DISCOURS ÉCRITS DE LA COMMUNAUTÉ D'INCELS.IS : LES BIAIS D'ASSOCIATIONS EXPLICITES DANS LES MODÈLES DE PLONGEMENT

Sandrine Sénéchal et Josiane van Dorpe
Université du Québec à Montréal

1. Introduction

Les *incels* sont reconnus pour la toxicité et le caractère haineux de leurs discours (Preston *et al.*, 2021 ; Pelzer *et al.*, 2021) et sont considérés comme des extrémistes violents depuis 2019 au Canada (SCRS 2020, 2021). Le terme *incel,* un mot-valise formé par l'amalgame des mots *involuntary* et *celibate*, réfère aux membres d'une communauté en ligne ayant évolué sur les forums de discussions au courant des deux dernières décennies (Pelzer *et al.*, 2021 ; Gothard *et al.*, 2021 ; Preston *et al.*, 2021). Les individus s'identifiant comme incels sont principalement des hommes explorant ensemble et revendiquant leurs besoins, leurs désirs, leurs blâmes, leurs reproches, leur haine, leur mépris ainsi que leurs frustrations à l'égard des femmes (Jaki *et al.*, 2019 ; Farrell *et al.*, 2019 ; Kassam, 2018). Leurs communications se caractérisent notamment par la revendication de préjugés et de stéréotypes envers les femmes et par l'utilisation de néologismes propres à cette communauté : p. ex. *femoid* et son abréviation *foid* qui ont le même référent que le terme *woman* (Jaki *et al.*, 2019). Cette tendance à associer consciemment des concepts négatifs aux femmes est un exemple de biais d'associations que nous qualifierons d'explicites. Inversement, il existe aussi des biais d'associations dits implicites ; ceux-ci se traduisent par la facilité inconsciente que l'on a à associer certains concepts entre eux plutôt qu'à leurs opposés, et ce, de façon automatique (p. ex. le biais d'associations implicites connu |*flowers-pleasant/insects-unpleasant*|, Greenwald *et al.*, 1998).

L'intelligence artificielle appliquée au langage a évolué au point de devenir un outil incontournable. Bien que les techniques de traitement automatique du langage (TAL) présentent des avantages significatifs, l'évolution constante de ces systèmes visant un produit homologue à la pensée humaine a amené des problèmes tels que les modèles vectoriels biaisés. Ceci est dû à la perpétuation des préjugés et des stéréotypes humains dans les données linguistiques non annotées sur lesquelles ces modèles d'apprentissage automatique sont entraînés (Bolukbasi *et al.*, 2016 ; Caliskan *et al.*, 2017 ; Jurafsky et Martin, 2021). Il s'agit précisément de ce dont nous traitons dans cet article, par l'entremise des communications de la communauté d'`incels.is`.

Considérant que les incels revendiquent leurs biais d'associations, il est légitime de se demander si ces biais explicites se distinguent des biais implicites connus dans les modèles de plongements, mais aussi si ces derniers sont présents dans les données textuelles produites par les incels. De la même manière, le caractère misogyne prononcé de la communauté est en lien avec l'usage fréquent de néologismes péjoratifs pour se référer aux femmes (*femoid/foid* ~ 62 k occurrences). Cela étant, des termes standards renvoyant aux femmes sont aussi utilisés (*woman/women* ~ 116 k occurrences). On se demandera alors dans quelles mesures on peut distinguer les biais associés à ces deux classes de termes chez les incels.

Si plusieurs méthodes existent pour détecter les biais d'associations implicites dans les données textuelles, peu d'études se sont en revanche penchées sur le statut des biais d'associations explicites

tels que nous les présentons, et sur la façon dont les modèles de représentation du sens lexical les capturent. Dans cet article, nous expliquons comment nous avons compilé un ensemble de données de plus de 25 millions de tokens (mots) en anglais extraits du forum `incels.is`, et comment nous avons testé la présence des biais d'associations latents projetés dans les données textuelles par les humains les ayant produites.

Cette étude pourra donner lieu, grâce au vaste corpus qui sera rendu disponible en libre accès, à une poursuite des travaux sur les communications écrites des incels, et contribuera par ailleurs à fournir de nouvelles pistes pour le repérage des communautés en ligne extrémistes et toxiques.

2. Contexte

2.1. Les incels

Les communautés incels et les interactions de leurs membres se trouvent en ligne, principalement sur des forums de discussion spécialisés tels que `incels.is`. Les incels n'ont pas d'idéologie claire ou unanime, mais plusieurs éléments les unissent et caractérisent leurs discours : le désire de relations hétérosexuelles et l'absence ou l'incapacité à en obtenir, l'antiféminisme, l'approbation et la justification de la violence envers les femmes, l'élimination des droits des femmes, le racisme, et un lexique qui leur est propre (Preston *et al.*, 2021 ; Pelzer *et al.*, 2021 ; Chang 2020 ; Jaki *et al.*, 2019). Ils communiquent sur ces plateformes d'échanges asynchrones de façon anonyme, par l'entremise de leur pseudonyme ; il nous est impossible de déterminer quelques informations personnelles que ce soient, outre ce qu'ils communiquent ouvertement. En général, ces plateformes sont divisées en plusieurs sections et sous-sections, catégorisées par sujets. Les discussions (communément appelées *threads*) prennent un format hiérarchique : un utilisateur publie un message (communément appelé *post*), puis une discussion s'ensuit sous forme de commentaires subséquents (Holtz *et al.*, 2012). Sur le forum `incels.is`, plus précisément, la langue principale de communication est l'anglais – seulement qu'à de très rares occasions d'autres langues sont utilisées ; ainsi, il nous est aussi impossible de déterminer la L1 des auteurs.

Dans leur analyse inductive qualitative de plus de 8000 publications recueillies sur deux forums incels, O'Malley *et al.* (2020) expliquent que la communauté incel gravite autour de cinq grandes catégories interdépendantes : le marché du sexe, les femmes étant naturellement mauvaises et la légitimation de la masculinité, de l'oppression masculine et de la violence. Pelzer *et al.* (2021) affirment que la communauté incel est reconnue pour l'utilisation d'un langage toxique – qu'ils définissent comme étant du contenu agressif et humiliant de manière flagrante – dont l'utilisation est omniprésente aussi sur le forum `incels.is`. Les propos nocifs tenus sur ce forum sont le plus souvent à caractère misogyne, sexiste, homophobe, raciste ; mais la liste ne s'arrête pas là. Le contenu toxique le plus courant est celui à caractère misogyne : évalué à 41,1 % du contenu toxique total du forum (Pelzer *et al.*, 2021). Ce qui distingue les incels d'autres communautés toxiques est le fait qu'on y retrouve aussi beaucoup de discours empreints de haine, de mépris et de dégoût de soi (Pelzer *et al.*, 2021 ; Preston *et al.*, 2021). Cette forme de toxicité plutôt atypique serait la deuxième plus courante et est évaluée à 23 % du contenu total du forum.

Le contenu toxique réside entre autres dans le langage propre à cette communauté ; nous parlerons de néologismes qui leurs sont propres, encodant en partie leurs discours. Leur utilisation est très répandue sur le forum. Par ce jargon ils expriment ouvertement, entre autres, leur haine, leurs

préjugés et leurs stéréotypes envers les femmes ; il s'agit de termes péjoratifs et souvent déshumanisants (Pelzer *et al.,* 2021). L'exemple le plus criant est celui du péjoratif *femoid* (et son abréviation *foid*), mot-valise provenant de *female humanoid*, que les incels présentent eux-mêmes comme un terme suggérant que les femmes ne sont pas pleinement humaines (Incels.is, 2022).

2.2. Biais d'associations

À la base des recherches sur les biais d'associations se trouve une méthode de détections des biais d'associations implicites chez l'humain. Comme plusieurs avant nous, c'est entre autres sur cette méthode et sur les principes l'accompagnant que nous basons notre recherche. Il est donc tout d'abord primordial de traiter des notions qui sous-tendent les biais d'associations en soi et cette méthode, d'élaborer sur l'interprétation la dichotomie implicite/explicite, pour finalement aborder les problématiques entourant les biais d'associations humains.

En psychologie, dans le but d'explorer ce que Nosek *et al.* (2002) décrivent comme l'incontrôlabilité des attitudes automatiques et des stéréotypes, Greenwald et Banaji (1995) ont élaboré l'*Implicit Association Test* (IAT), ou Test d'Associations Implicites en français (évalué ensuite dans Greenwald *et al.,* 1998). Ils décrivent ces attitudes et ces stéréotypes comme deux grandes catégories de cognition sociale implicite. Les attitudes et les stéréotypes peuvent être explicites – nous les exprimons ou nous y pensons de façon délibérée – ou implicites – ceux-ci sont beaucoup moins accessibles au niveau conscient et hors de notre contrôle.

L'IAT mesure donc la force des associations entre deux concepts cibles (p. ex. fleurs/insectes, femmes/hommes, etc.) et deux attributs : des évaluations (p. ex. termes agréables/bons, termes désagréables/mauvais, etc.) ou des stéréotypes (p. ex. arts libéraux/mathématiques, famille/carrière, etc.) en termes de temps de réaction. Le temps de réaction sera plus rapide pour l'association de deux concepts dits « compatibles ». Lorsque le temps de réaction aux associations entre un groupe de termes d'un concept cible et un groupe de termes d'un attribut est significatif, c'est qu'il y a ce que nous appelons un biais d'associations. Afin d'illustrer très simplement les biais d'associations, nous prenons un exemple célèbre de Greenwald *et al.* (1998) : les associations évaluatives connues impliquant les contrastes sémantiques des fleurs par rapport aux insectes (c.-à-d. le concept cible *flower* avec les attributs *pleasant* et le concept cible *insect* et les attributs *unpleasant*).

Ils ont démontré que les attitudes et les stéréotypes de groupes existent hors de notre capacité à exercer un contrôle conscient. Leurs résultats ont aussi reflété ceux exprimés de façon consciente. Nosek *et al.* (2002) explique cependant que les biais implicites se sont manifestés dans les données de façon nettement plus forte que leurs homologues explicites ; de tels résultats peuvent se traduire par notre tendance à nier ce genre d'attitudes ou de stéréotypes, de façon consciente ou non, en raison de normes personnelles (internes) ou de pressions sociales (externes). Ce qui est donc particulièrement intéressant, c'est que l'IAT décèle aussi les biais d'associations que l'on nie, de façon consciente ou non (Greenwald *et al.,* 1998 ; Nosek *et al.,* 2002).

2.2.1. Problématiques des biais d'associations

Bien que les biais d'associations implicites tels que |*flowers-pleasant/insects-unpleasant*| puissent sembler bénins, certains, comme les associations |femmes-famille/hommes-carrière| sont plus problématiques. La première partie de la problématique entourant les biais implicites est la

suivante : comme nous ne contrôlons pas ce genre de phénomène, les biais d'associations implicites que nous avons acquis peuvent affecter notre processus de décisions au quotidien, sans que l'on s'en aperçoive, et donc transparaître dans nos discours oraux comme écrits. *Implicit-explicit* capture un ensemble de distinctions qui peuvent aussi être qualifiées comme suit : *unaware*/*aware*, *unconscious*/*conscious*, *automatic*/*controlled*.

Conséquemment, la deuxième partie de la problématique entourant les biais d'associations implicites réside dans certains modèles TAL. Il a été démontré que ces derniers peuvent, par inadvertance, capturer, refléter ou amplifier une multitude de préjugés et de stéréotypes contenus dans les données textuelles non annotées sur lesquelles ils sont entrainés (Bolukbasi *et al.*, 2016 ; Caliskan *et al.*, 2017). En effet, ces modèles font émerger des données des informations intéressantes, dont des relations sémantiques : en plus des analogies appropriées (de genre dans le cas présent) telles que *king-queen/man-woman*, ces modèles font aussi des analogies stéréotypées telles que *doctor-nurse/man-woman* (Bolukbasi *et al.*, 2016). Ceci s'explique par le fait que l'humain projette ses préjugés et ses stéréotypes dans les textes qu'il produit (Lauscher *et al.*, 2019). La problématique réside précisément dans le fait que ces modèles ont vocation à être utilisés dans une multitude d'applications ; il devient problématique d'utiliser un modèle biaisé pour des tâches de traduction automatique, certaines liées à la linguistique légale, d'autres liées à l'analyse de sentiment, etc.

Ces notions nous amènent à aborder le sujet des biais véhiculés explicitement par les incels. Tel que mentionné dans la section 2.1., nous avons pu remarquer que, sur l'ensemble du forum `incels.is`, le contenu des discours, généralement parsemés de néologismes propres à la communauté, véhicule des préjugés et des stéréotypes de façon très explicite ; c'est la raison pour laquelle nous traiterons des biais d'associations liés aux incels comme *explicites.* Dans le cas présent, ces biais ne sont pas simplement produits de manière consciente, ils sont aussi fortement revendiqués et hautement négatifs. La propagation de tels préjugés et stéréotypes dans un modèle de représentation du sens lexical pourrait être grandement nocive pour les tâches subséquentes effectuées par celui-ci. C'est pourquoi nous cherchons à savoir de quelle façon ces biais se manifestent dans un certain modèle d'apprentissage automatique que nous expliquerons plus en détail dans la prochaine section. Considérant que les incels revendiquent leurs biais, nous croyons que les forces d'associations des biais revendiqués par les incels, et de ceux associés à l'usage des néologismes, se manifesteront de façon différente des biais implicites connus et/ou impliquant des termes standards.

2.3. Le modèle de plongement lexical

L'utilisation de techniques de TAL fait partie intégrante de cette étude, et ce, à plusieurs niveaux. Ces techniques utilisent l'intelligence artificielle (IA), un ensemble de systèmes informatiques qui imitent l'intelligence humaine. L'apprentissage automatique, aussi appelé apprentissage machine, est un domaine de l'IA qui permet, entre autres, de créer des modèles statistiques qui reconnaissent des motifs (*patterns*) dans les données et qui, en TAL, permet de faire des prédictions et d'extraire des informations pertinentes d'un texte à des fins d'analyse.

La sémantique vectorielle est une méthode d'apprentissage automatique qui se trouve à être l'approche standard de représentation du sens lexical en TAL (Jurafsky et Martin, 2021). L'idée générale étant que deux mots apparaissant dans des contextes très similaires ont des significations

similaires. Cette étude s'intéresse particulièrement au *word embedding* (ou « plongement lexical » en français), un type de modèle vectoriel qui fonctionne d'abord par l'apprentissage automatique de la représentation du sens des mots à partir de leurs distributions dans le corpus non annoté sur lequel le modèle est entraîné. Il existe deux approches pour le calcul des plongements, celle qui nous intéresse est de type statique. Dans ce processus, chaque mot est associé à un seul vecteur court et dense (appelé *embedding*, ou « plongement » en français) dont la valeur est un nombre réel pouvant être négatif ou positif, dans un espace vectoriel de 50 à 1000 dimensions (Jurafsky et Martin, 2021). Un espace vectoriel est un ensemble de vecteurs caractérisés par leur dimension. Il est impossible pour l'humain d'interpréter ou de visualiser ces vecteurs à *n* dimensions. Dans le but de mesurer la similarité sémantique entre deux termes vectorisés, il est nécessaire d'utiliser une métrique qui est capable de traiter deux vecteurs et de donner une mesure de leur similarité. La métrique de similarité la plus commune est le cosinus de l'angle entre ces deux vecteurs, une mesure de corrélation ; nous préciserons cela dans la section 3.3. Jurafsky et Martin (2021) présentent la façon la plus élémentaire de visualiser le sens d'un mot vectorisé : produire la liste des mots les plus similaires.

Dans la section 2.1., nous avons parlé de la propriété qu'ont les vecteurs des modèles de plongement lexical à capturer les biais humains latents dans les données textuelles sur lesquelles le modèle est entraîné et le fait que ce phénomène soit dû à leur capacité à capturer les relations sémantiques qui émergent des données (Jurafsky et Martin, 2021). Considérant que l'humain projette ses biais dans les données textuelles qu'il produit, en matière de cooccurrences de termes biaisées, il n'y a donc rien d'étonnant dans le fait que les stéréotypes et les préjugés humains émergent des données et se propagent aux modèles, et ce, en vertu du fonctionnement de ces derniers.

3. Méthodes

Cette section traitera de la méthode entourant le corpus, du modèle vectoriel utilisé et du test d'associations appliqué. Chaque étape implique l'écriture d'un code à l'aide du langage de programmation Python. Les bibliothèques et trousses à outils utilisées pour le code sont toutes gratuites et sont des sources ouvertes (open source).

2.1. Conception et description des corpus

Dans cette section nous décrivons les corpus et comment nous avons compilé plusieurs millions de tokens dans le but d'évaluer les forces d'associations des biais projetés par les incels dans leurs données textuelles en termes de distance entre les vecteurs.

Le corpus d'étude a été récupéré sur le site `incels.is` de façon automatique grâce à la bibliothèque python Beautiful Soup (Richardson, 2021). Les publications récupérées ont été mises en ligne sur le site entre le 4 novembre 2017 (date de création du forum) et le 14 novembre 2021 (date de création du corpus). Le corpus complet contient un total de 271 k publications originales, accompagnées de leurs titres et de leurs commentaires, et a été organisé dans un document XML. Notons que le corpus d'étude ne comporte que le titre et le contenu textuel des publications originales, sans les commentaires. Dans une première analyse quantitative du corpus d'étude, à l'aide de la bibliothèque python `spaCy` (Honnibal et Ines, 2017), nous avons pu établir le nombre

de tokens (qui s'élève à plus de 26,8 millions de tokens, dont 318 k sont distincts) et nous avons identifié les tokens et les bigrammes les plus fréquents.

En ce qui concerne le corpus contrôle, provient a quant à lui de Reddit et a été récupéré à l'aide de la trousse à outils python `ConvoKit` (Chang *et al.*, 2020) ; celle-ci met à disposition des outils et des corpus. Pour pouvoir faire une comparaison adéquate entre les deux corpus, nous avons recueilli les données de 10 subreddits[1] couvrant tous des sujets différents. Considérant qu'une grande quantité de subreddits consistent en des publications qui sont essentiellement des images et des vidéos, nous les avons sélectionnés au hasard parmi une liste regroupant des subreddits dont les publications sont principalement textuelles. Certaines de ces sous-sections de Reddit sont particulièrement volumineuses et contiennent une énorme quantité de données (fichier de plus de 2 gigaoctets). Afin d'éviter autant que possible tout biais potentiel, nous avons limité le nombre de publications par subreddit à 30 k. 280 k publications ont été extraites au total et, de la même manière qu'avec le corpus d'`incels.is`, nous avons effectué le compte des tokens ; celui-ci s'élève à 44,3 millions de tokens, dont 291 k tokens sont distincts.

3.1. Vectorisation des corpus

Pour la vectorisation des deux corpus, nous avons opté pour l'algorithme `Word2Vec` (Mikolov *et al.*, 2013). Il s'agit de la méthode de plongement lexical statique dont nous avons expliqué le fonctionnement à la section 2.3. Afin d'y avoir accès, nous avons utilisé la bibliothèque python Gensim (Řehůřek et Sojka, 2010) qui permet d'obtenir un modèle de vecteurs entraîné sur un corpus choisi. Dans le but de faciliter la comparaison, nous avons gardé les mêmes paramètres pour les deux corpus : chaque terme correspond à un vecteur de 300 dimensions et le terme doit apparaître un minimum de 10 fois dans le corpus pour qu'un vecteur lui soit attribué.

Pour nous assurer de la validité des représentations de chaque corpus, nous avons effectué quelques tâches de similarité sémantique, incluant plusieurs tests d'analogies sémantiques et syntaxiques. Ces tests consistent en une série d'analogies de quatre termes dans laquelle deux termes ayant une relation sémantique ou syntaxique sont présentés au modèle (p. ex. *possible, impossible*), puis un troisième terme (p. ex. *tasteful*) pour lequel le modèle doit identifier un terme partageant la même relation sémantique que la première paire (p. ex. *untasteful*). Le premier test de référence appliqué au corpus est le *Google Analogy Test Set*, qui comporte plusieurs tests d'analogies testant, entre autres, les relations capitale-pays, les termes familiaux (p. ex. *brother-sister/dad-mom*) et les transformations d'adjectif à adverbe (p. ex. *amazing-amazingly/typical-typically*). Ces tests fournissent un score selon l'habileté de l'ensemble de vecteurs à identifier le terme attendu/voulu. Le score d'exactitude des résultats du modèle provenant du corpus d'étude est de 0.38, ce qui est très faible considérant la taille du modèle. Toutefois, un tel résultat n'est pas surprenant puisque le test a été conçu pour un modèle entraîné sur des corpus beaucoup moins spécifiques et couvrant une variété de sujets beaucoup plus grande. Il est fort probable que certains termes utilisés dans les tests ne se retrouvent pas ou peu dans notre modèle. Le score du corpus de contrôle est similaire, soit 0.34. Encore une fois, il est possible que les termes ne fassent pas tous partie des sujets abordés dans les subreddits choisis.

[1] Titres des subreddit choisis : r/LetsNotMeet, r/PettyRevenge, r/AskHistorians, r/CasualConversation, r/FuckYou, r/TalesFromTechSupport, r/bodybuilding, r/skiing, r/backpacking, r/TalesFromRetail

Nous avons manuellement vérifié certaines analogies, soit les plus pertinentes pour notre corpus. Le Tableau 1 présente quelques exemples d'analogies réussies par le modèle.

Tableau 1. Résultats de certains tests d'analogies. Les définitions sont tirées du glossaire d'`incels.is` (2022).

Relation	**Sémantique – Genre**					
Termes	Chad	→	Man	Stacy	→	Woman
Définitions	Homme attirant physiquement, le plus souvent blanc.			L'équivalent féminin d'un C*had.*		
Relation	**Sémantique – Origine ethnique**					
Termes	Chadpreet	→	Indian	Chaddam	→	Arab
Définitions	*Chad* avec des origines et traits indiens.			*Chad* avec des origines et traits arabes.		
Relation	**Syntaxique – Dérivation d'un nom en Adjectif**					
Termes	Betabux	→	Betabuxxed	Cuck	→	Cuckold
Définitions	Un homme qui subvient aux besoins financiers d'une femme. Terme péjoratif qui sous-entend que la femme n'a aucune attraction pour lui et l'utilise que pour son argent.			Un homme qui est en relation avec une femme infidèle ou qui profite de lui. Insulte pour désigner un homme « faible ».		
Relation	**Syntaxique – Verbe conjugué et Verbe à l'infinitif**					
Termes	Ascend	→	Ascending	Looksmax	→	Looksmaxxing
Définitions	Un incel *Ascend* (s'élève) lorsqu'il perd sa virginité.			Une personne qui *looksmax* est en processus d'améliorer son apparence physique (soin de la peau, entrainement physique, chirurgie, etc.)		

3.2. Mesure des biais d'associations : le Word-Embeddings Association Test

S'il est possible de dire que les représentations vectorielles des modèles de plongements capturent les biais implicites qui proviennent des données textuelles, c'est qu'il existe des moyens de les déceler. Caliskan *et al.* (2017) sont parmi les premiers à avoir développé une telle technique. Afin de déterminer si les vecteurs ont la capacité de capturer les biais d'associations latents dans les données textuelles sur lesquelles ils sont entrainés, ils se sont inspirés des travaux effectués dans le cadre de l'IAT (Greenwald *et al.*, 1998 ; Greenwald et Banaji 1995 ; Nosek *et al.*, 2002 ; Project Implicit, s. d. a ; Project Implicit, s. d. b). Pour ce faire, ils ont élaboré un test statistique homologue à l'IAT : le Word-Embeddings Association Test (WEAT), pouvant être appliqué vecteurs d'un modèle de plongement lexical. Si l'IAT mesure le temps de réaction des sujets, le WEAT, pour sa part, mesure la distance, ou plus précisément l'angle entre deux vecteurs de termes cibles ; il s'agit de la métrique de similarité cosinus dont nous avons parlé dans la section 2.3. Il est important de spécifier que contrairement à l'IAT, le WEAT est teste des vecteurs de termes et non des personnes.

Le WEAT calcule l'association différentielle entre les termes des deux ensembles de concepts cibles en fonction de la moyenne de similarité qu'ils entretiennent avec les termes des deux ensembles d'attributs ; autrement dit, les forces d'associations.

À partir d'un corpus de Google News (environ 100 milliards de tokens) vectorisé avec `Word2Vec`, Caliskan *et al.* (2017) ont mesuré les forces d'associations entre les vecteurs des termes tirés des ensembles de concepts cibles et d'attributs de l'IAT. Ils ont ensuite comparé leurs résultats avec ceux de l'IAT. Les résultats étant tous significatifs, et concordant tous avec ceux de l'IAT, ils démontré que les modèles d'apprentissage automatique tels que `Word2Vec` sont en mesure de capturer les biais d'associations latents dans les données textuelles brutes sur lesquelles ils sont entraînés. Dans cette étude, comme nous cherchons à mesurer les forces d'associations des biais des incels, le WEAT correspond parfaitement à nos besoins.

Bien que le code créé par Caliskan *et al.* (2017) pour le WEAT n'est pas libre d'accès, plusieurs sont parvenus à reproduire les calculs de façon satisfaisante. La bibliothèque Python Responsibly (Hod, 2018) permet une implémentation facile du test pour obtenir rapidement les résultats dans le même format que les présentent Caliskan *et al.* (2017) dans leur étude.

La première étape pour appliquer le WEAT est de déterminer les concepts cibles (2) et les attributs (2) à évaluer, ainsi que les termes composant chaque ensemble (4). Un corpus de publications tiré d'un forum tel que `incels.is` comporte des différences majeures avec un corpus tel que celui de Google News ; notamment l'utilisation fréquente des nombreux néologismes, le format, les sujets abordés, etc. Cela signifie que même si nous avons conservé quelques tests et termes utilisés par Caliskan *et al.* (2017) dans leurs tests, nous avons dû substituer deux de leurs catégories de concepts cibles. Au Tableau 2 figurent les paires de concepts cibles et d'attributs, tirés de Caliskan *et al.* (2017), que nous avons retenus pour nos tests.

Comme les néologismes prennent une place importante au sein des communications écrites des incels, nous avons aussi créé quelques catégories dérivées de celle des tests de Caliskan *et al.* (2017) afin d'y inclure les néologismes et de vérifier si ceux-ci affectent d'une quelconque façon les forces d'associations des biais. Les catégories Termes féminins et Termes masculins (voir Tableau 3) sont des catégories dérivées des catégories *Male terms* et *Female terms* tirées de Caliskan *et al.* (2017) : à ces ensembles de termes standards ont été ajoutés deux termes, qui ne sont pas des néologismes, afin que ces ensembles aient le même nombre de termes que les ensembles des autres catégories de concepts cibles que nous avons créées (cela est nécessaire pour le bon fonctionnement du WEAT). Pour nos propres tests, nous nous sommes concentrées sur les attributs d'évaluation *pleasant/unpleasant.* Tous les tests que nous avons élaborés nous-mêmes ou dérivés sont présentés dans le Tableau 3.

Les résultats du WEAT sont calculés en vérifiant la compatibilité du premier concept cible donné avec le premier attribut donné (p. ex. fleurs/insectes–agréable/désagréable) (Caliskan *et al.*, 2017). L'ordre dans les Tableaux 2 et 3 a donc pour seul but de démontrer les catégories de concepts cibles et d'attributs. Certains tests présenteront les concepts cibles dans d'autres ordres, dans le but d'évaluer certaines associations.

Tableau 2. Catégories de concepts cibles et d'attributs tirés de Caliskan *et al.* (2017).

Concepts cibles		Attributs	
Instruments	Weapons	Pleasant	Unpleasant
Female names (remplacé par Female terms)	Male names (remplacé par Male terms)	Family	Career
Math	Arts	Male terms	Female terms
Science	Arts	Male terms	Female terms

Tableau 3. Catégories de concepts cibles et d'attributs élaborés selon notre corpus d'étude.

Concepts cibles	Attributs	
Termes masculins	Pleasant	Unpleasant
Termes féminins		
Incels		
Néologismes masculins		
Néologismes féminins		
Néologismes et Termes féminins		

4. Résultats

4.1. Analyses de fréquences

Dans le but d'offrir une description plus générale des communications écrites de la communauté étudiée, nous avons effectué des analyses de fréquences sur le corpus d'étude. Nous présentons les résultats dans les Tableaux 4 et 5.

Tableau 4. Fréquences des lemmes *woman*, *girl*, *femoid/foid*, *man* et *incel* dans le corpus d'étude.

#	Lemmes	Nombre d'occurrences	% des tokens du corpus
1	Woman	116 456	0.43
2	Girl	54 511	0.20
3	Femoid/Foid	62 171	0.23
4	Man	99 028	0.37
5	Incel	80 138	0.30

Tableau 5. Fréquences des lemmes *woman*, *girl* et *man* dans le corpus contrôle.

#	Lemmes	Nombre d'occurrences	% des tokens du corpus
1	Woman	24 276	0.055
2	Girl	24 346	0.055
3	Man	44 704	0.10

Il est difficile de dresser une liste exhaustive de termes et des néologismes faisant référence aux femmes, aux hommes et aux incels puisqu'ils sont très nombreux. Par exemple, il existe des termes référant aux incels qui sont spécifiques selon la couleur de peau, l'origine ethnique, les intérêts, la taille, le degré d'attirance physique, le statut socioéconomique de la personne, etc. Pour l'analyse

de fréquence nous nous sommes donc concentrées sur des termes apparaissant dans la liste des 200 premiers tokens les plus fréquents. Il est pertinent de noter que le terme *femoid,* bien qu'absent de la liste, est un synonyme direct de *foid* (rappelons que *foid* est l'abréviation de *femoid*).

Les résultats de ces analyses de fréquences nous permettent d'en apprendre un peu plus sur nos corpus, et plus particulièrement sur la dichotomie de genre dans les communications des incels. Nous pouvons constater que le lemme *woman* occupe un plus grand pourcentage du corpus d'étude (0.43 %) que le lemme *man* (0.37 %). Le corpus contrôle présente les proportions inverses : le lemme *woman* n'occupe que 0.055 % du corpus et le lemme *man* occupe 0.10 %.

4.2. Word-Embedding Association Test — Corpus d'étude

Tableau 6. Les lignes 1 à 4 présentent les résultats de l'application du WEAT sur les catégories tirées de Caliskan *et al.* (2017). Les lignes 5 à 9 présentent les résultats significatifs de l'application du WEAT sur nos catégories. Néo. = Néologismes ; Fém. = Féminin ; Masc. = Masculin.

Test	Concepts cibles			Attributs			Nc	Na	d	p
1	Male terms	vs	Female terms	Career	vs	Family	8x2	8x2	0.4263	0.21
2	Math	vs	Arts	Male terms	vs	Female terms	6x2	8x2	-0.2714	0.61
3	Science	vs	Arts	Male terms	vs	Female terms	6x2	8x2	-0.4304	0.76
4	Instruments	vs	Weapons	Pleasant	vs	Unpleasant	11x2	25x2	1.1969	0.00095
5	Female terms	vs	Male terms	Pleasant	vs	Unpleasant	8x2	25x2	1.0651	0.012
6	Néo. + Termes fém.	vs	Incels	Pleasant	vs	Unpleasant	10x2	25x2	0.772	0.046
7	Termes fém.	vs	Incels	Pleasant	vs	Unpleasant	10x2	25x2	0.8604	0.029
8	Termes fém.	vs	Néo. fém.	Pleasant	vs	Unpleasant	10x2	25x2	0.8762	0.023
9	Néo. masc.	vs	Néo. fém.	Pleasant	vs	Unpleasant	10x2	25x2	0.7484	0.051

Nous présentons les résultats obtenus à certains WEATs appliqués sur le corpus d'étude dans le Tableau 6 ; nous avons sélectionné les plus pertinents parmi plus de 40 tests effectués. Il est pertinent de réitérer le but de notre étude ainsi que nos questions de recherche. Nous souhaitions mettre en lumière le statut des biais d'associations de la communauté d'`incels.is` et la façon dont les modèles de plongement lexical les capturent. De ce fait, nous nous sommes questionnées à savoir si les biais implicites connus étaient partagés par les membres de la communauté d'`incels.is`, si les biais qu'ils revendiquent se distinguent des biais connus et si ceux véhiculés par l'entremise de néologismes se distinguent des biais véhiculés par l'entremise de termes standards, le tout, dans les modèles de plongement lexical.

Trois (**1**, **2** et **3**) des quatre tests des biais d'associations implicites connus (**1** à **4**) ne reproduisent pas les résultats obtenus par Caliskan *et al.,* (2017) ; aucun de nos résultats à ces tests n'est significatif : **1** $p = 0.21$; **2** $p = 0.61$; **3** $p = 0.76$. Seul le test **4** s'est avéré significatif ($p < 0.001$) confirmant la présence du biais |*instruments-pleasant/weapons-unpleasant*| dans les données textuelles.

Nous avons obtenu des résultats significatifs aux tests **5** (p = 0.012), **6** (p = 0.046), **7** (p = 0.029), **8** (p = 0.023) et **9** (p = 0.051). Il s'agit de tests que nous avons élaborés nous-mêmes et ceux-ci confirment la présence de biais que nous considérons comme propres à la communauté, pour le moment.

4.3. Word-Embedding Association Tests – Corpus contrôle

Dans le but de pousser notre analyse plus loin, nous avons appliqué trois WEATs au corpus contrôle Reddit. Nous présentons ces résultats dans le Tableau 7. Le corpus contrôle obtient lui aussi un résultat très significatif au test **4** ($p < 0.0001$) et des résultats non significatifs au test **5** (p = 0.32).

Tableau 7. Résultats de l'application du WEAT sur le corpus contrôle Reddit.

Test	Concepts cibles			Attributs			Nc	Na	d	p
1	Male terms	vs	Female terms	Career	vs	Family	8x2	8x2	0.4578	0.2
2	Maths	vs	Arts	Male terms	vs	Female terms	6x2	8x2	0.3338	0.3
3	Science	vs	Arts	Male terms	vs	Female terms	6x2	8x2	0.8872	0.07
4	Instruments	vs	Weapons	Pleasant	vs	Unpleasant	11x2	24x2	1.4687	0.0001
5	Female terms	vs	Male terms	Pleasant	vs	Unpleasant	8x2	24x2	0.2406	0.32

5. Discussion

Avant de discuter de nos résultats, un retour sur nos questions de recherche est de mise. Tout d'abord, il s'avère que seul un des biais implicites connus que nous avons testés se retrouve dans les communications écrites des incels. Considérant que les biais connus tirés de l'IAT étaient surtout reliés au genre, il était attendu que les tests (**1**, **2**, **3** et **4**) en lien avec les catégories *male terms* et *female terms* soient significatifs. Le type de communications à l'étude pourrait être à l'origine du manque de significativité de certains de ces tests d'associations. Dans le but de vérifier si la significativité serait semblable pour un autre corpus de type forum, nous avons appliqué ces mêmes tests sur le corpus contrôle de Reddit. Nous avons obtenu des résultats similaires à ceux du corpus d'étude : le seul test ayant produit des résultats significatifs est le test **4**, testant le biais |*instruments-pleasant/weapons-unpleasant*| (un des biais implicites connus). Il serait donc intéressant de tester cette hypothèse sur d'autres corpus tirés de forums.

Nous cherchions également à vérifier si les biais explicites se démarqueraient particulièrement des biais implicites connus dans un modèle de plongement. Nos résultats permettent une réponse claire à cette question : on ne peut distinguer les biais implicites des biais explicites (et vice versa). On peut donner l'exemple du biais d'association |*female terms-pleasant/male terms-unpleasant*| (des associations dont le test s'est avéré significatif) qui ne serait pas un biais explicite si on considère ce qui est connu du discours des incels d'`incels.is`, en termes de misogynie prononcée et revendiquée par exemple. En s'appuyant seulement sur les résultats obtenus, il ne nous est pas non plus possible d'affirmer que le biais est implicite.

En ce qui concerne la question de la distinction entre les néologismes et les termes standards dans le modèle de plongement, le fait que le test 8 (|termes féminins-*pleasant*/néologismes féminins-*unpleasant*|) soit significatif suggère une nette distinction entre les néologismes et les termes standards référant à la femme. À nouveau, on ne peut déterminer le caractère implicite ou explicite de ce biais. S'il existe une différence entre ces deux catégories de termes féminins, il faut toutefois noter que nos résultats ne démontrent pas de différence significative entre les termes standards masculins et les néologismes masculins.

Pour pousser notre recherche plus loin, il y aurait plusieurs éléments à considérer. La misogynie exprimée par les incels d'`incels.is` est explicite, on le sait déjà par les nombreux exemples de publications dans le forum et par les recherches précédentes qui ont été effectuées par d'autres sur les incels (Pelzer *et al.*, 2021 ; Gothard *et al.,* 2021 ; Preston *et al.,* 2021 ; Jaki *et al.,* 2019 ; Farrell *et al.,* 2019). En nous appuyant sur nos résultats et sur ce que nous savons de cette communauté, nous proposons que cette misogynie serait exprimée d'abord et avant tout par les néologismes féminins ; ceux-ci étant davantage associés à des termes déplaisants que les termes standards féminins. Du côté implicite, on ne peut que poser une hypothèse qui serait éventuellement à vérifier par des recherches subséquentes. Cela étant, les résultats en lien avec les termes standards féminins nous laissent envisager un possible biais implicite plus positif en ce qui concerne l'utilisation de cette catégorie de termes. Avancer de cette hypothèse requiert tout de même une certaine prudence, puisqu'il y a plusieurs facteurs en jeu. Par exemple, un biais positif envers les femmes pourrait être possible si on s'attarde à l'origine du mot *incel* : amalgame de *involuntary et celibate* (célibataire involontaire). Malgré une haine considérable pour les femmes, ils considèrent tout de même leur situation comme involontaire. Cette volonté de partager leur vie avec une femme pourrait être un élément explicatif de ce biais. Un autre facteur important à prendre en compte est leur sarcasme. Il est vrai que les termes standards féminins sont plus facilement associés à des termes plaisants que les néologismes féminins, mais les incels font preuve d'un grand sarcasme dans leurs propos ; on ne peut ignorer que l'utilisation de termes plaisants dans leur discours est très souvent empreinte de négativismes, par exemple, dans le but de ridiculiser ou mépriser.

Pour continuer sur les hypothèses reliées à nos résultats, nous pouvons aussi traiter des biais envers les termes masculins (incels, néologismes masculins et termes standards masculins). Encore une fois, la distinction explicite et implicite est purement spéculative selon ce que nos résultats et les recherches précédentes laissent transparaître. En se penchant sur les biais d'associations des termes référant aux incels, on voit que ces termes sont davantage associés à des termes déplaisants, lorsque comparés à des termes standards féminins. Cela semble concorder avec leur perception d'eux-mêmes, qui communique une forte haine de soi. Puisque ce phénomène a été observé dans le corpus, mais aussi par d'autres auteurs (Pelzer *et al.*, 2021 ; Gothard *et al.,* 2021 ; Preston *et al.,* 2021 ; Jaki *et al.,* 2019 ; Farrell *et al.,* 2019), nous suggérons que ce biais soit explicite et, évidemment, véhiculé par les termes qui les désignent. Pour les autres termes masculins (néologismes et termes standards), il est plus difficile de postuler quelconque hypothèse : mis à part le test significatif |néologismes masculins-*pleasant*/néologismes féminins-*unpleasant*|, aucun autre test traitant de ces catégories de termes ne s'est avéré significatif.

Finalement, parmi un peu plus de 30 tests effectués afin d'analyser les possibles biais « propres aux incels », seulement 5 tests se sont avérés significatifs. Nous sommes forcées de constater que le niveau de toxicité et le caractère très explicite de leurs communications écrites n'affectent pas les représentations du sens lexical du modèle de plongement tel que nous l'avions imaginé. Les

résultats ont généré une multitude de questionnements quant aux raisons derrière ceux-ci. Cela porte entre autres à se demander si leur haine et leur mépris sont si prononcés et si généralisés que cela engendrerait des forces d'associations négligeables pour la majorité des WEATs testant des biais que nous considérions comme prévisibles ; en d'autres mots, la majorité de ces biais s'« annuleraient ».

6. Conclusion

Malgré les habitudes de communications explicitement très négatives et généralisées des incels d'`incels.is`, le WEAT a seulement détecté la présence de biais ne semblant pas correspondre à ce qu'on sait et ce à quoi on s'attend du profil toxique de cette communauté. Six tests se sont avérés significatifs : l'un révèle la présence d'un biais implicite connu (|*instruments-pleasant/weapons-unpleasant*|) un autre révèle le biais |*female terms-pleasant/male terms-unpleasant*|, et quatre tests subséquents révèlent un biais envers les termes référant à la femme, selon lequel les termes standards sont plus plaisants que les néologismes péjoratifs propres à cette communauté. Le reste des quelque 40 tests ne sont pas significatifs.

Nos résultats ont fait émerger plusieurs hypothèses et offrent des pistes dignes d'intérêt et pertinentes qu'il serait intéressant d'approfondir dans des recherches subséquentes. Notre corpus ainsi que son modèle vectorisé sont disponibles en ligne pour permettre une continuation des recherches transdisciplinaires sur le sujet. Nous considérons, entre autres, que l'utilisation d'un autre type de modèle de plongement pourrait être une façon efficace d'obtenir davantage d'informations sur les relations sémantiques des termes du corpus. Rappelons que nous avons utilisé le modèle de plongement statique `Word2Vec`, qui associe chaque terme à un seul vecteur, selon sa distribution dans le texte. Des modèles plus dynamiques, permettant d'associer chaque occurrence d'un terme à un vecteur en prenant en considération les occurrences précédentes, le contexte et bien plus, pourraient probablement faire émerger des informations sémantiques plus subtiles qu'un modèle statique ne pourrait capturer.

Dans le même ordre d'idées, l'une des majeures limitations de l'étude réside dans de la méthode statistique employée. Comme les domaines du TAL évoluent à une vitesse fulgurante, le WEAT de Caliskan *et al.* (2017) est déjà, en l'espace de 5 ans, désuet. De plus, certains problèmes et limitations lui ont aussi été attribués par d'autres auteurs : p. ex. Ethayarajh *et al.* (2019) ont démontré que le test surestime systématiquement les biais. Peut-être que le WEAT n'est pas le test le plus adéquat pour traiter les vecteurs d'un modèle de plongement lexical entraîné sur un corpus hautement et explicitement toxique et biaisé tel que celui de la communauté d'`incels.is`. Comme pour les modèles de plongement plus dynamiques, il serait donc possible de trouver une technique identifiant les biais d'associations dans ces modèles de façon plus optimale.

L'impact de nos résultats se fait ainsi surtout sentir dans les multiples perspectives de recherches reliées au discours des incels. Puisque notre étude implique des éléments reliés à la psychologie ainsi qu'à la linguistique, nous soutenons que nos résultats apportent des pistes intéressantes pour les deux domaines, minimalement. Une meilleure compréhension du discours, des néologismes et des biais des incels pourrait permettre de faciliter les communications avec eux et possiblement de désamorcer certaines situations ou de les encourager à aller cherche de l'aide.

Références

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V. et Kalai, A. 2016. *Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings*. https://doi.org/10.48550/arXiv.1607.06520

Caliskan, A., Bryson, J. J. et Narayanan, A. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183-186. https://doi.org/10.1126/science.aal4230

Chang, W. 2020. The monstrous-feminine in the incel imagination: investigating the representation of women as "femoids" on /r/Braincels. *Feminist Media Studies*, 1-17. https://doi.org/10.1080/14680777.2020.1804976

Ethayarajh, K., Duvenaud, D. et Hirst, G. 2019. *Understanding Undesirable Word Embedding Associations*. https://doi.org/10.48550/arXiv.1908.06361

Farrell, T., Fernandez, M., Novotny, J. et Alani, H. 2019. Exploring Misogyny across the Manosphere in Reddit. Dans *Proceedings of the 10th ACM Conference on Web Science - WebSci '19* (p. 87-96). ACM Press. https://doi.org/10.1145/3292522.3326045

Gothard, K., Dewhurst, D. R., Minot, J. R., Adams, J. L., Danforth, C. M. et Dodds, P. S. 2021. The incel lexicon : Deciphering the emergent cryptolect of a global misogynistic community. *arXiv:2105.12006 [cs]*. http://arxiv.org/abs/2105.12006

Greenwald, A. G. et Banaji, M. R. 1995. Implicit social cognition : Attitudes, self-esteem, and stereotypes. *Psychological Review*, *102*(1), 4-27. https://doi.org/10.1037/0033-295X.102.1.4

Greenwald, A. G., McGhee, D. E. et Schwartz, J. L. K. 1998. Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, *74*(6), 1464-1480. https://doi.org/10.1037/0022-3514.74.6.1464

Hod, S. 2018. Responsibly : Toolkit for Auditing and Mitigating Bias and Fairness of Machine Learning Systems. http://docs.responsibly.ai/

Holtz, P., Kronberger, N. et Wagner, W. 2012. Analyzing Internet Forums. *Journal of Media Psychology*, *24* (2), 55-66. https://doi.org/10.1027/1864-1105/a000062

Honnibal, M. et Ines, M. 2017. *spaCy 2 : Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing*. https://spacy.io/

Incels.is. 2021, 23 novembre. Incel Term Glossary. Dans *Incel Wiki*. Récupéré le 5 décembre 2021 de https://incels.wiki/w/Incel_Term_Glossary

Incels.is - Involuntary Celibate. (s. d.). Incels.is - Involuntary Celibate. https://incels.is/

Jaki, S., De Smedt, T., Gwóźdź, M., Panchal, R., Rossa, A. et De Pauw, G. 2019. Online hatred of women in the Incels.me forum : Linguistic analysis and automatic detection. *Journal of Language Aggression and Conflict*, *7*(2), 240-268. https://doi.org/10.1075/jlac.00026.jak

Jurafsky, D. et Martin, J. H. 2021. Chapter 6 : Vector Semantics and Embeddings. Dans *Speech and Language Processing* (3 rd ed. draft). https://web.stanford.edu/~jurafsky/slp3/6.pdf

Kassam, A. 2018, 26 avril. Woman behind « incel » says angry men hijacked her word « as a weapon of war ». *The Guardian*, section World news. https://www.theguardian.com/world/2018/apr/25/woman-who-invented-incel-movement-interview-toronto-attack

Lauscher, A., Glavaš, G., Ponzetto, S. P. et Vulić, I. 2019. *A General Framework for Implicit and Explicit Debiasing of Distributional Word Vector Spaces*. https://doi.org/10.48550/arXiv.1909.06092

Mikolov, T., Chen, K., Corrado, G. et Dean, J. 2013. *Efficient Estimation of Word Representations in Vector Space*. https://arxiv.org/abs/1301.3781v3

Nosek, B., Banaji, M. et Greenwald, A. 2002. Harvesting Implicit Group Attitudes and Beliefs from a Demonstration Web Site. *Group Dynamics-theory Research and Practice - GROUP DYN-THEORY RES PRACT*, *6*. https://doi.org/10.1037//1089-2699.6.1.101

O'Malley, R. L., Holt, K. et Holt, T. J. 2020. An Exploration of the Involuntary Celibate (Incel) Subculture Online. *Journal of Interpersonal Violence*, 0886260520959625. https://doi.org/10.1177/0886260520959625

Pelzer, B., Kaati, L., Cohen, K. et Fernquist, J. 2021. Toxic language in online incel communities. *SN Social Sciences*, *1* (8), 213. https://doi.org/10.1007/s43545-021-00220-8

Preston, K., Halpin, M. et Maguire, F. 2021. The Black Pill: New Technology and the Male Supremacy of Involuntarily Celibate Men. *Men and Masculinities*, *24*(5), 823-841. https://doi.org/10.1177/1097184X211017954

Project Implicit. s. d.-a. *About the IAT*. https://www.projectimplicit.net/resources/about-the-iat/

Project Implicit. s. d. -b. *Project Implicit*. https://implicit.harvard.edu/implicit/canada/

Řehůřek, R. et Sojka, P. 2010. *Software Framework for Topic Modelling with Large Corpora*. University of Malta. https://is.muni.cz/publication/884893/en/Software-Framework-for-Topic-Modelling-with-Large-Corpora/Rehurek-Sojka

Richardson, L. 2021. *Beautiful Soup* (version 4.8.1). http://www.crummy.com/software/BeautifulSoup/
Service canadien du renseignement de sécurité. 2020, avril. *Rapport public du SCRS 2019*. https://www.canada.ca/fr/service-renseignement-securite/organisation/publications/rapport-public-2019.html
Service canadien du renseignement de sécurité. 2021, avril. *Rapport public du SCRS 2020*. https://www.canada.ca/fr/service-renseignement-securite/organisation/publications/rapport-public-2020.html