L’intelligence artificielle contre la haine en ligne

Repérer automatiquement sur Internet les discours incitant à la haine contre les musulmans, et aider les ONG à contre-attaquer de manière efficace, tel est le but du programme de recherche européen, Hatemeter, auquel participe un sociologue de l’université, Jérôme Ferret.

Les discours de haine fleurissent et circulent massivement en ligne, notamment sur les réseaux sociaux. En 2019, un rapport européen sur l'islamophobie montrait que les musulmans figuraient parmi les premières victimes de cette haine qui peut parfois se transformer en tragédie. Il suffit de se souvenir l’attentat raciste survenu à Hanau, en Allemagne, le 19 février dernier ; il visait deux bars à chicha et a provoqué la mort de neuf personnes.

Face à ce constat, quinze chercheurs, dont Jérôme Ferret, sociologue à l’Université Toulouse Capitole, ont mis au point un outil d'analyse en ligne et de lutte contre la haine antimusulmane, Hatemeter. Financé par la Commission européenne, ce projet, coordonné par le groupe de recherche eCrime de l’Université de Trente en Italie, associe des criminologues, des sociologues, des linguistes et des informaticiens.

> “ Ce cyberharcèlement provoque fréquemment de la dépression ”

La plateforme Hatemeter est avant tout destinée aux ONG et aux acteurs associatifs de la lutte contre les discriminations. Elle leur permet en effet de repérer et analyser automatiquement les discours de haine anti-musulmans circulant sur Internet et les réseaux sociaux. Elle identifie en temps réel les éléments caractéristiques de ces discours et permet de comprendre les schémas de pensée associés. Grâce à l'intelligence artificielle, la plateforme va même jusqu’à suggérer automatiquement des trames de réponses afin d’alimenter des contre-discours efficaces, utiles pour des campagnes de sensibilisation.

Très innovant, cet outil s’appuie sur une combinaison de langage naturel, d'apprentissage automatique et de visualisation. Il a été validé à la suite d’expériences menées en Italie, en France et au Royaume-Uni avec environ deux cents salariés et bénévoles des trois ONG partenaires du projet.

Contexte et témoignages

Dans un premier temps, il a fallu identifier les configurations d'expression de cette haine et ses spécificités historiques pour chacun des trois pays. En France, la construction de ce discours provient notamment de l'histoire coloniale, et de sa relation instrumentalisée en particulier avec l'Algérie : « Aujourd'hui, il subsiste des discriminations fréquentes envers toute personne qui présente une caractéristique pouvant avoir un lien avec la population issue de l'immigration d'Afrique du nord, que ce soit par le nom, les habitudes culturelles souvent fantasmées sans parler de la religion », précise le rapport.

La croissance de l'islamophobie a accompagné le développement de l'extrême droite. Puis les attaques terroristes récentes, revendiquées en partie par des groupes se réclamant de l'Islam, ont accru dans l'opinion publique l'association stéréotypée entre « musulmans » et dangerosité.

> “Ils ne font pas état des attaques subies”

Le rapport pointe par ailleurs le paradoxe entre, d'un côté, le caractère illégal de ces discours (une loi a été adoptée en juillet 2019 par l'Assemblée nationale, interdisant sur les réseaux sociaux les messages incitant à la haine, la discrimination ou la violence envers des populations), et d'autre part le principe de laïcité inscrit dans la constitution mais dont l'interprétation est sujette à forte controverse. N'importe quel signe religieux même s'il n'est pas accompagné d'un discours religieux, a tendance, en France, à être interprété comme du prosélytisme, observe le rapport.

Interviews et mots clés

Pour réaliser la base de données de cette plateforme, il a fallu identifier les mots clés des discours de haine anti-musulmans, sur Twitter et Youtube. Pour la France, ont été intégrés tous les tweets employant les mots clé et hashtags suivants : #EtatIslamique #IslamAssassin ; #IslamHorsdEurope #IslamDehors ; #Islamisation #Islamophobie ; #StopCharia #StopIslam ; #StopIslamisme #Invasion Musulmane ; #Musulmans. Il s'agit des mots clés considérés, au départ du projet, comme les plus pertinents au regard de leur utilisation sur Twitter, selon l'ONG partenaire. Mais de nouveaux hashtags et mots clés ont été ajoutés ensuite, incluant : #hijab #GrandRemplacement et #remigration #laïcité ». Au total, 151 738 tweets, 268 548 commentaires et 1 206 347 partages (retweets) ont été consignés entre septembre 2018 et mai 2019.

> “Alimenter des contre-discours”

Dans le cadre de cette recherche exploratoire, des victimes d'islamophobie en France ont aussi été interviewées longuement et plusieurs enseignements ont pu être tirés. Tous les musulmans ne connaissent pas les associations qui peuvent les soutenir, ni même leurs droits, si bien qu'ils ne font pas état des attaques subies alors même que ce cyberharcèlement provoque fréquemment de l'anxiété, de la dépression, les mêmes effets post-traumatiques que les agressions sexuelles. Porter en justice de telles affaires coûte cher et prend du temps. Il n'y a pas de célébrité ayant lutté avec succès contre l'islamophobie, comme il peut y avoir aujourd'hui des victimes d'attaques sexistes ou racistes.

Quelques résultats

Cet outil s'avère particulièrement utile pour former les nouveaux employés et volontaires des ONG. Il permet de communiquer sur le phénomène de l'islamophobie à partir de données et de faits concrets mais aussi de suggérer des trames de contre discours et formuler des réponses appropriées. Une estimation a montré que ceux qui l'utilisaient réduisaient leur temps de réponse de moitié, passant de huit à quatre minutes.

> “Combattre aussi l'homophobie, la transphobie ou l'antisémitisme”

Cette plateforme a également permis d'observer les pics et les tendances narratives islamophobes sur une durée choisie, d'enquêter sur les évolutions du débat public autour d'un événement spécifique, en observant les réactions les plus répandues ou les hashtags les plus fréquents. Elle peut ainsi donner une idée du contexte dans lequel un discours de haine se développe et aider à

décortiquer et comprendre ces constructions narratives. L'analyse des occurrences des hashtags peut enfin fournir des indications intéressantes sur l'affiliation politique ou idéologique de groupes d'utilisateurs.

Initialement, l'objectif de Hatemeter portait sur l'analyse et la lutte contre les discours de haine antimusulmans mais la plateforme a pu être utilisée pour combattre l'homophobie, la transphobie ou l'antisémitisme. Il a suffi pour cela de modifier les mots clés de la base de données. Il sera également possible de la dupliquer pour étudier les phénomènes de haine dans d'autres pays.

Jérôme Ferret

Sociologue, maître de conférences HDR (EHESS), membre de l'Institut du droit de l'espace, des territoires et de la communication (**IDETCOM**) et directeur-adjoint de la Maison des Sciences de l'Homme et de la Société de Toulouse (**MSHST**), Jérôme Ferret est spécialisé dans la sociologie de la violence et du maintien de la sécurité.

Hatemeter

D'une durée de deux ans (2018-2020**),** le projet **Hatemeter** est financé par la Direction Générale de la justice et des consommateurs de la Commission européenne. Il est coordonné par l'Université de Trente en Italie et associe l'Université Toulouse Capitole, l'Université Teesside (Royaume-Uni), la Fondation Bruno Kessler, et trois organisations non gouvernementales : Amnesty International (section italienne), le Collectif contre l'Islamophobie en France, et Stop Hate UK.

Pour aller plus loin

Présentation détaillée du projet Hatemeter

"Accélération de l'information et dérives en ligne : Internet, le bon coupable ?", Revue Exploreur