L’Est républicain, Le Républicain lorrain, Vosges matin, 9 juillet 2023

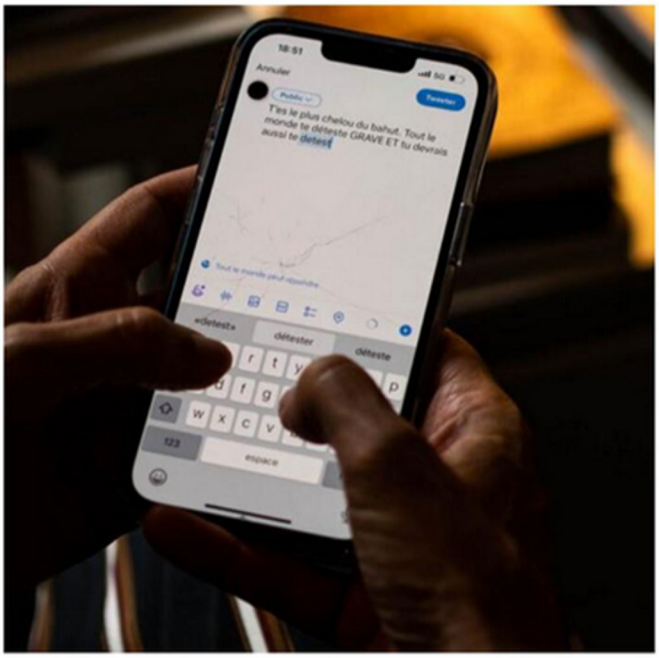
Grande Région

Un outil franco-allemand de modération contre la haine en ligne

Les chercheurs en sciences humaines et informatiques se sont associés, trois années durant, en Lorraine et en Allemagne, pour mettre au point un modérateur de messages de haine sur les médias et autres réseaux en ligne. Il est, pour l’heure, le meilleur sur le marché !

C’est la règle. À l’intérieur de nos frontières européennes, les médias et plateformes internet ont l’obligation de supprimer les messages haineux dans les 24 heures suivant leur mise en ligne… Une injonction forcément difficile à faire appliquer au quotidien, alors que le déversoir a depuis longtemps pris des allures de puits sans fond. Une problématique à laquelle se sont attaquées deux disciplines de l’Université de Lorraine, de Mayence et Sarrebruck en Allemagne : les « sciences de l’information et de la communication » et les « sciences de l’informatique ». Côte à côte, durant trois ans, les chercheurs sont parvenus à mettre au point un outil, sorte de modérateur, permettant de nettoyer aussitôt la toile de ses commentaires illicites.

« Jusqu’ici, aucun corpus similaire n’a existé », explique Angeliki Monnier, directrice du Crem (Centre de recherche en économie et management), enseignante en information et communication, pour l’Université de Lorraine.



Un modérateur mis au point par le programme M-Phasis pour lutter contre la haine sur les réseaux sociaux. Photo Hugo Azmani

« C’est assez exceptionnel, un tel dispositif binational et interdisciplinaire. Avec une collecte de messages (près de 10 000 au total, N.D.L.R.) sur des sites français et allemands. Cela nous a permis de comprendre comment cette haine s’exprime sur les deux territoires. Nous avons travaillé surtout sur les sites des médias « mainstream » (grand public, N.D.L.R.), sur Twitter ainsi que sur des sites politiques. Sur ces trois années de travail, nous avons constaté que dans le couloir des médias, il devenait de plus en plus compliqué de trouver ces messages, effacés rapidement. »

Définir la haine

Dans un premier temps, c’est la définition même d’un écrit haineux qu’il a fallu établir. « Effectivement, car selon l’endroit où nous le trouvons, sa définition diffère. Il prend un nouveau visage. On remarque des stratégies pour contourner l’interdit : l’ironie, le sarcasme qui offrent plusieurs niveaux de lecture. Ceux sans équivoque appellent à des agressions pures et simples. »

De leur côté, les informaticiens, comme Irina Illina, ont apporté leur écot en mettant au point un système d’annotations face à chaque message considéré comme haineux. « Cela nous a permis d’obtenir des résultats d’une grande finesse, justement pour cibler les discours de haine implicite. Nous avons ainsi créé un programme permettant de les détecter, avec des algorithmes, évidemment. Notre travail a porté essentiellement sur les messages textuels et toutes les données provenant des journaux en ligne, les articles et commentaires. »

En libre accès

Le modérateur né de ce programme franco-allemand baptisé M-Phasis est aujourd’hui en libre accès, à la disposition des entreprises, de toutes les plateformes mais également des chercheurs qui pourront l’améliorer, en ligne. « On dit qu’il est aujourd’hui le plus performant mais cela ne durera pas longtemps s’il n’est pas travaillé. Comme tout ce qui touche aujourd’hui au numérique. »

● **Saada-Gisèle Sebaoui**