SANDRINE SÉNÉCHAL
JOSIANE VAN DORPE
UQAM

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

Ce résumé présente une étude sur les communications écrites des *incels* sur le forum spécialisé `incels.is` (*4*). Nous explorons la façon dont les biais d'associations explicitement revendiqués par cette communauté linguistique se retrouvent dans les modèles de représentation du sens lexical produits automatiquement. Les *incels* (*involuntary celibates*) forment une communauté en ligne d'hommes s'identifiant comme hétérosexuels et considérés comme extrémistes violents (*10, 11*). Ils sont reconnus pour la toxicité et le caractère haineux de leurs discours (*2, 5, 8, 9*). Leurs communications se caractérisent notamment par la revendication de préjugés et de stéréotypes envers les femmes et par l'utilisation généralisée de néologismes péjoratifs propres à cette communauté. 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*| (*3*)).

Considérant que les incels revendiquent leurs biais d'associations, il est légitime de se demander si ce type de biais se distinguent des biais implicites connus dans les modèles de plongements, et 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 lié à l'usage généralisé de néologismes péjoratifs référant aux femmes. Cela étant, des termes standards renvoyant aux femmes sont aussi utilisés. On se demandera alors dans quelles mesures on peut distinguer les biais associés à ces deux classes de termes.

Nous avons d'abord extrait notre corpus d'étude du forum `incels.is` (271 000 publications, 26,7 millions de tokens en anglais). Nous avons ensuite extrait notre corpus contrôle de *Reddit* (280 000 publications, 44,3 millions de tokens en anglais). Si plusieurs méthodes existent pour détecter les biais d'associations implicites, 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. Pour ce faire, nous avons eu recours à un modèle de plongement lexical statique. Produit automatiquement sur la base d'un large corpus non annoté, ce modèle offre des représentations sémantiques des termes du corpus sous forme de vecteurs en se basant seulement sur leur distribution dans le texte (*6*). Nous avons produit deux modèles de plongement entraînés sur chaque corpus grâce à l'algorithme `Word2Vec` (*7*), puis, afin de déterminer si les vecteurs ont capturé des biais d'associations latents dans les données textuelles sur lesquelles ils ont été entraînés, nous avons eu recours au *Word-Embedding Association Test* (WEAT) (*1*). Le WEAT calcule le cosinus de l'angle entre deux vecteurs de termes cibles.

Nous avons soumis nos deux modèles à plus de 40 tests. Pour les biais d'associations implicites connus, un seul des quatre tests s'est avéré significatif, et ce, pour les deux corpus. En ce qui concerne la distinction des biais implicites/explicites dans le modèle, nous n'avons pas été en mesure de distinguer les deux types biais à partir de nos résultats. Cependant, nous avons observé une nette distinction entre les néologismes féminins et les termes standards féminins, suggérant la présence d'un biais d'associations |termes standards féminins–plaisant/néologismes féminins–déplaisant|. Inversement, nos résultats ne démontrent pas de différence significative entre les termes standards masculins et les néologismes masculins.

Notre corpus d'étude ainsi que son modèle vectorisé sont disponibles en libre accès afin de permettre une continuation des recherches transdisciplinaires sur le sujet. L'impact de nos résultats se fait ainsi surtout sentir dans les multiples perspectives de recherches liées aux discours écrits des incels. Puisque notre étude implique des éléments liés à la psychologie et à la linguistique, nous soutenons que nos résultats apportent des pistes intéressantes pour ces deux domaines, minimalement.

RÉFÉRENCES

1. 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

2. 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

3. 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

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

5. 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

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

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

8. 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

9. 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

10. Service canadien du renseignement de sécurité. (2020, avril). *RAPPORT PUBLIC DU SCRS 2019 : Des renseignements et des conseils fiables pour un Canada sûr et prospère*. https://www.canada.ca

11. Service canadien du renseignement de sécurité. (2021, avril). *RAPPORT PUBLIC DU SCRS 2020 : Des renseignements et des conseils fiables pour un Canada sûr et prospère*. https://www.canada.ca