemann, C., Goyal, P., & Mukherjee, A. (2021). Hatexplain: A benchmark dataset for explainable hate speech detection. *Proceedings of the AAAI conference on artificial intelligence* (Vol. 35, No. 17, p. 14867-14875). https://doi.org/10.48550/arXiv.2012.10289

Mohammad, S. M. (2025). The NRC Valence, Arousal, and Dominance (NRC-VAD) Lexicon (Version 2.1). National Research Council Canada. http://saifmohammad.com/WebPages/nrc-vad.html

Naeem, T. (2022). Muslims as Terrorists: Hate Speech against Muslims. *International Journal of Islamic Thoughts*, Vol. 22, p. 114 - 124. https://doi.org/10.24035/ijit.22.2022.245

Nockleby, J. T. (2000) Hate Speech. *Encyclopedia of the American Constitution*, sous la direction de Leonard W. Levy et Kenneth L. Karst, 2ème éd., vol. 3, Macmillan Reference USA, 2000, pp. 1277-1279. Gale eBooks, link.gale.com/apps/doc/CX3425001193/GVRL?u=biblioquebes&sid=bookmark-GVRL&xid=75e07ccb

Peeters, S. Hagen, S. (2022). The 4CAT Capture and Analysis Toolkit: A Modular Tool for Transparent and Traceable Social Media Research. *Computational Communication Research* (Vol. 4), Numéro 2, p.571 - 589. https://doi.org/10.5117/CCR2022.2.007.HAGE

Silver, L. Alper, B. A. Keeter, S. Lippert, J. Besheer, M. (2024) Majority in U.S. Say Israel Has Valid Reasons for Fighting; Fewer Say the Same About Hamas. Pew Research Center. https://www.pewresearch.org/wp-content/uploads/sites/20/2024/03/PRC_2024.3.21_Israel-Hamas_REPORT.pdf

Statista. (2025) Most popular social networks worldwide as of February 2025, by number of monthly active users. Social Media & User-Generated Content. https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

University Network for Human Rights. (2024). Genocide In Gaza: Analysis of International Law and Its Application To Israel's Military Actions Since October 7, 2023. https://static1.squarespace.com/static/66a134337e960f229da81434/t/66fb05bb0497da4726e125d8/1727727037094/Genocide+in+Gaza+-+Final+version+051524.pdf

Zannettou S., ElSherief M., Belding E., Nilizadeh S., Stringhini G. (2020) Measuring and Characterizing Hate Speech on News Websites. WebSci '20: Proceedings of the 12th ACM Conference on Web Science, 2020. https://doi.org/10.1145/3394231.3397902