IDENTIFIER LES CONTENUS TOXIQUES

Sur les plateformes sociales, la prévalence du langage haineux ou offensant en ligne a augmenté rapidement ces dernières années, et désormais le problème est grave. Dans certains cas, des commentaires toxiques en ligne ont entraîné des violences réelles, du nationalisme religieux au Myanmar à la propagande néonazie aux États-Unis.

Pourtant, des milliers de vérificateurs humains luttent pour modérer le volume toujours croissant de contenus préjudiciables. Mais, en 2019, il a été rapporté que les modérateurs de Facebook risquaient de souffrir de troubles de stress post-traumatique du fait d'une exposition répétée à des contenus aussi dérangeants. Confier au moins en partie ces activités de modération à des machines aiderait à gérer les volumes croissants de contenus préjudiciables, tout en limitant l'exposition humaine à ces derniers. Et en effet, de nombreux géants de la technologie intègrent des algorithmes dans leur modération de contenus depuis des années.

C'est le cas de Google Jigsaw. En 2017, cette entreprise a contribué à la création de *Conversation AI*, un projet de recherche collaboratif visant à détecter les commentaires toxiques en ligne. Cependant, un outil produit par ce projet, appelé Perspective, a fait l'objet de nombreuses critiques. Un reproche fréquent était qu'il créait un « score de toxicité » général qui n'était pas assez flexible pour répondre aux besoins variés des différentes plateformes. Par exemple, certains sites web souhaitent détecter les menaces mais pas les blasphèmes, tandis que d'autres ont des exigences opposées.

Un autre problème est que l'algorithme, après apprentissage, ne distinguait pas les commentaires toxiques des commentaires non toxiques contenant des mots liés au sexe, à l'orientation sexuelle, à la religion ou au handicap. Par exemple, un utilisateur a signalé que des phrases neutres et simples telles que « Je suis une femme noire lesbienne » ou « Je suis une femme sourde » donnaient des scores de toxicité élevés, tandis que « Je suis un homme » donnait un score faible.

À la suite de ces préoccupations, l'équipe de *Conversation AI* a invité les développeurs à entraîner leurs propres algorithmes de détection de toxicité et à les inscrire à trois concours (un par an) organisés sur Kaggle, une filiale de Google connue pour sa communauté de praticiens de l'apprentissage automatique, ses ensembles de données publiques et ses défis. Pour aider à entraîner les algorithmes, *Conversation AI* a publié deux ensembles de données publiques contenant plus d'un million de commentaires toxiques et non toxiques provenant de Wikipedia et d'un service appelé Civil Comments. Les commentaires ont été classés selon leur toxicité par des annotateurs, avec une étiquette « Très toxique » indiquant « un commentaire très haineux, agressif ou irrespectueux qui vous fera très vraisemblablement quitter une discussion ou renoncer à partager votre point de vue », et une étiquette « Toxique » signifiant « un commentaire grossier, irrespectueux ou déraisonnable qui est assez susceptible de vous faire quitter une discussion ou de renoncer à partager votre point de vue ». Certains commentaires ont été vus par beaucoup plus de dix annotateurs (jusqu'à des milliers), en raison de l'échantillonnage et des stratégies utilisées pour renforcer la précision des évaluateurs.

Detoxify pour purger le web

L'objectif du premier défi de Jigsaw était de construire un modèle de classification des commentaires toxiques avec des étiquettes telles que « toxique », « gravement toxique », « menace », « insulte », « obscène » et « haine identitaire ». Les deuxième et troisième défis se sont concentrés sur des aspects plus spécifiques : minimiser les biais involontaires concernant les mentions identitaires et ceux dus au fait que les modèles multilingues sont entraînés avec des données en anglais uniquement. Bien que les défis aient conduit à des moyens astucieux d'améliorer les modèles de langage toxique, notre équipe chez Unitary, une société d'intelligence artificielle modératrice de contenus, a constaté qu'aucun des modèles entraînés n'avait été rendu public.

C'est pourquoi nous avons décidé de nous inspirer des meilleures solutions de Kaggle et d'entraîner nos propres algorithmes dans l'intention de les rendre publics. Pour ce faire, nous nous sommes appuyés sur des modèles existants de traitement du langage naturel, comme Bert, de Google. Beaucoup de ces modèles ont leur code source en libre accès. C'est ainsi que notre équipe a créé Detoxify, une bibliothèque d'algorithmes de détection des

commentaires, ouverte et conviviale, pour repérer les textes inappropriés ou préjudiciables. Son utilisation est destinée à aider les praticiens à identifier les commentaires potentiellement toxiques.
Dans le cadre de cette bibliothèque, nous avons publié trois modèles différents correspondant à chacun des trois défis de Jigsaw. Alors que les meilleures solutions de Kaggle pour chaque défi utilisent des ensembles de modèles, qui font la moyenne des scores de plusieurs modèles entraînés, nous avons obtenu une performance similaire avec un seul modèle par défi. Chaque modèle est facilement accessible en une seule ligne de programme et tous les modèles ainsi que le programme d'apprentissage sont disponibles publiquement sur GitHub.

« Stupide », un mot à toxicité variable

Bien que ces modèles soient performants dans de nombreux cas, il est important de noter également leurs limites. Tout d'abord, ils fonctionneront bien sur des exemples similaires aux données sur lesquelles ils ont été formés, mais ils risquent d'échouer s'ils sont confrontés à des exemples moins familiers. Nous encourageons les développeurs à affiner ces modèles sur des ensembles de données toxiques représentatifs de leur cas d'utilisation.
De plus, nous avons remarqué que l'inclusion d'insultes ou de blasphèmes dans un commentaire textuel entraînera presque toujours un score de toxicité élevé, indépendamment de l'intention ou du ton de l'auteur. Par exemple, la phrase « Je suis fatigué d'écrire cet essai stupide » donnera un score de toxicité de 99,7 %, tandis que la suppression du mot « stupide » fera passer le score à 0,05 %.
Enfin, malgré le fait qu'un des modèles publiés ait été spécifiquement formé pour limiter les biais involontaires, les trois modèles sont toujours susceptibles de présenter certains biais, ce qui peut poser des problèmes éthiques lorsqu'on les utilise pour modérer des contenus.

Bien que des progrès considérables aient été accomplis en matière de détection automatique des propos toxiques, beaucoup reste à faire avant que les modèles puissent saisir le sens réel, nuancé, de notre langue – au-delà de la simple mémorisation de mots particuliers ou de phrases particulières. Bien sûr, investir dans des ensembles de données plus représentatifs et de meilleure qualité, pour l'apprentissage des systèmes d'intelligence artificielle, permettrait des améliorations progressives. Mais nous devons aller plus loin et commencer à interpréter les données dans leur contexte, un élément crucial pour comprendre le comportement en ligne. Un texte apparemment anodin publié sur les réseaux sociaux, mais accompagné d'un symbolisme raciste dans une image ou une vidéo, passerait facilement inaperçu si nous ne regardions que les mots écrits.
Il est bien connu qu'en ignorant ou en négligeant le contexte, nous commettons facilement des erreurs de jugement. Pour que l'intelligence artificielle remplace à grande échelle les humains dans la modération des contenus échangés sur les réseaux sociaux, il est impératif que nous donnions à nos modèles une image complète de ces contenus.

Laura Hanu
est ingénieure en vision par ordinateur dans la société britannique Unitary.

James Thewlis
est docteur en vision par ordinateur, cofondateur et directeur technique de Unitary.

Sasha Haco
est docteure en physique théorique, cofondatrice et PDG de Unitary.

la majorité), le biais d'ancrage (qui pousse à se fier à l'information reçue en premier dans une prise de décision)…

L'emploi de *nudges* à travers des *bots* pour influencer les internautes s'est développé ces dernières années. Il peut s'agir de *nudges* pour inciter ou pour freiner des comportements, que Richard Thaler appelle *sludges* (littéralement « boues », en anglais). L'un des défis consiste à évaluer l'apport ou la menace qu'ils représentent. Ainsi, le projet *Bad Nudge-Bad Robot* que je dirige au CNRS à l'université Paris-Saclay vise à mettre en évidence le danger des techniques de *nudging* pour les personnes vulnérables telles que les enfants ou les personnes âgées. Ce projet entre dans le cadre de ma chaire d'intelligence artificielle Humaaine (Human-Machine affective interaction & ethics) au CNRS, qui repose sur une forte collaboration interdisciplinaire entre l'informatique affective et la robotique émotionnelle, l'économie comportementale, la linguistique et le traitement du langage naturel.

DANS LA JUNGLE DU WEB

Par exemple, en 2019, dans une expérience de psychologie sociale conduite dans une école primaire, nous avons montré que les *nudges* des machines (robots, *chatbots*) étaient plus efficaces que les *nudges* d'humains. Dans cette expérience, la machine ou l'humain donne le même nombre de billes à chaque enfant. La première question posée à l'enfant est : combien il en garde pour lui et combien il en donne aux autres enfants. La deuxième question est un *nudge* pour le faire changer d'avis, en ajoutant par exemple que les autres en donnent plus (ce qui active le biais de conformisme). Plus les enfants étaient jeunes, plus ils étaient facilement manipulés par les machines. Cet exemple met en lumière le fort potentiel d'incitation associé à des agents capables de déployer un *nudge*. À n'en pas douter, de tels *bots* accroîtront les moyens de manipulation sur le web.

L'intelligence artificielle fournit ainsi des armes diverses, les unes permettant de traquer les propos haineux et repérer les *trolls* et les *bots*, les autres de propager des fausses informations et de manipuler les internautes, notamment grâce aux *nudges*. Ceux-ci sont aussi susceptibles d'être utilisés de façon positive. Imaginons par exemple qu'à la fin du mois on vous attribue des points en fonction de la qualité de ce que vous avez tweeté, et que vous puissiez comparer avec vos amis ou vos collègues. Ce *nudge* vous inciterait à moins relayer des informations non vérifiées et peut-être à adopter des pratiques plus respectueuses des autres.

Lutter contre la désinformation s'annonce donc comme un défi permanent. Le recours à l'intelligence artificielle dans ce combat n'en est certainement qu'à ses débuts. Si se doter de tels outils sera utile pour faire face aux propos toxiques ou aux *bots* malveillants, cela restera néanmoins insuffisant.

Pour protéger citoyens et démocraties, une régulation de cette jungle qu'est le web, par des lois et des garde-fous éthiques, est nécessaire. C'est assurément une question aussi difficile que cruciale pour l'avenir de nos sociétés. Elle suscite aujourd'hui de nombreuses réflexions. Notamment, le Comité national pilote d'éthique du numérique a publié en juillet 2020 un document portant sur les enjeux dans la lutte contre la désinformation et la mésinformation, lequel identifie les tensions éthiques relatives à la mise en œuvre, par les plateformes, d'outils de modération de contenus et aux mécanismes de lutte contre la désinformation. Ce texte interroge aussi l'autorité acquise par ces plateformes et les rapports qu'elles entretiennent avec l'État et la justice. En résumé, il convient de redéfinir la responsabilité des plateformes aux niveaux national et européen, tout en protégeant la liberté d'expression. De ces sujets de société, nous allons devoir débattre.

— L'autrice —

> Laurence Devillers est professeuse en intelligence artificielle à Sorbonne-Université et dirige une équipe de recherche au CNRS-LISN (Saclay), où elle est responsable de la chaire IA Humaaine (https://humaaine-chaireia.fr).

— À lire —

> L. Devillers, *Vague IA à l'Élysée*, L'Observatoire, 2022.

> H. Ali Mehenni, L. Devillers *et al.*, Nudges with conversational agents and social robots: a first experiment with children at a primary school, Conversational Dialogue Systems for the Next Decade, *Springer*, pp. 257-270, 2021.

> C. Schneider *et al.*, Digital nudging: guiding online user choices through interface design, *Communications of the ACM*, 2018.

> D. Kahneman, *Thinking, Fast and Slow*, Farrar, Straus and Giroux, 2011.