Comparaison de méthodes pour la détection du discours des *incels* sur Reddit

Camille Demers* — **Dominic Forest***

* *Université de Montréal, École de bibliothéconomie et des sciences de l'information*

RÉSUMÉ. *Les* incels *(célibataires involontaires) regroupent typiquement des hommes se trouvant dans l'incapacité de former des relations amoureuses ou intimes et partageant par conséquent des opinions négatives à l'endroit des femmes. Compte tenu de la gravité des attaques commises par des individus incels et de leur propension à se radicaliser sous l'effet de chambres d'écho, il s'avère plus que nécessaire de détecter le discours de ces communautés virtuelles. Cette étude compare la performance de différents systèmes de détection du discours incel utilisant une approche d'apprentissage par sacs de communautés. Les expérimentations menées permettent de comparer l'efficacité de diverses représentations vectorielles pour entraîner différents algorithmes d'apprentissage supervisé à détecter le discours incel dans un corpus de commentaires provenant de Reddit. Nos modèles les plus performants obtiennent une mesure-F globale variant entre 82,35 % en phase d'apprentissage et 79,70 % en phase de test.*

MOTS-CLÉS : *détection de propos misogynes,* incel*, classification supervisée, comparaison de méthodes, Reddit.*

TITLE. *Comparison of methods for detecting incel speech on Reddit*

ABSTRACT. *Incels (involuntary celibates) typically bring together men that are unable to form romantic or intimate relationships, and therefore share negative opinions about women. Given the seriousness of attacks committed by incel individuals as well as their propensity to radicalize under the effect of echo chambers, it is more than necessary to detect the discourse of these virtual communities. This study compares the performance of various incel speech detection systems using a bag-of-communities learning approach. The experiments carried out compare the effectiveness of various vector representations for training supervised learning algorithms to detect incel speech in a corpus of comments from Reddit. Our best-performing models achieve a macro F-score ranging from 82,35% in the learning phase to 79,70% in the test phase.*

KEYWORDS: *online misogyny detection, incel, supervised classification, comparison of methods, Reddit.*

1. Introduction

> Incel, /ˈɪn.sel/
> Nom masculin : « Membre d'une communauté virtuelle composée principalement de jeunes hommes ne se sentant pas attirant envers les femmes et partageant régulièrement des opinions négatives à leur endroit.»
>
> (Hornby, 2020)

Le 7 novembre 2017, suivant l'adoption d'une nouvelle politique interdisant toute forme de contenu prônant ou incitant à la violence sur sa plateforme, le réseau social Reddit bannit le forum *r/Incels*, regroupant alors plus de 40 000 membres (Hauser, 2017). Un an plus tard, le 15 octobre 2018, le forum en ligne *Incels.me* est suspendu du domaine .me pour violation de sa politique en matière d'abus et de promotion de la violence et des discours de haine (.ME, 2018). Malgré la fermeture de ces espaces numériques, les *incels*, ou « célibataires involontaires », désignent encore aujourd'hui un ensemble de communautés virtuelles formées majoritairement d'hommes partageant une incapacité à lier des relations avec les femmes et alimentant par conséquent une subculture caractérisée par la propagation d'idéologies aux tendances misogynes et antiféministes (Halpin, 2022 ; Meier et Sharp, 2024). Selon le Réseau de sensibilisation à la radicalisation de l'Union européenne (RSR, 2021) :

> Les *Incels* sont persuadés que leur incapacité à avoir des relations sexuelles est due à des facteurs génétiques, des processus évolutivement prédéterminés de sélection du partenaire, ainsi qu'aux structures sociales. Ils pensent que les femmes ne les trouvent pas séduisants et qu'elles ne s'intéressent qu'aux beaux « mâles alpha » (également appelés « chads »). Fréquemment mentionnée chez les incels, la « règle 80/20 » signifie que les 20 % des hommes les plus attirants ont monopolisé 80 % des femmes.

Bien que les espaces de discussion des *incels* se voient fréquemment censurés par les plateformes qui les hébergent, les activités de ces communautés ont franchi les frontières du numérique à un certain nombre de reprises au cours des dernières années, ayant donné lieu à des actes de violence extrême de la part d'individus s'identifiant comme *incels*. L'un des événements les plus médiatisés à cet égard fut la tuerie d'Isla Vista perpétrée en mai 2014, où Elliot Rodger a abattu six personnes après avoir annoncé son plan de représailles envers les femmes dans une vidéo publiée sur sa chaîne YouTube quelques heures avant l'événement : *« You forced me to suffer all my life, and now I'll make you all suffer »* (Associated Press, 2014). Cet événement a contribué par la suite à motiver l'attaque au camion-bélier ayant eu lieu en avril 2018 à Toronto, où Alek Minassian a happé dix personnes à bord d'un véhicule de location après avoir inauguré l'événement sur sa page Facebook, en saluant au passage son prédécesseur : *« The Incel Rebellion has already begun ! We will overthrow all the Chads and Stacys ! All hail the Supreme Gentleman Elliot Rodger ! »* (Nadeau, 2022). Plus récemment, en

juin 2023, Geovanny Villalba-Aleman s'est infiltré dans la salle d'un cours en études de genre à l'université de Waterloo et y a poignardé trois personnes (Shetty, 2023). Bien que l'auteur de cet acte ne se soit pas identifié comme un *incel*, son geste a fait l'objet de célébrations au sein de ces communautés (Halpin *et al.*, 2024).

Depuis la mise à jour de sa politique d'utilisation en 2017, la plateforme Reddit bannit régulièrement des forums de discussion occupés par des *incels*, cependant la fermeture de ces espaces de discussion ne semble pas véritablement arrêter la progression de ces communautés. Ainsi, après que le forum *r/Incels* a été supprimé en 2017, seuls quelques jours ont suffit pour que ses utilisateurs migrent vers de nouveaux espaces, notamment *r/Braincels* (Ribeiro *et al.*, 2021), qui a à son tour été banni en 2019, suivi de *r/AskAnIncel*, banni en 2020. Plus d'actions sont donc requises pour modérer la radicalisation des propos émis au sein des forums de discussion des *incels* et pour éviter la perpétration de nouveaux actes de violence par ces individus.

Différents projets de loi ont été proposés au cours des dernières années pour encadrer les pratiques d'utilisation des médias sociaux et pour favoriser la mise en place d'environnements numériques qui soient plus sécuritaires. En février 2024, le gouvernement du Canada a notamment présenté le projet de loi C-63 sur les préjudices en ligne (gouvernement du Canada, 2024), visant plus particulièrement sept types de contenu préjudiciable dont le contenu fomentant la haine et le contenu incitant à la violence. Ce projet de loi prévoit la mise en place de mesures de protection des publics ainsi que la responsabilisation des plateformes numériques à l'égard de ces contenus, notamment en obligeant la mise en œuvre de mesures pour identifier les contenus préjudiciables et en rendant ces contenus inaccessibles au public dans des délais raisonnables (gouvernement du Canada, 2024 ; Benmoussa *et al.*, 2024).

Étant donné la difficulté à modérer les propos des communautés *incels* en raison de l'anonymat de leurs membres et de leur propension à se radicaliser sous l'effet de chambres d'écho, le développement d'outils permettant de détecter le discours des *incels* constitue une piste d'action concrète visant à mettre en œuvre les mesures proposées par de tels projets de loi. Cet article s'inscrit dans cette perspective. Il rend compte d'expérimentations visant à évaluer différents systèmes de détection automatique permettant d'extraire les spécificités du discours des *incels* sur Reddit en le comparant aux autres formes de discours émises sur cette plateforme. Les retombées associées au développement de ce type d'outils visent à améliorer les fonctionnalités de modération déjà en place pour signaler aux utilisateurs de Reddit la sensibilité des contenus associés au discours véhiculé au sein des espaces de discussion des *incels*[1]. De telles fonctionnalités ont déjà fait leur place au sein des plateformes du groupe Meta [2], sous la forme d'écrans d'avertissement présentés à l'utilisateur avant d'afficher le contenu

1. https://www.reddit.com/r/help/comments/aayoxb/what_is_a_quarantined_subreddit/
2. https://transparency.meta.com/enforcement/taking-action/context-on-sensitive-misleading-content/

sensible. Ce type d'outils permet d'ailleurs de suivre les migrations des communautés produisant ces contenus afin d'en faciliter la suppression.

La suite de l'article est organisée de la manière suivante. Nous dressons d'abord un portrait de travaux portant sur la détection de discours misogynes et sexistes ainsi que sur l'analyse du discours *incel* à proprement parler. La section « Méthodologie » décrit les étapes de constitution d'un corpus de commentaires issus de forums de discussion *incels* et non-*incels* sur Reddit ainsi que le paramétrage de différents modèles de classification visant à détecter ce type de discours. La section « Résultats et discussion » compare la performance des modèles en phase d'apprentissage et de test, puis analyse les meilleurs paramètres de détection. Des recommandations liées aux développements futurs de ce type d'approche sont formulées en conclusion.

2. État de la question

Un important nombre de travaux en traitement automatique des langues, en lexicométrie et en fouille de textes ont récemment visé à caractériser et à identifier des discours sexistes et misogynes sur les réseaux sociaux. Ces travaux incluent ceux réalisés dans le cadre de plusieurs campagnes d'évaluation en traitement automatique des langues, notamment Evalita 2018 et Evalita 2020, *Automatic Misogyny Identification* (Fersini, 2018 ; Fersini *et al.*, 2020), SemEval-2022, *Multimedia Automatic Misogyny Identification* (Fersini *et al.*, 2022) et SemEval-2023, *Explainable Detection of Online Sexism* (Kirk *et al.*, 2023), ou encore les campagnes EXIST, *sEXism Identification in Social neTworks* (Rodríguez-Sanchez *et al.*, 2021 ; Rodríguez-Sánchez *et al.*, 2022 ; Plaza *et al.*, 2023 ; Plaza *et al.*, 2024).

Pour sa part, l'analyse du discours des *incels* a fait l'objet de travaux situés à l'intersection des disciplines de la communication, de l'étude des médias sociaux et des études de genre. Peu de travaux se sont cependant intéressés spécifiquement à détecter le discours des *incels* parmi d'autres formes de discours. Outre les travaux de Gemelli et Minnema (2024) portant sur la constitution et la description d'un corpus de propos *incels* en italien à l'aide de FrameNet, l'annotation de corpus misogynes pose généralement d'importants défis (Sheppard *et al.*, 2024). À cet égard, les travaux liés à la détection automatique de sexisme et de misogynie présentent néanmoins un intérêt pour le présent contexte, cette tâche bénéficiant d'une communauté de pratique plutôt bien établie. Plusieurs de ces travaux se sont penchés non seulement sur la détection, mais également la catégorisation de diverses formes de sexisme et de misogynie.

En l'occurrence, pour répondre à la tâche *Automatic Misogyny Identification* (AMI) de la campagne *Evalita 2018* (Fersini, 2018), Saha *et al.* (2018) ont développé un modèle d'apprentissage automatique permettant de détecter la présence de propos sexistes dans des publications provenant de Twitter et de catégoriser le caractère actif ou passif de la cible des propos. Leur modèle a été le plus performant en utilisant conjointement des méthodes de *sentence embedding* et de vecteurs pondérés par TF-IDF, avec une performance de 70,4 % d'exactitude. Frenda *et al.* (2019) ont réutilisé

le corpus de commentaires misogynes de la compétition AMI de 2018 ainsi qu'un corpus supplémentaire de tweets sexistes pour extraire automatiquement des analogies et des distinctions entre les propos misogynes et sexistes publiés sur Twitter. Ils ont entre autres employé des mesures de similarité lexicale pour comparer la richesse du vocabulaire des deux corpus. Ils ont également développé un système de détection des tweets sexistes et misogynes en utilisant une approche de machines à vecteurs de support dont les traits discriminants ont été pondérés selon leur valeur de TF-IDF, et ont obtenu une performance de 76,05 % d'exactitude.

Dans une étude longitudinale portant sur l'évolution des communautés de la manosphère sur le Web, Ribeiro *et al.* (2021) ont constitué un corpus s'échelonnant sur une période de 14 ans issu de nombreux espaces de discussion de la plateforme Reddit (appelés *subreddits*). Leurs travaux ont permis d'étudier les migrations et intersections des utilisateurs à travers ces plateformes en employant des mesures de similarité sémantique (l'indice de Jaccard et le coefficient de chevauchement). Ceux-ci ont également pu caractériser le niveau de toxicité des propos au sein de ces communautés en développant un système reposant sur un réseau neuronal convolutif (CNN).

Dans un article récent, Morales-Castro *et al.* (2023) ont cherché à détecter automatiquement les propos misogynes en extrayant des informations subjectives de textes non structurés à l'aide du jeu de données d'évaluation de la campagne Evalita. Pour ce faire, ils ont comparé les performances de plusieurs méthodes d'apprentissage supervisé dont les machines à vecteurs de support (SVM), le classifieur bayésien naïf, l'algorithme de régression logistique, les arbres de décisions et l'algorithme des K plus proches voisins (KNN). Les auteurs ont ensuite sélectionné les trois systèmes présentant la plus grande précision et les ont combinés en un métaclassificateur basé sur la régression logistique, atteignant une précision de 81,8 %.

Dans le même ordre d'idées, les travaux de Muti *et al.* (2024) ont abordé la détection des propos mysogynes en ligne en déployant une technique basée sur le raisonnement argumentatif à l'aide de grands modèles de langues. Les résultats obtenus sur un corpus de textes en anglais et en italien sont encourageants (certaines configurations permettent d'obtenir des mesures de rappel de 91,3 %, bien que leur approche basée sur le raisonnement se heurte à d'importantes limites.

En ce qui concerne spécifiquement le discours des *incels*, les travaux réalisés par Jaki *et al.* (2019) ont cherché à développer un système de détection des propos à caractère misogyne, homophobe et/ou raciste parmi les commentaires publiés sur le forum Incels.me. Ceux-ci ont permis de comparer un CNN et un perceptron entraînés à partir de n-grammes de différentes tailles de caractères et de mots. Les deux approches ont permis de détecter les propos misogynes, homophobes et/ou racistes parmi les commentaires issus du forum *incel* avec 95 % d'exactitude.

Pelzer *et al.* (2021) ont pour leur part développé un modèle de détection automatique du niveau de toxicité de trois des plus importants forums Incels connus en 2021, soit Incels.is, Lookisms.net et Looksmax.org. Ils ont utilisé une approche d'apprentissage par transfert (*transfert-learning*) basée sur le modèle de langue BERT. Leurs

analyses ont démontré que les propos de ces trois forums présentaient un taux de toxicité significativement supérieur à celui d'un corpus contrôle provenant de Reddit.

Une approche lexicométrique a été employée par Gothard *et al.* (2021) pour caractériser et analyser les spécificités des patrons de langage propres à trois forums incels sur la plateforme Reddit, soit *r/Braincels*, *r/Incels* et *r/Shortcels*. Ces travaux ont également comparé la richesse lexicale de *r/Braincels* à celui d'un corpus contrôle issu de différents *subreddits*, en mobilisant des mesures de statistiques lexicales. Les lexiques des deux corpus ont été comparés à l'aide d'une analyse rang-rang (*rank-rank*) des formes les plus fréquentes. Cette approche a démontré que le vocabulaire des commentaires issus de *subreddits Incels* a tendance à être moins riche que celui d'autres communautés de la plateforme.

Plus récemment, Yoder *et al.* (2023) ont utilisé des mesures de statistiques textuelles (fréquences, distribution, etc.) et une analyse de réseau sur un ensemble de données comprenant 6 248 234 commentaires postés sur le forum *incels.is* pour évaluer la construction d'une identité de groupe à travers le discours des incels. Parallèlement à l'utilisation de méthodes lexicales classiques, ils ont entraîné un modèle word2vec à partir de ces commentaires afin d'évaluer la diversité du vocabulaire associé à l'identité des incels. Leurs résultats suggèrent une forte importance accordée à l'apparence physique comme déterminante de la valeur individuelle des êtres humains ainsi qu'une prévalence de la question du genre au centre des discussions entre individus incels. Ces chercheurs ont également extrait des termes identitaires fortement péjoratifs à l'égard des femmes, ce qui corrobore les travaux mentionnés ci-dessus. Ceux-ci ont identifié une prévalence de 30 % de ces termes d'identité dans le corpus utilisé, et suggèrent que tout classificateur visant à détecter les discours haineux ou misogynes devrait inclure ces termes comme caractéristiques discriminantes.

Les travaux menés par Hajarian et Khanbabaloo (2021) sont les seuls à s'être spécifiquement intéressés à détecter les utilisateurs *incels* sur Facebook et sur Twitter. Pour ce faire, ces chercheurs ont combiné une méthode d'analyse de sentiments à un système de détection de propos injurieux qu'ils ont appliqué à un corpus de commentaires issus de ces deux plateformes. Leurs travaux ont été en mesure de détecter le discours *incel* dans 78.8 % des cas.

Finalement, dans une étude récente et très importante, Arango *et al.* (2022) ont souligné le contraste entre d'une part l'incapacité, malgré des investissements colossaux, des principaux réseaux sociaux à détecter automatiquement les contenus haineux et, d'autre part, les résultats de recherche, générés principalement par des chercheurs issus du secteur académique, indiquant que les approches de classification supervisée permettent d'atteindre des performances très appréciables. Selon ces chercheurs, l'écart entre les performances des systèmes issus de la recherche et ceux des concepteurs des réseaux sociaux s'explique par des problèmes d'ordre méthodologique et par un biais dans la conception des ensembles de données employés pour valider ces systèmes. Par conséquent, les performances des systèmes de pointe seraient, selon ces auteurs, considérablement surestimées. Ainsi, ceux-ci ont ré-évalué les résultats de certains travaux documentés dans la littérature et ont révélé une importante baisse

de performance de ces systèmes, passant dans certains cas de performances de plus de 90 % à des performances qui seraient plutôt de l'ordre de 50 % à 80 %, sur des ensembles de données plus représentatifs de la réalité des réseaux sociaux.

La recherche dont nous rendons compte dans le présent article vise à comparer rigoureusement différentes techniques de détection automatique en faisant varier divers paramètres afin d'en favoriser l'application à plus large échelle.

3. Méthodologie

3.1. *Constitution de corpus*

Les données servant à l'apprentissage des modèles proviennent de jeux de données existants en langue anglaise. Celles-ci représentent d'une part des propos représentatifs du discours *incel* (classe « *incels* ») et d'autre part des données représentatives des pratiques langagières propres à la plateforme Reddit, mais sans qu'une thématique particulière n'y soit associée (classe « neutres » ou « non-*incels* »). Cette dernière classe est également constituée de commentaires provenant de Reddit afin de limiter l'introduction de biais reflétant les spécificités langagières propres à cette plateforme plutôt que du discours d'intérêt à détecter (Tranchese et Sugiura, 2021).

L'annotation des données servant à entraîner les modèles de classification repose sur une approche appelée « sac de communautés » (*Bag-of-Communities* [BoC]) (Chandrasekharan *et al.*, 2017). Cette approche consiste à identifier une communauté entière comme étant haineuse ou toxique plutôt que d'annoter chacun des contenus publiés par ses utilisateurs individuellement. L'un des avantages de cette approche est qu'elle permet de pallier l'absence de données d'évaluation de même que le recours à un travail d'annotation manuelle auprès d'experts de domaine (Chandrasekharan *et al.*, 2017 ; Pelzer *et al.*, 2021). Cette méthode offre donc la possibilité d'exploiter les données de communautés sources pour classer les discours haineux dans une communauté cible, en se basant sur l'hypothèse que la communauté cible présente des similarités linguistiques avec les communautés sources (Muralikumar *et al.*, 2023).

Un enjeu potentiellement associé à l'annotation basée sur les sacs de communautés réside dans la configuration des données d'entraînement, laquelle rend difficile de distinguer si les propos détectés constituent véritablement des propos haineux ou s'ils reflètent uniquement des similitudes entre les communautés concernées (Berglind *et al.*, 2019). Or dans un contexte où les communautés sélectionnées sont reconnues pour la toxicité de leurs propos, cette approche pourrait offrir le potentiel d'améliorer l'adaptabilité des modèles de détection de discours haineux entre des communautés issues de différentes plateformes (p. ex. Reddit, Facebook ou X), ou encore de sous-communautés distinctes au sein d'une même plateforme (par exemple les *subreddits* sur Reddit) (Almerekhi *et al.*, 2020). Nous pensons donc que les sacs de communautés présentent un intérêt particulier pour la détection du discours des *incels*. D'une part, ces communautés partagent une vision du monde et une identité commune se manifestant d'une manière similaire dans les différents espaces de discussion où elles sont

actives ; d'autre part, la migration constante des utilisateurs vers de nouveaux espaces faisant suite au banissement de *subreddits incels* implique l'existence d'un alignement entre les espaces de discussion de ces communautés.

Le corpus constitué dans le cadre de cet article regroupe donc un ensemble de commentaires issus de *subreddits* reconnus comme des espaces de discussion occupés par des *incels*, par opposition à un ensemble de commentaires issus de *subreddits* non dédiés à ces communautés. Chaque commentaire est étiqueté comme étant *incel* ou non-*incel* en fonction du *subreddit* au sein duquel celui-ci a été publié.

3.1.1. *Données* incels

Deux jeux de données ont été mobilisés pour constituer la classe *incels*. Nous avons d'abord identifié un corpus rendu disponible par Ribeiro *et al.* (2020)[3] pour identifier un ensemble de *subreddits* caractéristiques du discours *incel*. Ce jeu de données regroupe des espaces de discussion de l'androsphère (*manosphere*) sur le Web, un ensemble de communautés aux revendications masculinistes incluant les communautés *incels*. Ce corpus compte plus de 28,8 millions de publications provenant de différents forums en ligne, dont 56 forums de discussion Reddit (*subreddits*). Dans le cadre de leurs travaux, les auteurs ont catégorisé chacun de ces 56 *subreddits* en fonction de sa communauté d'appartenance au sein de la manosphère (p. ex. *Incels*, *Men Get Their Own Way [MGTOW], Pick Up Artists [PUA], The Red Pill [TRP], Men's Rights Activist [MRA]*). Au regard de cette analyse, 23 des 56 *subreddits* ont été identifiés comme étant spécifiquement *incels* (p. ex. *r/Braincels, r/ForeverUnwanted, r/AskAnIncel*)[4]. Nous avons donc retenu ces 23 *subreddits* comme point de départ pour constituer les données d'apprentissage de la classe *incels*. Comme ce corpus couvre une période temporelle s'arrêtant en 2019, nous avons récupéré les données provenant des *subreddits incels* à partir des archives du projet PushShift (Baumgartner *et al.*, 2020), disponibles sur la plateforme The-Eye[5], et ce afin de représenter une période temporelle couvrant les 10 dernières années (janvier 2014 à décembre 2023). Nous avons choisi de ne retenir que les commentaires et publications émis depuis 2014 afin de refléter le jargon actuel de ces communautés ainsi que de tenir compte d'éventuelles évolutions dans ses pratiques discursives.

Parmi les 23 *subreddits* identifiés par Ribeiro *et al.* (2021) comme étant *incels*, nous en avons retenu 9 figurant parmi les 40 000 *subreddits* les plus populaires de Reddit à travers la totalité de son historique[6] : *r/AskAnIncel*, *r/BlackPillScience*, *r/Braincels*, *r/ForeverAlone*, *r/ForeverAloneDating*, *r/ForeverUnwanted*, *r/Incels*,

3. Ces données sont disponibles à l'adresse suivante : `https://zenodo.org/records/4007913`

4. La catégorisation effectuée par Ribeiro *et al.* (2021) est disponible à l'adresse suivante : `https://github.com/idramalab/manosphere_analysis/blob/master/data/subreddit_descriptions.csv`

5. `https://the-eye.eu/redarcs/`

6. Les données constituant la classe « *incels* » ont été récupérées à l'adresse suivante : `https://academictorrents.com/details/56aa49f9653ba545f48df2e33679f014d2829c10`

r/IncelSelfies, *r/IncelsWithoutHate*. Les publications ont été sélectionnées aléatoirement au sein de ces 9 *subreddits incels* pour une période s'échelonnant de 2014 à 2023, en maintenant égale la proportion de publications par année. Le tableau 1 illustre quelques exemples de données issues des *subreddits incels*.

Subreddit	**Commentaire**
r/Braincels	Women are the root. The reason we express our emotions through anger 90% of the time is because that's what women find attractive. Femoids, fuck you. You did this. fuck you !
r/ForeverUnwanted	Abandon all hope, gentlemen. Any woman you ever meet and fall in love with will be able to find some other dude to replace you quicker than all fuck.
r/Incels	Who gives a fuck. Bitch gave her word to be wife for life. No wonder nobody respects females. they can't hold up their end of such a sacred agreement. 80% of divorces initiated by females. Almost always for bullshit reasons too.

TABLEAU 1. *Exemples de commentaires issus de* subreddits incels

.

3.1.2. *Données non*-incels

Pour constituer la classe de données non-*incels*, nous avons extrait un échantillon de commentaires à partir des archives du projet PushShift (Baumgartner *et al.*, 2020) (via la plateforme The-Eye), lesquelles couvrent la totalité des publications et commentaires émis sur Reddit entre 2005 et 2023[7]. Nous avons extrait de ces données un échantillon de commentaires publiés entre 2014 et 2023 dans différents *subreddits* non reconnus comme *incels* (p. ex. *r/OffMyChest, r/Electronic_Cigarette, r/FinalFantasy*, etc.). Étant donné le volume important de cette archive, nous avons extrait, pour chaque année allant de 2014 à 2023, l'ensemble des publications et commentaires sur une période de 1 mois sélectionné aléatoirement (p. ex. 2014 = avril, 2015 = mai, etc.). Ces publications sont issues de 13 828 *subreddits* distincts. Le tableau 2 illustre quelques exemples de données provenant des *subreddits* non-*incels*.

3.1.3. *Partitionnement des données d'apprentissage et de test*

Le partitionnement des données a été effectué selon un ratio classique deux tiers un tiers : 40 000 commentaires ont été utilisés pour la phase d'apprentissage, tandis que 20 000 commentaires inédits ont été utilisés pour évaluer la performance des modèles paramétrés, c'est-à-dire tester leur capacité de généralisation à détecter les propos *incels* dans des commentaires inédits provenant de Reddit.

Pour les données d'apprentissage, nous avons constitué différents jeux de données en faisant varier le ratio de commentaires *incels* contenus dans ceux-ci. Neuf corpus

7. Les données constituant la classe « neutres » ont été récupérées à l'adresse suivante : `https://academictorrents.com/details/9c263fc85366c1ef8f5bb9da0203f4c8c8db75f4`

Subreddit	Commentaire
r/WhoWouldWin	He's on par with a lot of movie showings of Superman. When Superman shows up in long running series, like comics, he tends to gradually become significantly more powerful. This is true for most characters like him, as well.
r/SpinnCoffee	I ordered my Spinn in January. According to my order, I am also batch 6. I feel like June to January is a long batch period. I'm a bit worried now about when I'll actually get mine.
r/OldHagFashion	Years ago, a group of us had a "fun vest night." I wore this vest, made by my mom (apparently from a kit), when I was a kid. Showed up at my friend's door...he was wearing the exact same vest. It's a good one.

TABLEAU 2. *Exemples de commentaires issus de* subreddits *non*-incels

d'apprentissage ont ainsi été créés, contenant de 10 % à 90 % de commentaires *incels*. Cette stratégie a été employée dans le but de pallier le problème de déséquilibre des classes auxquelles font typiquement face les tâches de détection automatique (Ling et Sheng, 2010), où la tâche à détecter se trouve sous-représentée dans la réalité, rendant plus difficile l'extraction de ses spécificités par un modèle de classification. Bien que la proportion réelle du discours *incel* sur Reddit demeure inconnue, cette approche vise à évaluer l'effet d'une surreprésentation des données *incels* dans les données d'apprentissage sur les performances de détection résultantes. Cette approche s'oppose aux stratégies traditionnelles de suréchantillonnage, qui consistent à dédoubler des exemplaires des catégories de la classe à détecter dans les données d'apprentissage, lesquelles seraient à la source des problèmes méthodologiques associés à la surestimation des performances de certains systèmes selon Arango *et al.* (2022). Notre approche vise plutôt à faire varier la proportion de commentaires uniques associés à chaque classe dans les données d'apprentissage dans le but de contrer ce type de biais méthodologique. Ainsi, chaque corpus d'apprentissage totalise 40 000 documents, avec un ratio variable de données *incels* et non-*incels*. Pour chaque proportion testée, les données ont été sélectionnées au moyen d'une technique d'échantillonnage aléatoire stratifiée visant à maintenir égale la proportion de publications par année. Le tableau 3 illustre les caractéristiques des corpus d'apprentissage, comptant en moyenne 1 208 385 occurrences de mots (*tokens*) et 57 955 formes uniques (*types*).

Le corpus dédié à l'évaluation des modèles totalise pour sa part 20 000 commentaires, dont la proportion de commentaires *incels* a été fixée à 10 %. Comme la proportion réelle des propos *incels* sur Reddit est inconnue, ce ratio a été retenu en fonction des résultats obtenus par Hajarian et Khanbabaloo (2021), lesquels ont rapporté des proportions de propos *incels* de l'ordre de 9,1 % sur Facebook et de 8,5 % sur Twitter. Afin de tenir compte de ces proportions, le corpus d'évaluation des modèles compte donc 2 000 commentaires *incels* et 18 000 commentaires neutres, tel qu'illustré dans le tableau 4.

% *Incels*	Commentaires *incels*	Commentaires non-*incels*	Total	Occurrences de mots (tokens)	Mots uniques (types)
10	4 000	36 000	40 000	1 087 228	66 037
20	8 000	32 000	40 000	1 117 519	63 960
30	12 000	28 000	40 000	1 156 716	63 543
40	16 000	24 000	40 000	1 167 474	60 345
50	20 000	20 000	40 000	1 214 857	59 124
60	24 000	16 000	40 000	1 238 734	56 267
70	28 000	12 000	40 000	1 259 718	53 694
80	32 000	8 000	40 000	1 302 462	50 446
90	36 000	4 000	40 000	1 330 756	48 180

TABLEAU 3. *Variation de la proportion des commentaires* incels *et non-*incels *dans les données d'apprentissage, avec occurrences et types de mots*

% *Incels*	Commentaires *incels*	Commentaires non-*incels*	Total	Occurrences de mots (tokens)	Mots uniques (types)
10	2 000	18 000	20 000	461 414	39 554

TABLEAU 4. *Proportion des commentaires* incels *et non-*incels *dans les données d'évaluation (données de test)*

3.2. *Prétraitements*

Différents filtrages ont été appliqués de manière à supprimer les commentaires non pertinents. Les publications vides (« *[removed]* », « *[deleted]* ») ou ne contenant qu'un seul caractère ont été retirées, de même que les publications de robots modérateurs (auteur *« AutoModerator »*). L'ensemble du texte des commentaires a été minusculisé. Des patrons d'expressions régulières ont été employés pour retirer les URL ainsi que certains artefacts issus de l'API de Reddit tels que des entités HTML (p. ex. « *>* »). Les commentaires ont ensuite été segmentés en mots (*tokenisés*) avec la fonction *word_tokenize* de la librairie Python NLTK (Bird *et al.*, 2009). Les mots fonctionnels ont été filtrés au moyen d'un antidictionnaire de l'anglais, de même que les expressions contenant des chiffres ou des caractères spéciaux.

3.3. *Phase d'apprentissage*

Pour entraîner les modèles de classification, nous avons comparé trois approches permettant de représenter numériquement les commentaires provenant de Reddit : (1) des vecteurs basés sur une pondération TF-IDF des termes du lexique ; (2) des vecteurs basés sur le modèle de plongement lexical *Continuuous Bag-of-Words* (CBOW) (Mikolov *et al.*, 2013) ; (3) des vecteurs basés sur le modèle de plongement de phrases

Sentence Transformers (SBERT) (Reimers et Gurevych, 2019). Le choix de ces trois types de représentations a pour objectif d'évaluer la capacité de modèles de classification à exploiter différentes caractéristiques textuelles pour la tâche de détection de cette étude. Cette approche vise en l'occurrence à comparer des méthodes classiques issues de la statistique lexicale (TF-IDF) à des méthodes plus récentes exploitant les relations de dépendance contextuelle entre les mots pour produire des représentations sémantiques denses (Word2Vec, SBERT). Ces trois modèles ont été comparés compte tenu de l'existence de résultats contradictoires concernant l'emploi de plongements lexicaux pour des tâches de classification textuelle, notamment Word2Vec (Abubakar *et al.*, 2022 ; Truşcă, 2019), BERT ou SBERT (Jamshidian, 2023), en comparaison avec des vecteurs TF-IDF. Nos travaux cherchent donc à mettre en perspective ces résultats dans le contexte de la détection du discours *incel*.

Le *Term Frequency-Inverse Document Frequency* (TF-IDF) est une mesure de pondération statistique permettant de calculer un score de spécificité pour chaque terme d'un corpus en fonction de sa fréquence relative dans chaque document par rapport à sa proportion inverse sur l'ensemble des documents du corpus (IDF) (Ramos, 2003). Une motivation à employer cette approche dans le cadre d'une tâche de classification supervisée réside dans le fait que les termes présentant une forte fréquence dans un nombre restreint de documents obtiendront un score TF-IDF plus élevé pour ceux-ci, reflétant la pertinence de ces termes pour représenter ce groupe de documents. Les vecteurs TF-IDF ont été générés avec la fonction *Tfidfvectorizer* de la librairie Python Scikit-learn (Pedregosa *et al.*, 2011).

Le modèle de plongement lexical *Continuous Bag-of-Words* est pour sa part basé sur une architecture neuronale simple permettant de prédire la probabilité conditionnelle d'occurrence associée à chacun des mots d'un document compte tenu d'une fenêtre contextuelle donnée, laquelle est typiquement constituée des 5 mots entourant le mot à prédire (Azmy *et al.*, 2018 ; Mikolov *et al.*, 2013). Ce modèle permet de générer une représentation vectorielle de longueur fixe pour chaque mot d'un document, puis ces représentations de mots peuvent ensuite être combinées au moyen d'une fonction d'agrégation (par exemple la somme ou la moyenne des éléments des vecteurs) pour obtenir un seul vecteur par document. Les vecteurs CBOW ont été générés au moyen du modèle Word2Vec disponible dans la librairie Python Gensim, avec une fenêtre contextuelle de 5 mots. Pour chaque commentaire, un seul vecteur a ensuite été généré par l'agrégation des vecteurs de mots en utilisant la moyenne comme fonction d'agrégation.

Le modèle de plongement de phrases *Sentence Transformers* (SBERT) est un réseau neuronal basé sur l'architecture Transformer permettant de générer des représentations vectorielles de phrases ou de textes courts (Reimers et Gurevych, 2019). Ce modèle constitue une extension du réseau neuronal préentraîné BERT spécifiquement conçue pour la comparaison de paires de phrases. Il utilise une architecture bi-encodeur employant deux instances parallèles de BERT pour traiter indépendamment des paires de phrases, le rendant plus efficace que BERT pour des tâches nécessitant une comparaison de similarité textuelle (Reimers et Gurevych, 2019). Dans le

cadre de cette étude, les vecteurs SBERT ont été générés au moyen du modèle *all-MiniLM-L6-v2* de la librairie Python Sentence Transformers, qui permet d'encoder les commentaires issus de Reddit en un vecteur de longueur fixe de 384 dimensions.

Nous avons fait varier le nombre de dimensions des vecteurs TF-IDF et CBOW afin d'évaluer l'effet de ce paramètre sur la performance de détection résultante. Nous avons employé des valeurs allant de 1 000 à 15 000 éléments pour les modèles basés sur une pondération TF-IDF étant donné la haute dimensionnalité des vecteurs typiquement associés à cette approche. En contrepartie, l'une des caractéristiques des vecteurs générés au moyen d'approches par plongement comme CBOW et SBERT réside dans leur capacité à représenter l'information de manière dense, et ce avec une dimensionnalité relativement restreinte. Nous avons donc retenu des valeurs allant de 100 à 500 éléments pour les plongements associés aux modèles CBOW. Pour leur part, comme les plongements SBERT sont issus d'un modèle de langue préentraîné générant des vecteurs dont la dimension est fixée à 384 éléments (*all-MiniLM-L6-v2*), nous n'avons pas fait varier ce paramètre pour ce type de représentations.

3.4. *Classification*

Différents algorithmes de classification ont été évalués pour détecter la présence de propos *incels* dans les commentaires issus de Reddit. Pour ce faire, nous avons sélectionné quatre algorithmes implémentés dans la librairie Python Scikit-learn (Pedregosa *et al.*, 2011) : (1) la régression logistique (LR) ; (2) les machines à vecteurs de support (SVM) ; les forêts aléatoires (RF) ; (4) la classification bayésienne naïve multinomiale (MNB). Le choix de ces méthodes repose sur leur capacité à traiter des données dans les formats de représentation que nous faisons varier dans nos expérimentations, mais aussi sur le fait que plusieurs d'entre elles ont été fréquemment employées dans des tâches de détection du cyberharcèlement et, plus spécifiquement, des propos misogynes. Nous avons évalué la pertinence de ces méthodes sur des ensembles de données textuelles représentées numériquement à l'aide des trois méthodes de vectorisation mentionnées dans la section précédente (TF-IDF, CBOW, SBERT). D'un point de vue technique, ces approches présentent plusieurs avantages dans le contexte de nos travaux : elles ne nécessitent pas de grands ensembles de données (compte tenu de la taille relativement petite de notre corpus – 60 000 commentaires totalisant moins de 2 millions de mots) pour générer des résultats de qualité et sont relativement peu couteuses computationnellement.

La régression logistique (LR) est une approche de classification probabiliste permettant de déterminer l'importance de chaque attribut d'un document en termes d'un poids relatif (coefficient de régression) contribuant à une fonction de décision, ce poids pouvant être positif ou négatif. La classification consiste à appliquer une fonction logistique à la somme des poids associés aux attributs d'un document afin d'obtenir une probabilité que ce document appartienne à une classe d'intérêt, ce dernier étant classifié comme tel au-delà d'un certain seuil (Jurafsky et Martin, 2019).

Les machines à vecteurs de support (SVM) sont des modèles de classification binaire reposant sur l'identification d'un hyperplan séparant deux classes de données en maximisant la distance (marge) entre cet hyperplan et tout point de données dans un espace vectoriel (Manning *et al.*, 2008). L'hyperplan est identifié à partir de données d'apprentissage et peut ensuite être appliqué à de nouvelles données dans un même espace vectoriel.

Les forêts aléatoires (RF) sont issues d'approches d'apprentissage ensemblistes, c'est-à-dire effectuant la tâche de classification en combinant les décisions d'un ensemble de classifieurs à travers un processus de vote (Pal, 2005). Dans une forêt aléatoire, les classifieurs employés reposent sur un apprentissage par arbre de décision, où chaque arbre traite un sous-ensemble d'attributs relatifs à l'élément à classer, lesquels sont échantillonnés aléatoirement. Chaque arbre vote ensuite pour la classe sous laquelle l'élément devrait être classé en fonction de ces attributs (Breiman, 1999).

Les classifieurs bayésiens naïfs (NB) sont des algorithmes probabilistes modélisant la probabilité qu'un élément appartienne à une classe donnée en fonction d'un ensemble de caractéristiques considérées comme indépendantes les unes des autres (Feldman et Sanger, 2006). Des variantes de ces classifieurs dépendent du choix de modélisation des probabilités conditionnelles selon différentes distributions (Manning *et al.*, 2008). Nous avons employé un classifieur bayésien naïf basé sur une distribution multinomiale (MNB) pour les vecteurs générés par pondération TF-IDF. Nous n'avons pas testé ce classifieur pour les modèles CBOW et SBERT, comme ils génèrent des représentations vectorielles dont les éléments sont des valeurs continues.

Pour chacun des classifieurs ci-dessus, nous avons testé l'ensemble des configurations associées aux paramètres mentionnés dans les sections précédentes : ratio de données *incels* dans les données d'apprentissage, technique de vectorisation utilisée, nombre de dimensions des vecteurs, algorithme de classification employé. La figure 1 illustre l'ensemble des configurations testées. Chacun des modèles a fait l'objet d'une validation croisée à 5 plis. Les paramètres optimaux pour chaque modèle ont été identifiés au moyen d'une recherche en grille (*GridSearchCV*).

Les expérimentations ont été réalisées avec Python 3.11 sous le système d'exploitation Windows 11, sur une machine dotée de 16 Go de mémoire vive et d'un processeur à 2,9 GHz et 8 cœurs. L'entraînement des modèles utilisant SentenceTransformers ont été réalisés en utilisant une unité de calcul graphique (GPU) NVIDIA GeForce RTX 3070. Les scripts utilisés pour prétraiter les données, entraîner les modèles et générer les résultats sont disponibles à l'adresse suivante : `https://github.com/CamilleDemers/incels-detection-reddit`.

Les 20 configurations ayant généré les meilleures performances en phase d'apprentissage ont été évaluées sur un ensemble de test composé de 20 000 documents inédits. Ces performances sont présentées dans la section suivante.

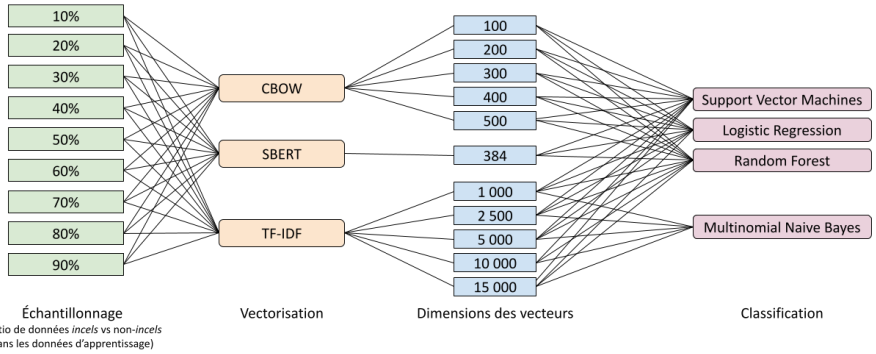

FIGURE 1. *Configurations testées pour l'entraînement des modèles de classification*

4. Résultats et discussion

4.1. *Performances de classification*

4.1.1. *Phase d'apprentissage*

Les résultats sont évalués au regard de l'exactitude prédictive, la précision et le rappel. La performance globale est évaluée au moyen de la mesure-F (macro f1). Cette mesure reflète d'ailleurs plus adéquatement la performance que l'exactitude prédictive dans un contexte de déséquilibre des classes (Zhao et Chen, 2014). Les résultats sont obtenus en calculant la moyenne arithmétique des métriques associées à chacun des plis de la validation croisée réalisée sur les données d'apprentissage.

Les tableaux 5a, 5b et 5c illustrent les 5 meilleures configurations en phase d'apprentissage pour les trois modèles de vectorisation sélectionnés. Globalement, les modèles de classification basés sur des vecteurs SBERT performent le mieux (mesure-F maximale = 84,60), suivis des vecteurs TF-IDF (mesure-F maximale = 81,98) et des vecteurs CBOW (mesure-F maximale = 78,58) (SBERT > TF-IDF > CBOW).

Pour les vecteurs TF-IDF, le classifieur bayésien naïf multinomial entraîné sur un corpus contenant 40 % de données *incels* avec des vecteurs de 15 000 dimensions obtient la meilleure performance de détection (mesure-F = 81,98). Autrement, un ratio plus élevé de données *incels* dans les données d'apprentissage donne lieu à de meilleures performances prédictives, pour un ratio allant jusqu'à 60 % de données *incels*. En ce qui concerne l'algorithme de détection employé, la classification bayésienne naïve multinomiale ainsi que la régression logistique obtiennent les meilleures performances. Finalement, les modèles retenant 10 000 à 15 000 dimensions performent mieux que ceux retenant un nombre plus faible de traits discriminants.

% Incels	Algorithme	Vecteurs	Dim.	Exactit.	Précision	Rappel	Mesure-F
40	MNB	TF-IDF	15 000	82,71	82,01	81,95	**81,98**
40	MNB	TF-IDF	10 000	82,69	82,05	81,76	81,89
50	LR	TF-IDF	15 000	81,77	82,07	81,77	81,73
50	LR	TF-IDF	10 000	81,59	81,86	81,59	81,55
60	LR	TF-IDF	15 000	82,24	81,48	81,62	81,54

(a) *Performances de classification des 5 meilleurs modèles basés sur des vecteurs TF-IDF, triés par mesure-F. MNB = Multinomial Naive Bayes, LR = Logistic Regression*

% Incels	Algorithme	Vecteurs	Dim.	Exactit.	Précision	Rappel	Mesure-F
50	RF	CBOW	500	78,59	78,61	78,59	**78,58**
50	RF	CBOW	200	78,55	78,56	78,55	78,55
50	RF	CBOW	450	78,50	78,51	78,50	78,49
50	RF	CBOW	350	78,46	78,48	78,46	78,46
50	RF	CBOW	100	78,43	78,44	78,43	78,43

(b) *Performances de classification des 5 meilleurs modèles basés sur des vecteurs CBOW (Word2Vec), triés par mesure-F. RF = Random Forest*

% Incels	Algorithme	Vecteurs	Dim.	Exactit.	Précision	Rappel	Mesure-F
40	SVM	SBERT	384	85,38	83,95	78,47	**84,60**
50	SVM	SBERT	384	84,50	85,32	83,34	84,50
40	LR	SBERT	384	85,26	83,64	78,51	84,47
50	LR	SBERT	384	84,37	85,14	83,29	84,37
60	SVM	SBERT	384	84,82	87,01	87,81	84,15

(c) *Performances de classification des 5 meilleurs modèles basés sur des plongements SBERT, triés par mesure-F. SVM = Support Vector Machine, LR = Logistic Regression*

TABLEAU 5. *Performances de classification associées aux trois approches de vectorisation testées, en phase d'apprentissage. Pour chaque modèle, la mesure-F maximale est indiquée en gras.*

Les modèles basés sur les vecteurs CBOW illustrent des tendances singulières. Ces modèles performent en effet systématiquement mieux en mobilisant des corpus d'apprentissage contenant 50 % de données *incels* et en exploitant un classifieur basé sur des forêts aléatoires (RF). Ces résultats ne permettent cependant pas de déterminer si la dimension des vecteurs est liée à la performance de détection pour les modèles basés sur des vecteurs CBOW. Bien que les meilleures performances soient associées à des vecteurs de haute dimension (500 dimensions), des vecteurs de plus faible dimension (100) se retrouvent également parmi les 5 meilleures configurations.

Les modèles associés aux plongements SBERT donnent pour leur part de meilleures performances en mobilisant des classifieurs basés sur les machines à vecteurs de support (SVM) et sur la régression logistique (LR). Ces modèles obtiennent par ailleurs de meilleurs résultats lorsqu'entraînés sur des corpus d'apprentissage contenant 40, 50 ou 60 % de données *incels*, ce qui est similaire aux modèles basés sur une pondération TF-IDF du lexique.

Autrement, l'écart entre les résultats relatifs aux vecteurs TF-IDF et CBOW semble cohérent avec les travaux de Wang *et al.* (2017), lesquels ont observé de meilleures performances de classification pour des textes courts avec des vecteurs TF-IDF plutôt qu'avec les modèles Word2Vec et Doc2Vec, en utilisant trois des classifieurs mobilisés dans cette étude (LR, SVM et MNB). Similairement, les travaux de Truşcă (2019) suggèrent que les vecteurs TD-IDF surpassent les vecteurs Word2Vec pour des problèmes linéairement séparables. En contrepartie, les différences de performance observées pour les vecteurs SBERT et TF-IDF font contraste avec les travaux de Jamshidian (2023), lesquels ont obtenu une meilleure performance de classification avec des vecteurs TF-IDF plutôt que des plongements SBERT pour entraîner des modèles d'analyse de sentiment basés sur des SVM.

4.1.2. *Phase de test*

Pour la phase de test, les 20 meilleurs modèles en phase d'apprentissage ont été retenus et évalués sur 20 000 commentaires inédits. L'ensemble de ces modèles sont basés sur des vecteurs TF-IDF ou des plongements SBERT, ceux-ci ayant illustré de meilleures performances d'apprentissage. Autrement, les meilleurs modèles sont issus des quatre algorithmes de détection testés : (1) la classification bayésienne naïve multinomiale ; (2) la régression logistique ; (3) les machines à vecteurs de support et (4) les forêts aléatoires. Pour les modèles TF-IDF, ces configurations sont associées à des vecteurs de 10 000 et de 15 000 dimensions et à des corpus d'entraînement contenant de 40 à 50 % de commentaires *incels*. Pour les modèles SBERT, les meilleures performances sont associées à des corpus d'entraînement contenant de 20 à 80 % de commentaires *incels*.

Les tableaux 6a et 6b présentent les métriques de performance des 5 meilleures configurations en phase d'apprentissage et de test pour chaque classe individuelle. Ces résultats indiquent une meilleure performance de détection pour la classe non-*incel*, qui atteint une mesure-F maximale de 96,13 avec une exactitude de 92,87. À l'opposé, la classe de données *incel* est l'objet de résultats nettement inférieurs, avec

une mesure-F maximale de 63,35. Cet écart entre les résultats des deux classes pourrait en partie s'expliquer par le débalancement des données dans le corpus d'évaluation (10 % *incel* / 90 % non-*incel*), lequel favorise la classe non-*incel*. Il est également possible que l'approche de sac de communautés employée pour l'annotation des données d'évaluation sous-estime la capacité réelle des modèles à extraire les caractéristiques permettant de détecter le discours *incel* parmi les autres formes de discours sur Reddit. En effet, la présence d'erreurs d'annotation est inhérente à cette approche puisqu'elle considère uniquement l'espace de discussion où un commentaire a été publié (le *subreddit*) pour lui attribuer une catégorie. Une évaluation sur un corpus annoté manuellement permettrait de mesurer plus justement la performance des modèles.

Le tableau 6c présente les performances des 5 modèles en phase de test. Le modèle le plus performant repose sur une régression logistique avec des vecteurs SBERT et un corpus d'apprentissage contenant 20 % de données *incels* (mesure-F = 79,70). Globalement, les meilleurs modèles sont associés aux ratios de 20 % et de 30 % de données *incels* en apprentissage (mesure-F = 79,70, 79,65 et 78,15), tandis que les pires sont ceux entraînés sur des corpus contenant 80 % de données *incels* (mesure-F = 43,76, 56,71 et 56,82). Cette tendance suggère un suréchantillonnage trop élevé de la classe *incel* en phase d'apprentissage, résultant en une proportion plus élevée de prédictions faussement positives pour la classe *incels* en phase de test. À titre indicatif, les modèles TF-IDF les plus performants arrivent aux dixième, onzième et douzième rangs parmi tous les modèles évalués en phase de test. Il s'agit des modèles suivants : classification bayésienne naïve multinomiale avec 10 000 traits discriminants, entraînée sur un corpus contenant 40 % de données *incels* (mesure-F = 71,67) ; régression logistique avec 15 000 traits discriminants et 50 % de données *incels* en apprentissage (mesure-F = 71,44) ; classification bayésienne naïve multinomiale, 40 % de données *incels* en apprentissage et 10 000 traits discriminants (mesure-F = 71,39).

4.2. *Analyse des traits prédictifs de chaque classe*

Malgré la meilleure performance de classification des modèles basés sur des vecteurs SBERT pour la détection du discours *incel*, les caractéristiques extraites par ces modèles ne permettent pas de rendre compte des propriétés textuelles contribuant au processus de décision de ces modèles. En contrepartie, les modèles issus de vecteurs basés sur une pondération TF-IDF du lexique offrent de manière générale une plus grande interprétabilité quant au processus de classification résultant. Puisque la performance de ces derniers modèles demeure comparable à celle basée sur des vecteurs SBERT, nous avons mené une analyse supplémentaire permettant d'examiner les termes revêtant une plus grande importance pour la frontière de décision des modèles de classification basés sur des vecteurs TF-IDF.

Le tableau 7 présente les caractéristiques les plus prédictives pour chaque classe. Ces caractéristiques ont été analysées à partir du modèle de régression logistique basé sur des vecteurs TF-IDF ayant obtenu les meilleures performances en phase de test (50 % de données *incels* en apprentissage, 15 000 traits discriminants).

incel											
				Apprentissage				Test (= 10 % *Incels*)			
% Incels	Algorithme	Vecteurs	Dim.	Exactit.	Précision	Rappel	Mesure-F	Exactit.	Précision	Rappel	Mesure-F
20	LR	SBERT	384	89,61	79,42	64,88	71,41	92,69	63,39	63,55	**63,47**
20	SVM	SBERT	384	89,53	79,35	64,38	71,08	92,72	63,75	62,95	63,35
30	SVM	SBERT	384	87,01	82,14	72,46	**77,00**	91,09	54,16	70,70	61,33
30	LR	SBERT	384	86,80	82,18	71,50	76,47	90,86	53,25	70,45	60,65
30	LR	TF-IDF	15000	84,91	85,54	59,83	70,40	91,99	59,84	60,50	60,17

(a) *Performances de détection de la classe de données* incel *pour les 5 meilleurs modèles de classification en phase de test*

non-*incel*											
				Apprentissage				Test (= 10 % *Incels*)			
% Incels	Algorithme	Vecteurs	Dim.	Exactit.	Précision	Rappel	Mesure-F	Exactit.	Précision	Rappel	Mesure-F
20	RF	SBERT	384	88,15	88,44	98,00	92,97	92,87	93,93	98,44	**96,13**
20	SVM	SBERT	384	89,53	91,50	95,81	93,60	92,72	95,89	96,02	95,96
20	LR	SBERT	384	89,61	91,60	95,80	**93,65**	92,69	95,95	95,92	95,94
30	RF	SBERT	384	84,53	84,76	94,96	89,57	92,16	95,05	96,30	95,67
30	SVM	SBERT	384	87,01	88,76	93,25	90,95	91,09	96,63	93,35	94,96

(b) *Performances de détection de la classe de données non-*incel *pour les 5 meilleurs modèles de classification en phase de test*

incel + non-*incel*											
				Apprentissage				Test (= 10 % *incels*)			
% Incels	Algorithme	Vecteurs	Dim.	Exactit.	Précision	Rappel	Mesure-F	Exactit.	Précision	Rappel	Mesure-F
20	LR	SBERT	384	89,51	79,22	64,50	82,35	92,69	79,67	79,74	**79,70**
20	SVM	SBERT	384	89,63	79,80	64,50	82,50	92,72	79,82	79,49	79,65
30	SVM	SBERT	384	87,05	82,50	72,15	83,98	91,09	75,39	82,02	78,15
30	LR	SBERT	384	86,89	81,95	72,21	83,82	90,86	74,92	81,79	77,74
40	SVM	SBERT	384	85,38	83,95	78,47	**84,60**	88,71	71,67	83,39	75,55

(c) *Performances macro (moyennes arithmétiques des métriques de chaque classe) pour les 5 meilleurs modèles de classification en phase de test*

TABLEAU 6. *Métriques de performance des 5 meilleurs modèles de classification en phase de test, pour les classes* incel ***(a)*** *et non-*incel ***(b)*** *respectivement. Les performances macro (moyenne pour les deux classes) sont présentées en (c). Les données sont triées par mesure-F en phase de test, les performances en phase d'apprentissage sont indiquées à titre indicatif. Les valeurs maximales (mesure-F) sont indiquées en gras. MNB = Multinomial Naive Bayes, LR = Logistic Regression, SVM = Support Vector Machine*

L'importance relative de chaque trait pour la fonction de décision du modèle est représentée par son coefficient de régression logistique. Ainsi, les traits associés à un plus haut coefficient pour une classe donnée contribuent plus fortement à la probabilité qu'un commentaire soit catégorisé dans cette classe. Ces coefficients sont accessibles par un attribut du modèle de régression logistique de *scikit-learn* (`model.coef`).

incel		**non-*incel***	
Trait	**Coefficient**	**Trait**	**Coefficient**
incel	6,5982	team	3,5702
chad	6,5319	kink	3,0854
woman	5,7303	player	2,8360
ugly	5,7011	host	2,7403
incels	5,5302	character	2,5176
normies	4,6894	use	2,4960
alone	4,6703	appreciated	2,4724
virgin	4,5697	trade	2,4536
loneliness	4,5553	season	2,4162
relationship	4,4181	item	2,4016
attractive	4,3775	using	2,3349
normie	4,2995	system	2,2578
social	4,2354	issue	2,2470
life	4,2053	server	2,2433
cope	4,0723	advance	2,2372
personality	4,0205	running	2,2040
dating	3,9777	horny	2,1822
girl	3,9139	version	2,1594
hobby	3,8762	suggestion	2,1390
lonely	3,7593	killed	2,1345

TABLEAU 7. *Traits prédicitfs des classes* incel *et non*-incel

Les termes prédictifs de la classe *incel* obtiennent de manière générale de plus hauts coefficients de régression que ceux associés à la classe de données non-*incel*. Malgré les faibles métriques de performance observées pour la classe *incel* individuellement, ces coefficients suggèrent l'existence d'un vocabulaire fortement relié à cette classe pour la fonction de décision du modèle. Ces termes incluent un ensemble d'expressions dénotant la subculture *incel* ou témoignant de la vision du monde qui y est associée (p. ex. *incel* = 6,5982, *chad* = 6,5319, *normies* = 4,6894). Un certain nombre de ces termes font référence aux femmes (p. ex. *women* = 5,7303, *girl* = 3,9138), tandis que d'autres relèvent de la solitude associée au célibat involontaire (p. ex. *alone* = 4,6703, *loneliness* = 4,5553, *lonely* = 3,7593). Certains traits reflètent finalement l'importance accordée à l'apparence physique dans ces communautés (p. ex. *ugly* = 5,7011, *attractive* = 4,3775) ou dénotent d'aspects relatifs aux relations intimes (p. ex. *virgin* = 4,5697, *relationship* = 4,4181, *dating* = 3,9777).

Ces traits sont représentatifs de la sous-culture *incel* dans une perspective générale. Il convient ici de rappeler que nous considérons dans cette étude l'ensemble des *subreddits incels* comme formant une communauté homogène, ce qui est attesté par le fait que plusieurs utilisateurs sont fréquemment abonnés aux mêmes subreddits (Ribeiro *et al.*, 2021). Une analyse plus fine pourrait mettre en lumière les thématiques de discussion de *subreddits* spécifiques, par exemple *r/gymscels*, dédié au fitness, ou *r/shortcels*, dédié aux difficultés relationnelles associées à la grandeur physique chez les *incels*.

En contrepartie, plusieurs des traits présentant les plus hauts coefficients pour la classe de données non-*incels* relèvent davantage de l'usage général de la langue sur Reddit sans évoquer de thématique particulière (p. ex. *use* = 2,4960, *appreciated* = 2,4724, *item* = 2,4016, *suggestion* = 2,1390). D'autres de ces termes pourraient relever de thématiques populaires sur la plateforme telles que le sport (p. ex. *team* = 3,5702, *player* = 2,8360, *trade* = 2,4536, *season* = 2,4162) ou la sexualité (p. ex. *kink* = 3,0854, *horny* = 2,1822). Néanmoins, les valeurs plus faibles associées aux coefficients de cette classe par rapport à la classe *incel* suggèrent que ces caractéristiques sont moins discriminantes pour la fonction de décision.

5. Conclusion

Cette étude évalué différents modèles de classification supervisée pour détecter le discours des *incels* sur la plateforme Reddit. Les expérimentations menées ont fait varier le ratio de la classe à détecter dans les données d'apprentissage, l'approche de vectorisation employée, le nombre de traits discriminants retenus par les modèles ainsi que l'algorithme de classification utilisé. Nos résultats indiquent de meilleures performances avec 20 % de données *incel* dans les données d'apprentissage pour des vecteurs SBERT et un modèle de régression logistique. Le système le plus performant obtient une mesure-F globale de 82,35. Les résultats obtenus sur les données test, lesquelles contiennent un ratio de 10 % de données *incel*, sont toutefois légèrement inférieurs (mesure-F globale de 79,70). Cela est possiblement explicable par le débalancement des classes à prédire, mais également par la manière dont les données sont annotées (en assignant une catégorie à des *subreddits* entiers) lors de la constitution des corpus. Un examen plus approfondi des erreurs de classification de nos modèles permettrait de faire la lumière sur ces résultats. Les performances obtenues sont cependant encourageantes. Notre démarche comparant rigoureusement différents modèles répond aux critiques d'ordre méthodologique parfois soulevées dans ce type de tâches. En nous inspirant des travaux de Klein et Golbeck (2024), nous entendons poursuivre cette recherche en explorant l'évolution des patrons langagiers les plus pertinents pour la détection de ce type de discours, en évaluant les performances d'autres méthodes de classification adaptées aux particularités de notre corpus.

6. Remerciements

Les auteurs remercient les évaluateurs dont les commentaires ont grandement contribué à la qualité de cet article. Nous remercions également Isabelle Fontaine et Audrée Frappier pour leurs contributions aux phases antérieures du projet.

7. Bibliographie

Abubakar H. D., Umar M., Bakale M. A., « Sentiment classification : Review of text vectorization methods : Bag of words, Tf-Idf, Word2vec and Doc2vec », *SLU Journal of Science and Technology*, vol. 4, n° 1 & 2, p. 27-33, 2022.

Almerekhi H., Jansen S. b. B. J., Kwak c.-s. b. H., « Investigating toxicity across multiple Reddit communities, users, and moderators », *Companion proceedings of the web conference 2020*, p. 294-298, 2020.

Arango A., Pérez J., Poblete B., « Hate speech detection is not as easy as you may think : A closer look at model validation (extended version) », *Information Systems*, vol. 105, p. 1-11, 2022.

Associated Press, « Le suspect d'une tuerie en Californie est le fils d'un réalisateur de Hollywood », *Radio-Canada*, May, 2014.

Azmy W. M., Moulahi B., Bringay S., Azé J., Servajean M., « Lirmm@ deft-2018–modèle de classification de la vectorisation des documents », *Actes de DEFT, Rennes, France*, 2018.

Baumgartner J., Zannettou S., Keegan B., Squire M., Blackburn J., « The Pushshift Reddit Dataset », *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 14, p. 830-839, May, 2020.

Benmoussa M., Chénier I., Keenan-Pelletier M., Mason R., Robichaud M., Tanguay L., Valiquet D., Walker J., « Résumé législatif du projet de loi C-63 : Loi édictant la Loi sur les préjudices en ligne, modifiant le Code criminel, la Loi canadienne sur les droits de la personne et la Loi concernant la déclaration obligatoire de la pornographie juvénile sur Internet par les personnes qui fournissent des services Internet et apportant des modifications corrélatives et connexes à d'autres lois », *Bibliothèque du Parlement du Canada*, 2024.

Berglind T., Pelzer B., Kaati L., « Levels of hate in online environments », *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, p. 842-847, 2019.

Bird S., Klein E., Loper E., *Natural language processing with Python : analyzing text with the natural language toolkit*, " O'Reilly Media, Inc.", 2009.

Breiman L., « Random Forests—Random Features [Technical Report 567] », 1999.

Chandrasekharan E., Samory M., Srinivasan A., Gilbert E., « The Bag of Communities : Identifying Abusive Behavior Online with Preexisting Internet Data », *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, Association for Computing Machinery, New York, NY, USA, p. 3175-3187, May, 2017.

Feldman R., Sanger J., *The Text Mining Handbook : Advanced Approaches in Analyzing Unstructured Data*, Cambridge University Press, Cambridge, 2006.

Fersini E., Gasparini F., Rizzi G., Saibene A., Chulvi B., Rosso P., Lees A., Sorensen J., « SemEval-2022 Task 5 : Multimedia Automatic Misogyny Identification », *in* G. Emerson,

N. Schluter, G. Stanovsky, R. Kumar, A. Palmer, N. Schneider, S. Singh, S. Ratan (eds), *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, Association for Computational Linguistics, Seattle, United States, p. 533-549, July, 2022.

Fersini E., Nozza D., Rosso P., « AMI @ EVALITA2020 : Automatic Misogyny Identification », *in* V. Basile, D. Croce, M. D. Maro, L. C. Passaro (eds), *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, vol. 2765 of *CEUR Workshop Proceedings*, CEUR, Online event, December, December, 2020.

Fersini Elisabetta e. a., *EVALITA Evaluation of NLP and Speech Tools for Italian*, Accademia University Press, chapter Overview of the Evalita 2018 Task on Automatic Misogyny Identification (AMI), p. 59-66, 2018.

Frenda S., Ghanem B., Montes-y Gómez M., Rosso P., « Online Hate Speech against Women : Automatic Identification of Misogyny and Sexism on Twitter », *Journal of Intelligent & Fuzzy Systems*, vol. 36, n° 5, p. 4743-4752, January, 2019.

Gemelli S., Minnema G., « Manosphrames : exploring an Italian incel community through the lens of NLP and Frame Semantics », *in* P. Sommerauer, T. Caselli, M. Nissim, L. Remijnse, P. Vossen (eds), *Proceedings of the First Workshop on Reference, Framing, and Perspective @ LREC-COLING 2024*, ELRA and ICCL, Torino, Italia, p. 28-39, May, 2024.

Gothard K., Dewhurst D. R., Minot J. R., Adams J. L., Danforth C. M., Dodds P. S., « The incel lexicon : Deciphering the emergent cryptolect of a global misogynistic community », *arXiv [cs]*, May, 2021.

gouvernement du Canada, « Loi édictant la Loi sur les préjudices en ligne, modifiant le Code criminel, la Loi canadienne sur les droits de la personne et la Loi concernant la déclaration obligatoire de la pornographie juvénile sur Internet par les personnes qui fournissent des services Internet et apportant des modifications corrélatives et connexes à d'autres lois », , Projet de loi no C-63 (dépôt et 1re lecture – 26 février 2024), 1e sess., 44e légis., 2024.

Hajarian M., Khanbabaloo Z., « Toward Stopping Incel Rebellion : Detecting Incels in Social Media Using Sentiment Analysis », *2021 7th International Conference on Web Research (ICWR)*, p. 169-174, May, 2021.

Halpin M., « Weaponized Subordination : How Incels Discredit Themselves to Degrade Women », *Gender & Society*, vol. 36, n° 6, p. 813-837, December, 2022.

Halpin M., Preston K., Lockyer D., Maguire F., « A solider and a victim : Masculinity, violence, and incels celebration of December 6th », *Canadian Review of Sociology/Revue canadienne de sociologie*, vol. 61, p. cars.12460, January, 2024.

Hauser C., « Reddit Bans 'Incel' Group for Inciting Violence Against Women », *The New York Times*, November, 2017.

Hornby A. S., « Incel », *Oxford Advanced Learner's Dictionary*, 2020.

Jaki S., Smedt T. D., Gwóźdź M., Panchal R., Rossa A., Pauw G. D., « Online hatred of women in the Incels.me forum : Linguistic analysis and automatic detection », *Journal of Language Aggression and Conflict*, vol. 7, n° 2, p. 240-268, November, 2019.

Jamshidian M., « Evaluation of Text Transformers for Classifying Sentiment of Reviews by Using TF-IDF, BERT (word embedding), SBERT (sentence embedding) with Support Vector Machine Evaluation », 2023.

Jurafsky D., Martin J. H., *Speech and language processing*, Stanford Univ, 2019.

Kirk H., Yin W., Vidgen B., Röttger P., « SemEval-2023 Task 10 : Explainable Detection of Online Sexism », *in* A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, E. Sartori (eds), *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, Association for Computational Linguistics, Toronto, Canada, p. 2193-2210, July, 2023.

Klein E., Golbeck J., « A Lexicon for Studying Radicalization in Incel Communities », *Proceedings of the 16th ACM Web Science Conference*, WEBSCI '24, Association for Computing Machinery, New York, NY, USA, p. 262–267, 2024.

Ling C. X., Sheng V. S., « Class Imbalance Problem », *in* C. Sammut, G. I. Webb (eds), *Encyclopedia of Machine Learning*, Springer, Boston, MA, p. 171-171, 2010.

Manning C. D., Raghavan P., Schütze H., *Introduction to Information Retrieval*, Higher Education from Cambridge University Press, July, 2008.

.ME, « The suspension of incels.me », *.ME blog*, November, 2018.

Meier M. L., Sharp K., « Death to Chad and Stacy : Incels and anti-fandom as group identity », *International Journal of Cultural Studies*, vol. 27, p. 349-367, January, 2024.

Mikolov T., Chen K., Corrado G., Dean J., « Efficient Estimation of Word Representations in Vector Space », *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, 2013.

Morales-Castro J. C., Hernandez-Rayas A., Ruiz-Pinales J., Guzmán-Cabrera R., « Automatic Identification of Misogynistic Sentiments on Social Networks », *Journal of Social Researches*, vol. 9, n° 23, p. 10-18, 2023.

Muralikumar M. D., Yang Y. S., McDonald D. W., « A human-centered evaluation of a toxicity detection api : Testing transferability and unpacking latent attributes », *ACM Transactions on Social Computing*, vol. 6, n° 1-2, p. 1-38, 2023.

Muti A., Ruggeri F., Khatib K. A., Barrón-Cedeño A., Caselli T., « Language is Scary when Over-Analyzed : Unpacking Implied Misogynistic Reasoning with Argumentation Theory-Driven Prompts », *in* Y. Al-Onaizan, M. Bansal, Y.-N. Chen (eds), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Miami, Florida, USA, p. 21091-21107, November, 2024.

Nadeau J.-P., « Attaque au camion-bélier : Alek Minassian condamné à 25 ans minimum », *Radio-Canada*, June, 2022.

Pal M., « Random forest classifier for remote sensing classification », *International journal of remote sensing*, vol. 26, n° 1, p. 217-222, 2005.

Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., Duchesnay E., « Scikit-learn : Machine Learning in Python », *Journal of Machine Learning Research*, vol. 12, p. 2825-2830, 2011.

Pelzer B., Kaati L., Cohen K., Fernquist J., « Toxic language in online incel communities », *SN Social Sciences*, vol. 1, n° 8, p. 213, August, 2021.

Plaza L., Carrillo-de Albornoz J., Amigó E., Gonzalo J., Morante R., Rosso P., Spina D., Chulvi B., Maeso A., Ruiz V., « EXIST 2024 : sEXism Identification in Social neTworks and Memes », *in* N. Goharian, N. Tonellotto, Y. He, A. Lipani, G. McDonald, C. Macdonald, I. Ounis (eds), *Advances in Information Retrieval*, Springer, Cham, p. 498-504, 2024.

Plaza L., Carrillo-de Albornoz J., Morante R., Amigó E., Gonzalo J., Spina D., Rosso P., « Overview of EXIST 2023 – Learning with Disagreement for Sexism Identification and Charac-

terization », *in* A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, A. Giachanou, D. Li, M. Aliannejadi, M. Vlachos, G. Faggioli, N. Ferro (eds), *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, Springer, Cham, p. 316-342, 2023.

Ramos J., « Using tf-idf to determine word relevance in document queries », *Proceedings of the first instructional conference on machine learning*, vol. 242, Citeseer, p. 29-48, 2003.

Reimers N., Gurevych I., « Sentence-BERT : Sentence Embeddings using Siamese BERT-Networks », *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 11, 2019.

Ribeiro M. H., Blackburn J., Bradlyn B., Cristofaro E. D., Stringhini G., Long S., Greenberg S., Zannettou S., « Dataset for : The Evolution of the Manosphere Across the Web », *Zenodo*, August, 2020.

Ribeiro M. H., Blackburn J., Bradlyn B., Cristofaro E. D., Stringhini G., Long S., Greenberg S., Zannettou S., « The Evolution of the Manosphere across the Web », *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 15, p. 196-207, May, 2021.

Rodríguez-Sanchez F. J., de Albornoz J. C., Plaza L., Gonzalo J., Rosso P., Comet M., Donoso T., « Overview of EXIST 2021 : sEXism Identification in Social neTworks », *Proces. del Leng. Natural*, vol. 67, p. 195-207, 2021.

Rodríguez-Sánchez F., Carrillo-de Albornoz J., Plaza L., Mendieta-Aragón A., Marco-Remón G., Makeienko M., Plaza M., Gonzalo J., Spina D., Rosso P., « Overview of EXIST 2022 : sEXism Identification in Social neTworks », *Procesamiento de Lenguaje Natural*, vol. 69, p. 229-240, 09, 2022.

RSR R. d. S. l. R., Le phénomène incel : exploration des problèmes internes et externes touchant les célibataires involontaires, Technical report, Commission européenne, 2021.

Saha P., Mathew B., Goyal P., Mukherjee A., « Hateminers : Detecting Hate speech against Women », *arXiv :1812.06700*, December, 2018.

Sheppard B., Richter A., Cohen A., Smith E., Kneese T., Pelletier C., Baldini I., Dong Y., « Biasly : An Expert-Annotated Dataset for Subtle Misogyny Detection and Mitigation », *in* L.-W. Ku, A. Martins, V. Srikumar (eds), *Findings of the Association for Computational Linguistics : ACL 2024*, Association for Computational Linguistics, Bangkok, Thailand, p. 427-452, August, 2024.

Shetty A., « Accused in University of Waterloo stabbings charged with attempted murder, added to other counts », *CBC News*, 2023.

Tranchese A., Sugiura L., « "I Don't Hate All Women, Just Those Stuck-Up Bitches" : How Incels and Mainstream Pornography Speak the Same Extreme Language of Misogyny », *Violence Against Women*, vol. 27, nº 14, p. 2709-2734, November, 2021.

Truşcă M. M., « Efficiency of SVM classifier with Word2Vec and Doc2Vec models », *Proceedings of the International Conference on Applied Statistics*, p. 496-503, 2019.

Wang Y., Zhou Z., Jin S., Liu D., Lu M., « Comparisons and selections of features and classifiers for short text classification », *Iop conference series : Materials science and engineering*, vol. 261, IOP Publishing, 2017.

Yoder M., Perry C., Brown D., Carley K., Pruden M., « Identity Construction in a Misogynist Incels Forum », *The 7th Workshop on Online Abuse and Harms (WOAH)*, Association for Computational Linguistics, Toronto, Canada, p. 1-13, 2023.

Zhao Y., Chen Y., *Data Mining Applications with R*, Elsevier, 2014.