Combattre la haine sur Internet : trois défis à relever

Angeliki Monnier

THE CONVERSATION
L'expertise universitaire, l'exigence journalistique

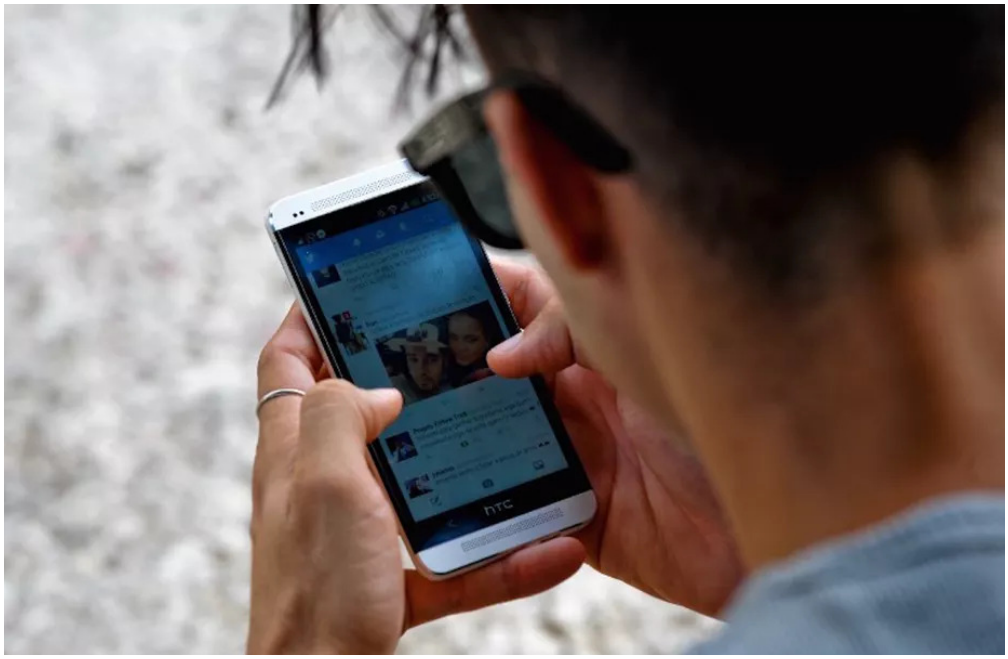
Combattre la haine sur Internet : trois défis à relever

20 mars 2019, 21:58 CET

Auteur

Angeliki Monnier

Professeure en Sciences de l'Information et de la Communication, chercheure au Centre de recherche sur les médiations, Université de Lorraine



La lutte contre la haine sur Internet et notamment sur les réseaux sociaux passe-t-elle par le vote d'une nouvelle loi ? (ici à Cuba, en mars 2019). Yamil Lage/AFP

Les discours de haine – aversion profonde de l'Autre – inondent Internet, et cela malgré les tentatives de réguler la parole en ligne. La société civile ainsi que les instances étatiques – européennes et extra-européennes (délégations, Commission européenne, Conseil de l'Europe, Unesco) – tentent de lutter contre la prolifération de ces propos haineux. Les plates-formes elles-mêmes – Facebook, Twitter, etc. – s'efforcent d'instaurer des politiques de surveillance des conduites haineuses, entre détection automatique et recours à des signalements par des personnes.

La volonté de « passer à une nouvelle étape » dans la lutte contre le racisme et la haine sur Internet a,

par ailleurs, été exprimée clairement dans un rapport soumis au premier ministre en septembre 2018. L'adoption d'une nouvelle loi est annoncée pour 2019.

Cependant, au-delà de la détermination des instances gouvernantes et de celle – parfois vue avec plus de scepticisme – des géants du web, l'éradication du discours de haine en ligne nécessite de pouvoir relever au moins trois défis.

Défi numéro 1 : maîtriser la parole affective sur le web

Aujourd'hui, exprimer ses états d'âme en ligne constitue un acte aussi banal que recherché, notamment par les plates-formes qui en tirent profit pour capter l'attention des publics. Entre divertissement, impression (illusion ?) d'émancipation, besoin de visibilité, etc., les raisons qui encouragent la désinhibition des individus en ligne sont diverses. Mais elles convergent toutes vers le même résultat : la légitimation de l'*affect* en tant que cadre naturel des échanges virtuels.

Bien évidemment, la dimension affective fait partie structurante de la nature humaine et ce ne sont pas les techniques qui l'ont inventée. Les médias dits de masse, la publicité, la communication politique, entre autres, se sont très tôt intéressés à l'instrumentalisation des émotions à des fins de propagande, sous-tendues par des préoccupations politiques ou mercantiles.

Néanmoins, la dimension affective se trouve aujourd'hui amplifiée par la portée du numérique, épaissie par l'anonymat, et érigée en ressource principale qui nourrit le web participatif et conditionne son fonctionnement. Au cœur des manifestations émotives, les discours de haine prolifèrent sur Internet : insultes, menaces, injures, rien ne semble retenir certains.

Des chercheurs voient dans ce phénomène une nouvelle forme de la lutte des classes. D'autres pointent davantage le rôle des dispositifs : l'immédiateté faciliterait la parole affective, l'anonymat réduirait les inhibitions. Les dispositifs « à forte dominante captatrice » que seraient les réseaux socio-numériques renforceraient des mises en scène « à visée pathémique », c'est-à-dire s'adressant à nos émotions.

Derrière l'apparence de diversité et de pluralisme, l'accumulation de répliques instaurerait un simulacre de démocratie et un relativisme subjectif. D'autres examinent les « affordances » affectives des dispositifs socio-numériques (petites phrases, storytelling, etc.), qui produisent à leur tour des « publics affectifs » : des ensembles d'individus disparates, liés juste par l'affect, aussi bien entre eux qu'avec le monde.

Défi numéro 2 : mieux comprendre le discours qui sous-tend la haine

Le discours de haine n'a pas de définition précise du point de vue des droits de l'Homme au niveau international. Selon le Comité des ministres du Conseil de l'Europe, il couvre toute forme d'expression qui répand ou justifie la haine raciale, la xénophobie, l'antisémitisme ou toute forme de haine basée sur l'intolérance, y incite ou en fait l'apologie.

En France, depuis 1972 et la loi Pleven, l'incitation à la haine *via* des propos tenus en public est une infraction pénale. Bien avant cela, en 1966, le Pacte international sur les droits civils et politiques prévoyait que « tout appel à la haine nationale, raciale ou religieuse qui constitue une incitation à la discrimination, à l'hostilité ou à la violence est interdit par la loi » (article 20).

Cependant, il faut souligner que le discours de haine n'est pas illicite parce qu'il est haineux, mais parce qu'il est dangereux : soit il débouche directement sur la discrimination et la violence, soit il y conduit indirectement.

En effet, combattre ce que l'on appelle couramment le discours de haine sur Internet signifie s'arrêter sur le sens même de la *haine*. Cette dernière doit être appréhendée sous une triple dimension.

Elle est d'abord une *émotion*, plus forte que la colère, vis-à-vis d'un *objet* (cible) ; mais la haine sous-entend aussi un *récit*, un scénario, une « argumentation », qui rend l'émotion légitime. Entre diagnostiquer les causes d'un « mal » et/ou suggérer des solutions, le récit de haine puise sa polyphonie dans un ensemble de jugements partageables, des « savoirs de croyance » ; il stigmatise la différence et refuse l'Autre.

Les actes de langage opérés dans les messages haineux sont souvent les mêmes : accuser, ordonner, décrire, associer, menacer, critiquer, faire taire, offenser, inciter à un acte (qui peut être violent)… Certains groupes sont les plus visés en raison de leurs caractéristiques spécifiques – notamment « raciales », ethniques, religieuses, sexuelles. Ce sont des groupes « protégés ») par la loi, puisqu'il s'agit de caractéristiques qui font partie structurante de l'identité d'un individu.

Mais les cibles de la haine vont au-delà de ces populations. Les propos haineux deviennent les noyaux d'espaces d'interaction en ligne où la disqualification de l'interlocuteur (son *ethos*) constitue l'objectif principal.

Cet univers émotionnel (de « pathémisation ») puise sa justification dans un fond plus général de « désordre social » qui vise aussi les « élites » : journalistes, médias, hommes et femmes politiques…

Défi numéro 3 : détecter automatiquement le discours de haine

L'absence d'une « norme de référence » dans la définition du discours de haine en ligne rend sa détection difficile, aussi bien pour les modérateurs humains que pour les machines (algorithmes). D'autant plus que le contenu en ligne à analyser est vaste et que le contexte – situationnel (actualité) et culturel (histoire, société, etc.) – affecte l'émergence et la nature de ce type de discours.

Dans le domaine de la linguistique informatique – notamment en anglais, danois, allemand, italien et finlandais –, la lutte contre le discours de haine passe souvent par le recours à des listes préétablies de mots « haineux ». Toutefois, ces listes arrivent difficilement à capter les formes les plus subtiles – et donc implicites – de haine (sarcasmes, euphémismes, stéréotypes, références contextuelles, par exemple historiques), ainsi que les « stratégies de masquage » mises en place par les internautes (jeux avec l'orthographe, crypto-langages réservés aux initiés, etc.).

Le recours au *deep learning* (« apprentissage profond »), qui présente quelques résultats impressionnants, apparaît comme une tendance actuelle. Des données étiquetées manuellement comme exemples « haineux » permettent ainsi à un ordinateur d'apprendre à étiqueter d'autres données de manière autonome.

Mais des progrès restent à faire et des recherches plus poussées sont nécessaires, afin de réussir à tenir compte des caractéristiques lexicales combinées à des paramètres syntaxiques et contextuels, pour repérer toutes les formes du discours de haine dans les contenus générés par les utilisateurs en ligne.

Malgré quelques avancées, une question de fond demeure pourtant sans réponse. Même si on parvenait à réduire, voire à réprimer les discours haineux sur Internet, peut-on espérer faire disparaître la haine elle-même ?

web **Internet** **racisme** **Facebook** **intelligence artificielle** **harcèlement** **Twitter** **violence** **antisémitisme** **cyberviolences** **langage**