IDENTIFIER LES CONTENUS TOXIQUES

Les plateformes sociales sur internet, grandes et petites, luttent pour protéger leurs communautés contre les discours de haine, les contenus extrémistes, le harcèlement et la désinformation. Très récemment, aux États-Unis, des agitateurs d'extrême droite ont ouvertement publié des informations sur des plans visant à prendre d'assaut le Capitole avant de le faire le 6 janvier. Une solution pourrait venir de l'intelligence artificielle (IA) : il s'agirait de développer des algorithmes qui détectent des commentaires toxiques et incendiaires et nous les signalent pour qu'ils soient retirés. Mais de tels systèmes sont confrontés à de grands défis.

La prévalence du langage haineux ou offensant en ligne a augmenté rapidement ces dernières années, et le problème est maintenant 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.

Les plateformes des réseaux sociaux, qui s'appuient sur des milliers de vérificateurs humains, luttent pour modérer le volume toujours croissant de contenus préjudiciables. En 2019, il a été rapporté que les modérateurs de Facebook risquent de souffrir de troubles de stress post-traumatique à la suite d'une exposition répétée à des contenus aussi dérangeants.

Confier au moins en partie ces activités de modération à des systèmes à apprentissage automatique 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, une entreprise qui s'efforce de rendre internet plus sûr. En 2017, elle 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.

L'objectif du premier défi Jigsaw était de construire un modèle de classification des commentaires toxiques avec des étiquettes

© Matyo

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 toxiques, notre équipe chez Unitary, une société d'IA modératrice de contenus, a constaté qu'aucun des modèles entraînés n'avait été rendu public.

Il reste encore beaucoup de chemin à parcourir avant que les modèles puissent saisir le sens réel, nuancé, de notre langue

C'est pourquoi nous avons décidé de nous inspirer des meilleures solutons 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 le BERT de Google. Beaucoup de ces modèles ont leur code source libre d'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 en ligne. Son utilisation est destinée à aider les chercheurs et 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 du 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 et le programme d'apprentissage sont disponibles publiquement sur GitHub.

Bien que ces modèles soient performants dans de nombreux cas, il est important de noter également leurs limites. Tout d'abord, ces modèles 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 peu familiers de langage toxique. Nous encourageons les développeurs à affiner ces modèles sur des ensembles de données 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 a 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, nous avons encore beaucoup de chemin à parcourir 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 le texte. 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[1], JAMES THEWLIS[2] et SASHA HACO[3]
[1]ingénieure en vision par ordinateur dans la société britannique Unitary
[2]docteur en vision par ordinateur, cofondateur et directeur technique de Unitary
[3]docteure en physique théorique, cofondatrice et PDG de Unitary

SUR LE WEB
Site de la société Unitary:
https://www.unitary.ai/
Le projet Conversation AI et son outil Perspective:
https://conversationai.github.io/
https://www.perspectiveapi.com/#/home
La bibliothèque Detoxify:
https://github.com/unitaryai/detoxify

> d'augmenter la prescription de médicaments génériques et réduire la prescription d'opioïdes.

Les *nudges* peuvent encourager des comportements ayant des effets à long terme pour l'individu ou la société. Par exemple, avant 1976, en France, l'individu devait déclarer explicitement son consentement à faire don, après sa mort, de ses organes à la médecine ; cette règle a été modifiée et le consentement est désormais le choix par défaut. Or une étude de 2003 portant sur plusieurs pays et réalisée par Eric Johnson et Daniel Goldstein, de l'université Columbia, a montré que lorsque le consentement est le choix par défaut, le pourcentage effectif de donneurs d'organes est notablement supérieur : il est presque le double.

QUAND LE « NUDGE » DEVIENT NUMÉRIQUE

Le concept de *nudge* s'intègre naturellement à l'informatique affective. Ce domaine regroupe des technologies d'intelligence artificielle comme la détection par les robots (qu'ils soient matériels ou purement logiciels) des émotions de l'interlocuteur humain, la génération d'émotions chez celui-ci, la modulation des décisions du robot à partir des informations d'ordre affectif détectées dans le comportement humain... Dans ce contexte des techniques d'intelligence artificielle, les *nudges* vont renforcer les pouvoirs d'incitation. On peut les déployer dans des bots pour orienter subrepticement les choix des individus en exploitant nos biais cognitifs tels que le biais de confirmation (tendance à privilégier les informations qui confirment nos idées préconçues), le biais de conformisme (tendance à penser et agir comme la majorité), le biais d'ancrage (qui pousse à se fier à l'information reçue en premier dans une prise de décision), le biais d'autorité (tendance à estimer plus juste l'opinion d'une figure d'autorité), etc.

Les *nudges* peuvent être utilisés dans le monde numérique à bon escient comme à mauvais escient. Ils pourraient notamment jouer un rôle positif dans la lutte contre la désinformation. Par exemple, la sensibilisation des internautes à l'existence des *nudges* numériques permettrait de leur faire prendre conscience des biais cognitifs auxquels ils sont soumis lorsqu'ils réagissent de façon quotidienne sur les réseaux sociaux en propageant des informations non vérifiées. On pourrait ainsi se servir de *nudges* pour contrebalancer certaines opinions radicales, dans un objectif pédagogique : faire comprendre les manipulations et renforcer l'esprit critique, lequel se distingue de l'opinion par le questionnement, l'argumentation, l'approche contradictoire et le souci d'approcher une certaine vérité.

L'utilisation de *nudges* à travers des bots pour influencer le comportement des internautes s'est développée ces dernières années. >