La lutte du cyberharcèlement sur Twitter grâce à l’intelligence artificielle

	La lutte du cyberharcèlement sur Twitter grâce à l’intelligence artificielle

Le réseau social va s'aider d'outils d'apprentissage automatique pour repérer plus vite les messages allant à l'encontre de ses règles d'utilisation

Un concert d'excuses et quelques mesures concrètes. Après plusieurs années de silence et d'hésitation, Twitter promet que 2017 sera l'année de la lutte contre le harcèlement. Le réseau social présente trois nouveaux outils pour limiter l'influence des discours de haine et des attaques ciblés contre ses utilisateurs. Ils seront déployés à partir de mardi. D'autres fonctionnalités devraient être présentées dans le courant de l'année. «Nous avons entendu vos critiques. Nous n'avons pas progressé assez l'année dernière», avait déclaré Ed Ho, vice-président de Twitter, fin janvier. «Nous continuerons à être attentifs à vos retours, d'apprendre des critiques et de sortir des nouvelles fonctionnalités jusqu'à ce que tous nos utilisateurs ressentent ces changements.» Twitter va se reposer sur une nouvelle arme pour l'aider dans sa modération: l'intelligence artificielle.

Repérer plus rapidement les agressions

Le nouveau plan de Twitter comporte trois mesures phares. La première doit lutter contre la création abusive de nouveaux comptes par des utilisateurs déjà bannis du réseau social. Il est difficile de repérer ces internautes. Il changent généralement d'adresse mail, de numéro de téléphone et d'adresses IP pour s'inscrire à nouveau. Twitter va s'appuyer sur un programme d'apprentissage automatique afin de repérer les resquilleurs. Tout compte banni définitivement du réseau social sera analysé afin de repérer des signaux permettant d'identifier une personne, comme une manière de parler, des sujets ou des victimes de prédilection, des hashtags préférés, etc. Si un nouveau compte Twitter correspond à cette analyse, il pourra être rapidement supprimé.

Twitter crée également une nouvelle option pour masquer les images choquantes dans les recherches de tweets. Sont concernées les photos pornographiques ou violentes. Elles seront repérées automatiquement. Par exemple, une personne tapant «Bataclan» dans la barre de recherche de Twitter devrait en théorie ne pas voir de photos de la tuerie. Cet outil devrait aussi être utile pour les personnes faisant l'objet d'une campagne de dénigrement, afin de ne pas voir son pseudo associé à des images pornographiques ou violentes. L'option sera enclenchée par défaut, mais pourra être désactivée dans les réglages Twitter.

Dernier outil lancé par le réseau social: les réponses à un tweet seront bientôt classées par ordre d'intérêt. Les messages automatiquement détectés comme «peu intéressants» par Twitter seront relégués en bas. Parmi les critères examinés par le réseau social: si le compte est nouveau et ne suit aucune autre personne, s'il a déjà été signalé pour abus ou qu'il emploie des insultes.

Accélerer le signalement

L'intelligence artificielle ne va pas remplacer les modérateurs de Twitter. Elle interviendra pour accélérer le signalement de contenus. Comme les autres réseaux sociaux, Twitter applique une modération *a posteriori*: les utilisateurs doivent lui signaler les contenus problématiques pour qu'ils soient contrôlés et éventuellement supprimés s'ils enfreignent les règles d'utilisation. Le réseau social collabore aussi avec les autorités qui peuvent lui signaler des contenus illégaux. En France, plus de 466 tweets ont fait l'objet d'une demande de retrait par la police ou le gouvernement entre janvier et juin 2016…[lire la suite]

Réagissez à cet article

Source : *Twitter s’appuie sur l’intelligence artificielle pour lutter contre le harcèlement*