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